Understanding User Behavior From Online Traces

Elad Kravi
Technion – Israel Institute of Technology
Haifa, Israel
ekravi@cs.technion.ac.il
Advisors: Yaron Kanza and Benny Kimelfeld
Expected Graduation: Dec. 2017

ABSTRACT

People nowadays share large amounts of data online, explicitly or implicitly. Analysis of such data can detect useful behavior patterns of varying natures and scales, from mass immigration between continents to trendy venues in a city in turn. Detecting these patterns can be used for improving online services. However, capturing behavior patterns may be challenging, since such patterns are often of a specialized essence, no benchmark or labeled data exist, and it is not even clear how to formulate them to enable computation. Moreover, it is often unclear how recognition of these patterns can be translated into concrete service improvement.

We analyzed major datasets of three common types of online traces: microbloging, social networking, and web search. We detected online behavior patterns and utilized them toward novel services and improvement of traditional services. In this paper we describe our studies and findings, and offer a vision for future development.

Keywords

Pattern detection, user behavior, online traces, log analysis

1. INTRODUCTION

Online activity of users is constantly growing, and is facilitated through various devices and applications. Users post texts on microblogs, social networks and other platforms. They pose queries in search engines and click on some of the results. Besides the data provided explicitly by users, different attributes are collected, including locations, timestamps of activities, the devices users are using, etc.

Collected data can be analyzed for detecting *online behavior patterns* of users, and to improve online services. Patterns can be detected at a varying granularity. Some patterns are related to large populations, for example, mass immigration between continents [18], change in life patterns (e.g., moving from rural to urban areas), detection of diseases and search behavior on the web [7]. Other patterns take place at a smaller scale, for example, analysis of popular events and traffic within a city [19, 2].

Given traces from online activity, detection of a pattern requires the definition of a formal *model* that can properly capture the behavior and allows its automatic detection. Such model should avoid detecting erroneous patterns, while properly portraying the diversity of human behavior. For example, consider a model for detecting a gathering of a crowd in a small place. This should not be confused with many people waiting in a traffic jam. However, it should also be sensitive enough to detect relatively small gatherings of, say, merely a few dozens of demonstrators.

Frequently, verifying a proposed model is difficult due to the lack of proper ground truth. For example, consider a model capturing the difference between queries posed in places that are familiar to the user and unfamiliar places [7]. Without knowing whether a place is familiar to a specific user, it is hard to verify the model.

Pattern detection can be used for improving online services, e.g., detection of crowds or of heavy traffic can improve route planning and navigation services. Yet, there may be many types of anomalies, events and fluctuations in a large data set. So, detection of significant events and defining patterns to specify notable events is challenging.

In this paper we present studies we conducted for detecting and utilizing online behavior patterns of users. We studied three major types of data: microblogs, social activity, for example, posting of messages on Facebook and web search. In particular we analyzed Twitter data, social networks in large corporates, and search queries that were posed in a commercial search engine. We focused on behavior patterns related to locations of users, content of their messages (or queries), the topology of the network and properties of the user, such as age, or of the related context of the search query, e.g., the type of device that was used or the time during the day when the query was posed.

The rest of this paper is organized as follows. In Section 2 we present an analysis of online activity in microblogs, in Section 3 we present our study of activity in social networking, and in Section 4 we present a study of phenomena detection in web search. We conclude in Section 5.

2. ANALYZING MICROBLOG TRACES

Microblogs, like Twitter, allow users to publish geotagged posts—short textual messages assigned to a geographic location. Users send posts from places they visit and discuss personal and general topics. We utilized geotagged posts to discover online behavior patters based on location tags and

SIGMOD'16 PhD Symp, June 26-July 01 2016, San Francisco, CA, USA

DOI: http://dx.doi.org/10.1145/2926693.2929901

textual content of posts. We focused on the following problems: *(1)* discovering geospatial similarity between users, and *(2)* detection of geospatial correlation between places in a city, that is, finding pairs of places such that many users who visited one of the places also visited the other place.

2.1 Geospatial Similarity between Users

Detecting the similarity between users is an important problem with many applications. For example, personal recommendations are commonly generated based on similarity between users [10]. Utilizing data from microblogs for this purpose has been extensively studied [11, 1, 8].

