

Detecting Congestion Patterns in Spatio Temporal Traffic Data using Frequent Pattern Mining

D.Gokula priya, Dr.C.Saravanabhavan, K.Suvitha
¹M.E-II Year, ²Associate professor, ³Assistant professor

Department of Computer Science and Engineering,
Kongunadu college of engineering and technology, Tamilnadu

Abstract

Data mining is a process of extracting valuable and invaluable information from the large data base. Congestion on road is the condition in which it is characterized as slow speed and long travel time. The detection of unusual traffic patterns is an important research problem in the data mining. In this research, the detection of unusual traffic patterns based on spatio-temporal traffic data is by constructing causal congested tree and then to find the frequent sub tree, FP-Growth algorithm is used. Frequent substructures of these causality trees reveal not only recurring interactions among spatial-temporal congestions, but potential bottlenecks or flaws in the design of existing traffic networks. The FP-Growth algorithm is an efficient and scalable method for mining the complete set of frequent patterns by pattern fragment growth, using an extended prefix-tree structure for storing compressed and crucial information about frequent patterns named frequent-pattern tree

Keywords

spatio-temporal, FP-Growth, Frequent Pattern.

I. INTRODUCTION

A) Data Mining:

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational database.

B) Sequential pattern mining

Sequential pattern mining is a topic of data mining concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence. Two important sequential pattern mining algorithms are Apriori and FP-Growth algorithm. It is usually presumed that the values are discrete, and thus time series mining is closely related, but usually considered a different activity. Sequential pattern mining is a special case of structured data mining. There are several key traditional computational problems addressed within this field. These include building efficient databases and indexes for sequence information, extracting the frequently occurring patterns, comparing sequences for similarity, and recovering missing sequence members. In general, sequence mining problems can be classified as string mining which is typically based on string processing algorithms and item set mining which is typically based on association rule learning. Local process models extend sequential pattern mining to more complex patterns that can include (exclusive) choices, loops, and concurrency constructs in addition to the sequential ordering construct.

II. RELATED WORK

A. Lozano in his research said that the spatial smoothness is enforced by an additional penalty term that encourages similarity between coefficients for spatial neighbors [1]. This formulation leads to a grouped version of the so-called “elastic net” problem, for which we devise an efficient solution.

A novel framework, called Trajectory- based Path Finding (TPF)[2], which is built upon a novel algorithm named Mining-based Algorithm for Travel time Evaluation (MATE) for evaluating the travel time of a navigation path and a novel index structure named Efficient Navigation Path Search Tree (ENS-Tree) for efficiently retrieving the fastest path. With MATE and ENS-tree, an efficient fastest path finding algorithm for single destination is derived.

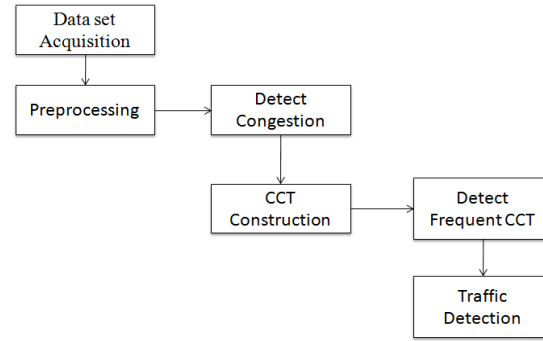
The increasing availability of large-scale trajectory data provides us great opportunity to explore them for knowledge discovery in transportation systems using advanced data mining techniques. [3]. Large scale traffic congestion often stems from local traffic jam in single road or intersection. A macroscopic method was used to explore the formation and propagation of local traffic jam[4]. The formalization of a semantic enriched KDD process is for supporting meaningful pattern interpretations of human behavior. It is based on the approach of the integration of inductive reasoning (movement pattern discovery) and deductive reasoning (human behavior inference)[5]. In order to effectively use the discovered semantic trajectory patterns, the associative classification-based event detection framework is adopted to discover the possibly occurred event[6]

Traffic flow modeled as high order Markov Chain.[7] And the transition probability from one state to the other state describes, given the current and recent values of the traffic flow, what the future value will be. Under the criteria of minimum mean square error, the optimal prediction is given as the conditional expectation according to the transition probability

Estimating and predicting traffic conditions in arterial networks using probe data have proven to be a substantial challenge. Sparse probe data represent the vast majority of the data available on arterial roads. This paper proposes a probabilistic modeling framework for estimating and predicting arterial travel time distributions using sparsely observed probe vehicles [10]

III. PROPOSED SYSTEM

A) System Architecture



B) Data Acquisition and Preprocessing

This module is used to upload the Spatial-Temporal Traffic data. The overall traffic network contains road segments where travel times of the vehicles are recorded at every fixed time interval. A spatiotemporal database is a database that manages both space and time information. Common examples include: Tracking of moving objects, which typically can occupy only a single position at a given time. A database of wireless communication networks, which may exist only for a short time span within a geographic region. An index of species in a given geographic region, where over time additional species may be introduced or existing species migrate or die out. Spatiotemporal databases are an extension of spatial databases. A spatiotemporal database embodies spatial, temporal, and spatiotemporal database concepts, and captures spatial and temporal aspects of data and deals with: geometry changing over time and/or location of objects moving over invariant geometry (known variously as moving objects databases or real-time locating systems). Preprocessing is done to remove any irrelevant data.

