Journal of Cases on Information Technology
Volume 24 • Issue 3

EMD-Based Semantic User Similarity Using Past Travel Histories

Sunita Tiwari, G. B. Pant Government Engineering College, India
https://orcid.org/0000-0003-1875-7928

Saroj Kaushik, Indian Institute of Technology, Delhi, India

ABSTRACT

The cost-effective and easy availability of handheld mobile devices and ubiquity of location acquisition services such as GPS and GSM networks have helped expedient logging and sharing of location histories of mobile users. This work aims to find semantic user similarity using their past travel histories. Application of the semantic similarity measure can be found in tourism-related recommender systems and information retrieval. The paper presents earth mover's distance (EMD)-based semantic user similarity measure using users' GPS logs. The similarity measure is applied and evaluated on the GPS dataset of 182 users collected from April 2007 to August 2012 by Microsoft's GeoLife project. The proposed similarity measure is compared with conventional similarity measures used in literature such as Jaccard, Dice, and Pearsons' correlation. The percentage improvement of EMD-based approach over existing approaches in terms of average RMSE is 10.70%, and average MAE is 5.73%.

KEYWORDS

Earth Mover Distance, GPS Log Mining, GPS Trajectories, Location Recommender System, Location-Based Tourist Recommender System, Pervasive Computing, Semantic Similarity, User Profile Similarity

INTRODUCTION

Handheld devices are the main platforms for communication as well as for information access in the current era. The handheld devices nowadays are mostly equipped with GPS systems that make it suitable to log users' travel histories in the form of GPS trajectories. These geo-spatial data have opened various opportunities to retrieve highly relevant information, which are useful in multiple location-based applications. Some of such applications include but not limited to identifying popular locations (Tiwari, 2013; Khetarpaul, 2011; Zheng, 2009), finding location correlation (Zheng & Xie, 2010), making location-based products and services recommendations (Tiwari, 2015; Zheng & Zhang, 2009) and many more (Logesh, 2018; Zheng & Xie, 2011).

The GPS trajectories are found to be an important source of information about the user (Zheng & Zhou, 2011). This extracted information can infer the movement pattern of user, retrieve the current context of user, identify the habits and likings of the user, and so on and so forth. Analysis of the trajectory data finds application in traffic planning, itinerary planning, advertising, disaster management, tourist spot recommendations, etc.

These trajectories are widely used in location-based services (LBS). The tourism-related location-based recommender systems are benefitted hugely by the information extracted by trajectories. Most of the currently available mobile tourist recommender systems generally recommend the point of

DOI: 10.4018/JCIT.20220701.oa2

interest (POI) in the user's vicinity. Such recommender systems are commonly based on a collaborative filtering approach. In any collaborative filtering recommender system, the user similarity plays a vital role (Herlocker, 1999). The prevailing tourist recommender systems use user profiles to identify similar users and do not consider their travel similarities. Some of those applications that uses travel histories only consider the spatial aspects of trajectories but not the semantic aspects. The availability of a semantic similarity measure based on past travel histories may further increase recommender systems application performance.

A recommender system helps the data scientists succeed in making good decision and similarity measure is one of the basic requirements of any recommender system. The proposed research work presents a novel EMD based semantic similarity measure using the past travel histories of users in form of GPS logs. Initially, the stay sequences for an individual user are computed from there GPS trajectories. The additional semantic tags are associated with every point in the obtained stay sequences. Finally, the similarity score of the semantic trajectories are computed using the EMD.

The rest of the paper is organized in sections named as related work, methodology, experimental results, discussions and conclusions, and future directions.

RELATED WORK

GPS trajectory mining has attracted researchers in the past decade due to the widespread application of GPS trajectories in location-based services. A significant contribution is made by several authors in this area which includes trajectory similarity (Sankararaman, 2013; Zheng & Xing, 2010), trajectory anonymization (Nergiz, 2008), information retrieval (Zheng & Ye, 2012), location-based recommender system (Tiwari, 2015; Zheng & Zhang, 2011; Zheng & Xing, 2011), user activity recognition & analysis (Bui, 2008; Liao, 2006), user activity prediction (Imai, 2020) and location correlation (Zheng & Xie, May 2010) etc. This section focuses on the related work in the field of user similarity based on travel histories.

Authors in (Abraham, 2012) has discussed an application of a spatiotemporal similarity measure with a given point of interest (POI) and time of interest (TOI). They explained the concept of trajectory similarity of moving vehicles along with the procedures used in processing information of similar vehicle trajectories which are changing dynamically over time. A hierarchical graph similarity measure is proposed and discovery of important patterns from the geographical data is discussed by authors in (Li, 2008). Other similarity measures based on past travel histories can be found (Lu, 2009; Li, 2008; Lee, 2007; Herlocker, 1999). These contributions emphasize on the similarity computation between users based on the geographical features. However, authors in (Ying, 2010) states that the GPS logs that are geographically similar may not be essentially similar semantically as the activities performed and experienced earned by the users at nearby locations may be different based on the semantics of that location. On the contrary, the geographically far trajectories may demonstrate a lot of similarity. Therefore, semantic aspects are important to contemplate while computing the user similarity based on the travel histories.

Research contribution in (Bogorny 2009; Alvares, 2007) introduces the idea of semantic trajectories. A semantic trajectory is a sequence that is annotated with the location semantics. The semantic annotation adds one or more relevant tags with every location such as temple, national heritage, monument, archaeological site, shopping, entertainment, amusement park, waterfall, etc. These attached tags augment semantic meaning to the geographical locations. Work in (Liu, 2012; Lu, 2009) discusses the semantic trajectory-based similarity measure.

The semantic trajectories may have thousands of GPS points, and evaluating the similarity based on these trajectories with a lot of points is difficult and computationally intensive. Also, annotating these trajectories is time-consuming. The proposed work emphasizes computing the similarity based on stay trajectories instead of GPS trajectories to reduce the computational cost. A geographical location where the user has spent a notable amount of time is considered as the user's stay point.