Existing methods for calculating the similarity make use of different properties of the data, like the textual or the spatial content. We studied whether it is possible to increase the accuracy of similarity measurement by combining geospatial attributes of posted messages with the textual content of messages the user posted. In our study [3] we tackled, among other problems, the following questions: 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 a similar textual content)? Do posts with similar content have a spatial distribution similar to that of any random set of posts? We provided statistical tests to examine the correlation between two methods for calculating similarity between users, one based on the textual content of the posts, and the other on the spatial content of the posts. We also considered a hybrid approach, combining the two aspects.

In our model, each user is associated with a set of *posts*. Each post has a location, specified by the geo-tag, and textual content. The similarity between two users is the inverse of the distance between the sets of their posts. Spatial distance was measured by the *nearest neighbor* distance [16], and textual distance by the standard *TF-IDF* formula [15].

We examined the correlation between textual similarity and spatial similarity in geotagged posts, and showed that although there is some correlation between them, they provide different similarity measures. Combining textual and spatial similarities is beneficial for calculating the similarity of users using their posts, and may be beneficial rather than methods using merely the locations or the textual content.

For example, in Fig. 1 we present a comparison between the similarity measures, for the task of user identification. In this task we randomly split the posts of a user u, and thereby transforming u into two distinct users u_1 and u_2, each associated with one part of the messages. We expect a good similarity measure to determine that u_2 is more similar to u_1 than most other users. Given one user, say u_1, we used the different similarity measures to search for the k most similar users to u_1 in a set of 1000 users including u_2. A success was considered if u_2 was found in the resulting set. It can be seen in the figure 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. We examined different combinations of content and locations, see details in [3].

In [3], we showed that in a place like Manhattan, there are terms that can be associated with specific places and terms that cannot be associated with geographic locations. For example, the term "observe" was associated with the area of the observation deck, in Rockefeller center. The term "hospital" was sent from a large surrounding, hence, was not associated with a specific place. Also, some areas were found to be characterized by specific terms, while others are not. We presented initial results of these properties, showing they might be beneficial for various applications, such as recommender systems.

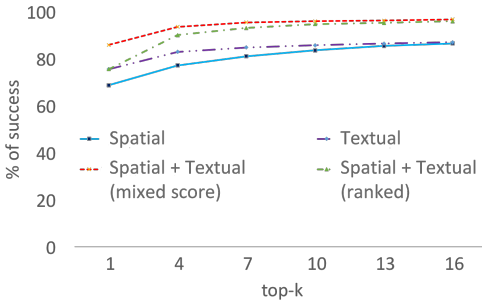

Figure 1: Identification test: Accuracy as a function of k

We are now extending and generalizing this study, aiming at the development of more accurate and efficient location-based similarity measures.

2.2 Correlated Locations

In [6], we discovered geosocial associations between places, that is, pairs of places in a city that were jointly visited by many users. Detecting these places can improve, for example, the planning of transportation routes, as these associations reveal a demand for transportation between the places. Another potential use case is in *recommendation systems*, providing tourists with recommendations for places to visit, based on joint-location history.

A major challenge in implementing such a system is the large volume of data. The data cannot be indexed by location since for a given pair of locations, we are looking for users who visited both places. Ignoring the location is also impossible since we are trying to group posts sent from nearby locations. We developed a system that collects, stores, clusters and processes geotagged posts efficiently to find jointly-visited places.

We have presented a demonstration [6] of our system. We have shown how a large number of messages can be collected, clustered, and analyzed using different parameters, for finding jointly-visited places in different cities in the world (New York, Los Angeles, London). The system illustrates an intuitive approach for presenting connected places on a map, while allowing novice users to control various parameters.

Fig. 2 presents the different parts of the system, including the back-end tier which consists of *crawlers*, a *database* and a *web server*. The front-end is presenting an intuitive interface for creating new data analysis tasks, detecting jointly-visited locations, and presenting the results of previous tasks. A video of the system is available via YouTube.[1]

3. ANALYZING SOCIAL NETWORKING

Enterprises often adopt social networks as an in-house communication tool between employees. Such social networks facilitate interaction within the enterprise. But, beyond that, the analysis of user behavior in these social networks can be used for improving the quality of other internal services, as we illustrates in this section.

