Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 63 (2015) 142 – 147

Procedia Computer Science

6th International Conference on Emerging Ubiquitous Systems and Pervasive Networks, EUSPN-2015

Defining Measures for Location Visiting Preference

Ha Yoon Song[1], Dong Yun Choi

Department of Computer Engineering, Hongik University, Seoul, Korea

Abstract

For better location based service or better analysis of human mobility pattern, measures for presenting frequently visiting locations are usually required. In this paper, we will establish related measures for specific meaningful locations. Location points as well as Location clusters are objects of the measurements. In order to represent the degree of a specific location visit, the degree of location visit called Position Frequency (PF), and Inverse Location Frequency (ILF) are defined. In order to represent the degree of location area (cluster) visit, Inverse Cluster Frequency (ICF) is established. Moreover, along with the frequency of location visit, the duration of location visit is also considered. Therefore Position Duration (PD), Inverse Location Duration (ILD), and Inverse Cluster Duration (ICD) are defined. Using R language, real positioning data set collected by volunteers are analyzed in order to demonstrate the usefulness of these measures. The definitions of measures and the application of measures will be presented.

Peer-review under responsibility of the Program Chairs

Keywords: Human Location Preference; Measures of Location Visit; Position Frequency; Inverse Location Frequency; Inverse Cluster Frequency; Position Duration; Inverse Location Duration; Inverse Cluster Duration; Positioning Data Analytics; Location Base Service

1. Introduction

Various academic fields and industrial area require to identifying human mobility pattern for better mobile services such as location based services. Recent advancement of mobile devices such as smartphones enable end-users to collect their positioning data. From the positioning data set including time and position information (e.g. latitude and longitude), it is possible to establish human mobility pattern into human mobility models.

Locations are usually classified into two categories: location point and location area. For example, one can drop by coffee shop just to take one cup of coffee out, and this coffee shop must be a location point or location position. On the contrary, one can visit a shopping mall for a while, and this shopping mall must be a location area and can be represented as a location cluster. In both location categories, one can visit with a certain frequency and for a certain duration. It does mean that frequency to visit a location and duration to stay a location will be both meaningful measures. Location position is actually a point represented by latitude and longitude pair, while location cluster is a set

[1] Corresponding author. Tel.: +82-02-320-1617; fax: +82-02-332-1653
E-mail address: hayoon@hongik.ac.kr

Peer-review under responsibility of the Program Chairs
doi:10.1016/j.procs.2015.08.324

of related location positions. Frequency of visit stands for the number of visits to a certain location point or a certain location cluster embedded in the total positioning data; i.e. the higher frequency means a large number of visits to a certain location. Duration of visit stands for the stay time at a certain location point or location cluster calculated from the total positioning data; i.e. the longer duration means a longer stay at a certain location. Therefore, two aspects are required in order to establish location related measures: location point versus location cluster and frequency of visit versus duration of stay. In this paper, we will establish measures to represent location visit in terms of these combined aspects. The purpose is to identify the preference of certain location among possible locations visited and to represent preference of locations quantitatively. For demonstrating the verification of measures, sets of positioning data are used. Eight volunteers collected their positioning data for several years. In section 2, the measures of location visiting will be defined. Section 3 shows the actual application of measures for positioning data set in order to identify the latent pattern of human mobility pattern as well as to verify the effectiveness of the measures, and to demonstrate the interpretation of our measures. Section 4 will conclude this research with possible future research topics.

2. Definitions of Measures for Location Visits

2.1. Related Works

There a very few previous research related to this topic. Human mobile trajectory is widely used for travel recommended system, wireless communication, location prediction and so on, however no clear measures can be found in the related documents. One of the interesting results can be found in a research titled Web Classification using Deep Belief Networks by Sun et. al[1]. In this research, a keyword in one web page must be identified by its importance by the frequency of the keyword. Term Frequency (TF) stands for the frequency of a term in a page and IDF stands for a frequency in whole pages. In this aspect, TF×IDF shows the importance of a keyword.

