Understanding User Behavior Geospatially

Xing Xie
Microsoft Research Asia
5F, Sigma Building, No.49, Zhichun Road, Beijing, 100190, P.R.China
xingx@microsoft.com

ABSTRACT

Understanding users is an essential task for providing personal Web experience and targeted advertisements. Current commercial or research systems try to understand users from their online behaviors, for example, how they search, read and write on the Web. However, this type of approaches missed a large part of people's everyday life, or called 'physical' behaviors. The physical behaviors include how people dine, shop, travel, or other activities happened in the real world. In our opinion, location is one of the most important aspects for people's everyday life. With the rapid growth of location sensing devices and Web based GIS tools, it becomes possible to track these physical behaviors from a geospatial view. In this paper, we present our recent work towards understanding users from a geospatial view. Particularly, we studied GPS trajectory transportation mode categorization and co-located query pattern mining problems.

Keywords

Geographic data mining, personalization, transportation mode, co-location pattern, log mining

1. INTRODUCTION

Understanding users is an essential task for providing personal Web experience and targeted advertisements. Current commercial or research systems try to understand users from their online behaviors, for example, how they search, read and write on the Web. However, this type of approaches missed a large part of people's everyday life, or called 'physical' behaviors. The physical behaviors include how people dine, shop, travel, or other activities happened in the real world.

In our opinion, location is one of the most important aspects for people's everyday life. Here location stands for the venue of user activities. It can be GPS coordinates, place names or names of point-of-interests. With the rapid growth of location sensing devices and Web based GIS tools, it becomes possible to track these physical behaviors from a geospatial view. People usually use search engines to plan their activities, for example, finding a higher rated seafood restaurant close to their home, searching for better driving directions to destinations, or reading user comments about an unfamiliar shop. During the activities, the trajectories of a user can be recorded if he/she brings a GPS enabled device. Last but not least, many people will write blogs or upload photos to share their experience to friends. In these scenarios, location information can be extracted from search queries, GPS data, blog posts or even photos.

Based on different types of data we have, we can mine different types of user interests. Search queries indicate places that users have interests in. Personal trajectories partly reveal users' life patterns, while blog posts and photos give more information about particular locations where users have been to. By analyzing data from multiple users, we can further know the statistical characteristics of places and also the relationship between different locations and people.

Just like all other systems that try to understand users, privacy is an important issue here. In real deployments, we need to design appropriate schemes to protect the data and make sure information is shard to the right person and in the right way.

2. UNDERSTANDING PERSONAL TRAJECTORIES

With the increasing prevalence of GPS devices, many communities that engaged in geographically related activities have been established. Most of these applications only use raw GPS data, e.g., GPS coordinates and timestamps, without much understanding, while the rest applications require people to manually label their GPS data.

As a kind of knowledge mined from raw GPS data, transportation modes such as walking, driving etc, and the transitions between them are valuable information for both users and application systems. In many research works aiming to understand user behavior from raw GPS data, the information of transportation mode is also important knowledge to predict an individual's movements outdoors, supervise cognitively-impaired person's activity, extract user's life pattern and discover the social pattern. In turn, all the knowledge learned from these works can be leveraged to boom many innovative local/mobile applications on the Web further.

Identification methods based on simple rules, such as velocity-based identification, cannot handle this problem well. The features of different transportation modes usually suffer from traffic conditions and weather. It is intuitive that in the congestion the mean velocity of driving would be as slow as walking while in a raining day a bus may moves more like a bike from the perspective of velocity. When users take more than one kinds of transportation modes along a trip, the problem becomes worse.

In [1], we propose an approach based on supervised learning to automatically learn the transportation modes including walking, taking bus, riding bike and driving from raw GPS data. The work is part of a research project called GeoLife, which focuses on visualizing, well organizing, fast retrieving and effectively mining GPS log data for both personal and public use. Figure 1 depicts the Web user interface of GeoLife prototype.

Our approach consists of three parts: a change point based segmentation method, an inference model and a post-processing

algorithm based on conditional probability. We evaluate our approach using GPS data collected by 45 people over a period of six months. As compared to uniform duration based and uniform length based segmentation methods, change point based method achieves higher accuracy in predicting transportation modes. It also obtains better precision in detecting transitions between different transportation modes. Over the change point based segmentation method, Decision Tree outperforms other inference models. However, based on the three segmentation methods mentioned above, CRF does not present its advantages in labeling sequence data.

