Understand Function of Location Entity Based on User Generated Content in Online Social Network

ABSTRACT

Microblogging, as a commonly used form of social network, has become a popular data source for research. Some work relative to the understanding of locations and location inference are done based on microblogging platform by text data, time series data and graph mining on social relations. However, there is yet not many studies that demonstrate how to do modeling and understand the location entities within a scale of city based on such data sets. In the work, based on purely user generated content, a method that

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications—Data mining; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Relevance feedback;

General Terms

Measurement, Documentation, Experimentation

Keywords

Text data mining, Location understanding, Social networks, Geo-location

1.INTRODUCTION

Traditional map search are mainly based on name matching. In order to satisfy the need of search by query about the location without the name, manually tagging and learning based automatic tagging are introduced [1]. More recently, reviews of locations generated by users on crowdsourcing based review website are hired as an extension of traditional method that allows users to find relative locations with queries that are not the name of locations. However, with the fast development of social network, there are now so many location related post generated everyday which would be much more helpful to a better understanding and modeling the function of locations.

Microblogs are short posts that generated and shared by users to reflect the life of them in a direct or indirect manner. In today's world, more and more microblogging posts are generated by users all over the world. People post huge amounts of "short messages" in microblogging platforms, making it possible for researchers to extract useful information related to social human behaviors [2]. Mining this data enables people to gain hidden knowledge and understand more about not only the physical world but also user behaviors and human dynamics. Considering marketing applications, the discovered knowledge about life of people can be beneficial to personalized recommendation and further consuming behavior [3].

Unfortunately, the information of location, as an important aspect for describing the life of a person, is not broadly adopted by the majority of users on microblogging platforms. Therefore, understand a location based on data without knowing exact longitude and latitude information which may make users feeling less uncomfortable because of some privacy concern is raised as an interesting problem. A better understanding of location would result in more accurate recommendation for news[4], advertising[5] and detection of events[6]. Online social network service providers would be willing to pay for such a system to enhance the performance of their existing ads system or news feed recommendation system.

Sina Weibo, as the largest Chinese microblogging website, allows users to post with 140 characters and interact similar to Twitter by quoting other people with "@", posting with a certain topic with "#Tag Name#" and repost like "//@UserName:". It also enables users to like a post and comment below a post like what users do on Facebook. Sina Weibo has a verification program for famous people and special organizations. Once an account is verified, a ‘V' badge and a verification description will be added next to the account name. Sina Weibo has more than 500 million users as of December 2012[7] .More than 100 million posts are generated everyday by users on web and mobile application.

In this paper, based on the posts data on Sina Weibo generated by users living in Beijing, a method for accurate detection of location related terms within a city is developed. By mapping location related terms to a certain longitude and latitude, a strategy to generate a bag of words model for each of the locations is suggest. Based on the model for locations An application of visualization is also developed, allowing users to input some query and doing map search based on the function and relative description of location instead of traditional name matching.

We make the following contributions:

- An algorithm to automatically detect location related terms within a city from contents of Weibo posts. The result dictionary data is available at URL
- Using the location related terms, a method that given the function description is introduced
- A visualization of the hot spot related to the user provided query is returned and displayed on a map when a input is given. Enabling people to explore and understand the function and programming of a city.
- The user profile data and their posts on Sina Weibo before June, 2012 is opened for public use. The URL
- The random sample data of locations in Beijing city is opened for public use.

There exist some weakness points in this study. Due to the limit of computing resource, a simple model is implemented instead of the original proposed one. Without inverted index, respond to a query could be time consuming and the cost of time is proportional to the number of locations we sampled.

The rest of this paper is organized as follows: Early related works are presents in Section 2. Section 3 introduced the data set and the definition of the problem. Plus, the processing of data and the generation of dictionary is presented. In Section 4, the detail about location model and searching implementation are introduced. The conclusion and some possible applications is introduced at last in Section 5.

2.RELATED WORK

With the fast development of Internet, in the last decade, researchers had some relevant study based on data of web pages, search queries and online social networks. People have down a lot of rsearch on location understanding and inference of locations. Beside the idea of location inference purely based on content, both gazetteer key words and probabilistic language model, some study also see into the social relation graph and pattern of activities of users.