The authors in (Xiao, 2010) proposes a user similarity using GPS logs based on stay points and is called a maximum travel match algorithm. Work presented in (Ying 2010) also proposed a measure of similarity known as maximal semantic trajectory pattern similarity, and it uses the longest common sequence (LCS) method. Authors in (Chen, 2013) proposes a mobility profile as traces of places that users frequently visit and use frequent sequence pattern mining techniques to extract them. Another user similarity measure was proposed in (Lv, 2013), and it focuses on mining the routine activity performed by users from GPS logs. The semantic user similarity approach is proposed in (Lee, 2011). A time-series analysis-based user similarity measure is presented in (Ossama,2009).

The work in (Zhao, 2020) presents a location recommender exploiting the sentimental and spatial context. The work presented in (Milton, 2019) identifies the reason of the user's movement. A method for predicting users' destination and searching for similar users of the same region of interest (ROI) is proposed by (Tang, 2019). Authors in (Logesh, 2018) proposes to utilize social network profiles and accurate GPS data for a personalized travel recommender system. An approach to recommend trip using real-life travel sequences based on geotagged photos is proposed in (Lim, 2018).

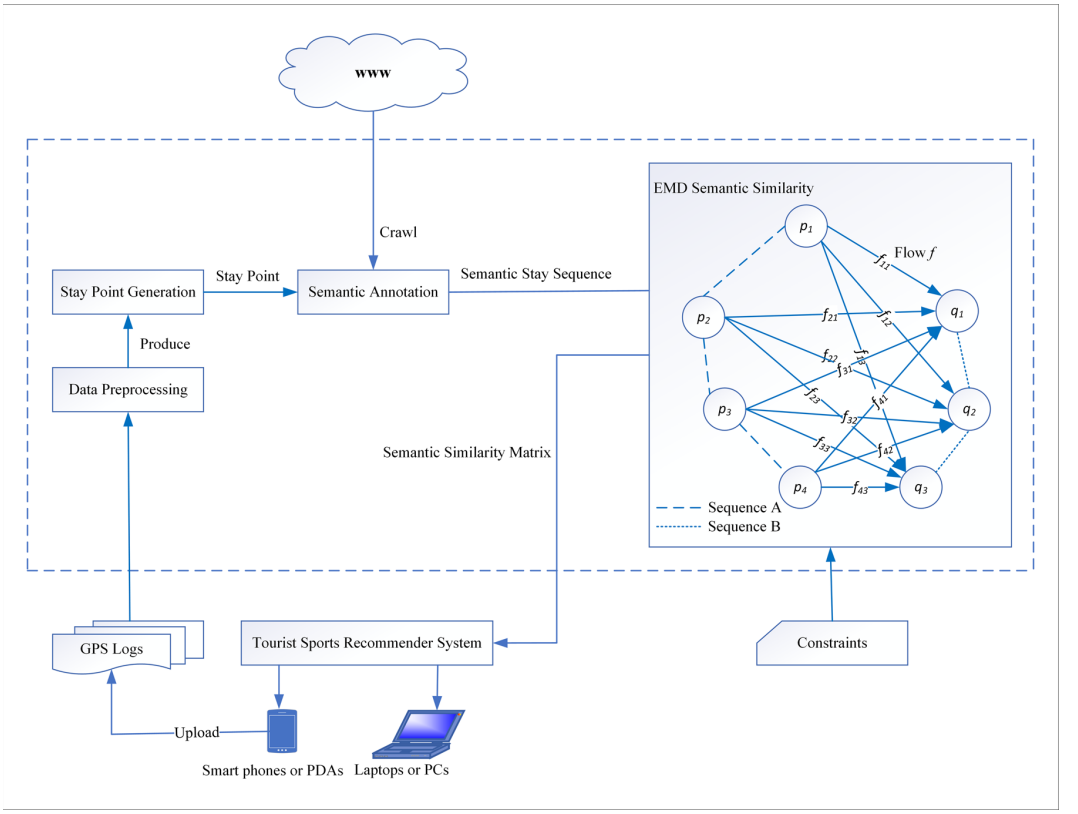
In difference to the approaches discussed above, the proposed work uses a stay point sequence-based method using EMD to evaluate the similarity between users. The semantic similarity problem is formulated as a transportation problem. It requires calculating the amount of work needed to transport one sequence of stay points into another sequence under consideration. EMD was first introduced in the year 2000 by Rubner et al. (Rubner, 2000), and since then, it is extensively used in image retrieval and document matching, etc. In general, the EMD evaluates the dissimilarity between two multidimensional probability distributions. For computing the many-to-many mapping between two semantic stay sub-sequences, EMD is useful. Additionally, the proposed approach permits partial matching and includes perceptual similarity in a simplistic manner. To the best of the knowledge of authors of this work, semantic user similarity measure using EMD is not found in the literature.

METHODOLOGY

The overall design of the proposed semantic similarity computation using past travel histories is shown in Figure 1. The major components of the proposed system are GPS trajectories pre-processing unit, stay sequence generator unit, semantic annotator unit, EMD based similarity computation unit, and tourist recommender system. These components are discussed in detail in the following sub sections. The terminologies used in the paper are discussed here in brief.

- **GPS point:** A GPS point is quadruple consisting of latitude, longitude, altitude, and time stamp in format (lat, long, alt, t).
- **GPS logs:** A GPS log is an ordered sequence of GPS points. These ordered points are represented as $\{P_1, P_2,\ldots P_n\}$ as shown in Figure 2.
- **GPS trajectories:** A connected sequence of GPS points represented in a 2-dimensional plane in relative order of timestamp are called trajectories. The example of GPS trajectories is shown in Figure 3.
- **Stay point:** It is a geographical location where the user spent a substantial amount of time more than a predefined threshold or a group of GPS points within a predefined radius where the user spent considerable time. The relation between GPS points and stay points is shown in Figure 4.
- **Semantic Tags:** The key terms attached to the (lat, long) pair such as temple, national heritage, monument, archaeological site, shopping, entertainment, amusement park, waterfall, etc. are called semantic tags as these terms or tags adds semantic meaning to the (lat, long) pair.
- **Semantic trajectories:** The GPS trajectories with attached semantic tags are called semantic trajectories.

