Similarity of Mobile Users Based on Sparse Location History

Pasi Fränti(✉), Radu Mariescu-Istodor, and Karol Waga

School of Computing, University of Eastern Finland, Joensuu, Finland
pasi.franti@uef.fi

Abstract. We propose a method to measure similarity of users based on their sparse location history such as geo-tagged photos or check-in activity of user. The method is useful when complete movement trajectories are not available. We map each activity point into the nearest location in a predefined set of fixed places. The problem is then formulated as histogram comparison. We compare the performance of similarity measures such as L_1, L_2, L_∞, ChiSquared, Bhattacharyya and Kullback and Leibler divergence using both crisp and fuzzy histograms. Results show that user can be recognized with fair accuracy, and that all similarity measures are suitable except L_2 and L_∞, which perform poorly.

Keywords: User similarity · Mobile activity · GPS data analysis
Histogram matching · Fuzzy pattern recognition · Location-based services

1 Introduction

Similarity of users has been widely used in recommender systems based on the assumption that similar users are interested in similar things. Collaborative recommender systems [1] estimate relevance of an item to a given user based on ratings given by similar users. To find similar users most of the recommendation systems searches for the common items that the users have rated [1]. This approach is used in online shops, movie databases and similar recommendation systems.

The knowledge of similar users has been applied to improve retail experience by finding correlation between buying and browsing behavior. Similarity of users has also been used for recommending events and friends in [9], and provide good initial guess for personalizing the recommendations for new users [21]. Similarity of users have also been measured based on how they tag for bookmarking purpose [13].

In [11], we studied whether social network can be used for improving location-aware recommendations. The results showed that user's own understanding of the similarity correlates more with the similarity of their page likings and less with the similarity of their location histories, for which only minor correlation was detected. The same order of importance was observed in [12]: people value most the things they have common, then the places where they are active, and least important was to know the same people.

L. Rutkowski et al. (Eds.): ICAISC 2018, LNAI 10841, pp. 593–603, 2018.
https://doi.org/10.1007/978-3-319-91253-0_55

In location-aware recommendations, however, opinions of local experts in the given area can be more valuable than just the similarity of the user [2]. This can be useful for improving rating of the services by utilizing users whose opinions matter most. In general, knowing the similarity of location history can provide additional information for improving recommendation. In this paper, we study how the similarity of users can be measured from a limited amount of location data. Sample location histories of three users are shown in Fig. 1.

One approach is to analyze complete trajectories of the user movements. Several similarity measures were compared in [16] based on the complete trajectories. In [21], potential friends are recommended based on users' movement trajectories. So-called *stay cells* are also created based on detected stops, which are considered important places because user stayed there a longer time. Similarity of trajectories are then measured based on their longest common subsequence giving higher weight to longer patterns. In [15], a revised version of the longest common subsequence is applied by partitioning the trajectories based on speed and detected turn points. The similarity score is based on both geographic similarity and the semantic similarity. In [8], similarity of the location and their temporal semantics were also taken into account.

In [4], similarity of a person's days is assessed based on the trajectory by discovering their semantic meaning. The data is collected from tracking users' cars and pre-processed by detecting stop points. Most common pairs of stops are assumed to be user's home and work locations. Dynamic time warping of the raw trajectories using geographic distances of the points is reported to work best. In [20], personalized search for similar trajectories is performed by taking into account user preferences of which parts of the query trajectory is more important.

Complete trajectories are not always available and the similarity must then be measured based on sparse location data such as visits, favorite places or check-ins. In [14], user data is hierarchically clustered into geographic regions. A graph is constructed from the clustered locations so that a node is a region user has visited, and an edge between two nodes represents the order of the visits to these regions. This method still relies on the visiting order.

In this paper, we study how the similarity of users can be measured based on their location history when the entire GPS trajectories are not available. We use single location points originating from geo-tagged photos, and the start and end points of movements. Other activity points that could be used are stay points, i.e. the places where the user stayed longer than 30 min, as in [23].

We propose to measure the similarity between users by taking each user activity as an observation into a histogram that represents places in the region. The problem then reduces to comparison of the two discrete distributions. We consider several measures based on normalized frequency vectors: L_1, L_2, L_∞, ChiSquared, Bhattacharyya and Kullback and Leibler divergence. Fuzzy histograms are also considered.

