Geographical and Temporal Similarity Measurement in Location-based Social Networks

Zhengwu Yuan Yanli Jiang
College of Computer Science and Technology
Chongqing University of Posts and Telecommunications
Chongqing 400065, China
yuanzw@cqupt.edu.cn, annie_jyl@126.com

Győző Gidófalvi
Department of Urban Planning and Environment
KTH Royal Institute of Technology
Stockholm, SE-100 44, Sweden
gyozo.gidofalvi@abe.kth.se

ABSTRACT
Using "check-in" data gathered from location-based social networks, this paper proposes to measure the similarity of users by considering the geographical and the temporal aspect of their geographical and temporal aspects of their "check-ins". Temporal neighborhood is added to support the time dimension on the basis of the traditional DBSCAN clustering algorithm, which determines the similarity among users at different scales using the classical Vector Space Model (VSM) with vectors composed of the amount of visits in different cluster area. The spatio-temporal similarity of the user behaviors are obtained through overlapping the different weighted user similarity values. The experimental results show that the proposed approach is effective in measuring user similarity in location-based social networks.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval Communications Networks**]: Clustering

General Terms
Algorithms

Keywords
Location-based Social Networks; User Similarity; Cluster; Temporal Scale

1. INTRODUCTION
With the wide use of intelligent terminal and the rapid development of Web2.0 technology, there is a large number of Location-based Social Networks (LBSN), e.g., Foursquare1, Gowalla2, Facebook3. In such LBSNs friends are three times more likely to share their locations with each other than non-friends and there is more check-in similarity among friends than non-friends. The study of user similarity in LBSN for excavated user behavior trajectories from the massive data and provided recommendation service has a very high reference value.

MobiGIS'13, November 05-08 2013, Orlando, FL, USA

http://dx.doi.org/10.1145/2534190.2534192.

In recent years, locations-based social networks have received a lot of attention from many researchers [13][11][17][10]. Ye et al. [13] considered the social and geographical characteristics of users to measure the user similarity, and provided recommendation service for users. Sang et al. [11] focused on excavating the check-in data that mobile users have generated, and measuring the user similarity by a probability based on Markov model. The points of interest location are sorted, and the Top-k locations are recommended to mobile users. Zhou et al. [17] used Jaccard coefficient and Pearson correlation as the similarity measures for the binary check-in utilization and the FIF check-in utilization, respectively. McKenzie et al. [10] proposed topic modeling to exploit sparse, unstructured data, e.g. tips and reviews, as an additional feature to compute user similarity.

Although the above methods have utilized the geospatial information of check-in data in LBSN and social relationship to compute user similarity, these researches only considered the social relationship and geographical property. However, a key property of check-in data of the LBSN-time information may not be considered enough. So, in this paper we aim to study the possibility of measuring the user similarity based on the geographical and temporal properties of check-in data from LBSN.

The rest of this paper is organized as follows. In Section 1, some related works on the user similarity in LBSN are introduced. In Section 2, the details about DBSCAN cluster algorithm are described. In Section 3, the method of user similarity is described in details. Section 4 presents the experiments and discusses of the results. Finally, the conclusion and future works are given in Section 5.

2. RELATED WORKS
2.1 User Similarity Measuring Methods
The Location-Based Social Networks have been prevalent on the Internet and have become a hot research topic that attracts many professionals from a variety of fields. The spatial dimension of location in LBSN helps bridge the gap between the physical world and online social networking services. A location-based social network does not only mean adding a location to an existing social network so that people in the social structure can share location-embedded information, but also consists of the new social structure made up of individuals connected by the interdependency derived from their locations in the physical world. The interdependency includes not only that two persons co-occur in the same physical location or share similar location histories but also the knowledge, e.g., common interests, behavior, and

[1]www. foursquare.com
[2]www. gowalla.com
[3]www. Facebook.com

activities, inferred from an individual's location (history) and location-tagged data [14][15].

