A Systemic Review on RFID Clustering and Similarity Measures: Issues and Challenges

Sagar Mahajan[1] Prof.R.H.Borhade[2]
[1,2]Smt.Kashibai Navale College of Engineering Pune-41

Abstract— recent advances in technologies such as radio-frequency identification (RFID) have made automatic tracking and tracing possible in a broad range of applications. Due to the significance of automatic tracking by RFID, trajectory clustering is an important topic in RFID data management and mining, which has a broad range of applications in various areas, such as traffic monitoring, video surveillance, cattle tracking and supply chain management. Trajectory clustering is the process of grouping similar trajectories according to a similarity distance. Depending on the task, given a set of trajectories, one may want to find clusters of objects that followed the same path or detect groups that moved together for given period. For trajectory tracking, literature presents different algorithms like Time-Focused Clustering (TFC) algorithm and Fuzzy C-Means algorithm. One of the recent method presented in the literature is Hierarchical RFID Trajectory Clustering [9] which considered a similarity measure called, Time-parameterized Edit Distance (TED). This type was used to find the similarity between to trajectory path by taking into consideration time dimension values in the calculation. Also, this model can deal with variants in both time and space dimensions, and the clustering algorithm is much less sensitive to noise and outliers than existing methods. But, similarity measure considered for clustering does not deem the weightage for the different parameters. TED grants the same weightage for all the parameters.

Key words: Internet of Things, radio frequency identification (RFID), Time-Focused Clustering (TFC), clustering algorithm, cloud computing, Time-parameterized Edit Distance (TED)

I. INTRODUCTION

Radio-frequency identification (RFID) is the wireless use of electromagnetic fields to transport data, for automatically identifying and tracking tags attached to objects. The tags contain electronically stored information. Some tags are examined by electromagnetic induction from magnetic fields generated near the reader. Some types collect energy from the interrogating radio waves and act as a passive transponder. Other types have a local power source such as a battery and may run at hundreds of meters from the reader. Unlike a barcode, the tag does not necessarily need to be within line of sight of the reader and may be embedded in the tracked object. RFID is one approach for Automatic Identification and Data Capture (AIDC). It can be referred to or consolidated into a product, animal, or person for the purpose of identification and tracking using radio waves. Some tags can be read from various meters away and beyond the line of sight of the reader. Most tags carry a plain text heading and a barcode as complements for direct reading and cases of any failure of radio frequency electronics. Most RFID tags contain at least two parts. One is an integrated circuit for storing and processing information, modulating and de-modulating a radio-frequency (RF) signal, and other specialized functions. The second is an antenna for collecting and transmitting the signal.

Due to an advancement of RFID technology, the tremendous amount of information has been triggered recently through Networked RFID is one of the crucial technological advances that help make RFID-enabled traceability possible. The increasing information requires the data abstraction techniques that can be simply achievable trough clustering. Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects. Clustering has been broadly employed in numerous applications such as market research, pattern recognition, data analysis, and image processing. Here, trajectory clustering is a novel and statistically well-founded method for clustering time series data from gene expression arrays and it has applications in many areas such as traffic monitoring, video surveillance, cattle tracking and supply chain management. Trajectory clustering uses non-parametric statistics and is hence not receptive to the particular distributions underlying gene expression data. Each cluster is clearly defined regarding direction of change of expression for successive time points, i.e., its trajectory. Recent developments in satellites and tracking facilities have made it possible to collect a large amount of trajectory data and there is increasing interest to perform data analysis over these trajectory data. Thus, an efficient clustering algorithm for trajectories is essential for such data analysis tasks.

II. LITERATURE SURVEY

To remunerate for the fundamental unreliability of RFID data streams, most RFID middleware systems employ a "smoothing filter", a sliding-window aggregate that interpolates for lost readings. In [1], SMURF is proposed for adaptive smoothing filter for RFID data cleaning. SMURF models the unreliability of RFID readings by viewing RFID streams as a statistical sample of tags in the physical world and exploits techniques grounded in sampling theory to drive its cleaning processes. SMURF contains two elemental cleaning mechanisms aimed at producing particular data streams for individual tag-ID readings (per-tag cleaning) and, providing accurate aggregate estimates over large tag populations (multi-tag cleaning). Additionally, SMURF incorporates two modules that apply to both data-cleaning techniques: a sliding-window processor for fine-grained RFID data smoothing, and an optimization mechanism for improving cleaning effectiveness by detecting mobile tags. Through the use of tools such as binomial sampling and π-estimators, SMURF continuously modifies the smoothing window size in a principled manner to produce accurate RFID data to applications. However, it does not propose the optimal smoothing filter for the readings of single tag and an aggregate signal.