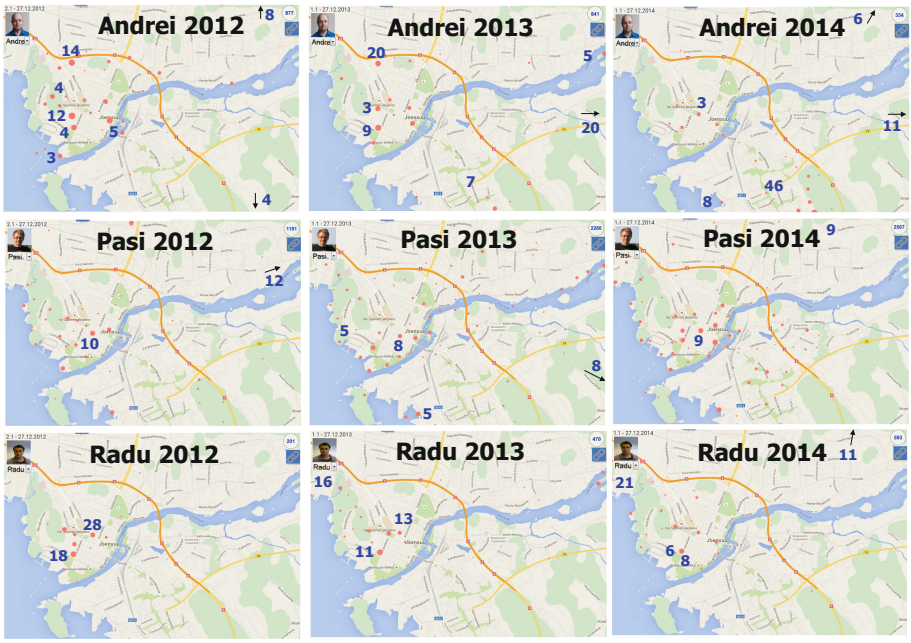

Fig. 1. Activity data of three users during 2012–2014.

2 Sparse Location Histograms

Mopsi (http://cs.uef.fi/mopsi) is a prototype media sharing platform in which users can collect geo-tagged photos, track their routes, and recommend places of interest [19] to other users by upgrading them to *services*. These services include hotels, restaurants, cafeterias and many others that the users consider relevant to themselves and to others. The locations of these services serve as the histogram bins, which we denote as *places*. User similarities are calculated based on the locations where the users have collected data. We call these as *activity points*.

2.1 Locations, Distance and Places

Calculating similarity of two users starts by constructing histograms of the users. We process the activity points of a user by mapping them to their nearest place. Every activity point adds the count of the corresponding histogram bin by one. Distance calculation is based on *haversine* distance of the two locations given as latitude (φ) and longitude (λ) coordinates. The formula to calculate the haversine distance (in kilometers) is defined as:

$$hav = 2 \cdot R \cdot \arcsin(\psi) \tag{1}$$

$$\psi = \sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos\left(\varphi_1\right)\cos\left(\varphi_2\right)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)} \tag{2}$$

where $R = 6372.8$ km, φ_1 and φ_2 are the latitudes, and λ_1 and λ_2 are the longitudes of the two points. User has n activity points mapped into m histogram bins $h(i)$ so that:

$$\sum_{i=1}^{m} h(i) = n \tag{3}$$

The histograms are normalized as follows:

$$p(i) = \frac{h(i)}{\sum_{j=1}^{m} h(j)} \forall i \in [1, m] \tag{4}$$

where $p(i)$ represent the probability that an activity point belongs to the bin i. Example of the histogram construction is shown in Fig. 2, where three users have $n_1 = 9$, $n_2 = 7$ and $n_3 = 7$ activity points (small icons on map). They are mapped to $m = 8$ places (hot spots) shown as the thumbnail images. The values of the bins (visit frequencies) are shown below each place.

Fig. 2. Converting location history of three users to histogram of $m = 8$ bins.

The choice of using existing places is just one possibility. Another alternative might be to cluster all activity data, and in this way, automatically determine the hot spot places. Semantics of the places are also not considered. For example, McDondald's and its local competitor, Hesburger, are considered different places. We are only interested whether the user can be distinguished based on the location only without considering what is there.

2.2 Histogram Matching

The problem is to calculate similarity (or distance) between two distributions, represented as histograms. The histogram is usually a one-dimensional array consisting of numerical values, for example, pixel intensities of an image. In this case, there is an explicit ordering of the bins, and the values of the neighboring bins highly correlate with each other. Mapping an observation into the histogram is straightforward.