Figure 1. Overall Design of Proposed Semantic Similarity Computation



GPS Trajectories Pre-Processing

To develop and evaluate the proposed system, the GPS trajectory dataset of 182 users made publicly available by Microsoft Research Asia's GeoLife project (Zheng & Xie, 2010; Zheng & Lie, 2008) has been used. This dataset is collected from April 2007 to August 2012 (over four years). The dataset trajectories are recorded using different GPS loggers and GPS enabled mobile devices and logged every 1~5 second or every 5~10 meters per GPS point. The dataset has more than 17.6k trajectories from 182 users, which are about 1.2 million km long, and they are recorded for a duration of around 48000 hours. Every data point contains seven fields, as shown in the following example.

Example: "*39.906631, 116.385564, 0, 492, 40097.5864583333, 2009-10-11, 14:04:30*".

Field numbers 1, 2, and 4 respectively represent latitude, longitude, and altitude of a geographical location. Field 3 is set to 0. Field number 5 represents the number of days (with a fractional part) that have passed since 12/30/1899. Fields numbers 6 and 7 respectively represent the date and time as a string. The GPS trajectories from the GeoLife dataset are transformed into an easy to use format and represented as a tuple (user id, latitude, longitude, and the timestamp).

Stay Sequence Generator

The stay point of a user is generated as a result of two situations-

Figure 2. GPS Logs

	Latitude	Longitude	.....	Time
P_1:	Lat_1	$Lngt_1$	.....	T_1
P_2:	Lat_2	$Lngt_2$	.....	T_2
...	.....	.....	.....	...
P_n:	Lat_n	$Lngt_n$	.....	T_n

Figure 3. GPS Trajectories

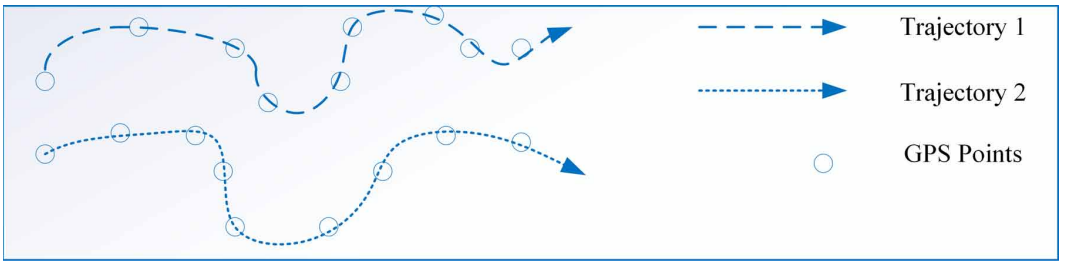
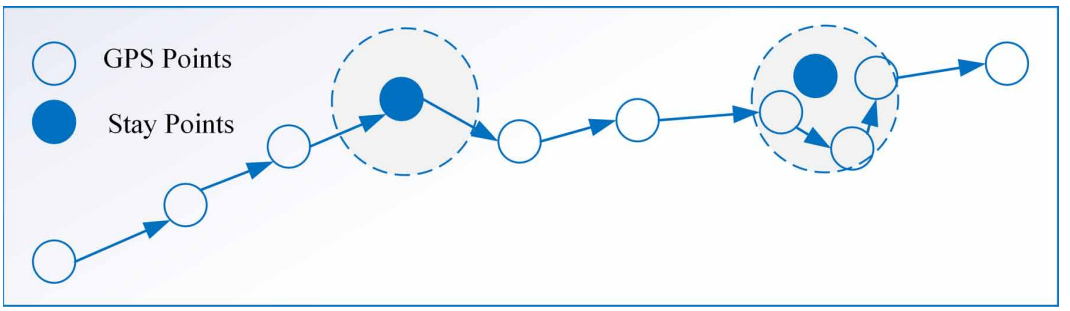


Figure 4. Stay Points



1. The user spent a significant amount of time greater than a predefined threshold (say τ) at some geographical point. This may happen when the user enters a building and remains there or because of the GPS signal loss. An example of this is the first stay point in Figure 4. Formally, a stay point $s = (s_{lat}, s_{long}, t_{arr}, t_{dep})$ where s_{lat}, s_{long} are latitude, and longitude of point s and t_{arr} and t_{dep} are arrival and departure time at point s.
2. The user moves around within a predefined radius (say δ) for a substantial amount of time. In this time, several GPS points are generated in a region. An example of this is the second stay point is Figure 4. In this case point s is a virtual location represented as a subsequence $P_i \rightarrow \ldots \rightarrow P_j$ taken from a trajectory $P_1 \rightarrow P_2 \rightarrow \cdots \rightarrow P_n$, such that:

$$\text{Distance}\left(P_k, P_{k+1}\right) < \delta \ , \forall k \in \left[i, j\right), \text{and } Interval\left(P_i.t_{arr}, P_j.t_{arr}\right) > \tau \quad (1)$$

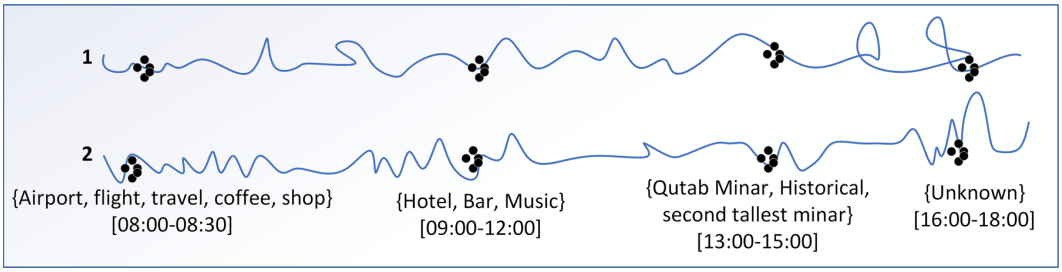
The arrival time t_{arr} at point P_i is represented by $P_i.t_{arr}$. The distance between two consecutive GPS points P_k and P_{k+1} is computed by function *Distance*. The function *Interval* is used to find the time difference between the GPS point P_i and P_j. The average of latitude and longitude of all the points in stay sub-trajectory are represented by s_{lat} and s_{long}.

The stay points are computed using the algorithm presented in (Tiwari, 2013). The time threshold τ and distance threshold δ are set to 25 minutes and 200 meters for experiments in this work.

