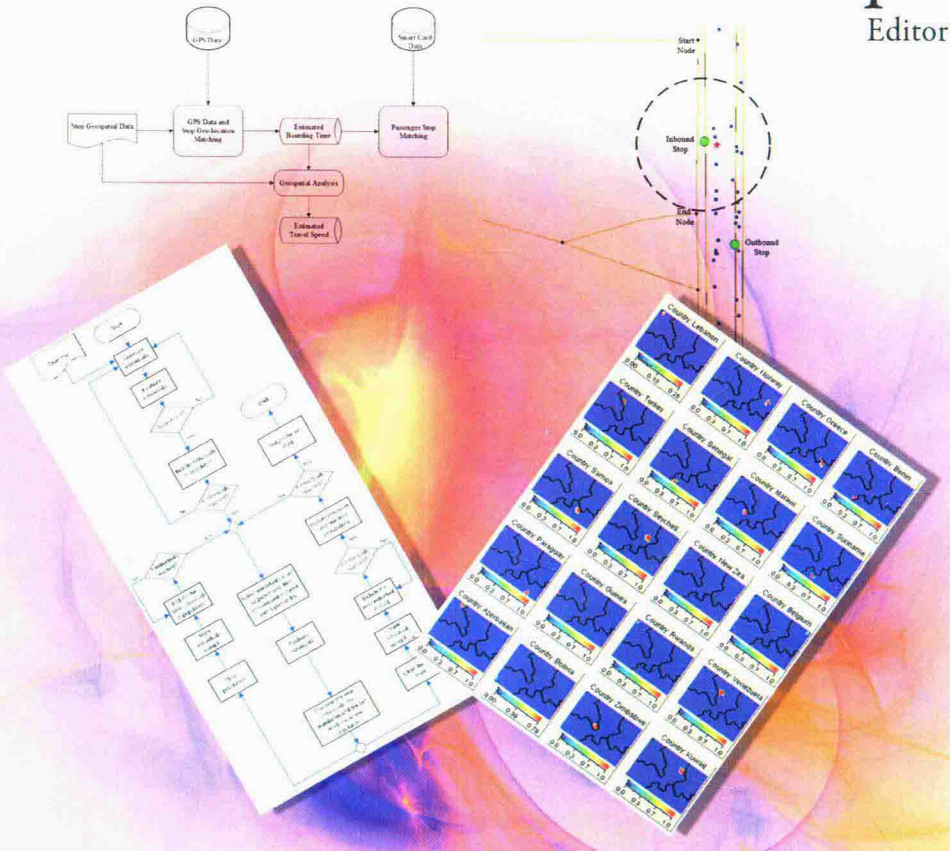
Computer Science, Technology and Applications

Data Mining

Principles, Applications and Emerging Challenges

Harold L. Capri
Editor



NOVA

COMPUTER SCIENCE, TECHNOLOGY AND APPLICATIONS

DATA MINING

PRINCIPLES, APPLICATIONS AND EMERGING CHALLENGES

HAROLD L. CAPRI.
EDITOR

nova publishers
New York

Library of Congress Cataloging-in-Publication Data

Data mining (Nova Science Publishers)
Data mining : principles, applications and emerging challenges / [edited by] Harold L. Capri.
pages cm. -- (Computer science, technology and applications)
Includes bibliographical references and index.
ISBN 978-1-63463-738-1 (softcover)
1. Data mining. I. Capri, Harold L., editor. II. Ma, Xiaolei, 1985- Transit passenger origin inference using smart card data and GPS data. III. Title.
QA76.9.D343D365 2014
006.3'12--dc23
2014047181

Published by Nova Science Publishers, Inc. † New York

COMPUTER SCIENCE, TECHNOLOGY AND APPLICATIONS

DATA MINING

PRINCIPLES, APPLICATIONS AND EMERGING CHALLENGES

COMPUTER SCIENCE, TECHNOLOGY AND APPLICATIONS

Additional books in this series can be found on Nova's website under the Series tab.

Additional e-books in this series can be found on Nova's website under the e-books tab.

PREFACE

Data mining is an area of research where appropriate methodological research and technical means are experienced to produce useful knowledge from different types of data. Data mining techniques use a broad family of computationally intensive methods that include decision trees, neural networks, rule induction, machine learning and graphic visualization. This book discusses the principles, applications and emerging challenges of data mining.

