10th International Conference on Recent Trends in Engineering Science and Management
Newton's Institute of Science & Technology, Guntur Dist, Andhra Pradesh, India (ICRTESM-17)
12th-13th August 2017, www.conferenceworld.in ISBN: 978-93-86171-56-6

A Study of Events Logs from Elevated Volume of Logistics Information in Spatio-Temporal Data Mining

Dr.M.Mohammed Ismail[1], S.Vaitheswari[2] P.Rizwan Ahmed[3]

[1]*Associate Professor & Head, Research Department of Computer science Mazharul Uloom College, Ambur*

[2]*Research Scholar, Mazharul Uloom College, Ambur*

[3]*Asst. Professor & Head of Computer Applications and IT, Mazharul Uloom College, Ambur*

ABSTRACT

In logistics, software aids for transportation planning and scheduling are often based in approximations and abstractions that do not take into account real-world data. The aim of this work is to provide an analysis and methodology, based on real-world data, on how to obtain probability density functions for prediction of activity duration. Such information can be used in planning algorithms, like vehicle routing problem, capable of dealing with stochastic time-windows.

Given a large spatio-temporal database of events, where each event consists of the fields event ID, time, location, and event type, the aim is to extract valuable information about activities duration. The process is not straightforward since the log is human-influenced creating uncertainty related with the time at which the events are logged. In order to overcome this, a novel framework is proposed: it uses the spatiotemporal trajectories to identify regions-of-interest based on speed, and builds an ROI activity time-line using the activities extracted from event logs. The framework's ability to estimate activities durations was tested in three different environments: the Amsterdam Airport Schiphol, the Port of Rotterdam and a single vehicle scenario. Experimental results validate the usefulness of the approach at finding probability density functions for prediction of activity durations at specific locations.

Keywords: *Spatio-Temporal, Event Logs, Logistics, Event Mining, Trajectories.*

I. INTRODUCTION

Transportation companies often find that their day-to-day transportation execution does not conform to the transportation plan that they made in advance. To a large extent this is caused by the fact that the software that aids in the creation of transportation plans, does not take into account the real-world complexity of transportation and logistics [1]. Rather, it uses approximations and abstractions that do not do justice to that complexity. As a consequence, the transportation plans that are generated by transportation planning software often lead to violated time windows [2], unnecessary delays, underutilized transportation capacity, etc. The real-world complexity of transportation planning is caused by the high level of detail that is required to get executable plans, the size of the instances as found in reality, and the large volumes of data that must be collected and processed to gather the information required to create the planning [3].

ISBN: 978-93-86171-56-6

The aim of this work is to provide an analysis and methodology on how to estimate the duration of process related activities, such as load and unload, based on spatio-temporal data and event logs with uncertainty related to the human behaviour. Further, the acquired information is categorized according the location to obtain probability density functions of activity durations at specific locations. Such information can be used in software applications for transportation planning such as vehicle routing problems that can handle stochastic time-windows.

Currently, mobile communications and positioning systems are well-established technologies. GPS equipped devices are able to provide valuable spatio-temporal data with increasingly finer spatial resolutions.

The use of GPS-enabled devices allow us to describe the movement of an object (i.e. trajectory)

as a sequence of spatial locations sampled at consecutive time-stamps. Spatio-temporal patterns in trajectories, which represent movement patterns of objects, can provide useful information for high quality location based services, such as traffic flow control, location-aware advertising, etc. [4] [5] [6]. In addition, many operating systems, software frameworks, and programs include a logging system. Event logs are able to record events taking place in the execution of a system.

This provides a chronological record of a sequence of activities that is essential to gather information about complex systems. The problem is that, in many cases, events are introduced by humans, the system users, leading to the existence of uncertainty in the data. When system logs are human dependent, it is not assured that event records happened in coherence with reality. For example, if an activity is characterized by a start and an end event, its duration can be calculated as the time difference between such events. However, if the events are recorded before, or after, such occurrences the previous statement no longer holds true.