The bin values can also be nominal, or multivariate as in our case, so that there does not exists any natural ordering of the bins. However, since the observations appear in a metric space, the observations (activity points) can still be mapped to the histogram by simple distance calculations. The problem therefore reduces to histogram comparison [5, 6].

Figure 3 demonstrates the process using *Bhattacharyya coefficient* originally proposed as a similarity measure between statistical populations. First, product $p_i \cdot q_i$ of two frequencies are calculated, and their square roots are then summed over the all histogram bins. The higher the frequencies, the higher the product. The result is converted to a distance in range [0, 1] by applying logarithmic scaling. This provides natural bounds on the Bayes misclassification probability.

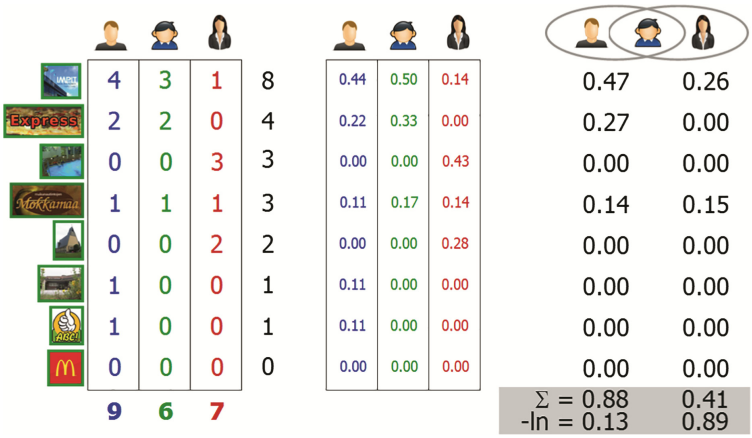

Fig. 3. Distance calculations of two histograms using Bhattacharyya distance.

We also consider various L-norms such as L_1, L_2 and L_∞, and classical Chi Squared distance. Kullback and Leibler distance [22] generalizes Shannon's concept of probabilistic uncertainty called entropy by calculating minimum cross entropy of two probability distributions [15]. These and the Bhattacharyya coefficient are defined below.

$$L_1 = \sum_i |p_i - q_i| \tag{5}$$

$$L_2 = \sum_i (p_i - q_i)^2 \tag{6}$$

$$L_\infty = \max_{1 \le i \le n} |p_i - q_i| \tag{7}$$

$$d_{ChiSq} = \sum_i \frac{(p_i - q_i)^2}{p_i + q_i} \tag{8}$$

$$d_{KLD} = \sum_i \left(p_i \cdot \log \frac{p_i}{q_i} + q_i \cdot \log \frac{q_i}{p_i} \right) \tag{9}$$

$$S_{BC} = \sum_i \sqrt{p_i \cdot q_i} \tag{10}$$

where p_i and q_i are the relative frequencies of the histogram bins i, and the summation is done over all m places.

All the above techniques calculate the distance of each bin independently. In case of sparse observations, it may happen that strongly peaked histograms would become mismatched due to slight translation. So-called *earth mover distance* (EMD) [17] aims at solving this by transforming surplus from one bin to the bins that have deficit. In case of one-dimensional numeric data this is straightforward to calculate by processing the histogram sequentially from left to right. However, in multivariate case the optimal moving of the surplus becomes more complicated problem. It was noted in [7] that the problem could be solved as transportation problem but faster algorithms would be needed.

In [18], the peaks of the histograms were considered as more important. Improved performance of L_1, L_2 and EMD was demonstrated in case of time-series analysis by calculating the sum of the peak weights multiplied by their closeness factors.

Sparseness of the data may also cause problems when there are too few observations compared to the number of available bins. Fuzzy histograms were proposed in [10] motivated by its successful application in image processing field. In case of one-dimensional histograms, observations are divided into several neighboring bins. We generalize the idea into multivariate case by utilizing the *k-nearest neighbors* (kNN) concept as follows.

For each location activity, we find its k nearest places and calculate fuzzy counts similarly as done in the well-known fuzzy C-means algorithm [3]:

$$w_i = \left(\sum_{j=1, j \ne i}^{k} \frac{d(x - h_i)}{d(x - h_j)} \right)^{-1} \tag{11}$$

Otherwise, the histogram comparison can be done in the same way as with the crisp variant.