Chapter 1 - To improve customer satisfaction and reduce operation costs, transit authorities have been striving to monitor their transit service quality and identify the key factors to attract the transit riders. Traditional manual data collection methods are unable to satisfy the transit system optimization and performance measurement requirement due to their expensive and labor-intensive nature. The recent advent of passive data collection techniques (e.g., Automated Fare Collection and Automated Vehicle Location) has shifted a data-poor environment to a data-rich environment, and offered the opportunities for transit agencies to conduct comprehensive transit system performance measures. Although it is possible to collect highly valuable information from ubiquitous transit data, data usability and accessibility are still difficult. Most Automatic Fare Collection (AFC) systems are not designed for transit performance monitoring, and additional passenger trip information cannot be directly retrieved. Interoperating and mining heterogeneous datasets would enhance both the depth and breadth of transit-related studies. This study proposed a series of data mining algorithms to extract individual transit rider's origin using transit smart card and GPS data. The primary data source of this study comes from the AFC system in Beijing, where a passenger's boarding stop (origin) and alighting stop (destination) on a flat-rate bus are not recorded

on the check-in and check-out scan. The bus arrival time at each stop can be inferred from GPS data, and individual passenger's boarding stop is then estimated by fusing the identified bus arrival time with smart card data. In addition, a Markov chain based Bayesian decision tree algorithm is proposed to mine the passengers' origin information when GPS data are absent. Both passenger origin mining algorithms are validated based on either on-board transit survey data or personal GPS logger data. The results demonstrate the effectiveness and efficiency of the proposed algorithms on extracting passenger origin information. The estimated passenger origin data are highly valuable for transit system planning and route optimization.

Chapter 2 - Differentiation between learners, adapted and personalized learning are interesting research directions on technology for human learning today. This issue leads to the design of educational systems integrating strategies for learners' monitoring to assist each by evaluating his knowledge and skills in one hand and detecting and analyzing his errors and obstacles in the other hand. In this respect, formative evaluation is the process used to capture data on the strengths and weaknesses of a learner. These data, to be useful, must be objectively analyzed so that it can be used to manage the following sessions. There are different data mining tools using different algorithms for data analysis and knowledge extraction. Can we use these tools in computer based systems? In such cases, is it possible to directly use a variety of general-purpose algorithms for learning data analysis? The authors discuss in this paper a learning cycle that can be a learning session with feedback loop integrating formative evaluation followed by knowledge extraction process using data mining algorithms. The author's experiments, presented in this work, shows a set of tests, about the exploration of learners' errors, obtained from a self e-learning by doing tool for the algorithmic domain. The authors used the data mining algorithms implemented in the Weka tool: the C4.5 algorithm for classification, A Priori one for association rules deduction and K-Means for clustering. The results given by these experiments have proved the interest of classification and clustering as implemented in Weka. However, the A priori algorithm gives in some cases results difficult to interpret so that it needs a specific optimization to get adequate frequents detection.

Chapter 3 - Since the concept of 'failed states' was coined in the early 1990s, it has come to occupy a top tier position in the international peace and security's agenda. This study uses data mining techniques to examine the effect of various social, economic and political factors on states' failure at the global level. Data mining techniques use a broad family of computationally

intensive methods that include decision trees, neural networks, rule induction, machine learning and graphic visualization. Three artificial neural network models: multi-layer perceptron neural network (MLP), radial basis function neural network (RBFNN) and self-organizing maps neural network (SOM) and one machine learning technique (support vector machines [SVM]) are compared to a standard statistical method (linear discriminant analysis (LDA). The variable sets considered are demographic pressures, movement of refugees, group paranoia, human flight, regional economic development, economic decline, de-legitimization of the state, public services' performance, human rights status, security apparatus, elites' behavior and the role played by other states or external political actors. The study shows how it is possible to identify various dimensions of states' failure by uncovering complex patterns in the dataset, and also shows the classification abilities of data mining techniques.

