# Information Filtering Meets Mobility[1]

Mario A. Nascimento
Dept. of Computing Science
University of Alberta, Canada
mario.nascimento@ualberta.ca

## ABSTRACT

The underlying premise in this position paper is that mobility patterns have been relatively underexplored from a data management perspective. More specifically, we discuss how to explore mobility patterns from a large set of users in order to allow proactive discovery and/or suggestion of spatiotemporal events of interest to users. Towards that goal, and considering the semantics of such an application in particular, we suggest three lines of work: (1) how to cluster trajectories, (2) how to find a representative trajectory for trajectory clusters, and (3) how to index trajectories.

## 1. MOTIVATION

It should be fairly safe to assume that people often move about according to patterns. For instance, the way from home to school/work (and vice-versa) likely follows a few established routes and schedules. A relative large number of people already carry smart-phones with embedded GPS. Ignoring relevant privacy issues for the sake of argumentation, one can envision these devices being used to proactively record one's movements (virtually) all the time. Let us call this type of data Spatio-Temporal Trajectory Patterns (STTPs).

Assuming the above, one can envision the following applications, to mention only a few. Knowing the users' interests (e.g., from their online activities) and their STTPs it becomes feasible to plan beforehand which, when, and where specific ads would be more interesting to be offered, based on the users' preferences and likely location, and without the need to perform (expensive) real-time tracking. "Collective" offers, e.g., instant e-coupons, could be offered to groups of users based on their STTPs, in order to take advantage of their probable proximity to the businesses offering the coupons. Along the same lines, in the context of social networks, targeted e-coupon offers could be sent to groups of online friends who share the same, or reasonably close, STTPs. Another potential application could be to proactively help friends in sharing car rides and/or optimizing them. Finally, users could also be advised, on a need-to-know basis, of road constructions, detours or accidents that may affect their typical daily drive or commute to work, school, etc. It is also conceivable to use STTPs to derive typical behavior of roads (e.g., driving speed as a function of the time of the day), which may be statistically more reliable and up-to-date than potentially stale historical data.

To realize the applications above, let us be inspired by traditional (text-based) information filtering [1]. In information filtering users have a standing query profile that is continuously checked against new data items arriving in a data stream. Once a data item is deemed relevant with respect to a user's profile, that user is somehow alerted about the data item, e.g., a news article. In this case, the user is rather passive as opposed to the typical case where the user pro-actively issues queries. Considering the context just discussed above, one can imagine the situation where the "query" is the user's trajectories and the data stream is a series of spatiotemporal events. This calls for efficient means to efficiently uncover events of interest for users in the context of their STTPs. In the next section we deal with a set of tasks that, collectively, can lead to a solution towards this problem.

We note that an important piece needed for materializing the idea above is learning a person's interest. If we consider the amount and quality of existing work on targeted advertisement based on one's online activities, we are certain that very effective and efficient ways to solve this problem do exist and could, eventually, be re-used. Hence we do not discuss this issue any further in the remainder of this paper.

## 2. A PROPOSED ROAD MAP

A critical task in addressing the STTP-based information filtering application stated above is to check every event of potential interest against the stored STTPs of all users. A naïve approach, i.e., comparing all STTPs against all events is clearly not practical. To perform this task efficiently we envision the need to (1) determine clusters of trajectories, (2) find representative trajectories for the determined clusters, and (3) index those representative trajectories. In this way, events can be checked against a smaller set of indexed trajectories, thus improving efficiency while maintaining effectiveness. Each of those three components is discussed in turn next.

### 2.1 Clustering Trajectories

Finding clusters requires one to have a notion of distance between the objects being clustered. In the context of STTPs, it is not sufficient to consider only the distance with respect to the spatial component of a trajectory; the temporal component is also just as relevant. Thus a first task is to determine a suitable distance function that accounts for both temporal and spatial components of trajectories. For that, sophisticated distance functions for time series could to be considered, keeping in mind the need for superior robustness with respect to scaling and shifting, e.g., extending the work presented in [4].

There has been work done in trajectory mining/clustering that we can use as a starting point, e.g., the work by Giannotti et al [5]. In that work however, the authors assume the existence of "locations of interest" which are relevant for all trajectories collectively, not on a per-user basis, as we would need to do in the context of users' STTPs. C.S. Jensen and his team/colleagues

[1] This work has been partially supported by NSERC Canada.

have done other works that may inspire good ideas. For instance, one deals with discovering convoys [6]. A convoy is defined as "a group of objects that have traveled together for some time"; in our case we are looking for STTPs induced by single individuals. Another work deals with constructing accurate routes from GPS data [2], but does not address issues necessary in the context of this proposal, such as identifying patterns of trajectories.