3 Experiments

We used data from Mopsi that has been collected using native mobile application available in all platforms with the following user distribution: WindowsPhone (55%), Android (28%), iOS (14%) and Symbian (3%). There are 36243 photos and 8963 trajectories, and by 9.5.2015, there have been 909 registered mobile users. Of these, 94 users have collected more than 5 photos and 5 routes. In the following, we focus on a small subset of users whose location activity we are familiar with.

3.1 Data Extraction

For creating the histogram bins, we selected 293 services from Mopsi (http://cs.uef.fi/mopsi). These include cafes, restaurants, holiday resorts, shops, parks and other services within the bounding box that covers the Joensuu sub-region, see Fig. 4. This region covers Joensuu downtown, its suburban, neighboring municipalities and large sparsely inhabited rural areas. We use here only the location of the services. The coverage of the bins is dense in the downtown area but sparse in rural area. For example, one service is the only one within 20 km radius whereas there are about 75 services within the 3×3 blocks (300 m $\times$ 500 m area) around the market place.

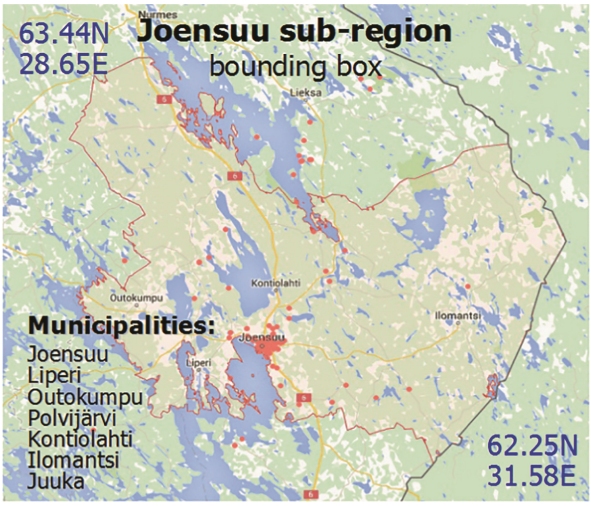

Fig. 4. Distribution of the places (histogram bins).

We then recorded activities of the users from the years 2011–2014 as follows: (1) places where they took photos, (2) places where tracking a route was started or ended. Each activity is counted to the frequency of its nearest service (histogram bin). Our first test consists of three active users (A = Andrei, P = Pasi, R = Radu) called A-P-R trio in the following. These users have 5831 location points in total. The most popular places with the corresponding visit frequencies are listed in Table 2.

The data is divided per year resulting into four subsets for each user, see Table 1. In total, we will have 12 subsets (pseudo users) denoted as: A11, A12, A13, A14, P11, P12, P13, P14, R11, R12, R13, R14. By default, the subsets A11–A14 should be similar to each other, and dissimilar to the other subsets. In practice, the situation is more complicated. The biggest frequency is in the bin corresponding to the location of user's home. However, both Andrei and Radu moved in 2014 causing significant changes in their histograms. All users have the same work place (#8). This and the homes are emphasized in Table 2.

Table 1. Three test users and the summary of their location data.

	2011	2012	2013	2014
Andrei	206	757	432	329
Pasi	1263	545	636	751
Radu	37	292	324	259

Table 2. Ten most popular locations (histogram bins) and the corresponding frequencies.

	Andrei				**Pasi**				**Radu**				
	2011	2012	2013	2014	2011	2012	2013	2014	2011	2012	2013	2014	**Total**
1	20	0	29	**150**	47	7	6	8	1	2	2	3	**275**
2	13	11	87	36	64	17	16	7	0	1	12	2	**266**
3	0	0	0	0	**51**	**54**	**54**	**69**	0	1	0	0	**229**
4	**12**	**107**	**87**	0	0	0	1	2	0	0	2	0	**211**
5	1	0	0	1	34	9	20	11	0	0	52	**54**	**182**
6	6	29	10	3	6	2	1	3	7	54	35	16	**172**
7	7	92	6	0	18	4	7	15	0	13	4	2	**168**
8	**22**	**6**	**5**	**4**	**36**	**9**	**6**	**18**	1	**12**	**11**	**21**	**151**
9	0	3	4	0	0	0	0	1	**12**	**82**	**41**	7	**150**
10	0	0	0	0	73	6	48	13	0	0	0	0	**140**

3.2 Test Setup and Results

We calculate the similarity (or distance) value between all pairs of the 12 subsets. The task is then to decide which of the subsets belong to the same user by thresholding. The expected result is that $3{\cdot}4{\cdot}4 = 48$ pairs (33%) should be recognized as the same user, and $3{\cdot}4{\cdot}8 = 96$ pairs (67%) should fail the test.