With rapid development, the LBSN has attracted more than two billion users in the world, who in turn have generated hundreds of millions of check-in data in the process of using LBSN. There are many traditional methods to measure the user similarity from the immense amounts of check-in data in LBSN, there are many traditional methods, such as Cosine Similarity, Adjusted Cosine Similarity and Pearson Correlation Coefficient [1].

(1) Cosine Similarity

$$sim(A,B) = \cos(U_A, U_B) = \frac{U_A \bullet U_B}{\|U_A\| \|U_B\|}$$

$$sim(A,B) = \frac{\sum_{i \in I_{AB}} R_{A,i} \bullet R_{B,i}}{\sqrt{\sum_{i \in I_{AB}} R_{A,i}^2} \sqrt{\sum_{i \in I_{AB}} R_{B,i}^2}}$$

where $I_{AB} = \{ i \in I \,|\, R_{A,i} \neq \varnothing \text{and } R_{B,i} \neq \varnothing \}$ (I is all item space) is a set of user A, B commonly rated items. The vectors U_A, U_B indicate the user A, B rating on I_{AB}, respectively. $R_{A,i}$, $R_{B,i}$ indicate the users A, B rating on item i.

(2) Adjusted Cosine Similarity

$$sim(A,B) = \frac{\sum_{i \in I_{AB}} (R_{A,i} - \overline{R_A})(R_{B,i} - \overline{R_B})}{\sqrt{\sum_{i \in I_A} (R_{A,i} - \overline{R_A})^2} \sqrt{\sum_{i \in I_B} (R_{B,i} - \overline{R_B})^2}}$$

where $I_{AB} = \{ i \in I \,|\, R_{A,i} \neq \varnothing \text{and } R_{B,i} \neq \varnothing \}$ (I represents all items space) is a set of the user A , B commonly rated items. $R_{A,i}$, $R_{B,i}$ indicate the user A, B rating on item i. I_A, I_B means the rated item sets of users A, B respectively. $\overline{R_A}$, $\overline{R_B}$ represent the average score of all items that user A, B has rated.

(3) Pearson Correlation Coefficient

$$sim(A,B) = \frac{\sum_{i \in I_{AB}} (R_{A,i} - \overline{R_A})(R_{B,i} - \overline{R_B})}{\sqrt{\sum_{i \in I_{AB}} (R_{A,i} - \overline{R_A})^2} \sqrt{\sum_{i \in I_{AB}} (R_{B,i} - \overline{R_B})^2}}$$

where $I_{AB} = \{ i \in I \,|\, R_{A,i} \neq \varnothing \text{and } R_{B,i} \neq \varnothing \}$ (I respected all item space) is a set of the user A,B commonly rated items. $R_{A,i}$, $R_{B,i}$ indicate the users A, B rating on item i. $\overline{R_A}$, $\overline{R_B}$ is the average score of all item that user A, B has rated.

2.2 Analysis of User Similarity Methods

From the Section 2.1 it can be learned that the traditional methods are mostly based on users' rating on items for calculating the user similarity. Due to the increasing availability of location-based services and GPS-enabled devices, more and more user locations and activities have been collected and used in several studies [2][7][8]. However, there is no users' rating on items to be used in the user similarity methods (such as Cosine Similarity, Adjusted Cosine Similarity and Pearson Correlation Coefficient) but only check-in data in LBSN. A variety of approaches for measuring the user similarity from LBSN data have been proposed, such as Candillier et al. [5] studied the user similarity by using Jaccard's similarity coefficient with the traditional similarity methods. Li et al. [9] calculated the similarity of user activity trajectories using GPS log, and firstly identified the geographical location of the user access, then clustered these locations. Next, the user similarity is calculated by matching the geographical location sequence clustered.