Chapter 4 - This paper presents a novel self-adaptive grammar-guided genetic programming proposal for mining association rules. It generates individuals through a context-free grammar, which allows of defining rules in an expressive and flexible way over different domains. Each rule is represented as a derivation tree that shows a solution (described using the language) denoted by the grammar. Unlike existing evolutionary algorithms for mining association rules, the proposed algorithm only requires a small number of parameters, providing the possibility of discovering association rules in an easy way for non-expert users. More specifically, this algorithm does not require any threshold, and uses a novel parent selector based on a niche-crowding model to group rules. This approach keeps the best individuals in a pool and restricts the extraction of similar rules by analysing the instances covered. The author's compare our approach with the G3PARM algorithm, the first grammar-guided genetic programming algorithm for the extraction of association rules. G3PARM was described as a high-performance algorithm, obtaining important results and overcoming the drawbacks of current exhaustive search and evolutionary algorithms. Experimental results show that the author's new proposal obtains very interesting and reliable rules with higher support values.

CONTENTS

In: Data Mining
Editor: Harold L. Capri
ISBN: 978-1-63463-738-1

Chapter 1

TRANSIT PASSENGER ORIGIN INFERENCE USING SMART CARD DATA AND GPS DATA

Xiaolei Ma*[*1], Ph.D. and Yinhai Wang[2], Ph.D.
[1]School of Transportation Science and Engineering,
Beihang University, Beijing, China
[2]Department of Civil and Environmental Engineering,
University of Washington, Seattle, WA, US

ABSTRACT

To improve customer satisfaction and reduce operation costs, transit authorities have been striving to monitor their transit service quality and identify the key factors to attract the transit riders. Traditional manual data collection methods are unable to satisfy the transit system optimization and performance measurement requirement due to their expensive and labor-intensive nature. The recent advent of passive data collection techniques (e.g., Automated Fare Collection and Automated Vehicle Location) has shifted a data-poor environment to a data-rich environment, and offered the opportunities for transit agencies to conduct comprehensive transit system performance measures. Although it is possible to collect highly valuable information from ubiquitous transit data, data usability and accessibility are still difficult. Most Automatic Fare Collection (AFC) systems are not designed for transit performance monitoring, and additional passenger trip information cannot be directly

* Email: xiaolm@uw.edu

retrieved. Interoperating and mining heterogeneous datasets would enhance both the depth and breadth of transit-related studies. This study proposed a series of data mining algorithms to extract individual transit rider's origin using transit smart card and GPS data. The primary data source of this study comes from the AFC system in Beijing, where a passenger's boarding stop (origin) and alighting stop (destination) on a flat-rate bus are not recorded on the check-in and check-out scan. The bus arrival time at each stop can be inferred from GPS data, and individual passenger's boarding stop is then estimated by fusing the identified bus arrival time with smart card data. In addition, a Markov chain based Bayesian decision tree algorithm is proposed to mine the passengers' origin information when GPS data are absent. Both passenger origin mining algorithms are validated based on either on-board transit survey data or personal GPS logger data. The results demonstrates the effectiveness and efficiency of the proposed algorithms on extracting passenger origin information. The estimated passenger origin data are highly valuable for transit system planning and route optimization.

Keywords: Automated fare collection system, transit GPS, passenger origin inference, Bayesian decision tree, Markov chain

INTRODUCTION

According to the Census of 2000 in the United States, approximately 76% people chose privately owned vehicles to commute to work in 2000 (ICF consulting, 2003). Recent studies conducted by the 2009 American Community Survey indicate 79.5% of home-based workers drive alone for commuting (McKenzie and Rapino, 2009). Many developing countries, e.g., China, also rely on privately owned vehicles to commute. For example, more than 34% of the Beijing residents chose cars as their primary travel mode while only 28.2% chose transit in 2010 (Beijing Transportation Research Center, 2012). Public transit has been considered as an effective countermeasure to reduce congestion, air pollution, and energy consumption (Federal Highway Administration, 2002). According to 2005 urban mobility report conducted by Texas Transportation Institute (2005), travel delay in 2003 would increase by 27 percent without public transit, especially in those most congested metropolitan cites of U.S., public transit services have saved more than 1.1 billion hours of travel time. Moreover, public transit can help enhance business, reduce city sprawl through the transit oriented development (TDO). During certain emergency scenarios, public transit can even act as a