Semantic Annotator

To gather the semantic information for every stay sequence, a basic semantic annotator is developed. However, any semantic crawler can be plugged in here. The example of annotated semantic stay sequences is shown in Figure 5. The basic flowchart of the semantic annotator developed for experimentation is shown in Figure 6. This semantic information will further simplify and enhance the analysis of users' movement from geographical and semantic perspectives.

Figure 5. Annotated Stay Sequences



The stay sequences of the individual user received from the stay sequence generator are feed to the semantic annotator. The world wide web is the richest source of information, and the source is used to crawl the semantic information.

The reverse geocoding application is used to find the actual address of the stay point. If the actual address of the stay point is not identified, then it is discarded. A seed URL is a feed to the semantic crawler, and the page downloader, in turn, downloads the page. The link extractor extracts the hyperlinks' links in the page and adds in the crawl queue if it is found to be relevant. The relevance of each link is calculated using a semantic ontology. The semantic ontology for tourism is adapted from the one discussed in (Tiwari, 2013) for testing purposes. The following equation (2) is used to compute the relevance of each link. The URL is added in the queue only if it is relevant in the context of tourism. The keyword extractor extracts the keywords from the relevant page. These keywords, if not yet already attached, are added to the semantic tag set of the stay point. The top 5 most frequent words are annotated with the stay point.

$$relevance\,Score = \frac{\left|Keywords\,in\,page\,macthing\,concept\,C\right|}{\left|Keywords\,in\,page\right|} \quad (2)$$

Semantic Similarity Computation

The semantic similarity for the semantically annotated stay sequences is computed using EMD.Given two distributions, EMD computes the work done to transform one sequence to another under a set of constraints. This is similar to the transportation problem.

The steps followed in similarity computation is explained as follows-

Step 1: Consider semantic stay sequences of user U as follows, where s_i represents a set of semantic tags associated with stay point i. then the semantically annotated stay sequence will look like the following.

$U = <s_1, s_2….,s_n>$

Construct new representations for U as a sequence of pairs where the first element is a set of semantic tags, and the second element is associate weight as follows-

$U'=<(s_1, w_1),…,(s_n,w_n)>$, where $w_i=1/|s_i|$

For example, set of semantic tags for user A's stay sequences are

A: <{school, park, monument}, {shop, mall},{temple}>

New representation of A is a pair of semantic tags and associated weights given as-

A':<({school, park, monument}, .33), ({shop, mall},.5), ({temple},1)>

Some other weight assignment method may also be used here.

Step 2: Formulation of problem for EMD application is as follows-

A directed weighted graph is constructed with stay points as nodes. The directed labelled links are from nodes of one sequence to the nodes of another sequence based on the common tags. Labels are minimum work required to transfer one node to another. Let us explain this concept with respect to the proposed problem through a pair of sample stay sequences for users A and B.

Consider the new representations of stay sequences of users A and B as follows:

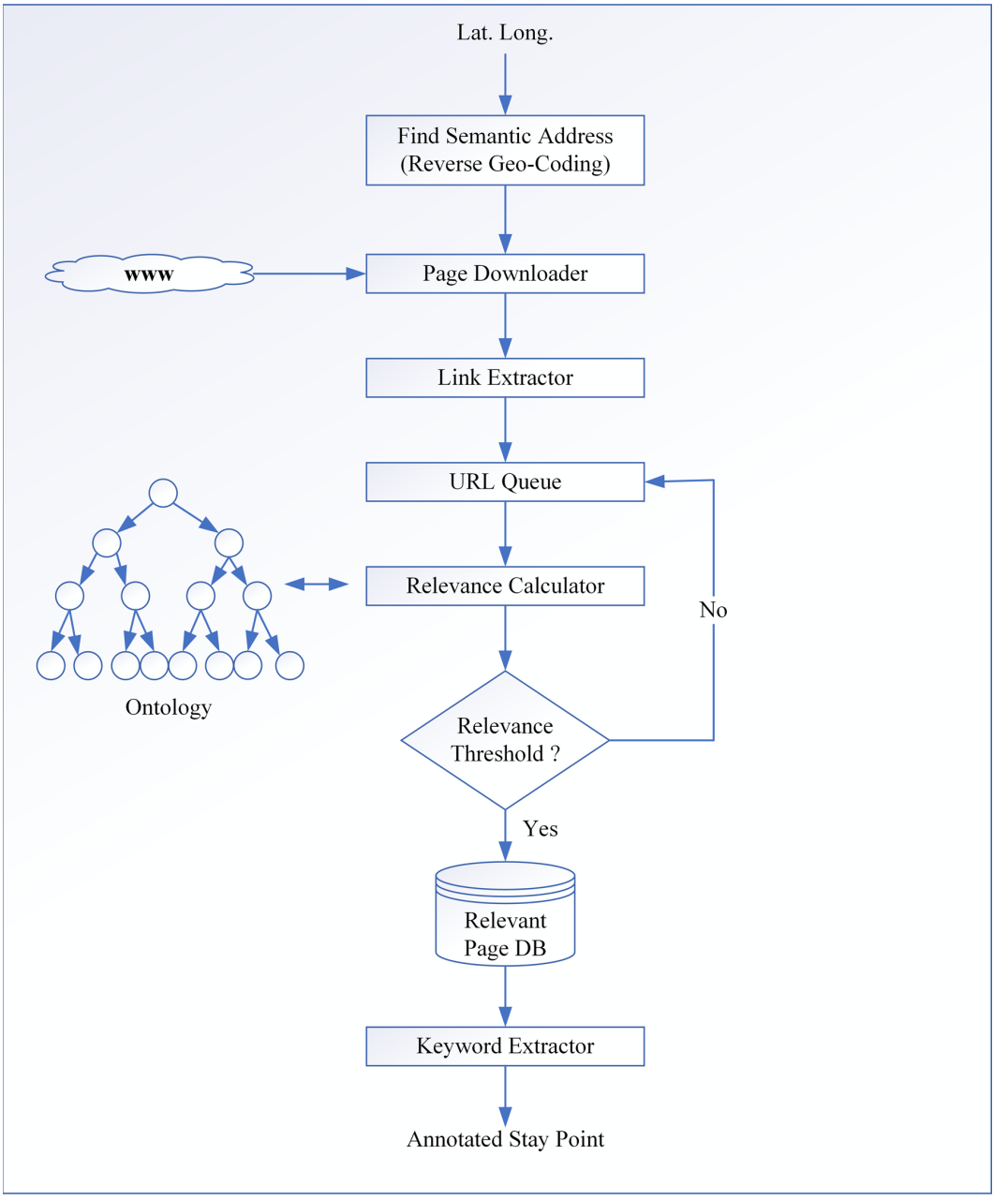
$A=<(s_{a1},w_{a1}),(s_{a2},w_{a2})….(s_{am}, w_{am})>$
$B=<(s_{b1},w_{b1}),(s_{b2},w_{b2})….(s_{bn}, w_{bn})>$