[1]http://www.youtube.com/watch?v=nUbc4uqsprs

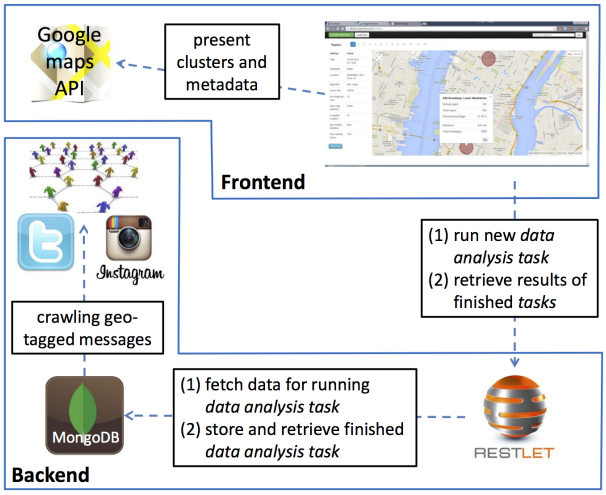

Figure 2: System Architecture of City Nexus [6]

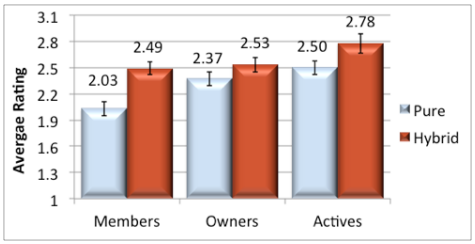

Figure 3: Average ratings for pure vs. hybrid profiles

In online social networks, users can preform activities in the context of groups, sometimes called communities. Studies [9] show that communities became central in social networks. For example, our measurements show that over half of the activities in the social network are performed in the context of a community. One service offered in networks, is of recommendations for content items. For example, users may be suggested with relevant posts, links and communities that might be interesting to them. Differently from recommendations for individuals, generating recommendations for a group of users sharing a mutual interest introduces new challenges [5]. For example, one of these challenges is how to aggregate profiles of users in the community, generating one profile representing the community as a whole.

The key role in the community is that of the *owner*. Owners are responsible for establishing the community, adding members and content, and keeping the community alive and engaged. Thus, we developed a recommender system for community owners. The system is designed for suggesting relevant content items to community owners, allowing them to rank the items, and to share them with their community.

Now, we further detail about generating group recommendations in our recommender system. One approach for doing so is by creating a single profile for the entire community as we explain below. This *community profile* is then the basis for generating recommendations. Recommendations were generated, in our system, by issuing a query containing the profile elements, including people and tags/terms, to a social search system [14]; resulting in a list of items from the social network. For example, our system can recommend a Wiki page about a new Java version to community of developers.

We considered three approaches for creating community profiles. The first approach, generates the *Member-Based* Profile (MBP), and it follows previous work aggregating users profiles [17]. We refined the past technique by distinguishing different types of users. The second approach produces the *Content-Based* Profile (CBP), by considering textual content of the community description. The third approach is a hybridization of both.

In MBP, we considered restrictions to several populations—the owners, a random subset of members, and the active members. In our study, "active" was based on the user's past activity in the community. To create a MBP profile we aggregated profiles from individuals from the corresponding population. In CBP, we extracted terms from the title, summary and tags of the community. The hybrid profiles, combined MBP and CBP. Overall, we considered 7 different profiles (3 MBP, 1 CBP, and 3 hybrid profiles).

In order to evaluate the generated profiles, we assigned a random profile to each community, and generated recommendations based on this profile. We then sent an email to random owners from each community, asking them to participate in a survey.

The survey was composed of two parts. In the first, described in Section 3.1, we asked the owners to characterized their community, and to rank a set of recommendations as described below. In the second, presented in Section 3.2, we continued the experiment for three more rounds of recommendations, and measured the engagement level inspired in the communities that participated in the survey. We now further detail about the results of each part.