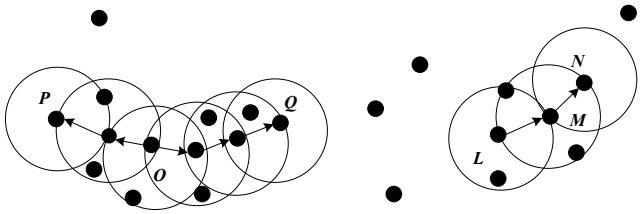


Figure 1. An example of the DBSCAN algorithm

The user behaviors have been formulated as a spatio-temporal similarity from LBSN in this paper. A lot of previous studies on the spatial or temporal aspects based on the check-in behavior of users have been existed. In [16], temporal aspect of the check-in data has been used in feature extraction of places to develop a semantic annotation technique for location-based social networks to automatically annotate all places with category tags. In [3], a location-based and preference-aware recommender system based on the users' location category information has been presented. In [12], an approach of user similarity measurement with a semantic location history has been proposed. Birant et al. [4] presented a new density-based clustering algorithm, ST-DBSCAN, which is based on DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [6], and has the ability of discovering clusters according to non-spatial, spatial and temporal values of the objects.

3. BASIC CONCEPTION

DBSCAN is a density based clustering algorithm that can efficiently discover clusters of arbitrary shape and can effectively handle noise. The main idea is: (1) DBSCAN is growing a cluster from each point which contains at least a minimum number of other points (*MinPts*) with a given radius (*Eps*). (2) When the density of the adjacent region is more than the minimum number of threshold, the clustering is continued. (3) The aim of DBSCAN algorithm is to divide mass raw data into separate groups (clusters) with enough high density.

Some key definitions of DBSCAN are as follows:

- Directly density-reachable (DDR): q is DDR from p if $p \in N_{Eps(q)}$ and $|N_{Eps(q)}| \geq Minpts$.
- Density-reachable (DR): if there is a chain of points $\{p_i \,|\, i = 1, \cdots, n\}$, $p_1=q$, $p_n=p$, and each p_i is DDR from p_{i+1}, then p is DR from q.
- Density-connected (DC): if o is DR from p and o is DR from q, then p is DC from q.

As shown in Figure 1, an example of DBSCAN algorithm is given. Each point in the graph represents the geographical position. It is assumed that *Eps* has been given and *MinPts*=3. It can be learned from the above definitions:

a) L is directly density-reachable to M, M is DDR to N;

b) L is density-reachable to N; N is not density-reachable to L; O is density-reachable to P, Q;

c) P is density-connected from Q.

4. METHODS BASED ON SPATIAL AND TEMPORAL PROPERTIS

When the users' temporal and geographical check-in data are similar, this indicates that both of users are of similarity in some sense. In the history of users' check-in data, if two users share their position in the same city during the day, these two users have some similarities. If two users access the same museum in the morning, it can be considered that the two users have higher similarity in the behavior trace. Moreover, if the two users always check-in in the morning at the same museum, they have even a higher similarity than the occasional users. Therefore, the user's geographical and temporal similarity is based on the following principles:

(1) The closer is the time and the geographical location that users access the more similar are the users to each other;

(2) The larger is the number of check-ins of two users in nearby locations at similar times, the more similar the two users are to each other.

4.1 A Hierarchy Clustering Based On Space and Time

The traditional DBSCAN algorithm has only two inputs, neighbor radius (*Eps*) and minimum threshold (*MinPts*). A method of hierarchical density clustering based on the scale of spatial and temporal is proposed in this paper. This method requires three inputs, including spatial scale (*Eps_space*), temporal scale (*Eps_time*) and the minimum number of POI (*MinPts*). The POI (Point Of Interest) refers to each location that is visited by users in LBSNs. Around a location there may be many different POIs that generate different check-in POI in the same location. So, the calculation of the similarity only depends on whether the user visiting the same POI is not accurate. For this reason, the new hierarchical cluster approach based on the different spatial and temporal scale is used to measure the user similarity. It is required that both spatial and temporal cluster scale meet the minimum number of POI simultaneously, and then the POI continually collects other POI as the cluster grows. In the whole process of clustering, the cluster scale is changed gradually from large to small, and allowed to obtain similar cluster at different hierarchy. As a result, it can be applied in LBS platform because the number of hierarchy is determined by the wealth of data resources. It is obvious that such a result with the new approach is more stable and more accurate than the single clustering.