safe and efficient transportation mode for evacuation (Federal Highway Administration, 2002). Based on the aforementioned reasons, it is of critical importance to improve the efficiency of public transit system, and promote more roadway users to utilize public transit. To fulfill these objectives, transit agencies need to understand the areas where improvements can be further made, and whether community goals are being met, etc. A well-developed performance measure system will facilitate decision making for transit agencies. Transit agencies can evaluate the transit ridership trends with fare policy changes and identify where and when better transit service should be provided. In addition, transit agencies are also required to summarize transit performance statistics for reporting to either the National Transit Database (Kittelson & Associates et al., 2003), or the general public who are interested knowing how well transit service is being provided. Nevertheless, developing a set of structured performance measures often requires a large amount of data and the corresponding domain knowledge to process and analyze these data. These obstacles create challenges for transit agencies to spend time and effort undertaking. Traditionally, transit agencies heavily rely on manual data collection methods to gather transit operation and planning data (Ma et al., 2012). However, traditional data collection methods (e.g., travel diary, survey, etc.) are fairly costly and difficult to implement at a multiday level due to their low response rate and accuracy. Transit agencies have spent tremendous manpower and resource undertaking manual data collections, and consumed a significant amount of energy and time to post-process the raw data. With advances in information technologies in intelligent transportation systems (ITS), the availability of public transit data has been increasing in the past decades, which has gradually shifted public transit system into a data-rich paradigm. Automatic Fare Collection (AFC) system and Automatic Vehicle Track (AVL) system are two common passive data collection methods. AFC system, also known as Smart Card system, records and processes the fare related information using either contactless or contact card to complete the financial transaction (Chu, 2010). There exist two typical types of AFC systems: entry-only AFC system and distance-based AFC system. In the entry-only AFC system, passengers are only required to swipe their smart cards over the card reader during boarding, while passengers need to check in and check out during both their boarding and alighting procedures for the distance-based AFC system. AVL and AFC technologies hold substantial promise for transit performance analysis and management at a relative low cost. However, historically, both AVL and AFC data have not been used to their full potentials. Many AVL and AFC systems do not archive data in a readily

utilized manner (Furth, 2006). AFC system is initially designed to reduce workloads of tedious manual fare collections, not for transit operation and planning purposes, and thereby, certain critical information, such as specific spatial location for each transaction, may not be directly captured. AVL system tracks transit vehicles' geospatial locations by Global Positioning System (GPS) at either a constant or varying time interval. The accuracy of GPS occasionally suffers from signal loss due to tall building obstructions in the urban area (Ma et al., 2011). Both of the AFC system and AVL system have their inherent drawbacks in monitoring transit system performance, and require analytical approaches to eliminate the erroneous data, remedy the missing values, and mine the unseen and indirect information.

The remainder of this paper is organized as follows: transit smart card data and GPS data are described in the section 2. Based on these data sets, a data fusion method is initially proposed to integrate with roadway geospatial data to estimate transit vehicles arrival information. And then, a Bayesian decision tree algorithm is presented to estimate each passenger's boarding stop when GPS data are unavailable. Considering the expensive computational burden of decision tree algorithms, Markov-chain property is taken into account to reduce the algorithm complexity. On-board survey and GPS data from the Beijing transit system are used to test and verify the proposed algorithms. Conclusion and future research efforts are summarized at the end of this paper.

RESEARCH BACKGROUND

Data from AFC system and AVL system are the two primary sources in this study. Beijing Transit Incorporated began to issue smart cards in May 10, 2006. The smart card can be used in both the Beijing bus and subway systems. Due to discounted fares (up to 60% off) provided by the smart card, more than 90% of the transit riders pay for their transit trips with their smart cards in 2010 (Beijing Transportation Research Center, 2010). Two types of AFC systems exist in Beijing transit: flat fare and distance-based fare. Transit riders pay at a fixed rate for those flat fare buses when entering by tapping their smart cards on the card reader. Thus, only check-in scans are necessary. For the distance-based AFC system, transit riders need to swipe their smart cards during both check-in and check-out processes. Transit riders need to hold their smart cards near the card reader device to complete transactions when entering or exiting buses. Smart card can be used in Beijing subway system as well, where passengers need to tap their smart card on top of fare gates during

entering and existing subway stations. Both boarding and alighting information (time and location) are recorded by the fare gates. Although transit smart card exhibits its superiority on its convenience and efficiency, there are still the following issues to prevent transit agencies fully taking advantages of smart card for operational purposes:

