Spatio–Temporal Data Mining for Location–Based Services

Ph.D. Thesis by Győző Gidófalvi (© 2007)

Abstract: Location–Based Services (LBS) are continuously gaining popularity. Innovative LBSes integrate knowledge about the users into the service. Such knowledge can be derived by analyzing the location data of users. Such data contain two unique dimensions, space and time, which need to be analyzed. The objectives of the thesis are three–fold. First, to extend popular data mining methods to the spatio–temporal domain. Second, to demonstrate the usefulness of the extended methods and the derived knowledge in promising LBS examples. Finally, to eliminate privacy concerns in connection with spatio–temporal data mining by devising systems for privacy–preserving location data collection/mining.

To this extent, first, a general methodology, pivoting, is described to extend a popular data mining method, namely rule mining, to the spatio–temporal domain. By considering the characteristics of a number of real–world data sources, a taxonomy of spatio–temporal data is derived, and the usefulness of the rules that the extended spatio–temporal rule mining method can discover is demonstrated. The proposed spatio–temporal extension is applied to find long, sharable patterns in trajectories of moving objects. Empirical evaluations show that the extended method and its variants, using high–level SQL implementations, are effective tools for analyzing trajectories of moving objects.

Since real–world trajectory data about a large population of objects moving is difficult to obtain, to aid the development in spatio–temporal data management and data mining, next the development of a Spatio–Temporal ACTivity Simulator (ST–ACTS) is described. ST–ACTS uses a number of real–world geo–statistical data sources and intuitive principles to effectively generate realistic spatio–temporal activities of mobile users.

Motivated by the concept of "sharable patterns," next an LBS in the transportation domain, namely cab–sharing is proposed. A unique spatio–temporal grouping algorithm is presented and implemented as a sequence of SQL statements. To eliminate a scalability bottleneck in the grouping algorithm, the grouping algorithm is expressed as a continuous stream query in a data stream management system, and computation is parallelized through simple but effective spatio–temporal partitioning methods for streams. Experimental results show that parallelization through adaptive partitioning methods leads to speed–ups of orders of magnitude without significantly affecting the quality of the grouping.

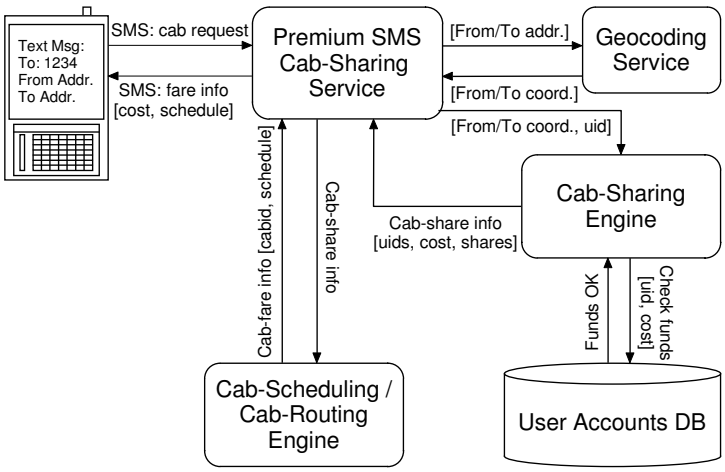

Figure 1: Cab–Sharing Service Components and Process.

When data about users is collected and analyzed, privacy naturally becomes a concern. To eliminate the concerns, two systems for the privacy–preserving collection and mining of trajectories are described. The first system uses a grid–based framework to anonymize location data through spatio–temporal generalization, and extracts probabilistic patterns from anonymous location data. Experimental results show that the privacy-preserving data mining component discovers patterns that, while probabilistic, are accurate enough to be useful for many LBSes. To eliminate any uncertainty in the mining results, the second system collects exact trajectories of moving objects in a privacy-preserving manner. In the collection system there are no trusted components and anonymization is performed in a distributed fashion by the clients in a P2P network via data cloaking and data swapping. Realistic simulations show that under reasonable conditions and privacy/anonymity settings the collection system is effective.