For thresholding, we study the effect of different alternatives. First choice (average) is to use the average of all similarity values. This is the simplest non-parametric threshold that attempts to adapt the method to the data. Second choice (apriori) is obtained by selecting the threshold value (for the particular method) that passes 48 pairs (or as close to this as possible) based on a priori information that 33% values should pass the similarity test, and 96 should fail. The last choice (oracle) is the threshold that provides best accuracy for the particular method.

The results in Table 3 show that L_1, Chi^2, BHA and KLD provide good results (8%, 8%, 10%, 11%) using the average as threshold, and only slightly better if the optimal

threshold (Oracle) was known (7%, 7%, 8%, 10%). The two other methods, L_2 and L_∞, perform poorly (35%, 43%). The a priori information does not help, and even if the optimal threshold is known their results are inferior (15%, 18%).

Table 3. Classification accuracy for the APR trio.

	Threshold (crisp)			Accuracy (crisp)			Accuracy (fuzzy)		
	Mean	Apr.	Oracle	Mean	Apr.	Oracle	Mean	Apr.	Oracle
L_1	0.31	0.27	0.28	8%	8%	7%	10%	10%	10%
Chi2	1.22	1.24	1.18	8%	7%	7%	17%	11%	10%
BHA	0.46	0.46	0.48	10%	10%	8%	15%	14%	11%
KLD	0.82	0.89	0.88	11%	11%	10%	36%	21%	15%
L_2	0.84	0.80	0.88	35%	47%	15%	35%	49%	14%
L_∞	0.79	0.72	0.87	43%	43%	18%	38%	47%	21%

Fuzzy histograms were also considered with neighborhood of fixed size $k = 3$. The classification errors systematically increased without clear reason. Especially KLD works significantly worse when using the fuzzy counts than with the crisp counts. Another observation is that the choice of the threshold becomes now critical. Using average as the threshold no long works with most measures.

The BHA and Chi^2 measures made classification error with users A13 and R13. In this case, the users reported to have recorded lots of joint bicycle trips at that year. In the data, this shows as increased counts in the bins representing rural areas.

Analysis of the other classification errors reveals the following details. All methods recognize A14 as different user than A11, A12 and A13. The same happens also between R11 and R14 with all methods except KLD. Both users changed their homes in 2014 causing different histogram bins to become dominant within the user. We therefore hypothesized that the methods might be affected too much by the dominant values. To test this, we performed additional tests by removing the top-10 location values. The changes are summarized in Table 4. The revised experiments show even weaker performance. The changes did not make A14 match with A11–A13. On the contrary, it caused more mismatch results and showed, that the methods somewhat rely on the detection of user's most popular places like work and home.

Table 4. Classification accuracy for the APR trio when 10 most popular bins have been eliminated from the calculations.

Method	All data	Excluding Top-10	Observation
L_1	8%	24%	Loses its ability
Chi^2	8%	13%	A11 becomes similar with P11, P13, P14
BHA	10%	13%	A11 becomes similar with P11, P13, P14 P11–R14 no longer matches
KLD	11%	13%	A11 and R14 become similar, no other effects
L_2	35%	40%	Works slightly worse
L_∞	42%	43%	Works slightly worse

4 Conclusion

Locations of people show their preferences and interests. In this paper, we have performed detailed study how histogram matching can be used to measure similarity of users based on their location history. We have shown that L_1, ChiSquared, Bhattacharyya and Kullback and Leibler divergence all are applicable to the problem. However, L_2 and L_∞ work poorly. Fuzzy histograms also worked worse than the corresponding crisp variant.