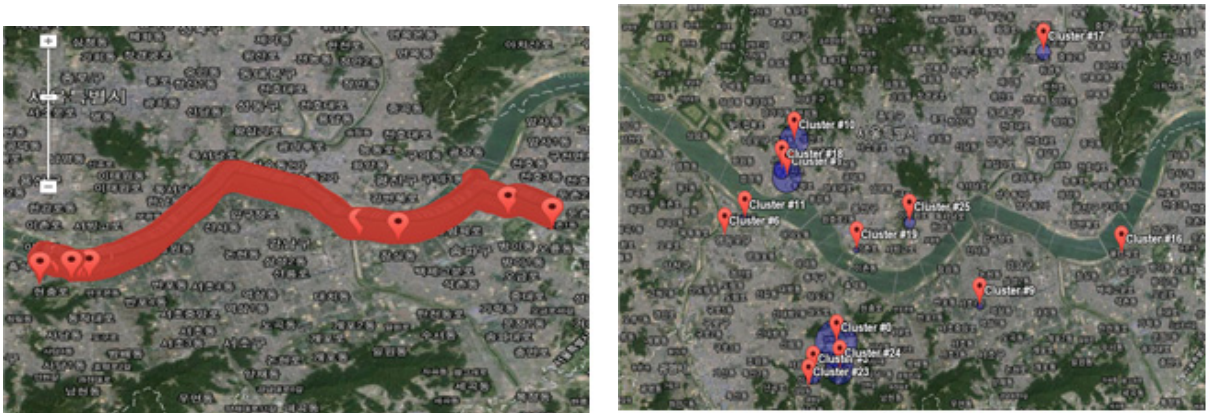
Based on this motivation, we developed measures for representing the tendency for location visit. Apart from the web keyword frequencies, the measures of visiting to a location are far complicated. As aforementioned, location point, location cluster, staying duration and visiting frequency are all related to the measures. Therefore, we introduced position frequency (PF) which represents frequency of visit to a certain location point (position) and Inversed Location Frequency (ILF) where position stands for latitude and longitude pair. As well, Position Duration (PD) and Inversed Location Duration (ILD) are established. The duration of a certain position will be reflected in these two measures. Inversed Cluster Frequency (ICF) and Inversed Cluster Duration (ICD) are measures related with location area or location clusters, while ICF is a measure of frequencies of points in a cluster and ICD is a measure of duration of stay in a cluster. The definitions of measures will be presented in this section.

Table 1. Sample of Raw Positioning Data Set.

Date	Time	UNIX time	Latitude	Longitude
2013—05—15	19:26:43—000	1368613603—000	37.55561833	126.9233483
2013—05—15	19:26:44—000	1368613604—000	37.55561833	126.9233483
2013—05—15	19:26:45—000	1368613605—000	37.55563333	126.9233367
2013—05—15	19:26:46—000	1368613606—000	37.55564667	126.923325
2013—05—15	19:26:47—000	1368613607—000	37.55565833	126.9233117

2.2. Positioning Data Collection and Location Clustering

The nature of positioning data is in a form of triple as $< time, latitude, longitude >$. From the set of raw positioning data, is it possible to extract human mobility model as shown in Kim and Song[2] which extracts meaningful location clusters of positions separately from transient positioning data with Expectation Maximization algorithm[3] from all positioning data. The location cluster extracted from the raw positioning data set will be used as criteria of location area in this paper. It does mean only the meaningful location data are residing in clusters while transient location data are excluded. Raw positioning data of a volunteer can be collected by smartphone using apps such as Sports Tracker[4]



(a) Example Mapping of Raw Positioning Data Set. (b) Mapping of Clusters distilled from Raw Positioning Data Set.

Fig. 1. Mapping of Location Position and Location Clusters.

or by dedicated devices such as Garmin[5,6]. Table 1 shows a small part of raw positioning data, including date, wall clock time, universal time in the form of UNIX time, latitude, and longitude.

The positioning data set can be projected on a map as shown in Fig 1 (a) and the extracted clusters also can be drawn on a map as shown in Fig 1 (b). The positioning data set has been collected by a female university student in the age of twenties. This positioning data set will be used to show the result of measures in section 3.