Amitay et al.[8], Zhong et al.[10] and Fink et al.[9], propose methods to give a location for web pages according to addresses, postal codes and special names of locations. Serdyukov et al.[11] hired a probabilistic language model based on the tags on Flickr to determine the location of photo posted. Cheng et al.[12] modeled the spatial distribution of words in Twitter's user-generated content to predict the user's location. Peregrino et al. [13] combined informal text data from twitter and formal data from wikipedia to detect geographical focus at city level.

Backstrom, Sun and Marlow [15] analyzed the distance between Facebook users' social relation and predict the physical location based on the location of a user's friend. Yardi and Boyd [14] characterize network properties in relation to locations. Backstrom, et al. [16] presented the improvement of geographical prediction when relationships between users are taken into account. McGee et al. [17] and Li et al. [18] suggest accurate method of location prediction based on strong ties of users and the possible enhancement when additional unlabeled users are considered.

Mahmud et al. [19] developed a classification method based on an ensemble of statistical and heuristic to predict locations at different scale, for example city, state and time-zone, based on twitter data. Yu et al. [20] used simple statistical methods on the activities pattern to predict the location and change of location at time-zone level. Ehrlich and Shami [21] examined the use of microblog in and out workspaces and highlighted the significant differences between contents generated in workspace and out of workspace.

3.PRELIMINARIES

3.1. Data Set

3.1.1 Weibo user profile and posts data

User data and the posts data of users are crawled by a spider program. The set of users this study aimed at are verified users who have more than 500 fans and 100 posts. In order to predict locations within the city and their alternation, all users selected have a label of location in their profile that is filled up as "Beijing".

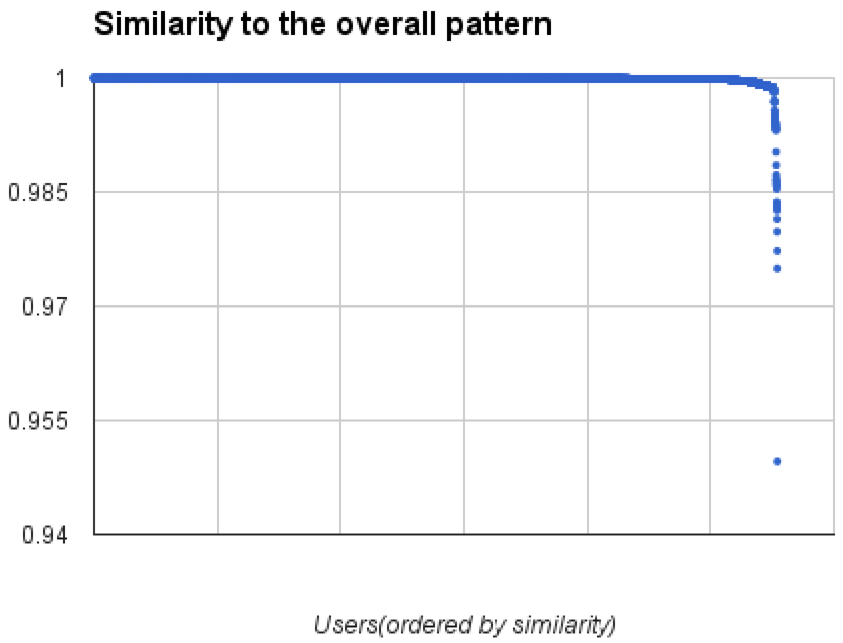
There are in total 11,082 profiles of users and 19,914,297 posts of these users generated before June, 2012 crawled at the beginning. The data of posts include the unique ID of user who posted the content, the timestamp of the post, the content, and also whether the post is a repost.

