博士論文 (要約)

Mobility Data Analysis Using Latent Variable Models

(潜在変数モデルを用いた空間移動データ分析)

Akira Kinoshita

木下 僚

A Doctoral Dissertation (An Abridged Version)

Submitted in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy (Information Science and Technology)

Graduate School of Information Science and Technology
The University of Tokyo

December 2016

The development of information and communication technologies has introduced several massive continuous real-world data streams. Faced with such varied and enormous data, social cyber–physical systems (social CPSs), which discover real-world situations from data and give precise feedback to the real world through information services and precise equipment control, are a desirable way to realize efficient, advanced urban functions and human societies. It is essential for a social CPS to acquire the real-world situation or latent semantics from the data. Because the real world consists of people, information about their movements is quite fundamental and useful in understanding what is happening in the real world. Mobility data, such as probe-car data (PCD) and smartphone location data, record the real-world movements of people and things, and are crucial for social CPSs. It is sometimes infeasible to observe such situations objectively; however, information about a situation should be latent in the observed data because the movements of people are influenced by real-world situations such as traffic, weather, and events. Situation awareness using mobility data will allow applications such as intelligent information services, automatic incident detection (AID) and alert, and optimization of road management. From a data analysis perspective, social CPSs may be regarded as allowing situation-aware anomaly detection and its applications.

Realistically, the analyses for new real-world tasks are conducted in an unsupervised, data-driven, and exploratory manner. This kind of analysis processes the given data comprehensively and applies methods of analysis by trial and error, allowing the properties of the analysis subject and the essence of the problem to become clear and higher-level information is extracted. In this study, real-world knowledge is acquired using a data-driven approach, in contrast with the knowledge-driven approach that exploits expert knowledge about the subject of the analysis. Analysis of latent information that does not appear in data explicitly has been studied as latent semantic analysis (LSA) in natural language processing: a document is simply textual data that consists of word sequences, and the semantic content or topic is latent and must be extracted. Topic modeling, which involves latent variables, has achieved success in LSA because it allows flexible model extensions and gives interpretable results. This dissertation presents a methodology for latent information analysis using latent variable models and applies it to three empirical analysis tasks to show the effectiveness of this approach for data-driven analyses of mobility data for traffic situation awareness and social CPS applications.

In recent years, much effort has been invested in discovering patterns and semantics of mobility data and their application. Chapter 2 presents an overview of the research field and the mobility data analysis tasks, which are related to social CPS and are tackled in this dissertation: AID on expressways, weather–traffic relationship analysis, and smartphone location data enrichment, and reviews the related work. Whereas existing work

exploited heuristics based on careful observations of real-world traffic, this study takes a data-driven approach using latent variable models for the discovery of latent situation information. This approach avoids heuristics and will adapt automatically to changes in traffic circumstances, which should reduce the costs of preparing data, minimize human error, and find knowledge that may otherwise be overlooked. The methodology of mobility data analysis using latent variable models is presented in Chapter 3. After topic modeling studies are reviewed, the methodology is formulated by generalizing the concept of topic modeling for data other than texts. This methodology is consistently applied throughout three empirical case studies in Chapters 4–6.

Chapter 4 describes the first case study: the analysis of PCD on the Shuto Expressway system (Tokyo Metropolitan Expressway, Tokyo, Japan). Although we can identify traffic states such as "smooth" and "congested," we cannot measure or observe such states directly or objectively. A traffic state model based on latent Dirichlet allocation (LDA), the simplest topic model, is introduced to describe the traffic situation of monitored traffic. In the model, the traffic states are considered to be latent information and to correspond to topics in LDA, whereby traffic states are determined automatically based on the training data and usual traffic behavior is learned. The proposed AID method evaluates the difference between a current traffic state and the usual one based on the model with a streaming algorithm, whereby sudden or unusual traffic events, i.e., incidents, are detected automatically. Given an observed data point, the system estimates the usual and current traffic states based on the learned model, and then evaluates the difference between the two states quantitatively to test whether it is anomalous. The experiment using real PCD showed that the proposed model represented the distribution of large amount of data compactly and revealed potential chronic bottleneck segments. The results also showed that the proposed AID method could quickly discriminate trajectories affected by incidents from other trajectories. The design and implementation of data processing systems that can perform the AID method in real time are discussed. They enable monitoring of an incident from beginning to end.

Chapter 5 describes the second case study: the analysis of PCD on arteries in the city of Sapporo, Hokkaido, Japan, which experiences heavy snowfalls during winter that severely affect vehicular traffic. As a solution, a weather-aware traffic state (WATS) model is proposed, which extends the traffic state model from the first case study and describes traffic observation data using both traffic and weather conditions. An analysis method to find "weather-sensitive" road segments based on the estimated model and predictive distributions is proposed. An empirical, fully data-driven, exploratory analysis of winter traffic in the city of Sapporo was conducted by applying the WATS model to real-world weather data and PCD. The experimental results showed that the relationship between weather and traffic as well as the potential bottleneck segments

and weather conditions could be extracted from the data.

Chapter 6 describes the third case study: the analysis of smartphone location data. This kind of data is different from PCD in that the mode of transportation is not specified. In this analysis, a latent variable model is proposed, which describes the relationship between transportation modes and observed movement data values. The mode of transportation is regarded as latent information and is extracted from the data; the flow of people is then modeled in view of the mode, and a latent variable model is developed to describe trajectories with their modes. Based on this information, an interpolation method for trajectories from the data of their origin, destination, and travel time was developed. The experiment using real-world data from Beijing, China, showed that different movement patterns could be extracted for each transportation mode. The actual trajectories traveled were correctly reproduced in more than three-quarters of the locations, which suggests the significance of the latent information for effective interpolation and prediction of trajectory data.

The three empirical case studies are reviewed and discussed in Chapter 7, concluding this dissertation.

The social world was studied using latent information analysis and latent variable models, and this study explored the extraction and utilization of latent, semantic, contextual, and situational information behind mobility data. The analysis methodology was applied consistently throughout the case studies to develop latent variable models according to the analysis objectives. This method analyzes given data statistically and develops probabilistic mixture models for them. The model can be developed flexibly with or without expert knowledge by considering its graphical representation, enabling unsupervised, data-driven, exploratory latent information analysis. The progress of technology and the sophistication of human society will lead to real-world data and analysis tasks that are more diverse and complex. Even in such situations, latent variable models should facilitate our data-driven comprehension of the social world.