2.3. Position Frequency (PF)

PF has arguments of position and cluster. It shows the frequency of the occurrence of *location data* (position) inside a cluster including the position and can be calculated as (1). PF is defined on a specific position. The higher PF stands for frequent visits to the location.

$$PF_i = \frac{\text{Count of the specific location } position_i \text{ with the same latitude, longitude}}{\text{Total number of positioning data in a cluster containing } position_i} \tag{1}$$

2.4. Inverse Location Frequency (ILF)

ILF has arguments of positions, and stands for the rank of occurrence for a *location position* with the same latitude and longitude among the total location point inside all clusters and can be calculated as (2). ILF is defined for a specific position. The smaller the ILF is, the higher the visiting frequency for the location is. Transient locations are excluded and only the location positions inside clusters are regarded in this measure.

$$ILF_i = log\frac{\text{Total count of location position (for total cluster)}}{\text{Total count of the specific location } position_i \text{ with the same latitude and longitude}} \tag{2}$$

2.5. Inverse Cluster Frequency (ICF)

ICF has arguments of position and cluster, and stands for the rank of visiting frequency to a cluster, and can be calculated as (3). ICF is defined for a specific cluster. The smaller the ICF is, the higher the frequency of visit to the cluster is.

$$ICF_i = log\frac{\text{Total count of location positions in all clusters}}{\text{Count of position in } cluster_i} \tag{3}$$

2.6. Position Duration (PD)

PD is defined on a specific location position, and has parameters of the position and cluster. PD stands for a ratio of staying duration for a specific location position with the same latitude, longitude inside a cluster and can be calculated as shown in (4). The higher the PD is, the more time spend at the specific location in a cluster.

$$PD_i = \frac{\text{Stay time at a specific location } point_i \text{ with the same latitude, longitude}}{\text{Total stay time of every location position in the cluster having the specific location } point_i} \tag{4}$$

2.7. Inverse Location Duration (ILD)

ILD is defined on a specific position, and has argument of position. ILD stands for rank of staying time for a specific location position with the same latitude and longitude over total staying time of all location position inside all clusters, and can be calculated as shown in (5). ILD actually shows the rank of the position, i.e. the smaller the ILD is, the longer the staying time is. Only location position inside clusters is considered since the transient locations are meaningless for this measure.

$$ILD_i = log\frac{\text{Total staying time of all location position inside all clusters}}{\text{Total staying time at location } position_i \text{ with the same latitude, longitude}} \tag{5}$$

2.8. Inverse Cluster Duration (ICD)

ICD stands for the rank of staying duration to each cluster, and can be calculated as (6). ICD is defined for a specific cluster. The smaller the ICD is, the higher the duration of stay to the cluster is.

$$ICD_i = log\frac{\text{Total staying duration at all location positions in all clusters}}{\text{Total duration at location positions in } cluster_i} \tag{6}$$

3. Measuring on Real Positioning Data Set

In this section, we will demonstrate the effect of the measures by applying each measures on real positioning data sets. The positioning data set shown in Fig 1 has been collected by a female university student in the age of twenties. This positioning data set will be used to show the result of measures in section 3.

3.1. Measuring Position Frequency and Inverse Location Frequency

Table 2 shows information on clusters and measuring results of PF and ILF. Two positions, marked as $*$, $< 37.47601, 126.9384133 >$ and $< 37.47601, 126.9384117 >$ can be found in the same cluster 4. The count of location position is 24 for $< 37.47601, 126.9384133 >$, and 20 for $< 37.47601, 126.9384117 >$. In this case, PF and PF×ILF are proportional to the count of location while ILF tends to retrograde to the count of location. For the positions with the same count of location of 14 marked as †, such as $< 37.47601, 126.93841 >$ and $< 37.47601167, 126.9383967 >$ in the same cluster 4 must have the same PF, ILF, and PF×ILF. On the contrary, two positions of $< 37.49601, 126.9384067 >$ and $< 37.52202333, 126.9616783 >$, marked as ‡, with the same count of Location of 9, must have different PF and PF×ILF since the latter is in cluster 20 while the former is in cluster 4.