Considering the phenomenon mentioned by Yu., Sun. and Lv, the location label in the profile are not necessarily true and users may have significant change of locations and Li et al. [32] proposed that user may have different relevant locations for relationship with different person. In this study, we hired similar idea to eliminate those users with fake location and their posts. The timestamp of each post grouped by user is firstly mapped to a vector of 1440 dimensions corresponding to each minute in a day. Then the dimension of each vector is reduced from D_{old}=1440 dimensions to D_{new}=24 dimensions under the assumption that the probability of the happening of an action *a* within the ε-neighborhood follows the uniform distribution:

In the equation above, d'_m denote the estimated counts for the m^{th} dimension of the vector after dimension reduction. And d_n correspond to the n^{th} dimension of the vector before dimension reduction. In this case, we assume the probability of an action a that occur within the 30-neighborhood of d is uniformly distributed.

After that, a user-pattern matrix *A* with lower dimension is formed. Next, vector x_{all} that represent the overall pattern of a user is constructed by summing up all the vectors of each of the users.

After that, all row space vectors of A and the x_{all} vectors mentioned above are normalized by dividing each dimension by the sum of counts in each of the dimensions. Then we could define the similarity of active pattern between each user and overall set of users by simply applying cosine similarity to row vectors of A and x_{all}. The results are shown as the following diagram.



It is not too hard to point out that most patterns of users are quite similar. However, there exist some users whose activity pattern differs a lot. Users with result of similarity less than 0.9997 are removed from the dataset and result in a cleaned data set with 11,032 profiles of users and 19,802,146 posts.

3.1.2 Location data

The location data of Beijing City if crawled by using the place api provided by Baidu[22]. In total 483,191 data are crawled. The data are of the form:

Lid	lname	Laddr	Lat	lon
Location id	Location name	Location address	Latitude	Longitude

In this study, only lid, lname, lat, lon are used. The information of laddr can be used for future study.

3.2. Post Segmentation

The basic idea of segmentation is maximum matching with some specific rules (Chih-Hao 1996). Since all the words are combined together to form a sentence with no spaces between any two characters, we cannot apply traditional NLP methods on Chinese. Based on a relatively complete dictionary of common Chinese words, we can use simply match the most likely character combinations in the sentences. However, there are many ambiguities in Chinese sentences that need to be resolved. Therefore, Chih-Hao introduced 4 rules to modify the algorithm to achieve a better segmentation result: maximum matching, largest word length, smallest variance of word length (Chen & Liu, 1992) and largest sum of degree of morphemic freedom of one-character words.

The segmentation results of each post are then stored in different files with names of each user’s uids for further usage.

4.METHOD

4.1. Geographical Term Finding in Posts

Based on the segmentation results of posts, the main purpose of this part is to find those location-related terms. At first, a greedy algorithm is applied that only matches the first n characters of each location with each term that has been segmented from original posts (n is equal to the number of the target term). This method generates a very high recall (i.e. a huge amount of locations) but relatively many unrelated locations (i.e. location names that consists of common words). This method generates much useless information (or, noises) that largely increases size of the whole dataset (space of O(n^2)). Thus, an efficient and precise method is necessary so a rule-based location-related word matching is introduced. Rules are based on POS (part of speech) of each word to eliminate those noisy words that act as non-nouns.

Rule 1: Location + (be(‘是’)) + adjective. Location-related term serves as the subject of a sentence.

Rule 2: verb + location. Location-related term serves as an object.

Rule 3: prep(‘在’ for most cases) + location + (verb). Location-related term is a part of a prepositional phrase.

Rule 4: No two consecutive location-related terms without any separating characters between them.

Through the matching method with these four rules, matched location-related terms become more precise and the amount of them dropped to a reasonable number. These terms are stored in a file with the wid of its original post and the segmented term list of that post to be used later.

4.2. Location Model

By the method introduced in 4.1, the geographical term *t* in posts of Weibo data was found. Further, under the assumption that a post is relative to a location if there exist at least one term in the post is relative to the location, each posts containing the location relative term is mapped to relative locations that contain the term and hence result in locations with set of posts.