The inherent uncertainty in RFID signals requires an RFID middleware system to clean the input data after capturing. Typically these systems employ a low pass filter

for reducing errors. In [2], an approach is proposed for data cleaning that exploits primary characteristics of RF signals as well as maximum likelihood operations. With this filter, proximity detection of RFID tags is improved. This permits reasoning about the position of RFID tags in the reader's range without measuring the signal strength of tag responses. It is, therefore, applicable on top of standard reader interfaces. Also, it improves data cleaning wherever the tag to reader distance is relevant. For instance, this qualifies correct ordering of items that pass by a reader on a conveyor or enhances tracking scenarios with RFID-equipped forklifts. However, the scheme considers simplified properties of RF signals, rather than general properties arising in applications.

A major problem in recognizing events in streams of data is that the data can be imprecise (e.g. RFID data). However, some state-of-the-art event detection systems assume that the data is accurate. Noise in the data can be captured using methods such as hidden Markov models. Inference on these models creates streams of probabilistic events that cannot be directly queried by existing systems. To address this challenge Lahar, an event processing system for probabilistic event streams is proposed. By using the probabilistic nature of the data, Lahar yields a much higher recall and precision than deterministic techniques operating over only the most feasible tuples. By using a static analysis and novel algorithms, Lahar processes data orders of magnitude more efficiently than a naïve approach based on sampling. Unfortunately, it does not provide a comprehensive analysis on the performance of these algorithms [3].

Recent discoveries in RFID technology are facilitating large-scale, cost-effective deployments in retail, healthcare, pharmaceuticals and supply chain management. The advent of mobile or handheld readers adds meaningful new challenges to RFID stream processing due to the inherent reader mobility, increased noise, and incomplete data. In [4], they addressed the problem of translating noisy, incomplete raw streams from mobile RFID readers into clean, perfect event streams with location information. Specifically, it proposes a probabilistic model to capture the mobility of the reader, object dynamics, and noisy readings. This model can self-calibrate by automatically estimating key parameters from observed data. Based on this model, it employed a sampling-based technique called particle filtering to gather clean, precise information about object locations from raw streams from mobile RFID readers. Since inference based on standard particle filtering is neither scalable nor efficient in our settings, three enhancements such as particle factorization, spatial indexing, and belief compression were proposed for scalable inference over large numbers of objects and high volume streams. Our experiments show that our approach can offer 49% error reduction over a state-of-the-art data cleaning strategy such as SMURF while also being scalable and efficient. However, it does not discuss query processing over inferred data for various monitoring applications.

Uncertain data streams, where data is incomplete, imprecise, and even misleading, have been observed in many environments. Feeding such data streams to existing stream systems produces results of unknown quality, which is of paramount concern to monitoring applications. In [5], the PODS system is aimed, that supports stream processing for random data naturally captured using continuous random variables. PODS employ a unique data model that is flexible and allows efficient computation. Built on this model, evaluation techniques are improved for complex relational operators, i.e., aggregates and joins, by exploring advanced statistical theory and approximation. However, the works on ranking in probabilistic databases give simplistic solutions to handling continuous distributions.

The data manipulated in emerging applications like location-based services, sensor monitoring systems, and data integration, are often inexact in nature. In [6], the significant problem of extracting frequent item sets from a large uncertain database, interpreted under the Possible World Semantics is studied. This matter is technically challenging, since an ambiguous database contains an exponential number of possible worlds. By observing that the mining process can be modeled as a Poisson binomial distribution, an approximate algorithm is developed, which can efficiently and accurately discover frequent item sets in a large uncertain database. Specifically, an incremental mining algorithm is proposed, which enable probabilistic regular item set results to be renewed. This diminishes the need of re-executing the whole mining algorithm on the new database, which is often more expensive and unnecessary. The proposed approach is useful for a probabilistic database with Boolean attributes, but not applicable for probabilistic, quantitative data.

Clustering trajectories as a whole could miss regular sub-trajectories. Discovering common sub-trajectories is very useful in many applications, especially if we have regions of special interest for analysis. In [7], a partition-and-group framework for clustering trajectories is proposed, which partitions a trajectory into a set of line segments, and then, groups similar line segments together into a cluster. The primary advantage of this framework is to discover common sub-trajectories from a trajectory database. Based on this partition-and-group framework, a trajectory clustering algorithm, TRACLUS is developed. This algorithm consists of two phases: partitioning and grouping. For the first phase, symmetrical trajectory partitioning algorithm using the minimum description length (MDL) principle is presented. For the second phase, a density-based line-segment clustering algorithm is manifested. A problem with this approach is that not all potential stops can be discovered during the clustering process.