- Passenger boarding and alighting information missing

Due to a design deficiency in the smart card scan system, the AFC system on flat fare buses does not save any boarding location information, whereas the AFC system stores boarding and alighting location, except for boarding time information on distance-based fare buses. Key information stored in the database includes smart card ID, route number, driver ID, transaction time, remaining balance, transaction amount, boarding stop (only available for distance-based fare buses), and alighting stop (only available for distance-based fare buses).

- Massive data sets

More than 16 million smart card transactions data are generated per day. Among these transactions, 52% are from flat-rate bus riders. These smart card transactions are scattered in a large-scale transit network with 52386 links and 43432 nodes as presented in figure 1:

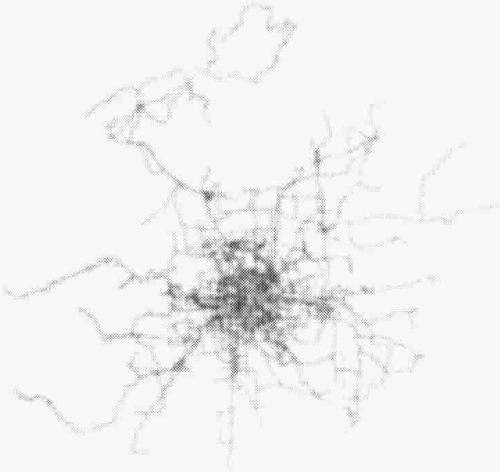


Figure 1. Beijing Transit GIS Network.

- Limited external data with poor quality

Only approximate 50% of transit vehicles in Beijing are equipped with GPS devices for tracking. GPS data are periodically sent to the central server at a pre-determined interval of 30 seconds. However, the collected GPS data suffer from two major data quality issues: (1) vehicle direction information is missing; (2) GPS points fluctuation (Lou, et al., 2009). Map matching algorithms are needed to align the inaccurate GPS spatial records onto the road network. In addition, most of transit routes are not designed to have fixed schedules because of high ridership demands, and only certain routes with a long distance or headway follow schedules at each stop (Chen, 2009). The above characteristics of the Beijing AFC and AVL systems create more challenges to process and mine useful information.

It is noteworthy that the AFC system used in Beijing is not a unique case. Most cities in China also employ the similar AFC system where passengers' origin information is absent, such as Chongqing City (Gao and Wu, 2011), Nanning City (Chen, 2009), Kunming City (Zhou et al., 2007). In other developing countries, such as Brazil, AFC system does not record any boarding location information as well (Farzin, 2008). Therefore, a solution for passenger boarding and alighting information extraction is beneficial to those transit agencies with imperfect SC data internationally.

TRANSIT PASSENGER ORIGIN INFERENCE

Because smart card readers in the flat-rate buses do not record passengers' boarding stops, it is desired to infer individual boarding location using smart card transaction data. In this section, two primary approaches are presented to achieve this goal. Approximately 50% transit vehicles are equipped with GPS devices in Beijing entry-only AFC system. Therefore, a data fusion method with GPS data, smart card data and GIS data is firstly developed to estimate each bus's arrival time at each stop and infer individual passenger's boarding stop. And then, for those buses without GIS devices, a Bayesian decision tree algorithm is proposed to utilize smart card transaction time and apply Bayesian inference theory to depict the likelihood of each possible boarding stop. In order to expand the usability of proposed Bayesian decision tree algorithm in large-scale datasets, Markov chain optimization is used to reduce the algorithm's computational complexity. Both two transit passenger origin

inference algorithms are validated using external data (e.g., on-board survey data and GPS data).

Passenger Origin Inference with GPS Data

In the first step, a GPS-based arrival information inference algorithm is presented to estimate the arrival time for each transit stop, and then, the inferred stop-level arrival time will be matched with the timestamp recorded in AFC system. The temporally closest smart card transaction record will be assigned with each known stop ID. The logic flow chart is demonstrated in Figure 2. The major data processing procedure will be detailed below.