3.2. Measuring Position Duration and Inverse Location Duration

Table 3 shows PD and ILD similarly to Table 2. Cluster 9 has location position of the longest stay, marked as $*$, showing 1,131 seconds at $< 37.67260833, 126.7928733 >$ with PD of 0.27747792. However, location position in cluster 24, marked as † has the highest PD of 0.514986376 with less stay duration of 567 seconds. Even though staying time at $< 37.67260833, 126.7928733 >$ in cluster 9 is bigger than staying time at $< 37.470875, 126.93598 >$ in cluster 24, PD at $< 37.470875, 126.93598 >$ is bigger since most of the stay in cluster 24 is at the location point $< 37.470875, 126.93598 >$ as calculated in (4).

Table 2. Frequency: Count of Position Point, PF, ILF, PF×ILF.

Cluster Number	latitude	longitude	Count of Location Position	PF	ILF	PF×ILF
20	37.52202167	126.9616767	25	1.32E-02	8.273948968	0.108982468
∗ 4	37.47601	126.9384133	24	1.30E-02	8.314770963	0.108276996
∗ 4	37.47601	126.9384117	20	1.09E-02	8.49709252	0.09220936
5	37.73751833	126.8630883	10	0.007558579	9.1902397	0.069465153
† 4	37.47601	126.93841	14	7.60E-03	8.853767464	0.067255966
† 4	37.47601167	126.9383967	14	7.60E-03	8.853767464	0.067255966
4	37.47601333	126.9383533	10	5.43E-03	9.1902397	0.049865652
‡ 4	37.47601	126.9384067	9	4.88E-03	9.295600216	0.045393598
‡ 20	37.52202333	126.9616783	9	4.74E-03	9.295600216	0.044078189

Table 3. Duration: Count of Position Point, PD, ILD, PD×ILD.

Cluster Number	latitude	longitude	Duration at Position (sec)	PD	ILD	PD×ILD
† 24	37.470875	126.93598	567	0.514986376	6.810828626	3.507483952
14	35.26804	129.0786717	501	0.387470998	6.934581828	2.68694934
15	35.163005	129.1623517	552	0.31011236	6.837639883	2.120436638
22	34.734055	127.7204267	145	0.235772358	8.174454187	1.927310337
∗ 9	37.67260833	126.7928733	1,131	0.27747792	6.120330453	1.698256561
17	37.52116667	127.1012117	549	0.232037194	6.843089488	1.58785128
21	36.42675	127.418225	193	0.169595782	7.888497741	1.337855944
21	36.426765	127.41829	166	0.145869947	8.039200141	1.172677701
‡ 16	34.894445	127.5161217	191	0.146697389	7.898914502	1.15875013
‡ 16	34.895235	127.516205	191	0.146697389	7.898914502	1.15875013

Cluster 16 has two different location points having the same duration of 191, marked as ‡ with the same ILD of 7.898914502. It means that the volunteer stays mostly at two location points equally likely but very short when the volunteer visits cluster 16. Since two points are in the same cluster, PD for two different location positions are the same of 0.146697389, and PD×ILD are the same of 1.15875013.

3.3. Measuring Inverse Cluster Frequency and Inverse Cluster Duration

ICF and ICD are for location area, as called as cluster and usually represents the rank of location visit. Table 4 shows ICF and ICD with cluster information. The smallest ICF, or the highest rank, can be found for cluster 1 with 37,121 visits and the smallest ICD also can be found for cluster 1 with 190,004 seconds of stay. The smaller ICD represents the larger cluster size or longer stay duration at the cluster. The higher ICD represents the smaller cluster size or shorter stay duration at the cluster. Results for cluster 8, 10, 13, 14, 15, 16, 17, 18, 21, 22, 23, and 24 have not been presented since their count of cluster point is less than 500.

4. Conclusions

In this research, we define six measures to represent for a human visiting preference. The frequency of visits as well as duration of visit are considered and the visit to a specific location as well as the visit to a location area, named as location clusters, are also considered. For the frequency of visit PF, ILF, ICF are defined and for the duration of visit PD, ILD, ICD are defined. The PF, PD, ILF, and ILD are for location position while ICF and ICD are for location cluster. We utilized positioning data set from eight volunteers and demonstrate the usefulness of measures. Among the positioning data set, the result from a female university student at the age of her twenties are presented in this paper.