II. SEQUENTIAL PATTERN MINING

Data mining techniques, also known as knowledge discovery tools in databases, are used in order to find valid, novel, potentially useful and understandable patterns in data [7]. Sequential pattern mining, as sub-field of data mining, is a topic concerned with finding statistically relevant patterns, from frequent sub-sequences, between data examples where the values are delivered in a sequence. It's strongly motivated by it's utility as a tool to obtain knowledge from customer purchase database [8], DNA sequences [9], web logs [10], event logs and medical time series [11]. Sequences are common, occurring in any metric space that facilitates either total or partial ordering [12]. Acquiring knowledge about them is an important data mining research problem with broad applications since the detection of frequent (totally or partially ordered) sub-sequences might be extremely useful to support decision making problems. Sequential pattern mining has arisen as a technology to discover such sub-sequences.

Over the last years there has been a substantial increase in temporal, spatial and spatio-temporal data mining publications due to the continuous growth of this sub-field of data mining. The high volume of available data, through internet mainly, and the prominent advantages provided by the data mining analysis to the market, are some of the principal reasons for its development. Since there are many application domains that have a temporal or spatial context, time and space are components that must be taken into account in data mining processes. In the following paragraph temporal, spatial and spatiotemporal data mining bases are going to be briefly explained.

ISBN: 978-93-86171-56-6

III. APRIORIALL ALGORITHM

Sequential pattern mining was firstly introduced by Agrawal and Srikant [8] in 1995, over transaction databases (known as basket data). The aim was to find frequent item sets bought by costumers in order to obtain typical behaviours according to the user's point of view. This wouldz support the decision making problem faced by most large retail organizations. The sequential pattern mining problem was defined as follows:

"Given a database of sequences, where each sequence consists of a list of transactions ordered by transaction time and each transaction is a set of items, sequential pattern mining is to discover all sequential patterns with a user-specified minimum support,where the support of a sequential pattern is the percentage of data-sequences that contain the pattern." [8]

To find all sequential patterns Agrawal and Srikant divided the mining problem into five phases:

(i) Sort phase - consists in creating costumer-sequences, from the original database, by finding the transactions with the same transaction-id and ordering them according the transaction time;

(ii) Large item set phase - the item sets with the user-specified minimum support, litemsets, are found;

The support for an item set i is defined as the fraction of costumers who bought the items in i in a single transaction;

(iii) Transformation phase - each transaction in the costumer-sequences is replaced by the set off all item sets contained in that transactions;

(iv) Sequence phase - the sequence phase is the actual mining phase, where the set of litemsets is used to fin\d the desired sequences (the ones that satisfy the minimum support constraint – large sequences);

(v) Maximal phase - is to find the maximal sequences among the set of large sequences. A sequence s is said to be maximal if, in a set of sequences, s is not contained in any other sequence.

IV. GENERALIZED SEQUENTIAL PATTERNS ALGORITHM

In 1996, Srikant and Agrawal [15] generalized the problem of sequential pattern mining. Time constraints that specify a minimum and/or maximum time period between adjacent elements in a pattern were added and, to overcome the rigid definition of transaction, the restriction that the items in an element of a sequential pattern must come from the same transaction was relaxed by allowing the items to be present in a set of transactions whose transaction-times are within a user-specified time window. Hierarchy was also introduced, given a user-defined taxonomy on items, sequential patterns are allowed to include items across all levels of the taxonomy. Since many applications require all patterns and their supports, the count-some algorithms from [8], that find only maximal sequential patterns, were abandoned.

Despite of being possible to extend the AprioriAll algorithm to handle time constraints and taxonomy, incorporate sliding windows was not feasible. Apart from that, the performance of the algorithm was poor since it had to preform the data transformation on-the-fly during each pass while finding sequential patterns. Generalized Sequential Patterns (GSP) algorithm was presented and problem definition reformulated:

Given a database D of data-sequences, a taxonomy T, a user-specified min-gap and maxgap time constraints and a user-specified sliding-window size, the problem of mining sequential patterns is to find all sequences whose support is greater than the user-specified minimum support. Each sequence represents a sequential pattern, also called a frequent sequence." [15]

ISBN: 978-93-86171-56-6

The GSP algortihm, as the AprioriAll, assume a horizontal database layout, which means that the database is formed by a set of input-sequences. Each input-sequence has a set of events, along with the items contained in the event. GSP works in a multiple pass fashion over the data. The first pass determines the support of each item, that is, the number of data-sequences that include the item. The items that respect the minimum support are the frequent items. Each such item yields an element frequent sequence and on each subsequent pass the previous frequent sequences, seed set, are used to generate new potentially frequent sequences, called candidate sequences, with one more item than the seed sequences. During the pass over the data the algorithm computes the support for each one of the candidate sequences and determines which ones of them are actually frequent. These frequent candidate sequences become the seed set for the next pass. When there are no frequent sequences or no candidate sequences generated, the algorithm terminates. GSP is a complete algorithm in that it guarantees finding all sequences that have a user-specified minimum support. Further more is up to twenty times faster than the previous presented algorithm AprioriAll.