And then, each the set of posts relative to a coordinate can be represented as

$$S_L = \{P_1, P_2 \ldots P_n\}$$

where L is a specific location, and S_L denote a set of posts and $P_1, P_2 \ldots P_n$ are posts relative to the coordinate (long means longitude and lat means latitude). And each post P in S is composed of pairs of terms and its number of counts.

$$P_j = \{(t_{ji}, c_i): t_{ji} \text{ is the } i^{th} \text{ term in the } j^{th} \text{ post and } c_{ij} \text{ is the count of the } i^{th} \text{ term appeared in the } j^{th} \text{ post}\}$$

where P_j denote the j^{th} post in S.

Thus, we have a set of pairs denote the term and its number of counts in S. In this paper, the set is named Location Point Term Set and denoted by LP:

$$LP_L = \{(t_i, c_i): c_i = \text{sum over all P in S for term t}\}$$

Considering the difference between the popularity of locations, normalization for the count to a score is reasonable. And then we generated a score to replace the count in each Location Point Term Set:

$$LP'_L = \{(t_i, s_i): s_i = c_i/c_{all}\}$$

where c_{all} is the sum of count of all terms.

Besides, smoothing is needed in this case since some words frequently appeared in all the data set will do no good to the performance of further exploration. With the observation and intuition, it is not hard to find that some words and phrases appeared with different frequency in different term set are more meaningful than other ones appeared equally in all the sets.

$$LP''_L = \{(t_i, s_i): s_i = \text{alpha} * c_i/c_{all}) + (1\text{-alpha})\ c'_i\}$$

Here, c'_i is the total count of the term at global scale.

4.3. Search and ranking

Based on the segment algorithm introduced in part 3.2, the query inputs are segmented into terms. Similar to the method of building the model for location that hired in 4.2, a vector representation of query is generated. Using the count of each term in the query to get the query vector of the terms with theirs scores normalized by the total number of terms in query.

$$Q= \{(t_i, s_i): s_i = c_i/c_{all}\}$$

Then, to get the final ranking of locations, we do dot product of query vector and vectors of location point term sets (location model) with terms relevant to the query (has term matches exactly). Thus the score of the relevance of the query with a location point is:

$$QLP= \{sum\ (s_{iq}*s_{ilp}): t_i \text{ in both LP and Q}\}$$

And rank the final locations points by the score from the highest to lowest to give to the front interface.

4.3. Future work on search and ranking

Based on the segment algorithm introduced in part 3.2, the query inputs are segmented into terms. Similar to the method of building the model for location that hired in 4.2, a vector representation of query is generated. Using the modified version of BM25, the ranking scores are generated. The following four scores are the added features to BM25 when computing the ranking score of a term qi at location D based on all Weibo posts Wi:

- scoreA: (qi, D).count1 / qi.count
- scoreB: ∑ (for all Wi) Wi.location_count / Wi.term_count
- scoreC: (qi,D).count2
- scoreD: qi.score_sum

Thus, we can compute score' by:

score'=a*log(scoreA)+b*log(scoreB)+c*log(scoreC)+d*log(scoreD)

The original BM25 formula is modeified such that:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot\left(1-b+b\cdot\frac{|D|}{\text{avgdl}}\right)},$$

$$\text{IDF}(q_i) = \log\frac{N-n(q_i)+0.5}{n(q_i)+0.5},$$

f(qi, D)=Term Frequency + score’

5. CONCLUSION

Based on the microblogging data training and the Chinese word segmentation, the group is able to rank relevant locations based on user’s input in weibo form. The data for now is on Beijing for training and testing but it is scalable to larger geographical regions with more storage capability and algorithms.

In summary, the data-storage is in sparse bag or words form and we are able to:

1. Have an algorithm to detect location related terms within Beijing.
2. Have a way to visualize of locations relative to querying weibo.
3. Be scalable to be trained and tested by other data in other cities or in same cities.

For this paper, the group limits the training posts to be from users living in Beijing, thus we are having authentic training data. If we mask the information of the location of the posts for training, then the group need to be more careful for the outliers on the data. Many other improvements are possible in the future and the best case would be that the model would be able to be trained and rank locations accurately in the whole country or world-wide.

6. REFERENCE

[1] Hollenstein, L., & Purves, R. (2014). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, (1), 21-48.