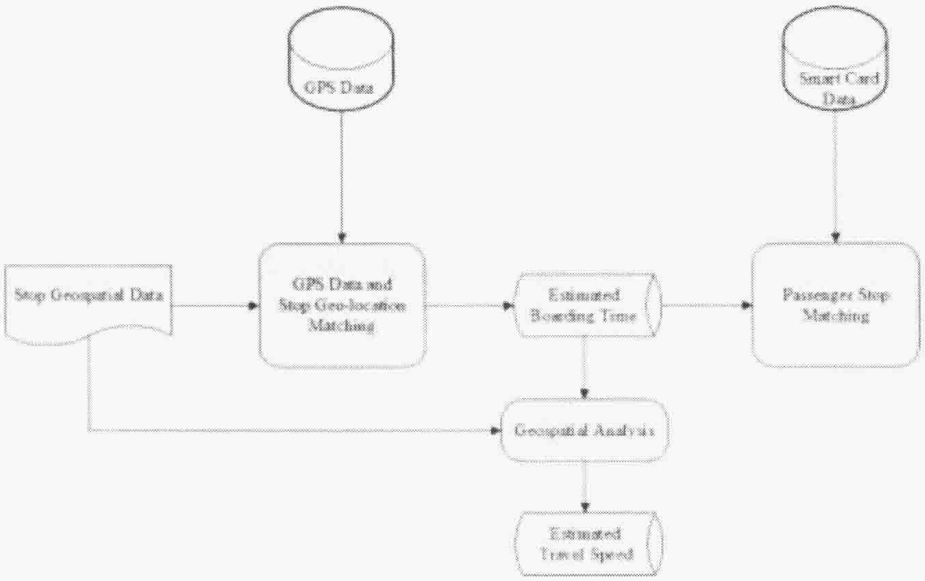


Figure 2. Flow Chart for Passenger Origin Inference with GPS Data.

Bus Arrival Time Extraction

Three primary data sources are involved in the passenger information extraction: vehicle GPS data; transit stop spatial location data; and flat-fare-based smart card transaction data. A transit GIS network contains the geospatial location of each stop for any transit routes. The GPS device mounted in the bus can record each bus's location and timestamp every 30 seconds, but the data quality of collected GPS records is not satisfying: No directional information is recorded in Beijing AVL system; GPS points are off

the roadway network due to the satellite signal fluctuation. Data preprocessing is required prior to bus arrival time estimation. A program is written to parse and import raw GPS data into a database in an automatic manner. Key fields of a GPS record are shown in Table 1.

Table 1. Examples of GPS raw data

Vehicle ID	Date time	Latitude	Longitude	Spot speed	Route ID
00034603	2010-04-07 09:28:57	39.73875	116.1355	9.07	00022
00034603	2010-04-07 09:29:27	39.73710	116.1358	14.26	00022
00034603	2010-04-07 09:29:58	39.73592	116.1357	19.63	00022
00034603	2010-04-07 09:30:28	39.73479	116.1357	0	00022
00034603	2010-04-07 09:30:58	39.73420	116.1357	3.52	00022

The first step is to estimate the bus arrival time for each stop by joining GPS data and the stop-level geo-location data. A buffer area can be created around each particular stop for a certain transit route using the GIS software. Within this area, several GPS records are likely to be captured. However, identifying the geospatially closest GPS record to each particular stop is challenging since there could be a certain number of unknown directional GPS records within the specified buffer zone. Thanks to the powerful geospatial analysis function in GIS, each link (i.e., polyline) where each transit stop is located is composed of both start node and end node, and this implies that the directional information for each GPS record is able to infer by comparing the link direction and the direction changes from two consecutive GPS records. With the identified direction, the distance from each GPS point to this particular stop can be calculated, and the timestamp with the minimum distance will be regarded as the bus arrival time at the particular stop. Figure 2 visually demonstrates the above algorithm procedure. Inbound stop represents the physical location of a particular transit stop, and this stop is snapped to a transit link, whose direction is regulated by both a start node and an end node. By comparing the driving direction from GPS records with the link direction, the nearest GPS records to this particular stop can be identified, and marked by the red five-pointed star on the map. The timestamp associated with this five-pointed star will be considered as the arrival time for this inbound stop. The