Here each s_{xi} is the set of tags associated with i^{th} stay point of a sequence of user X, and w_{xi} is the corresponding weight, where X is either A or B.

Construct a directed weighted graph G=< V, D > where V=A ∪ B is the set of vertices and D=[d_{ij}], where d_{ij} is the distance between tags t_i and t_j and is computed using weighted dice similarity method (Srinivas, 2010) as follows-

$dij= (1-wt\_sim\_dice(t_i, t_j))$

Figure 6. Flowchart of Semantic Annotator



Weighted dice similarity is computed using-

$$wt_{sim_{dice(t1,t2)}} = \frac{\sum_{r \in T_1 \cap T_2} w_{t_{1r}} * w_{t_{2r}}}{\sum_{r_1 \in T_1} w_{t_1 r_1} + \sum_{r_2 \in T_2} w_{t_2 r_2}} \quad (3)$$

Here T_i and T_j are the set of stays, tagged with t_i and t_j, respectively.

Consider the following specific example to explain the results obtained from the above equation (3). Let A and B are two sequences of stay points.

A=<({park, school, playground},0.33), ({temple, park},0.5)>
B=<{park, school },0.5), ({temple, school, park, canteen},0.25)>

T_{park}, T_{school} and $T_{park} \cap T_{school}$ are given below-

T_{park}=*{({**park**, school, playground},0.33), ({temple,**park**},0.5), ({**park**, school },0.5),({temple, school,**park**, canteen},0.25)}*

T_{school}=*{({park,**school**, playground},0.33),({ park,**school**},0.5),({temple,**school**, park, canteen},0.25)}.*

*T1∩T2=={({**park, school**, playground},0.33),({**park, school**},0.5),({temple,**school, park**, canteen},0.25)}*

Similarity of park and school is obtained using equation (3) as follows-

wt_sim_dice(park,school) = (0.33*0.33+0.5*0.5+0.25*0.25) /
((0.33+0.5+0.5+0.25)+(0.33+0.5+0.25))

Thus, wt_sim_dice(park,school)=0.158. Similarly, wt_sim_dice(temple,school) =0.0341.

Intuitively it may be verified that park and school has higher similarity as compared to temple and school. Thus D =[dij] is pre-computed for all tag pairs in the collection of semantic stay sequences. These values are further normalized.

In next step the flow F is computed for the above constructed weighted graph as F=[fij]. Here the flow f_{ij} is computed between t_{ai} and t_{bj} such that it minimizes the overall cost. Let *W (A, B, F)* represents the minimum work done to transform A to B computed as follows:

$$W\left(A,B,F\right)=\sum_{i=1}^{m}\sum_{j=1}^{n}f_{ij}d_{ij} \tag{4}$$

Under the constraints given as follows (m and n are the lengths of stay sequences A and B, respectively).

$$f_{ij}\geq 0;\quad 1\leq i\leq m and 1\leq j\leq n \tag{5}$$

This equation allows the movement of tags from sequence A to sequence B but not in reverse direction.

$$\sum_{j=1}^{n} f_{ij} \leq w_{a_i}\text{ , 1≤i≤m} \tag{6}$$

$$\sum_{i=1}^{m} f_{ij} \leq w_{bj}\text{ , 1≤j≤n} \tag{7}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min\left(\sum_{i=1}^{m} w_{a_i}, \sum_{j=1}^{n} w_{b_j}\right) \tag{8}$$

By solving the above transportation problem, we find flow matrix F=[f_{ij}]. EMD (dissimilarity between two sequences) is computed as follows:

$$EMD(A,B) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} d_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} \tag{9}$$

Step 3: Finally, the Semantic Similarity between users A and B is given as below.

$$Sim_{EMD} = 1 - EMD\left(A,B\right) \tag{10}$$

The semantic similarity between all the pairs of user sequences is computed and used to evaluate accuracy of the proposed approach. The results of the experiments are discussed in the next section.

Tourist Recommender System

A collaborative filtering-based recommender system is developed in which the user similarity matrix is supplied as input. Any other recommender system can be plugged in here, which requires user similarity based on past travel histories. The authors have developed a recommender system in (Tiwari, 2015) and it is used here. The recommender system developed here takes the user details and similarity matrix for that user obtained from the similarity computation module. On the basis of current location of user and the interest of similar users in the nearby locations are used to generated the meaningful recommendations for the user. The precision and recall of the recommender system are 93.65% and 89.2%, respectively.

EXPERIMENTAL RESULTS

The effectiveness of the proposed approach is verified using various experimental settings, as discussed here. The proposed similarity measure is compared with various other existing similarity measures, such as Jaccard, Dice, Cosine, and Pearson's methods. Due to the lack of actual similarity values, the cosine similarity values are considered as base results for comparisons as it is one of the most popular and commonly used similarity measures found in recommender systems. The metrics used for evaluation are the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). These evaluation matrices are frequently used in information filtering. A lower value of RMSE and MAE corresponds to better performance. From the dataset of 182 users, the semantic stay sequences of all the users are computed. Five groups of randomly selected fifteen users are picked for experiments from these users, and these groups are named as G1, G2, G3, G4 and G5, respectively. In a group set the users are selected randomly to remove the bias in selection. The datasets are divided in groups to evaluate the proposed similarity measure with in a different set of users. For evaluation purpose a smaller group of users is used. However, the experiments are also performed to find semantically similar users from the set of 182 users. The results of experiments show that the RMSE and MAE

of proposed approach outperforms the other similarity approaches. The experimental scenarios are discussed as follows.