C) Detecting Congestions

In this module the travel time between segments are not comparable because each segment has different characteristics such as length, number of lanes and speed limits. To identify the congestions, the evaluation of real travel time was done by different percentiles of travel time range from each segment. The i th percentile is the value below which i percent of the observations may be found. In other words, a segment is considered as congested at a specific snapshot if its average travel time is longer than i th (i range from 50 to 95) percentile of overall travel time distribution. In finally the list of all congested segments in the network for each snapshot is identified and they will be used as inputs to construct the Congestion Trees.

D) Constructing Causal Congestion Trees (CCT)

It finds congestion dependencies by looking at the relationships of congestions in the datasets. The congestions in consecutive datasets were linked together to form the Causal congested trees. The next step was discovering the most frequent sub trees from the entire forest.

E) Detection of Frequent CCT

It is used to detect the frequent sub tree from the CCT (Causal congested Tree). The FP-Growth algorithm can be applied directly to identify the frequent sub tree.

F) FP- Growth Algorithm

FP-tree algorithm is used to identify frequent patterns in the area of Data Mining. In the first pass, the algorithm counts occurrence of items (attribute-value pairs) in the dataset, and stores them to header table. In the second pass, it builds the FP-tree structure by inserting instances. Items in each instance have to be sorted by descending order of their frequency in the dataset, so that the tree can be processed quickly. Items in each instance that do not meet minimum coverage threshold are discarded. If many instances share most frequent items, FP-tree provides high compression close to tree root.

IV. ALGORITHM: FP-GROWTH

Input: A database DB, represented by FP-tree constructed according to algorithm and a minimum support threshold

Output: The complete set of frequent sub tree

Method: call FP-Growth(FP-tree, null).

Procedure FP-Growth(Tree, a) {

if Tree contains a single prefix path then { // Mining single prefix-path FP-tree

 let P be the single prefix-path part of Tree;

 let Q be the multipath part with the top branching node replaced by a null root;

 for each combination (denoted as β) of the nodes in the path P do

 generate pattern $\beta \cup a$ with support = minimum support of nodes in β ;

 let freq pattern set(P) be the set of patterns so generated;

 }

else let Q be Tree;

for each item a_i in Q do { // Mining multipath FP-tree

 generate pattern $\beta = a_i \cup a$ with support = a_i .support;

 construct β 's conditional pattern-base and then β 's conditional FP-tree Tree β ;

if Tree $\beta \neq \emptyset$ then

 call FP-growth(Tree β , β);

 let freq pattern set(Q) be the set of patterns so generated;

 }

return(freq pattern set(P) \cup freq pattern set(Q) \cup (freq pattern set(P) \times freq pattern set(Q)))

}

V. CONCLUSION

The FP-Growth is more efficient when mining frequent sub tree. This technique greatly reduces the time spent on traversing FP-trees, and works especially well for sparse data sets. Furthermore, the frequent sub tree algorithm can be used to reveal recurrent in the road network. Based on the causal congested tree and FP growth algorithm we can identify real and valid instances of congestions propagations in network traffic data. In the future work, finding congestion propagation and alternate path to reduce congestion is suggested.

REFERENCE

- [1] C. Lozano, et al., "Spatial-temporal causal modeling for climate change attribution," in Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2009, pp. 587–596.
- [2] E. H.-C. Lu, W.-C. Lee, and V. S. Tseng, "Mining fastest path from trajectories with multiple destinations in road networks," Knowl. Inf. Syst., vol. 29, no. 1, pp. 25–53, 2011.
- [3] L. X. Pang, S. Chawla, W. Liu, and Y. Zheng, "On detection of emerging anomalous traffic patterns using GPS data," Data Knowl. Eng., vol. 87, pp. 357–373, 2013.
- [4] H.-S. Qi, D.-H. Wang, and P. Chen, "Formation and propagation of local traffic jam," Discrete Dynamics Nature Soc., vol. 2013, 2013, Art. no. 748529.
- [5] Renso, M. Baglioni, J. A. F. de Macedo, R. Trasarti, and M. Wachowicz, "How you move reveals who you are: Understanding human behavior by analyzing trajectory data," Knowl. Inf. Syst., vol. 37, pp. 331–362, 2012.
- [6] X. Wang, G. Li, G. Jiang, and Z. Shi, "Semantic trajectory-based event detection and event pattern mining," Knowl. Inf. Syst., vol. 37, pp. 305–329, 2011.
- [7] Zhang, S. Sun, and G. Yu, "A Bayesian network approach to time series forecasting of short-term traffic flows," in Proc. 7th Int. IEEE Conf. Intell. Transp. Syst., 2004, pp. 216–221.
- [8] Hofleitner, R. Herring, P. Abbeel, and A. Bayen, "Learning the dynamics of arterial traffic from probe data using a dynamic Bayesian network," IEEE Trans. Intell. Transp. Syst., vol. 13, no. 4, pp. 1697–1693, Dec. 2012.
- [9] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," ACM Trans. Intell. Syst. Technol., vol. 5, no. 3, 2014, Art. no. 38.
- [10] B. Zhang, C. Zhang, and X. Yi, "Competitive EM algorithm for finite mixture models," Pattern Recog., vol. 37, no. 1, pp. 131–144, 2004.
- [11] Hoang Nguyen, Wei Liu and Fang Chen, "Discovering congestion propagation in spatio-temporal traffic data", IEEE Trans. On Big Data, vol.3, No.2, April-June 2017