4.2 User Similarity Measurement

The Vector Space Model (VSM) is used to calculate the user similarity in this paper. The clustering region of each user check-in is considered as vector. In order to reflect the user's repeated access to the same cluster region and the similar user behavior the number of visiting users' the cluster region is defined. In spatial and temporal dimension clustering region of all users consists of the users' check-in location Matrix within a certain period of time. The cosine similarity approach is used to calculate the similarity between users.

Definition 1: We define the user-location matrix within a certain period as

$$V_{l(m\times n)}=\begin{bmatrix} V_{1,1} & V_{1,2} & \cdots & V_{1,n-1} & V_{1,n} \\ V_{2,1} & V_{2,2} & \cdots & V_{2,n-1} & V_{2,n} \\ \vdots & \vdots & & \vdots & \vdots \\ V_{m,1} & V_{m,2} & \cdots & V_{m,n-1} & V_{m,n} \end{bmatrix}$$

where m is the number of visiting users, n is the number of clusters discovered by *Eps_space* and *Eps_time* in the improved DBSCAN method. V_{ij} is the number of check-ins by the user i in the clustering region j. l is the level l after partition clustering hierarchy.

The check-in location can be considered as a vector in an n-dimensional space. The cosine angle between two vectors is used to measure the user similarity. Suppose that the user A and the user B are represented as vector U_A and U_B in n-dimensional check-in location space, the similarity between user A and the user B is defined as follows:

$$sim(A,B)=\cos(U_A,U_B)=\frac{U_A \bullet U_B}{\|U_A\|\|U_B\|}$$

The points of interest locations in the check-in data are hierarchically divided at varying levels into density based clusters. The overall similarity of users must be calculated essentially under different hierarchies of clustering regions. The similarity of across cluster hierarchy levels is measured as follow:

$$sim_{overall}=\sum_{i=1}^{N}\mu sim_i$$

$$\mu=\beta_i \Big/ (\sum_{i=1}^{N}\beta_i)$$

where N is the total number of levels, and sim_i is the similarity in the level i, β_i is the weigh at the level i weights of clustering hierarchy levels increase as the level of detail of the clustering increases, giving more importance to similarity at the detailed levels in the overall similarity score.

5. EXPERIMENT

5.1 Dataset

The experiments are based on the check-in datasets from Gowalla which is a large location-based social network. The experiments dataset is provided by Stanford Network Analysis Project from three US cities, such as Austin, New York, San Francisco. The real-world check-in dataset consist of 7219 users, 16326 locations. All users in the three cities generate a total of 291,161 check-in records. Table 1 provides description of the check-in data in the three US cities.

Table 1. The description of check-in data in three US cities

region	number of users	Number of locations	Number of check-ins
Austin	6644	5932	160524
New York	3068	3802	22662
San Francisco	5276	6033	96083

5.2 Parameter Tuning

The method of DBSCAN cluster based on spatial and temporal dimensions requires three inputs, *Eps_space*, *Eps_time* and *MinPts*. Because DBSCAN cluster algorithm has sensitive parameters, the large *MinPts* will cause the low density cluster. Furthermore, the too low number of *MinPts* can lead to high density of clusters by low density of adjacent clusters containing. So, the selection of threshold for the three inputs determines the results of clustering. In the experiment, the average POI density of the three cities (cities Austin, New York, San Francisco) is calculated. Table 2 shows the four hierarchical clustering parameters. The geographic area is grown with n^2 by clustering. The minimum number of POI grows up with the corresponding proportion. The weight of different hierarchy (β_i) is 2^i. Subsequent experiments can effectively, reasonably find similar users, which means that the clustering parameter of DBSCN hierarchy is reasonable. Table 2 shows the thresholds division.