As mobile devices with positioning capabilities continue to proliferate, data management for so-called trajectory databases that capture the historical movements of populations of moving objects becomes essential. In [8], the querying of such databases for convoys is examined. More specifically, it formalizes the concept of a convoy query using density-based notions, to capture groups of arbitrary extents and shapes. Convoy discovery is relevant for real life applications in throughput planning of trucks and carpooling of vehicles. Although there has been extensive research on trajectories in the literature, none of this can be applied to retrieve correctly exact convoy result sets. Motivated by this, three efficient algorithms were developed for convoy discovery that adopt the well known filter-refinement framework. In the filter step, lines amplifications techniques on the trajectories are applied and set distance bounds between the simplified trajectories. This authorizes efficient convoy discovery over the reduced trajectories without missing any actual convoys. In the refinement step, the

candidate convoys are further processed to obtain the actual convoys. The proposed work focuses on raw trajectories, therefore missing the related semantic information contained in the background geographic and application databases.

In [9], an efficient clustering algorithm is proposed that is much less sensitive to noise and outliers than the existing methods. To better promote the emerging cloud computing resources, algorithm is designed cloud-friendly so that it can be easily utilized in a cloud environment. They proposed a software approach that handles uncertainties in distributed RFID-enabled traceability applications. They formalize the problems as trajectory clustering and propose an efficient clustering algorithm with a novel similarity measurement model. But, similarity measure considered for clustering does not deem the weightage for the different parameters. TED considers the same weightage for all the parameters. In [10], RFID technology has been employed to the moving object tracking system, by clustering the trajectory of moving objects and extract based on the motion trajectory clustering and extract the moving characteristic patterns and predict the object's motion behavior. In[12] Due to the efficiency for the existing rail clustering algorithm is not large, therefore facing the low granularity and large orbital data set, the algorithm usually cannot work effectively .

In [11] this paper is based on a pattern recognition approach to figure out the RFID-enabled trajectory which is classified so as to find the characteristics and predict the moving leads of the objects. The intention is to fill the gap of the inefficiency of classification approach on RFID-enabled trajectory.

In this paper, we have performed a systematic study on mining of trajectory patterns in huge trajectory databases and developed a tree approach for efficient and scalable mining of trajectory patterns. Instead, of refinement of the apriori-like,candidate generation-and-test approach a P-tree structure is proposed. The experimental results we have reported here show that the Tree Partitioning method described is remarkably effective in limiting the maximal memory requirements of the algorithm while its execution time scales only slowly and linearly with increasing data dimensions .In [13] Data mining on spatiotemporal data and in particular on Data mining on spatiotemporal data, trajectory data, in particular, is a largely unexplored area. In this paper, we presented several classes of problems that so far have been studied very little. The problems considered have been organized along a classical categorization of data mining tasks inherited from standard contexts, which include clustering, classification, and local pattern tasks.

III. PROPOSED METHODOLOGY

This work aims to devise a Weighted Time-parameterized Edit Distance-based trajectory clustering in RFID environments. In the proposed method, this similarity function will be enhanced with weighted values. The Weighted Time-parameterized Edit Distance (WTED) will be then applied for finding the similarity between two trajectories in trajectory clustering. At first, every RFID points are projected to single cluster by putting all the trajectory points into single plane. In the second step, every point data points are merged to get the required number of trajectory clusters. During the merging operation, the RFID points are joined into sub-clusters that represent the "branches" at that node. Here, merging operation is used for merging two sets of trajectories into one set using the proposed similarity function that find the suitable RFID points to join into a single cluster. This process is continued until we get the required number of clusters. The proposed trajectory clustering will be implemented using JAVA programming, and the performance of the proposed clustering will be validated using clustering quality.

A. Mathematical Formulation:

Traceability Network: A traceability network is a directed graph G = (V, E). V represents the set of nodes where RFID readers are deployed and E represents the set of possible connections between nodes. A node vi is represented by its unique identifier, and a connection is represented by (vs, ve) where vs and ve are two nodes. It should be noted that a node refers to a location where more than one reader might be installed. Unlike other RFID systems where each reader is treated as a location, we aggregate the readers at the same location as one node. This is a reasonable abstraction in distributed RFID systems.

Trajectory: A trajectory of a given RFID tagged object oi is a polyline in a three-dimensional space (V, Ts, Te), where Ts is the time space for arrival readings and Te is the time space for leaving readings. A trajectory TRi of oi is represented as a sequence of points, accompanied by a unique ID of the object: TRi = {oi, {(v1, ts1, te1), (v2, ts2, te2), . . . , (vn, tsn, ten)}}. Its V-axis values, ordered by Ts values, form a path P in G. The set of all trajectories is denoted as ST. The deployment of RFID readers may affect the timestamps of arrival/leaving readings: (i) If readers are deployed at the entrance and exit of a node, ts and te (captured by entrance reader and exit reader respectively) are different, i.e., ts < te. (ii) If only one reader is deployed at a node and only one reading of each object is captured, ts = te. (iii) If only one reader is deployed at each node, but the first and last readings of each object are captured, ts and te (captured by the same reader) may be different, i.e., ts ≤ te.