[2] Grace, J. H., & Zhao, D. (2010, April). Microblogging: what and how can we learn from it?. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems* (pp. 4517-4520). ACM.

[3] Michman, R. D., Mazze, E. M., & Greco, A. J. (2003). *Lifestyle marketing: reaching the new American consumer*. Greenwood Publishing Group.

[4] Phelan, O., McCarthy, K., & Smyth, B. (2009, October). Using twitter to recommend real-time topical news. In *Proceedings of the third ACM conference on Recommender systems* (pp. 385-388). ACM.

[5] Dao, T. H., Jeong, S. R., & Ahn, H. (2012). A novel recommendation model of location-based advertising: Context-Aware Collaborative Filtering using GA approach. *Expert Systems with Applications*, *39*(3), 3731-3739.

[6] Sakaki, T., Okazaki, M., & Matsuo, Y. (2010, April). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (pp. 851-860). ACM.

[7] Yuan, W., & Liu, Y. (2014). Empirical Analysis of User Life Span in Microblog. In *Advanced Technologies, Embedded and Multimedia for Human-centric Computing* (pp. 855-862). Springer Netherlands.

[8] Amitay, Einat, et al. "Web-a-where: geotagging web content." *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2004.

[9] Fink, C., Piatko, C. D., Mayfield, J., Finin, T., & Martineau, J. (2009). Geolocating Blogs from Their Textual Content. In *AAAI Spring Symposium: Social Semantic Web: Where Web 2.0 Meets Web 3.0* (pp. 25-26).

[10] Zong, W., Wu, D., Sun, A., Lim, E. P., & Goh, D. H. L. (2005, June). On assigning place names to geography related web pages. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*(pp. 354-362). ACM.

[11] Serdyukov, P., Murdock, V., & Van Zwol, R. (2009, July). Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 484-491). ACM.

[12] Cheng, Z., Caverlee, J., & Lee, K. (2010, October). You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*(pp. 759-768). ACM.

[13] Peregrino, F. S., Tomάs, D., & Llopis, F. (2013, November). Every move you make I'll be watching you: geographical focus detection on Twitter. In*Proceedings of the 7th Workshop on Geographic Information Retrieval* (pp. 1-8). ACM.

[14] Yardi, S., & Boyd, D. (2010, May). Tweeting from the Town Square: Measuring Geographic Local Networks. In *ICWSM*.

[15] Backstrom, L., Kleinberg, J., Kumar, R., & Novak, J. (2008, April). Spatial variation in search engine queries. In *Proceedings of the 17th international conference on World Wide Web*(pp. 357-366). ACM.

[16] Backstrom, L., Sun, E., & Marlow, C. (2010, April). Find me if you can: improving geographical prediction with social and spatial proximity. In*Proceedings of the 19th international conference on World wide web* (pp. 61-70). ACM.

[17] McGee, J., Caverlee, J., & Cheng, Z. (2013, October). Location prediction in social media based on tie strength. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management* (pp. 459-468). ACM.

[18] Li, R., Wang, S., Deng, H., Wang, R., & Chang, K. C. C. (2012, August). Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1023-1031). ACM.

[19] Mahmud, J., Nichols, J., & Drews, C. (2012, May). Where Is This Tweet From? Inferring Home Locations of Twitter Users. In ICWSM.

[20] Yu, H., Sun, G., & Lv, M. (2012, September). Users sleeping time analysis based on micro-blogging data. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (pp. 964-968). ACM.

[21] Ehrlich, K., & Shami, N. S. (2010, May). Microblogging Inside and Outside the Workplace. In *ICWSM*.

[22] Baidu LBS Cloud APIs, http://lbsyun.baidu.com/

[23] Chih-Hao Tssi. (1996, April). *MMSEG: A Word Identification System for Mandarin Chinese Text Based on Two Variants of the Maximum Matching Algorithm*

[24] Chen, K. J., & Liu, S. H. (1992). Word identification for Mandarin Chinese sentences. *Proceedings of the Fifteenth International Conference on Computational Linguistics*.