As future work, we consider the following open questions. Automatic method for selecting the threshold could be constructed based on expected distribution, as well as optimizing the number of bins. Possible use of double normalization, log-scaling of frequencies, cosine distance, and fuzzy modeling for sparse data can also provide further improvement. The size of test data vs. recognition accuracy should be studied further. Generalization of earth mover distance for the multivariate case is also worth to consider. Finally, histogram-based comparison might be better to replace prototype based comparison using clustering. These are all points of future studies.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. **17**(6), 734–749 (2005)
2. Bao, J., Zheng, Y.H., Mokbel, M.F.: Location-based and preference-aware recommendation using sparse geo-social networking data. In: International Conference on Advances in Geographic Information Systems (SIGSPATIAL), Redondo Beach, CA, pp. 199–208 (2013)
3. Bezdek, J.C., Ehrlich, R., Full, W.: FCM: the fuzzy *c*-means clustering algorithm. Comput. Geosci. **10**(2–3), 191–203 (1984)
4. Biagioni, J., Krumm, J.: Days of our lives: assessing day similarity from location traces. In: Carberry, S., Weibelzahl, S., Micarelli, A., Semeraro, G. (eds.) UMAP 2013. LNCS, vol. 7899, pp. 89–101. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38844-6_8
5. Cha, S.-H.: Comprehensive survey on distance/similarity measures between probability density functions. Int. J. Math. Models Methods Appl. Sci. **4**(1), 300–307 (2007)
6. Cha, S.-H.: Taxonomy of nominal type histogram distance measures. In: American Conference on Applied Mathematics, Harvard, MA, USA, pp. 325–330 (2008)
7. Cha, S.-H., Srihari, S.N.: On measuring the distance between histograms. Pattern Recogn. **29**(13), 1768–1774 (2008)
8. Chen, X., Pang, J., Xue, R.: Constructing and comparing user mobility profiles for location-based services. In: ACM Symposium on Applied Computing, pp. 261–266 (2013)
9. De Pessemier, T., Minnaert, J., Vanhecke, K., Dooms, S., Martens, L.: Social recommendations for events. In: ACM Conference on Recommender Systems, Hong Kong, China (2013)
10. Fober, T., Hullermeier, E.: Similarity measures for protein structures based on fuzzy histogram comparison. In: IEEE International Conference on Fuzzy Systems, Barcelona, pp. 1–7 (2010)
11. Fränti, P., Waga, K., Khurana, C.: Can social network be used for location-aware recommendation? In: International Conference on Web Information Systems & Technologies (WEBIST 2015), Lisbon, Portugal (2015)

12. Guy, I., Jacovi, M., Perer, A., Ronen, I., Uziel, E.: Same places, same things, same people?: Mining user similarity on social media. In: ACM Conference on Computer Supported Cooperative Work, Savannah, GA, USA, pp. 41–50 (2010)
13. Li, X., Guo, L., Zhao, Y.: Tag-based social interest discovery. In: Conference on World Wide Web, Beijing, China, pp. 675–684 (2008a)
14. Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., Ma, W.-Y.: Mining user similarity based on location history. In: ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Paper #34, Irvine, CA, USA (2008b)
15. Liu, H., Schneider, M.: Similarity measurement of moving object trajectories. In: International Workshop on GeoStreaming, pp. 19–22 (2012)
16. Mariescu-Istodor, R., Fränti, P.: Grid-based method for GPS route analysis for retrieval. ACM Trans. Spat. Algorithms Syst. **3**(3), 8:1–8:28 (2017)
17. Rubner, Y., Tomasi, C., Guibas, L.J.: A metric for distributions with applications to image databases. In: IEEE International Conference Computer Vision, pp. 59–66 (1992)
18. Strelkov, V.V.: A new similarity measure for histogram comparison and its application in time series analysis. Pattern Recogn. Lett. **29**(13), 1768–1774 (2008)
19. Waga, K., Tabarcea, A., Fränti, P.: Recommendation of points of interest from user generated data collection. In: IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), Pittsburgh, USA (2012)
20. Wang, H., Liu, K.: User oriented trajectory similarity search. In: International Workshop on Urban Computing, pp. 103–110 (2012)
21. Yang, X., Steck, H., Guo, Y., Liu, Y.: On Top-k recommendation using social networks. In: ACM Conference on Recommender Systems, Dublin, Ireland, pp. 67–74 (2012)
22. Ying, J.J.C., Lu, E.H.C., Lee, W.C., Wen, T.C.M., Tseng, V.S.: Mining user similarity from semantic trajectories. In: International Workshop on Location Based Social Networks, San Jose, CA, USA (2010)
23. Zheng, V.W., Zheng, Y., Xie, X., Yang, Q.: Collaborative location and activity recommendations with GPS history data. In: ACM International Conference on World Wide Web, Raleigh, NC, USA, pp. 1029–1038 (2010)