3.1 Recommending to Community Owners

The first study [12] started by asking the owners to characterize their community. Then we presented them with 11 recommended items, 10 were generated by the chosen profile, and the $11th$ item was randomly selected as a control. Each recommended item included an icon that represented its type, its title with a link to the original entry in the enterprise's social network, the names of the authors, the last-update date, and up to 5 related tags and 5 related people, if existed. The owners were asked to rank the recommendation on a 5-point scale. We received 907 responses to our $7,592$ survey invitations (12%). These responses covered a total of 851 distinct owners of 796 different communities.

Comparing the profiles, we have found that hybrid profiles (MBP with CBP) yield better recommendations than non-hybrid profiles. Among the MBP profiles, the profile that is based on active members was found to be the most effective. Fig. 3 shows the average ratings for the three MBP compared to the hybrid version. The average of the pure CBP was 2.48. The differences between the profiles were all found to be statistically significance (one-tailed unpaired t-test, $p < 0.001$). Hence, *active members*, emerge as the most effective group for producing interesting recommendations, specifically, outperform the set of owners despite the fact the recommendations were evaluated by owners.

We also analyzed other behavior patterns such as differences in ratings between large and small communities, and between importance of items to the owner and to the community. See details in [12].

3.2 Increasing Activity By Recommendation

In the second part of the experiment, we examined the impact of sharing content items on the community's engagement [13]. The problems we studied were twofold. First, we explored the challenge of regenerating recommendations over a short period of time, and secondly, finding a method for measuring engagement by activities

We extended the previous experiment to four rounds of recommendations. In each of the last three rounds the owners were presented with 5 new recommendations, based on the profile assigned to the community in the first round. For each recommended item, the owners were asked about their willingness to share the item with the community. They were also able to actually share it with the community, by creating a new content item in the community. Overall, 1033 sharing actions were carried out in our survey across all four rounds, over 7.23% of all recommended items. These actions were performed for a total of 340 communities (37.65%) by a total of 354 owners (33.62%).

Activity is a widely used measure for community success (see e.g., [4]). To measure the difference in activity due to the shared items, we focused on the eight weeks preceding our survey (the survey started on August 8^{th}, 2013) versus the eight weeks that followed. We disregarded any activity that was performed by using the sharing action in our survey.

Fig. 4 shows the average activity level for the communities that took part in round 1 (i.e., all survey communities for which at least one owner responded) and round 4, compared to the control group (a set of communities that did not participate in the experiment). It can be seen that before the survey started, all three groups had a rather similar level of activity, while after the beginning of the survey, the activity level of both round-1 and round-4 communities became substantially higher than the control group. This difference was consistent and stable across all eight weeks that followed the beginning of the survey.

4. ANALYZING TRACES OF WEB SEARCH

In previous sections we discussed the analysis of traces from a microblog and a social network. Next, we discuss the analysis of user activity in web search engines.

Search engines are important arena of online activity. Users interact constantly with them, searching for information and while doing so reveal special behavior patterns. We focused on two patterns; the first examines whether search is affected by the level of familiarity of the user of its geographical environment at the time of the search. Specifically, do users have different search intent in familiar locations and in unfamiliar locations. Our second study explores a set of queries in which users multi-click on several results of the search engine. We now further describe the two studies.

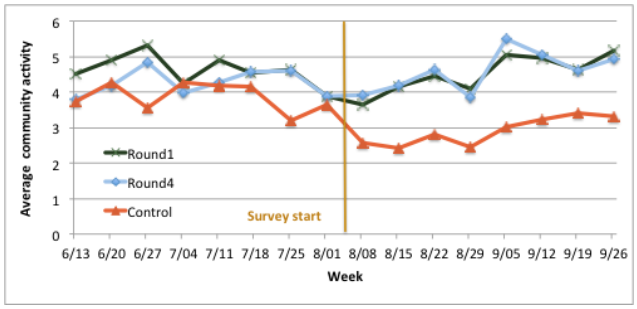

Figure 4: Community activity before and after the survey.

4.1 Location Sensitive Search

Do users have different information needs in a familiar surrounding and in unfamiliar surrounding? Can this be utilized by search engines to improve information retrieval? Our hypothesis, in this study, was that information need is indeed affected by the familiarity of the environment.

To examine our hypothesis, we defined a formal model for familiar and unfamiliar locations. As no benchmark exists, we verified our model using several approaches elaborated below. We characterized the differences between searches in familiar and in unfamiliar locations, pointing out the differences. Finally, we showed an indication for improvement of query auto-completion using our model [7].