Table 2. Thresholds division

Level	*Eps_space*/km	*Eps_time*/hour	*MinPts*
4	4	3	40
3	8	6	160
2	16	9	640
1	32	12	2560

5.3 Experiments and Results

The experiment calculates the similarity between 500 users and finds out the most similar user for each user. Moreover, we compare the distribution of POI for each pair of users that have the most similarity. We observe that the more common check-in locations a user pair generates, the higher is the accuracy of the algorithm. Therefore, this paper is assumed that the number of users that check-in the geographic proximity is similar, and compared the *Top-N* locations that users have visited.

The performances of the user similarity are evaluated by two evaluation metrics, namely, *Precision* and *Recall*, which are defined as follow.

(1) The *Precision* of the level *i* is defined as:

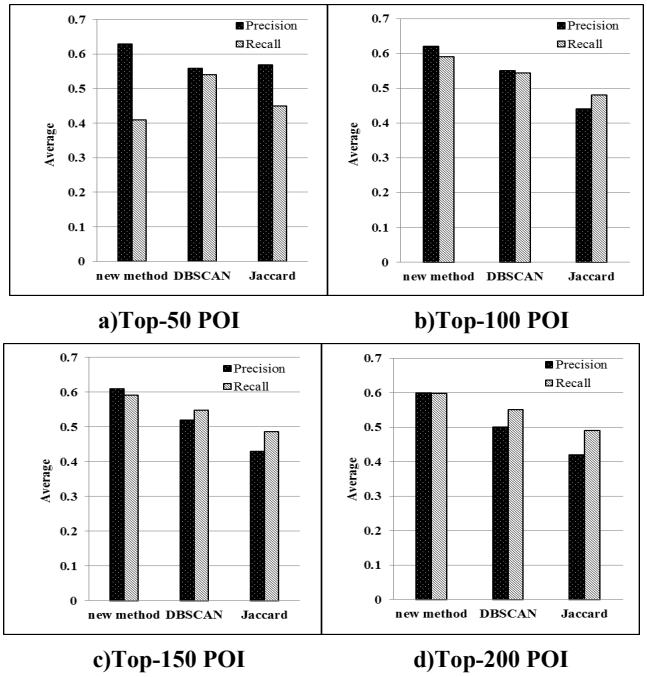
$$Precision = \frac{|TopN(u) \cap TopN(u_r)|}{|TopN(u_r)|}$$

The *Precision* in overall hierarchy is defined as:

$$Precision_{overall} = \sum_{i=1}^{H} \mu Precision_i$$

$$\mu = \beta_i \Big/ (\sum_{i=1}^{H} \beta_i)$$

where u_r is the user that is most similar with the user u. *TopN(u)* is the set of cluster regions contained top-N locations that possess maximum number of check-in data. H is the number of levels in the hierarchy, and β_i is the weigh at the level i.



a)Top-50 POI **b)Top-100 POI**

c)Top-150 POI **d)Top-200 POI**

Figure 2. ***Precision*** **and** ***Recall***

(2) Similarly, a formula for *Recall* is given:

The *Recall* of the level *i* is defined as:

$$Recall = \frac{|TopN(u) \cap TopN(u_r)|}{|TopN(u)|}$$

The *Recall* in overall hierarchy is defined as:

$$Recall_{overall} = \sum_{i}^{H} \mu Recall_i$$

$$\mu = \beta_i \Big/ (\sum_{i=1}^{H} \beta_i)$$

In the experiment the method proposed in this paper is evaluated using the average Precision and Recall. From Figure.2, we observe that the method proposed in this paper has higher precision than the method that only judge the user similarity at the highest hierarchical clustering in DBSCAN. It shows that the similarity measurement method in different spatial and temporal scale has more accuracy than the fine-grained methods. Compared with the Jaccard coefficient method the proposed method has higher *Precision* and *Recall*. Because the Jaccard coefficient considers only the similarity of user's check-ins location, the number of the location could not be calculated.