Scenario-1 (RMSE)

RMSE is a measure of the difference between actual and predicted values. RMSE is computed using equation (11). As already discussed, due to the lack of actual values for user similarity, results of cosine are considered as actual values.

$$RMSE = \sqrt{\frac{\sum\left(\Pr edicted - Actual\right)^2}{N}} \tag{11}$$

The value of RMSE for Jaccard, Dice, Pearson, and the proposed approach are computed and compared. Table 1 shows the RMSE values for all the similarity measure. The relative percentage improvement of the proposed approach is shown in Figure 7. When compared with Jaccard, the percentage improvement of the proposed EMD-based approach is 15.9%. The percentage improvement in the case of Dice and Pearson's methods is 13.9% and 2.6%, respectively. Hence the overall percentage improvement of EMD based approach is 10.7%.

Scenario-2 (MAE)

Table 1. The comparision of root mean sqiuared error (RMSE)

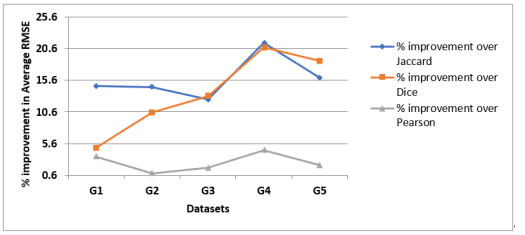
Dataset	Jaccard	Dice	Pearson	EMD Based
G1	0.88	0.88	0.82	**0.73**
G2	0.85	0.83	0.84	**0.82**
G3	0.83	0.83	0.84	**0.82**
G4	0.91	0.90	0.89	**0.88**
G5	0.81	0.79	0.79	**0.77**

The average magnitude of difference between the predicted and actual values without consideration of direction is called Mean Absolute Error (MAE). It is a measure of accuracy and computed using equation (12).

$$MAE = \frac{\sum\left|\left(Predicted - Actual\right)\right|}{N} \tag{12}$$

MAE is computed for the similarity computed from Jaccard, Dice, Pearson's, and EMD based approach with respect to Cosine similarity measure, and results are shown in Table 2. The percentage improvement of EMD based approach is presented in Figure 8. The average percentage improvement of EMD based approach over Jaccard, Dice, and Pearson's method is 7.31%, 5.57%, and 4.31%, respectively. The overall average percentage improvement comes out as 5.73%.

Figure 7. Average improvement of RMSE of proposed method over existing methods



Scenario-3 (EMD Based Method On Trajectories)

The experiments are performed using EMD based method directly on semantic trajectories instead of semantic stay sequences to verify the effective information loss while using semantic stay sequences instead of semantic trajectories. It is observed that in both cases, the top similar users found for the test users are the same. However, the strength of similarities varies for different users in the two cases. The results of the comparison are shown in Figure 9. The semantic trajectory-based method takes more time in similarity computation as compared to the semantic stay sequence-based method. Therefore, it is claimed that the semantic stay-based method is more effective as the results are not affected, and time is reduced.

Consider the user 4, top four similar users by stay based method are u8, u6, u2, and u9, and using an individual trajectory-based approach, similar users are again u8, u2, u6, and u9. The strength of similarity is different, but the set of similar users is the same. Based on the above result, it is concluded that the stay-based approach is equally effective as trajectory-based approach.

Scenario-4 (EMD based method Vs. Least Common Sub-Sequence based method)

The proposed method is compared with the longest common subsequence (LCSS) and proposed trajectory similarity approach. The list of top ten similar users found using both the methods. These experiments are performed 20 times for different users, and when results are compared, they show that on an average, 7.6 similar users are common in both the list for every user. This proves the accuracy of the proposed approach.

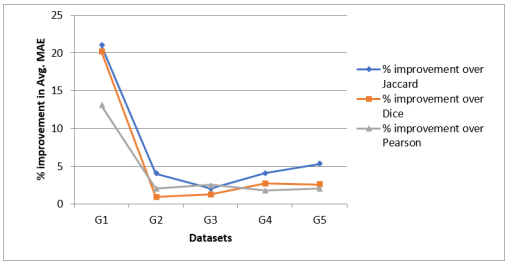
DISCUSSION

In all the experimental settings above, the cosine similarity measure is considered as actual values. This assumption that cosine values are true values need to be verified. To verify the results on actual data, the experiments are conducted on a set of 10 real users. Out of ten users seven were male and three were female users in age group of 19-23 years. Most of the experimental data is collected in

Table 2. The comparision of mean absolute error (MAE)

Dataset	Jaccard	Dice	Pearson	EMD Based
G1	1.00	0.91	0.90	**0.87**
G2	0.93	0.90	0.82	**0.81**
G3	0.88	0.88	0.79	**0.78**
G4	0.99	0.99	0.85	**0.82**
G5	0.83	0.85	0.73	**0.72**