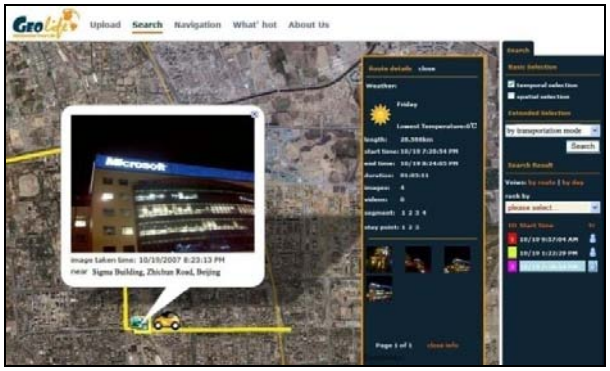

Figure 1. Web user interface of GeoLife.

3. UNDERSTANDING GEOGRAPHIC SEARCH LOGS

Recently, geographic search engines, such as Live Local Search (local.live.com), Google Maps (maps.google.com) and Yahoo! Local (local.yahoo.com), are attracting more and more traffic on the Web. People frequently use them to decide driving directions, find dining places and make travel plans. This results in large amounts of logs saved on search engines. Though general query log mining has been studied for years, little work has been done to utilize these geographic search logs.

Basically, a geographic search request consists of two fields: 1) a query consisting of one or more keywords, and 2) a location that associates with the query to specify the geographic search area, which we call search-location, or location for short in this paper. Note that, search-locations may be different from the users' location that they are located at.

In [2], we study one particular type of log mining tasks: finding co-located query patterns. One example co-located query pattern is {'Children's Museum', 'Experience Music Project'}. The search-locations of the two queries are close to each other in Seattle. Actually, both of them are museums in downtown Seattle and they are located within 250 yards from one another. Another example is {'shopping mall', 'parking'}. This pattern indicates that 'shopping mall' and 'parking' tend to have nearby search-locations.

In summary, a co-located query pattern consists of a set of queries that are associated with nearby search-locations frequently. In Figure 2, different symbols denote different queries. Different coordinates of the same symbol represent search-locations of a query. As shown in the figure, queries {+, ♦} and {Δ, O} are two co-located query patterns since each pair is often searched for locations within proximity.

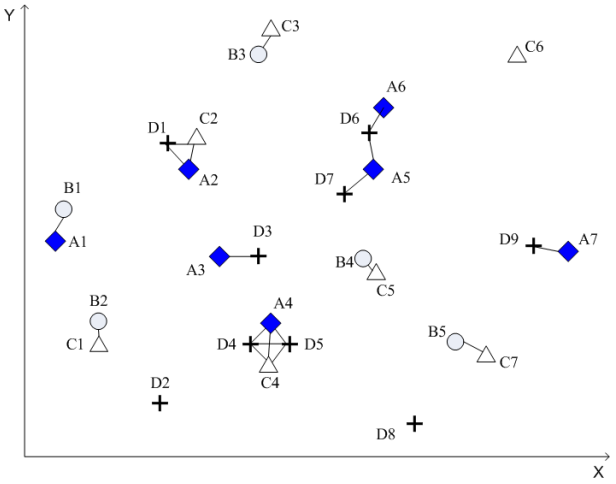

Figure 2. Co-location pattern discovery.

Mining co-located query patterns has broad applications, such as query suggestion, location recommendation, and local advertisement. In [2], we compare a basic approach that uses an existing algorithm, and a lattice based approach that mines co-located query patterns in regions and categorize patterns into local and global by calculating locality degrees of patterns. In our user study, the participants give an average score of quality (between 0 and 1) of 0.76 for the patterns discovered by the basic method and 0.9 for those by the lattice-based method. They also categorize 64% and 94% of the patterns discovered by the basic and lattice based methods respectively as local ones. In addition to higher quality of patterns, the lattice based method consistently discovers more patterns than the basic method. With these results, we conclude that the lattice based approach is more effective than the basic one in terms of mining more patterns, larger percentage of local patterns, and patterns with higher quality.

4. CONCLUSIONS

In this paper, we have introduced our recent work on understanding users from a geospatial view. Particularly, we studied GPS trajectory transportation mode categorization and co-located query pattern mining problems. We will investigate more types of data and user interests in our future work.

5. REFERENCES

[1] Y. Zheng, L. Liu, L. Wang, and X. Xie. Learning Transportation Mode from Raw GPS Data for Geographic Applications on the Web. 17th International World Wide Web Conference (WWW 2008), Beijing, China, Apr. 2008.

[2] X. Xiao, L. Wang, X. Xie, and Q. Luo. Discovering Colocated Queries in Geographic Search Logs, First International Workshop on Location and the Web (LocWeb 2008), Beijing, China, Apr. 2008.