Improving the Quality of GPS-based Personal Gazetteers

Mark M. HALL, Ahmed N. ALAZZAWI, Alia A. ABDELMOTY, Christopher B. JONES

Abstract

The increasing prevalence of location-aware mobile devices such as car-navigation systems, phones and cameras provides a wealth of location information about a person, which can be used to build up a personal gazetteer for the device owner. The main technology employed for determining device location is GPS and classically the loss of the GPS signal at the device is interpreted as identifying places to include in the personal gazetteer. However, in densely built-up areas GPS data is often very dirty, with frequent loss of signal due to an insufficient number of satellites visible, and which do not indicate an actual place. In this paper heuristics for improving the quality of GPS log data are investigated and their impact on the personal gazetteer derived from the GPS data is evaluated.

1 Introduction

The number of location-aware, mobile devices is steadily increasing, with Phones, cameras, PDAs being enhanced with GPS receivers. The fact that the device always knows where it is, can be used to provide location based services to the user as described in Schiller and Voisard (2004). The main problem with this is stated by Nivala and Sarjakoski (2004) that all data-providers are interested in pushing their data an the user can easily be overwhelmed by the plethora of information.

To reduce the information overload, the device needs to be able to select those service advertisements that are relevant to the user. This selection would be based on a personal gazetteer, which stores the places that are important to the user. Such personal gazetteers are either defined directly by the user, automatically extracted from spatio-temporal logs provided by the device, or sometimes hybrid solutions of the two methods are used as in Froehlich et al (2006). Classically in spatio-temporal logs based on GPS data, places are detected based on when the GPS signal is lost, as in Marmasse and Schamndt (2000) or Zhou et al (2004). This is sometimes extended by other machine-learning techniques such as in Patterson et al (2003) and Liao et al (2004). Unfortunately in heavily built-up areas, such as city centres, all these techniques share their susceptibility to large numbers of incorrectly detected places, as the urban canyons lead to frequent loss-of-signal events that do not indicate actual, relevant places. While clustering as in Ashbrook and Starner (2002) reduces the number of such places, we feel that a more thorough data-cleaning approach is required.

2 Data cleaning

The goal of the initial data cleaning step is to improve the quality of the GPS data, which then reduces the error rate of the place detection. Non-place-related loss-of-signal events need to be removed, and to clean up these split tracks two spatio-temporal heuristics have been developed, that can detect those tracks that actually represent one trip and then merge them. These loss-of-signal events in urban canyons can be split into two cases, where the signal is lost briefly due to buildings, and longer signal-loss events caused by tunnels or other enclosed pathways.

For the short loss-of-signal events, a simple heuristic based on the distance (less than 100m) and time (less than 60 seconds) between the last point of the first track and the first point of the second track is used to determine whether the two tracks should be merged. To detect longer loss-of-signal events, a second heuristic based on the method-of-transport (MOT) is employed. The MOT is calculated for both tracks and then also for the space between the two tracks. If the MOT for all three parts is the same and the time between the the two tracks is less than 30 minutes, the two tracks are merged.

4 Evaluation

The results of the cleaning heuristics' application to a test data-set can be seen in Table 1. The test data-set was collected over four weeks in a heavily built-up urban environment and the evaluation is based on a manually logged list of places.

Place detection within both the raw and the cleaned data-sets is performed using two metrics. Loss-of-signal indicated by the end of a GPS track is the main method employed. Additionally places are detected where the movement of the GPS device is restricted to a certain area for a given amount of time. The places detected by these two methods are then clustered using a slightly modified DBSCAN algorithm (see Ester et al (1996) and Sander et al (1998)) that takes these two place-types into account. The result of the clustering is then compared to the user-generated evaluation data-set.

Table 1: Evaluation results for the three detection strategies. The test set has 36 places specified by the user.

Detection Strategy	True Positive	False Positive	False Negative	Precision	Recall	F-score
No cleaning	25	41	11	0.38	0.69	0.49
Cleaning	25	23	11	0.52	0.69	0.6

Table 1 shows that adding the cleaning step to the place detection algorithm, almost halves the number of false-positives and increases precision to a point where it is possible to use the results for further processing. The remaining false-positives and false-negatives are currently being analysed to determine why they are not either filtered or not detected at all.

5 Discussion

Personal gazetteers mined from GPS logs, using loss-of-signal as a place indicator, often contain a large number of incorrectly identified places, especially when used in heavily built-up urban areas, due to frequent non place-related loss-of-signal events. The heuristics presented in this paper make it possible to cut down on the number of incorrectly identified places, but are not able to remove all incorrectly identified places. Further analysis of these incorrectly identified places has revealed that in many cases inaccuracies in the original GPS data are the cause and that further heuristics are required to improve the results. Future work and discussion will thus focus on how the existing heuristics can be improved to remove more loss-of-signal events, and also on what new heuristics could be applied to filter out those cases where inaccuracies in the original data are the source. Additionally we are in the process of acquiring GPS logs for more users to improve the reliability of our evaluation process and also to analyse whether it is necessary to tailor the various algorithms to the individual users or to groups of users.

References

Marmasse, N., Schmandt, C. (2000), Location-aware information delivery with comMotion. – In: Proc. HUC 2000

Ashbrook, D., Starner, T. (2002), Learning significant locations and predicting user movement with GPS, – In: Proc. IEEE 6th Intl. Symp. On Wearable Comp. 2002

Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., Terveen, L. (2004), Discovering Personal Gazetteers: An Interactive Clustering Approach, – In: ACM GIS'04

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. (1996), A density-based algorithm for discovering clusters in large spatial databases with noise, – In: Proc. KDD 1996

Froehlich, J., Chen, M.Y., Smith, I.E., Potter, F. (2006), Voting With Your Feet: An Investigative Study of the Relationship Between Place Visit Behaviour and Preference, – In: UbiComp 2006: Ubiquitous Computing, 333 – 350

Liao, L., Fox, D., Kautz, H. (2004), Learning and inferring transportation routines, – In: Proc. AAAI 2004

Nivala, A-M., Sarjakoski, L. T. (2004), Preventing Interruptions in Mobile Map Reading Process by Personalisation, – In: Proceedings of The 3rd Workshop on 'HCI in Mobile Guides', in adjunction to: MobileHCI04, 6th International Conference on Human Computer Interaction with Mobile Devices and Services, September 13-16, 2004, Glasgow, Scotland.

Patterson, D., Liao L., Fox, D., Kautz, H. (2003), Inferring high-level behaviour from low-level sensors, – In: Proc. UbiComp 2003

Sander, J., Ester, M., Kriegel, H.-P., Xu, X (1998), Density-based clustering in spatial databases: The algorithm gdbscan and its application, – Data Mining and Knowledge Discovery 2: 169 – 194

Schiller, J. H., Voisard, A. (2004), Location-based services, – Morgan Kaufmann Publishers