Figure 8. Average improvement of MAE of proposed method over existing methods

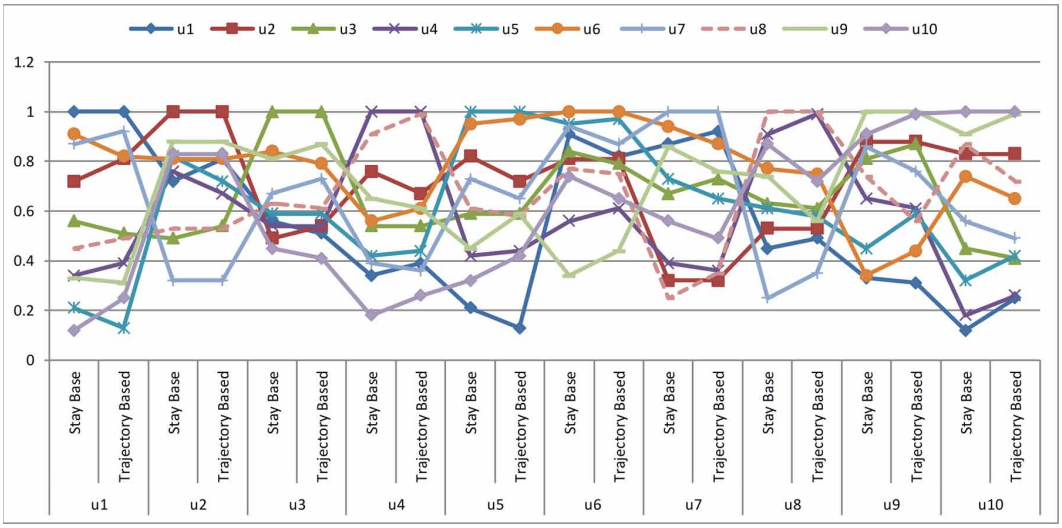


National Capital Territory of Delhi, India. The GPS data is collected from these 10 volunteer users for a period of 2 months. They used GPS logger app for android to collect the data. These users are all students, and therefore data is collected during the summer vacation period when the travel is most frequent. Once the data is gathered for all ten users, each one of them is requested to rate other remaining nine users based on their perception of similarity in travel taste on the scale of 0 to 1 where 0 is least similar, and 1 is most similar. Users were allowed to discuss their travel preferences with each other to understand the travel preference of each other before rating. The similarity is symmetric in nature, but in this case, user A and user B may not rate each other with the same rating values. To remove the ambiguity average of the rating given by both users is considered as the ground value. Considering user U1 as a test user and now similarity, this user is computed with the remaining nine users using EMD based proposed approach. The MAE is computed for EMD based similarity using the ground truth take from users themselves. The average of the MAE, in this case, comes out to be below 5%. This supports the validity of the proposed approach.

CONCLUSION

The contribution of this work is the semantic user similarity approach based on earth mover distance for the tourist sport recommender system. The proposed approach considers the semantic features

Figure 9. Comparison of stay sequence-based and trajectory-based approach



while computing similarity along with the geographical features. The experiment has shown that the approach outperforms all the conventional similarity measures when compared in terms of RMSE and MAE. The proposed approach shows an average percentage improvement of 10.7% in RMSE and 5.73% in MAE. It has also been demonstrated experimentally that the semantic stay sequence-based approach is computationally beneficially and equally effective when compared with the semantic trajectory-based approach. Also, the proposed approach is comparable with the existing trajectory similarity method LCSS. The precision and recall of the RS developed are also motivating. Further, more experiments may be performed for the proposed method using social network check-in data instead of GPS logs. The location sharing on social networks is becoming popular nowadays and therefore the proposed approach may be tested on the check-in data of users collected from the social networks. Also, the proposed work may be extended to deal with sparse trajectories.

ACKNOWLEDGMENT

Acknowledgment to Dr. Sushil Kumar for his support in typesetting and reviewing.

REFERENCES

Abraham, S., & Lal, P. S. (2012). Spatio-temporal similarity of network-constrained moving object trajectories using sequence alignment of travel locations. *Transportation Research Part C, Emerging Technologies*, *23*, 109–123. doi:10.1016/j.trc.2011.12.008

Alvares, L. O., Bogorny, V., Kuijpers, B., de Macedo, J. A. F., Moelans, B., & Vaisman, A. (2007, November). A model for enriching trajectories with semantic geographical information. In *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems* (pp. 1-8). doi:10.1145/1341012.1341041

Bogorny, V., Kuijpers, B., & Alvares, L. O. (2009). ST-DMQL: A semantic trajectory data mining query language. *International Journal of Geographical Information Science*, *23*(10), 1245–1276. doi:10.1080/13658810802231449

Bui, H. H., Phung, D. Q., Venkatesh, S., & Phan, H. (2008, July). The Hidden Permutation Model and Location-Based Activity Recognition. In AAAI (Vol. 8, pp. 1345-1350). Academic Press.

Chen, X., Pang, J., & Xue, R. (2013, March). Constructing and comparing user mobility profiles for location-based services. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing* (pp. 261-266). doi:10.1145/2480362.2480418

Herlocker, J. L., Konstan, J. A., Borchers, A., & Riedl, J. (1999, August). An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 230-237). ACM.

Imai, R., Tsubouchi, K., & Shimosaka, M. (2020, March). Next Place Prediction Using GPS traces and Web Search Queries. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)* (pp. 1-4). IEEE.

Khetarpaul, S., Chauhan, R., Gupta, S. K., Subramaniam, L. V., & Nambiar, U. (2011, March). Mining GPS data to determine interesting locations. In *Proceedings of the 8th International Workshop on Information Integration on the Web: in conjunction with WWW 2011* (pp. 1-6). doi:10.1145/1982624.1982632

Lee, J. G., Han, J., & Whang, K. Y. (2007, June). Trajectory clustering: a partition-and-group framework. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data* (pp. 593-604). doi:10.1145/1247480.1247546

Lee, M. J., & Chung, C. W. (2011, April). A user similarity calculation based on the location for social network services. In *International Conference on Database Systems for Advanced Applications* (pp. 38-52). Springer. doi:10.1007/978-3-642-20149-3_5

Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., & Ma, W. Y. (2008, November). Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems* (pp. 1-10). doi:10.1145/1463434.1463477

Liao, L., Fox, D., & Kautz, H. (2006). Location-based activity recognition. In Advances in neural information processing systems (pp. 787-794). Academic Press.

Lim, K. H., Chan, J., Leckie, C., & Karunasekera, S. (2018). Personalized trip recommendation for tourists based on user interests, points of interest visit durations and visit recency. *Knowledge and Information Systems*, *54*(2), 375–406. doi:10.1007/s10115-017-1056-y

Liu, H., & Schneider, M. (2012, November). Similarity measurement of moving object trajectories. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on GeoStreaming* (pp. 19-22). doi:10.1145/2442968.2442971

Logesh, R., Subramaniyaswamy, V., & Vijayakumar, V. (2018). A personalised travel recommender system utilising social network profile and accurate GPS data. *Electronic Government, an International Journal, 14*(1), 90-113.