In our model, a place P is considered familiar to a user u if *(1)* u was active (i.e., posed search queries) in P at least 10% of the days she was active, *and (2)* u returned to P at least twice (i.e., searched in P, then in another location, and then in P again). Another case, in which we consider P as familiar to u, is when all the search activities of u were performed in P. While the first condition can be interpreted as the portion of time, in days, spent in P, the second condition reassures the visit of u in P was not a unique event. By requiring these conditions we filtered 51.4% of the places. The parameters of the model (10% of the days and 2 returns) were tuned after an examination of different values, further details appear in the paper. A *place*, in this study, was defined in the granularity of a postal code.

We verified our model by comparing it to the home location, as declared by the users. We found that for 53% of the users, the minimal distance between the user's home and the places considered familiar (in our model), was smaller than 20 kilometers, and for 75.4% of the users this distance was smaller than 100 kilometers. Overall, this confirms that in many cases the home of the user is a familiar place, but not always. Note that the declared home is not always the actual home of the user (e.g., a fake or obsolete address).

We presented an initial results of an auto-complete application that offers to use different completions for familiar versus unfamiliar places. In Table. 1, we reported the initial word of the query, together with excerpts from the query completion lists, ranked in decreasing order by completion frequency. For example, in the "Food" category, searches from familiar locations focus on aspects more likely to be performed in one's home, such as marinading a steak or buying a coffee table.

4.2 Understanding Queries with Multiple Clicks

Clicking on search results is considered a key signal for search engines, relating between queries and offered links. Previous analysis of such data does not include queries followed by multiple clicks performed by the same user. Observing this subclass of queries reveals an interesting intent of engaged users that explore several results, and indicates on a complex information need that requires special handling by search engines. By automatic detection of multiple clicked queries search engines can improve retrieval models e.g., by re-ranking the results. They can also offer enhanced user experience to support this set of queries.

Table 1: Sample word completion patterns reporting the first word (prefix), and example completions, excluding stop words, with rank, sorted by decreasing completion frequency for the familiar and unfamiliar settings, respectively.

Category	Prefix	Familiar	Unfamiliar
Food	coffee	table:2, shop:4	shop:2, table:5
	steak	marinade:1, house:2	house:1, marinade:4
	pizza	dough:3, places:5	places:3, dough:5
Travel	gas	fireplace:3, station:6	station:3, fireplace:8
	train	crash:1, schedule:7	schedule:2, crash:6
	car	games:2, wash:6	wash:2, games:7
Leisure	wild	rice:2, horse:5	horse:2, rice:4
	soccer	drills:4, scores:13	scores:5, drills:7
	piano	sheet:2, bar:8	bar:2, sheet:4

However, defining the class of Multiple Clicked Queries (MCQs) requires to aggregate search activities associated with each query, including varying number of clicks. We formally define MCQs, and characterize the differences from its complementary set, SCQ——sparse click queries.

We modeled MCQs as queries that most of the search activities (query, and a set of followed by clicks) associated with them include two or more clicks.

Our dataset includes 31.42 million search activities sent to a popular commercial search engine, sampled at random between May 1^{st} and May 21^{st} 2015. The dataset included 2.23 million search activities associated with MCQ (according to our modeline), which account for 6.52% of the data.

We found MCQs to differ from SCQs by various parameters, among them is the semantics of the queries. Queries related to *Questions*, *Health*, *Reference* and *Adult Content* are more common in MCQs rather than in SCQs. *Science*, *Shopping* and *Business* are more common in non-MCQs. The results can be explained by the hypothesis that MCQ encapsulates a need for exploration, that is more common in the domains found. Seeking for an entry in Wikipedia, or searching for a merchandise has a navigational behavior, hence being more common in SCQ.

Automatic detection of MCQs was found challenging, since the data is imbalanced; MCQs are about 6.5% of the overall data, and the search activities associated with them have various number of clicks. Our contribution also includes an indication that classification works with precision of 72.5% and recall of 86% over a balanced dataset. More details will be published in a future paper.

5. SUMMARY AND OUTLOOK

We presented several studies of problems related to detection and utilization of online behavior patterns of users. The studies were conducted over traces of three major platforms: microblogs, social networks, and logs of web search.

