DASHTrails: An Approach for Modeling and Analysis of Distribution-Adapted Sequential Hypotheses and Trails

Martin Atzmueller and Andreas Schmidt and Mark Kibanov
University of Kassel, Research Center for Information System Design
Wilhelmshöher Allee 73, D-34121 Kassel, Germany
{atzmueller, schmidt, kibanov}@cs.uni-kassel.de

ABSTRACT

The analysis of sequential trails and patterns is a prominent research topic. However, typically only explicitly observed trails are considered. In contrast, this paper proposes the *DASHTrails* approach that enables the modeling and analysis of distribution-adapted sequential trails and hypotheses. It presents a method for deriving transition matrices given a probability distribution over certain events. We demonstrate the applicability of the proposed approach using real-world data in the mobility domain, i. e., car trajectories and spatio-temporal distributions on car accidents.

Keywords

behavioral analytics, sequential hypothesis, human trails, mobility, social media, social network analysis

1. INTRODUCTION

The analysis of human behavior is a prominent topic in web and network science, e. g., for analyzing human navigation trails on the web or for exploring movement patterns in mobile and spatio-temporal applications. The HypTrails approach [18], for example, allows the comparison of hypotheses with such trails, for identifying the hypotheses that show the largest evidence concerning the observed data. However, the approach considers *explicitly observed* trails, e. g., navigational trails in social online systems. In contrast, this paper outlines an approach for the extended modeling and analysis of sequential hypotheses and trails, i. e., by deriving according transition matrices in a distribution-adapted approach. Then, we can analyze, e. g., geo-tagged datasets or (social) network data, with a probability distribution assigned to the data points and nodes of the network, respectively. There is a vast range of possible application areas. These include, e. g., the derivation and analysis of paths in mobile applications and social software, as well as analyzing complex heterogeneous networks in industrial plants, where e. g., connections between sensors (assets), events in alarm logs, and human (operator) actions can be investigated.

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890082.

Objectives. Understanding the influence factors for trail-related events is important, e. g., for predictive modeling. This paper provides a systematic approach for modeling sequential trails and hypotheses – as a set of transitions between discrete states represented as transition matrices – given a probability distribution on these states. We can then model, e. g., (human) location-based indicators, geo-tagged datasets, or time-stamped events on complex networks, using the respective distributions. Specifically, we can model both *observed* and *derived* transition matrices and analyze them in a unified framework, given by the DASHTrails approach. In this paper, we exemplify the approach presenting first experiments that focus on the real world problem of understanding trail-related effects on car accidents, which can, e. g., be beneficial for resource planning and load balancing in health care, dynamic pricing for insurance companies as well as road planning and traffic control.

Contribution. Our contribution is summarized as follows:

1. We propose a systematic method for the modeling and analysis of sequential trails and hypotheses that are *derived* from *observed data*, i. e., estimated given a probability distribution of certain events. This enables a data-driven approach for analysis, comparison and assessment of sequential trails and hypotheses.
2. We present a modeling method that applies a certain process interpretation for deriving transition matrices, e. g., based on flow-based mechanisms: The transition matrices incorporate thus a certain interpretation of the data towards a sequential representation. This enables a comprehensive analysis approach, e. g., by estimating and comparing evidence for the hypotheses induced by a set of influence factors, statistically grounded utilizing Bayesian estimation techniques.
3. We demonstrate the applicability of the proposed approach and provide first results using real-world data in the context of the Telecom Italia Big Data Challenge 2015[1]: We model human sequential trails given a spatio-temporal distribution on car accident claim data in seven different cities in Italy. Hypotheses include one given observed car trajectories.

The remainder of the paper is organized as follows: Section 2 discusses related work. After that, we introduce the proposed approach in Section 3. Section 4 presents the results using real world mobility data from the Big Data Challenge 2015. Finally, Section 5 concludes the paper with a discussion and some interesting directions for future work.

[1] `http://www.telecomitalia.com/tit/en/bigdatachallenge/`