Because TopN(u) is the set of cluster regions contained N locations that possess maximum number of check-in data. The N in Top-N refers to the locations that possess maximum number of check-in data. Involved in the Recall and Precision calculation is a collection of cluster regions contained Top-N locations. With the increase of the N, a set of cluster region does not necessarily increase. So, in theory, the Recall and Precision have no fixed tendency.

6. CONCLUSIONS AND FUTURE WORK

The new method is proposed to calculate the user similarity on LBSN based on the spatial and temporal properties of the user check-in data. Through increasing the temporal scale in the traditional DBSCAN algorithm the check-in data is clustered on the spatial and temporal scale. The spatio-temporal distribution characteristics of the check-in data in the Gowalla are analyzed. The vector space model and the cosine similarity computing method are introduced. The new geographical and temporal similarity measurement method on location-based social networks is proposed. The evaluation results show that the proposed method is effective. The method can be applied to recommend location or friends in LBSN, because the key of a recommendation system is the similarity measurement of user or item.

In the future, the social relationships, the geographical relationships and the temporal relationships will be integrated into research of the similarity in LBSN. More empirical study will be conducted in terms of location-based social networks.

7. ACKNOWLEDGMENTS

This paper is supported by the National Natural Science Foundation of China (Project No. 61075019), the Natural Science Foundation of Chongqing (Project No. cstc2012jjA40064) and the China Scholarship Council.

8. REFERENCES

[1] Adomavicius, G. and Tuzhilin, A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering*. IEEE, 2005, 17(6): 734-749.

[2] Ashbrook, D. and Starner, T. Using gps to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 2003, 7(5): 275-286.

[3] Bao, J., Zheng, Y. and Mokbel, M. Location-based and preference-aware recommendation using sparse geo-social networking data. *In Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, 2012, 199-208.

[4] Birant, D., Kut, A. ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data & Knowledge Engineering*, 2007, 60(1): 208−221.

[5] Candillier, L., Meyer, F., Fessant, F. Designing specific weighted similarity measures to improve collaborative filtering systems. In *ICDM*, 2008: 242-255.

[6] Ester, M., Kriegel, H-P., Sander, J., Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*. 1996, 96: 226-231.

[7] Liao, L., Patterson, D., Fox, D., Kautz, H. Learning and inferring transportation routines. *Artificial Intelligence*, 2007, 171(5-6): 311-331.

[8] Lin, J., Xiang, G., Hong, J., Sadeh, N. Modeling people's place naming preferences in location sharing. *In ACM International Conference on Ubiquitous Computing (UbiC omp)*, 2010, 75-84.

[9] Li, Q., Zheng, Y., Xie X, Chen Y., Liu, W., Ma, W. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*. ACM, New York, NY, 2008: 34.

[10] McKenzie, G., Adams, B., and Janowicz, K. A thematic approach to user similarity built on geosocial check-ins. *Geographic Information Science at the Heart of Europe*. 2013: 39-53.

[11] Sang, J., Mei, T., Sun, J., Xu, C., Li, S. Probabilistic sequential POIs recommendation via check-in data. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*. ACM, 2010: 402-405.

[12] Xiao, X., Zheng, Y., Luo, Q., Xing, X. Finding similar users using category-based location history. *In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2010, 442-445.

[13] Ye, M., Yin P., and Lee, W-C. Location recommendation for location-based social networks. In *Proceedings of the 18th SIGSPATIAL International Conferrnce on Advances in Geographic Information Sydtems*. *ACM*, 2012: 458-461.

[14] Yu, Z. Location-based social networks: users. *Computing with Spatial Trajectories*. 2011: 243-276.

[15] Yu, Z. Tutorial on location-based social networks. *WWW2012*. 2012.

[16] Ye, M., Shou, D., Lee, W-C., Yin, P. On the semantic annotation of places in location-based social networks. *In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, 520-528.

[17] Zhou, D., Wang, B., Rahimi, S., Wang, X. A study of recommending locations on location-based social network by collaborative filtering. In *Proceedings of the 25th Canadian Conference on Artificial Intelligence*. *Canadian AI*, 2012: 255-266.