V. CONCLUSIONS

The actual estimation process undergoes some assumptions that have to be made considering the application. For the multiple activity estimations, where there is a lot of uncertainty due to the amount of activities, assumptions have to be made looking at possible scenarios that lead to such activity timelines.

In this work, four hypothesis focus in the short activities were formulated. They mainly differ in two properties: to have in consideration, or not, the existence of short activities, and to estimate, or not, the duration of long activities. The major difference is results comes when short activities are considered and long activity durations are not estimated (hypothesis 2) leading to higher probabilities in short time ranges. The reasoning for this is given to a specific phenomenon of the event log used: users commonly log a long activity followed by one or two short activities at the end of the ROI, causing the short activities to have no room to ”expand” and thus leading to a higher number of activities with estimated duration up to 20 minutes.

Apart from that, the obtain results were very consistent, revealing that the time constrains applied by time-windows and the other activities (whose duration was not estimated) create a well conditioned problem leading to similar output results independently of the hypothesis used to estimate the duration of multiply activity situations.

REFERENCES

[1] DINALOG. DAIPEX, 2015. URL http://www.dinalog.nl/en/projects/r_d_projects/daipex/.

[2] Martin Desrochers, Jacques Desrosiers, and Marius Solomon. A new optimization algorithm for the vehicle routing problem with time windows. Operations research, 40(2):342–354, 1992.

[3] Zolt ´an Fazekas, P´ eter G´asp´ ar, and Roland Kov´acs. Determining truck activity from recorded trajectory data. Procedia-Social and Behavioral Sciences, 20:796–805, 2011.

[4] T. Pedersen G. Gidofalvi. Mining long, sharable patterns in trajectories of moving objects. Proceedings of STDBM, pages 49–85, 2006.

ISBN: 978-93-86171-56-6

[5] Juyoung Kang and Hwan-Seung Yong. Spatio-temporal discretization for sequential pattern mining. In Proceedings of the 2nd international conference on Ubiquitous information management and communication, pages 218–224. ACM, 2008.

[6] Juyoung Kang and Hwan-Seung Yong. Mining trajectory patterns by incorporating temporal properties. In Proceedings of the 1st International Conference on Emerging Databases, pages 63–68, 2009.

[7] P. Smyth U. M. Fayyad, G. Piatetsky-Shapiro and R.Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.

[8] R. Agrawal and R. Srikant. Mining sequential patterns. In Proceedings of the 11th International Conference on Data Engineering, pages 3–14, Taipei, Taiwan, March 1995.

[9] Mohammed J Zaki. Mining data in bioinformatics. Handbook of Data Mining, pages 573–596, 2003.

[10] Adam Perer and Fei Wang. Frequence: Interactive mining and visualization of temporal frequent event sequences. In Proceedings of the 19th international conference on Intelligent User Interfaces, pages 153–162. ACM, 2014.

[11] Debprakash Patnaik, Patrick Butler, Naren Ramakrishnan, Laxmi Parida, Benjamin J Keller, and David A Hanauer.

[12] CARL H. MOONEY and JOHN F. RODDICK. Sequential pattern mining – approaches and algorithms. ACM Computing Surveys, 2013. doi:10.1145/2431211.2431218.

[13] K Venkateswara Rao, A Govardhan, and KV Chalapati Rao. Spatiotemporal data mining: Issues, tasks and applications. International Journal of Computer Science & Engineering Survey (IJCSES) Vol, 3, 2012.

[14] John F Roddick and Myra Spiliopoulou. A bibliography of temporal, spatial and spatio-temporal data mining research. ACM SIGKDD Explorations Newsletter, 1(1):34–38, 1999.

[15] Ramakrishnan Srikant and Rakesh Agrawal. Mining sequential patterns: Generalizations and performance improvements.