Lu, E. H. C., & Tseng, V. S. (2009, May). Mining cluster-based mobile sequential patterns in location-based service environments. In *2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware* (pp. 273-278). IEEE. doi:10.1109/MDM.2009.40

Lv, M., Chen, L., & Chen, G. (2013). Mining user similarity based on routine activities. *Information Sciences*, *236*, 17–32. doi:10.1016/j.ins.2013.02.050

Milton, S., & McCall, D. (2019). *U.S. Patent No. 10,235,683*. Washington, DC: U.S. Patent and Trademark Office.

Nergiz, M. E., Atzori, M., & Saygin, Y. (2008, November). Towards trajectory anonymization: a generalization-based approach. In *Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS* (pp. 52-61). doi:10.1145/1503402.1503413

Ossama, O., & Mokhtar, H. M. (2009). Similarity search in moving object trajectories. In *Proceedings of the 15th International Conference on Management of Data* (pp. 1-6). Academic Press.

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, *40*(2), 99–121. doi:10.1023/A:1026543900054

Sankararaman, S., Agarwal, P. K., Mølhave, T., & Boedihardjo, A. P. (2013). *Computing similarity between a pair of trajectories.* arXiv preprint arXiv:1303.1585.

Srinivas, G., Tandon, N., & Varma, V. (2010, October). A weighted tag similarity measure based on a collaborative weight model. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents* (pp. 79-86). doi:10.1145/1871985.1871999

Tang, L., Cai, D., Duan, Z., Ma, J., Han, M., & Wang, H. (2019). Discovering travel community for POI recommendation on location-based social networks. *Complexity*, *2019*, 2019. doi:10.1155/2019/8503962

Tiwari, S., & Kaushik, S. (2013, March). Mining popular places in a geo-spatial region based on GPS data using semantic information. In *International Workshop on Databases in Networked Information Systems* (pp. 262-276). Springer. doi:10.1007/978-3-642-37134-9_20

Tiwari, S., & Kaushik, S. (2015, March). Modeling personalized recommendations of unvisited tourist places using genetic algorithms. In *International Workshop on Databases in Networked Information Systems* (pp. 264-276). Springer. doi:10.1007/978-3-319-16313-0_20

Xiao, X., Zheng, Y., Luo, Q., & Xie, X. (2010, November). Finding similar users using category-based location history. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems* (pp. 442-445). doi:10.1145/1869790.1869857

Ying, J. J. C., Lu, E. H. C., Lee, W. C., Weng, T. C., & Tseng, V. S. (2010, November). Mining user similarity from semantic trajectories. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks* (pp. 19-26). doi:10.1145/1867699.1867703

Zhao, G., Lou, P., Qian, X., & Hou, X. (2020). Personalized location recommendation by fusing sentimental and spatial context. *Knowledge-Based Systems*, *196*, 105849. doi:10.1016/j.knosys.2020.105849

Zheng, Y., Li, Q., Chen, Y., Xie, X., & Ma, W. Y. (2008, September). Understanding mobility based on GPS data. In *Proceedings of the 10th international conference on Ubiquitous computing* (pp. 312-321). doi:10.1145/1409635.1409677

Zheng, Y., & Xie, X. (2010, May). Learning location correlation from gps trajectories. In *2010 Eleventh International Conference on Mobile Data Management* (pp. 27-32). IEEE. doi:10.1109/MDM.2010.42

Zheng, Y., & Xie, X. (2011). Learning travel recommendations from user-generated GPS traces. *ACM Transactions on Intelligent Systems and Technology*, *2*(1), 1–29. doi:10.1145/1889681.1889683

Zheng, Y., Xie, X., & Ma, W. Y. (2010). GeoLife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, *33*(2), 32–39.

Zheng, Y., Ye, Y., & Xie, X. (2012). *U.S. Patent No. 8,275,649*. Washington, DC: U.S. Patent and Trademark Office.

Zheng, Y., Zhang, L., Ma, Z., Xie, X., & Ma, W. Y. (2011). Recommending friends and locations based on individual location history. *ACM Transactions on the Web*, *5*(1), 1–44. doi:10.1145/1921591.1921596

Zheng, Y., Zhang, L., Xie, X., & Ma, W. Y. (2009, April). Mining interesting locations and travel sequences from GPS trajectories. In *Proceedings of the 18th international conference on World wide web* (pp. 791-800). doi:10.1145/1526709.1526816

Zheng, Y., & Zhou, X. (Eds.). (2011). *Computing with spatial trajectories*. Springer Science & Business Media. doi:10.1007/978-1-4614-1629-6

Sunita Tiwari is Asst. Professor in the Department of Computer Science and Engineering at G B Pant Govt. Engineering College, Delhi. She received her M. Tech and Ph.D degree from IIT Delhi. She has more than 15 years of teaching experience in various engineering colleges. Her research interests are in the fields of Soft Computing, Location Based Services, GPS log Mining, Recommender Systems, Web Service Recommendations and Ad hoc networks. She has more than 20 research papers in refereed journal and international conferences. Couple of her publications were selected as best paper in international conferences. She has also published three books which includes Soft Computing with McGraw Hill (2018). She has supervised several thesis at M.Tech and B.Tech level. She held various administrative posts such as Dean for student welfare, Head of Training Placement Cell.

Saroj Kaushik is currently Distinguished Professor and HoD, Computer Science & Engineering at Shiv Nadar University. She is retired Professor in the Department of Computer Science and Engineering at Indian Institute of Technology. She obtained her Ph.D degree in Computer Science from IIT Delhi in 1980 and joined as faculty in the same year. She has about thirty-nine years of teaching and research experience in IIT Delhi. Her interests are in all fields of Artificial Intelligence including Natural Language Processing, Agents, Genetic algorithms, Knowledge Representation. Recently she is working in Location Based Services and Social Networks Analysis. She has successfully supervised many Ph.D students and currently supervising few more Ph.D. students. She has collaborated on 7 research projects and completed one consultancy project. She published three books: one on Logic and Prolog Programming with New Age International (2000), second on Artificial Intelligence with Cengage Learning (2010) and third on Soft Computing with McGraw Hill (2018). She has more than 100 refereed papers in journals and international conferences. She has guided about 100 theses at MTech and BTech level. She has been examiner for many theses at Ph.D and MS level. She is on the board of studies of various universities and Institutions of repute and was senate member of NIT Hamirpur. She is a member of faculty selection committees for various Institutes and universities.