## 2.2 Determining Representative Trajectories

Once trajectory clusters have been determined, the next task is to determine a representative trajectory for each cluster. At the information filtering stage these representative trajectories alone, thus a relatively small set, will be used to determine the relevance of an event with respect to the STTPs of a group of users (clusters). There are a few possibilities to accomplish this, from computing an "average" trajectory to re-using an existing one that minimizes the error with respect to the others.

In both tasks above one must consider that STTPs are bound to change over time, for instance, in case of road constructions or collective change in behavior (e.g., during a long holiday weekend). Therefore, cost-effective means to keep large clusters of STTP, as well as their representatives, up-to-date and consistent, needs to be investigated as well. We are not aware of any previous research on how to solve this problem.

## 2.3 Indexing Trajectories

The third task is to index the obtained representative trajectories. The potentially large number of events, actually a stream thereof, needs to be checked against every representative trajectory, which is bound to be a very large dataset. Hence, this task must be performed very efficiently using an index. Again, a considerable amount of research has been done in the topic of trajectory indexing, including some more suitable for DBMS integration, e.g., [8, 9], based on the well-known R-tree.

All those need to be considered, but it is likely that a new indexing structure, based on the semantics of the application at hand, need to be designed. By semantics of the application we note that not only high scalability, a typical requirement, is important, but sustainable high throughput is also essential. A possible venue for work, that is relatively underexplored, is related to the use of Flash-based Disks (SSDs) [7].

# 3. ENTER THE CLOUD …

All tasks above can be accomplished in a "typical" centralized computing model. However, it is only natural to (re)consider all of them within the realm a computing cloud. For instance, one possible research venue is the massive parallelization of the proposed indices, e.g., using the Map-Reduce paradigm, e.g., [3].

Cloud storage itself is another issue to be considered. One cannot ignore that, underlying the development of all tasks above, one must explicitly address the fact that fairly large and dynamic trajectory datasets will need to be stored, as well as manipulated in order to facilitate the clustering (mining) and indexing tasks. Although one can initially envision most of the processing being done offline, a practical system needs to be able to ingest streaming data to better accommodate changes in the identified patterns and/or in the events of interest in a timely and likely distributed manner.

# 4. FURTHER WORK

There are several directions in which the research outlined here can be extended. For instance, it would be also interesting to determine common sub-trajectories among STTPs. This would be of use, for instance, for city officials interested in minimizing traffic disruption in case of road construction or in order to maximize the return of expansion investments. Another application would be choosing time and location for road-side electronic advertisement. These types of applications would greatly benefit from the framework developed for the research being suggested here.

Another direction for work, although not directly related to the above application, is the use of STTPs for hop-wise routing of queries and their answers; some works have already considered using user-carried mobile devices as sensors, e.g., [9], which can be seen as either data sources and sinks. For instance, one can issue a query about a given location, and obtain readings from a sensor that has just been (or will shortly be) at the location of interest. This would ensure a fresh answer at relatively low cost provided that the user is able to tolerate some query latency. We are currently using STTPs for finding encounter patterns and then use those patterns for optimizing energy cost or data latency in hop-wise query/data routing.

Finally, a perhaps more visionary idea is to use mobile devices, e.g., smartphones, themselves as components of a cloud. In fact, considering for instance that Cisco estimates that there will be "more than 7 billion mobile devices globally by 2015" and considering that the "total internet traffic will more than quadruple by 2014"[2], this is a path to be considered. Certainly much more progress is needed towards energy management, privacy and data security, but with the ever-increasing capabilities of such devices, this possibility should not be overlooked.

# 5. REFERENCES


[1] N.J. Belkin et al: Information Filtering and Information Retrieval: Two Sides of the Same Coin? CACM 35(12): 29-38 (1992).

[2] A. Brilingaite et al: Enabling routes as context in mobile services. GIS 2004: 127-136.

[3] A. Cary et al: Experiences on Processing Spatial Data with MapReduce. SSDBM 2009: 302-319

[4] Y. Chen et al: SpADe: On Shape-based Pattern Detection in Streaming Time Series. ICDE 2007: 786-795.

[5] F. Giannotti et al: Trajectory pattern mining. KDD 2007: 330-339.

[6] H. Jeung et al: Discovery of convoys in trajectory databases. PVLDB 1(1): 1068-1080 (2008).

[7] M. Sarwat et al: FAST: A Generic Framework for Flash-Aware Spatial Trees. SSTD 2011: 149-167.

[8] D. Pfoser, C.S. Jensen: Trajectory Indexing Using Movement Constraints. GeoInformatica 9(2): 93-115 (2005).

[9] S. Rasetic, et al: A Trajectory Splitting Model for Efficient Spatio-Temporal Indexing. VLDB 2005: 934-945.

[10] Sasank R. et al: MobiSense - mobile network services for coordinated participatory sensing. ISADS 2009: 231-236.


[2] http://www.telegraph.co.uk/technology/internet/9051590/50-billion-devices-online-by-2020.html