For analyzing microblogs, we used the textual content and the locations of messages (tweets, posts). In one study we measured the similarity between users, and showed that it is possible to improve similarity detection by combining the location and textual content. In another study we utilized the posts for finding pairs of jointly visited locations.

To improve the utilization of corporate social networks, we compared between several methods for generating recommendations for communities. The methods are based on different populations of each community and on its content. Then, we studied the effect of these recommendations on community engagement.

We used a large repository of search queries to examine search patterns of users on the Web. In one study, we analyzed the differences between queries posed in places that are familiar to the user and queries posed in unfamiliar places. In a second study, we analyzed queries associated with multiple clicks. We proposed a model to capture this behavior, presented statistical analysis of the phenomenon and showed an initial indication that automatic classification of the model is possible.

We consider two main directions for extending our work. The first direction is building tools for finding complex patterns that combine traces from different datasets. For example, improving web search by using social activity, such as queries posed by friends of the current user.

Another direction is developing infrastructure for allowing users to define their models of online behavior patterns, and later on detecting these patterns on real datasets.

6. REFERENCES

[1] K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, and Y. Yu. Collaborative personalized tweet recommendation. In *SIGIR*, 2012.

[2] J. Cranshaw, R. Schwartz, J. I. Hong, and N. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *AAAI*, 2012.

[3] I. Grabovitch-Zuyev, Y. Kanza, E. Kravi, and B. Pat. On the correlation between textual content and geospatial locations in microblogs. In *GeoRich*, 2014.

[4] A. Iriberri and G. Leroy. A life-cycle perspective on online community success. *ACM Computing Surveys (CSUR)*, 2009.

[5] A. Jameson. More than the sum of its members: Challenges for group recommender systems. In *AVI*, AVI '04, 2004.

[6] Y. Kanza, E. Kravi, and U. Motchan. City nexus: Discovering pairs of jointly-visited locations based on geo-tagged posts in social networks. In *SIGSPATIAL*, SIGSPATIAL '14, 2014.

[7] E. Kravi, E. Agichtein, I. Guy, Y. Kanza, A. Mejer, and D. Pelleg. Searcher in a strange land: Understanding web search from familiar and unfamiliar locations. In *SIGIR*, 2015.

[8] S. M. Kywe, E.-P. Lim, and F. Zhu. A survey of recommender systems in twitter. In *Social Informatics*. 2012.

[9] T. Matthews, S. Whittaker, H. Badenes, B. A. Smith, M. Muller, K. Ehrlich, M. X. Zhou, and T. Lau. Community insights: helping community leaders enhance the value of enterprise online communities. In *SIGCHI*, 2013.

[10] P. Melville, R. J. Mooney, and R. Nagarajan. Content-boosted collaborative filtering for improved recommendations. In *AAAI/IAAI*, 2002.

[11] O. Phelan, K. McCarthy, and B. Smyth. Using twitter to recommend real-time topical news. In *RecSys*, 2009.

[12] I. Ronen, I. Guy, E. Kravi, and M. Barnea. Recommending social media content to community owners. In *SIGIR*, 2014.

[13] I. Ronen, I. Guy, E. Kravi, and M. Barnea. Increasing activity in enterprise online communities using content recommendation. To appear in TOCHI, 2016.

[14] I. Ronen, E. Shahar, S. Ur, E. Uziel, S. Yogev, N. Zwerdling, D. Carmel, I. Guy, N. Har'El, and S. Ofek-Koifman. Social networks and discovery in the enterprise (sand). In *SIGIR*, 2009.

[15] G. Salton and M. J. McGill. Introduction to modern information retrieval. 1986.

[16] H. Samet, J. Sankaranarayanan, and H. Alborzi. Scalable network distance browsing in spatial databases. In *SIGMOD*, 2008.

[17] C. Senot, D. Kostadinov, M. Bouzid, J. Picault, and A. Aghasaryan. Evaluation of group profiling strategies. In *IJCAI*, 2011.

[18] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392), 2012.

[19] C. Xia, R. Schwartz, K. Xie, A. Krebs, A. Langdon, J. Ting, and M. Naaman. Citybeat: Real-time social media visualization of hyper-local city data. In *WWW*, 2014.