The major consideration on locations is to divide location into a location point (micro location) and location area (macro location; cluster). For example, students can take a course in a classroom (location point) or can walk across the university campus (location area). The Preference of a location inside a cluster can be measured by PF and PD. As well, ILF and ILD can represent the rank of preference of a location absolutely all across the locations. In addition,

Table 4. Location Area: Count of cluster point, ICF, Stay time for cluster, ICD.

Cluster Number	Count of Cluster Point	ICF	Stay Duration at Cluster (sec)	ICD
1	37,121	0.970886667	190,004	0.996387526
2	16,715	1.768762994	112,959	1.51640773
3	923	4.665195559	6,357	4.393876083
4	1,843	3.973674836	10,161	3.924875788
5	1,323	4.305167629	7,684	4.204292406
6	874	4.719744418	7,955	4.169631989
7	962	4.623810343	5,381	4.56055842
9	538	5.204966233	4,076	4.838316535
11	1,813	3.990086583	5,806	4.484540785
12	607	5.084296002	1,748	5.684960373
19	16,279	1.795193581	91,017	1.732386348
20	1,898	3.944268814	2,383	5.370884842
25	14,683	1.898379152	47,255	2.387874182
26	912	4.677184803	664	6.65290578

the importance of a cluster can be found by ICF and ICD. The values of ICF and ICD show the similar tendency as expected. In other words, ICD is preferred unless the frequency of visit need to be specially treated. The other major consideration for location visiting is to separate visiting frequency from visiting duration. For example, one can frequently visit a coffee shop just for coffee take out whilst one can stay at a restaurant for a certain time. Visiting duration and visiting frequency are found clearly different in our measures. For example, we found a distinguished location having Count of Position Point of 3,433 with PF×ILF of 0.427661939, which is large enough whilst PD×ILD for the corresponding location is minute to consider. In addition, for cluster 26, ICF and ICD show different ranks as shown in Table 4. I.e., visiting frequency can have different meaning from visiting duration. Maybe these various measures can be used depending upon the situation of human mobility, solely or together from the aspect of applications. For example, an advertisement can be made by the measures of visiting frequency for a certain location using PF, ILF, and ICF.

For the actual calculation of location measures, execution time to find measures grows exponentially according to the number of position data which is one of the major problems to be solved. Calculation of measures on each positioning data set have been made on a computer system with six core Xeon CPU. For example, calculation time for measures on positioning data set of volunteer 7 is 810.67 minutes for frequency and 4,279.59 minutes for duration. In sum, for the biggest positioning data set, it took five days for frequency measures and sixteen days for duration measures. One of the possible solutions is to use GPGPU technique in order to reduce calculation time of measures for better application of these measures, even though realtimeness is not a major stuff to be accomplished.

The next and more sophisticated measures will consider time of a day in order to reflect the effect of time on the measures presented in this research.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (NRF-2012R1A2A2A03046473).

References

1. Sun, S., Liu, F., Liu, J., Dou, Y., Yu, H.. Web classification using deep belief networks. In: *Computational Science and Engineering (CSE), 2014 IEEE 17th International Conference on*. 2014, p. 768–773. doi:10.1109/CSE.2014.158.
2. Kim, H., Song, H.Y.. Daily life mobility of a student: From position data to human mobility model through expectation maximization clustering. In: *Multimedia, Computer Graphics and Broadcasting*. Springer; 2011, p. 88–97.
3. Dempster, A.P., Laird, N.M., Rubin, D.B.. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society Series B (methodological)* 1977;:1–38.
4. Sportstracker. http://www.sports-tracker.com; 2015.
5. Garmin gpsmap 62s. https://buy.garmin.com/en-US/US/on-the-trail/discontinued/gpsmap-62s/prod63801.html; 2015.
6. Garmin EDGE500. https://buy.garmin.com/en-US/US/into-sports/cycling/edge-500/prod36728.html; 2015.