Trajectory Cluster: A trajectory cluster TC is a sequence of node-range pairs, which describes the common temporal-spatio relationship of a group of objects. Formally, TC = {(v1, (ts1, te1)), (v2, (ts2, te2))... (vn, (tsn,ten))}. Now we define the research problems of managing uncertainties for traceability as follows.

1) Outlier Detection:

Suppose TR is the real trajectory of an object and TR' is the one captured by the readers, the first task is to determine whether TR'= TR, i.e., to determine whether there are missing readings in the process of obtaining the object. Evidently, there is no deterministic way to do so in the software layer. We define the probability of TR' ≠ TR as outlier (TR'). The reason we classify this problem as outlier detection is that when objects move together and are correctly captured, these objects should follow the same path P during the same time range. When TR' misses at least one segment of the trajectory, it can be treated as an outlier. TR' is an outlier if poutlier (TR') ≥ ϵoutlier, where ϵoutlier is a predefined threshold.

2) Classification:

If TR' is detected as an outlier (missing readings exist), the next task is to recover the missing readings. Similar to the outlier detection, we can assume that most objects'

trajectories are captured correctly. Suppose we have the correct and complete trajectory clusters SC = {TC1, TC2... TCn}, the recovery task can be transformed to a classification problem.

3) Clustering:

In most cases, the set of trajectory clusters SC = {TC1, TC2... TCn} is not known beforehand. Moreover, it may occasionally change. As the result, for the classification to work, it is necessary to generate SC by clustering the existing trajectories. A similarity measurement model called Weighted Time-parameterized Edit Distance (WTED) is proposed where the time dimension values are also used in the calculation.

$$WTED(TR_1, TR_2) = \begin{cases} |TR_1| & ;if\ |TR_1| = 0 \\ |TR_2| & ;if\ |TR_1| = 0 \\ \min \begin{cases} dist_1(TR_1(1), TR_2(1)) + WTED(R_{est}(TR_1), R_{est}(TR_2)), \\ WTED(R_{est}(TR_1), (TR_2) + 1, WTED(TR_1, R_{est}(TR_2)) + 1 \end{cases} \end{cases}$$

In WTED's we introduce a new function called Weighted Time-parameterized Distance (WTPD) for two elements e1 and e2 in two trajectories, namely,distt(e1, e2). Where,

$$dist(e_1, e_2) = \begin{cases} \sqrt{\alpha(e_1.t_e - e_2.t_e)^2 + \beta(e_1.t_s - e_2.t_s)^2} & ;if\ e_1.v = e_2.v \\ \sqrt{\alpha(e_1.t_e - e_2.t_e)^2 + \beta(e_1.t_s - e_2.t_s)^2} + 1 & ;if\ e_1.v \neq e_2.v \end{cases}$$

Where, $\alpha = \dfrac{\text{var}(t_e)}{mean(t_e)}$; $\beta = \dfrac{\text{var}(t_s)}{mean(t_s)}$

B. Pseudocode:

RFID Trajectory Clustering and cluster merging algorithm
Input: The set of trajectories: TR = {TR1, TR2, . . . , TRn}.
Output: The set of trajectory clusters: TC.
The set of outliers: OL.
The set of merged trajectory clusters (in place merging): TC.
Procedure:
TC = Φ, OL = Φ
hrtc (TC, OL, 1)
function hrtc (clusters, outliers, depth)
for each cluster TCpi in clusters with |pi| = depth
for each trajectory TRi belongs to TCpi
if|TRi| > depth
p'i = pi + TRi(depth)
ifTCp'iexists, assign TRi to TCp'i
otherwise add TCp'i to clusters
end if
end for
for each newly added TCp'i
replaceTCp'i with OPTICSt(TCp'i)
outliers+ = outlier(OPTICSt(TCp'i))
end for
removeTCpfrom clusters
end for
ifdepth ≤ MAX_DEPTH
hrtc(clusters, outliers, depth + 1)
end function
set of trajectory clusters: TC = {TC1, TC2, . . ., TCn}.
Sort TC by the lengths of the paths
for each TC in sorted TC
find the set of candidates to merge to: {TC1, TC2, . . . , TCm}
where |TCi| − |TC| = 1 and the difference of nodes is 1
for each TC' in candidates
ifWTED(RT (TC), RT (TC')) is the minimum and WTED(RT (TC), RT (TC')) <ϵ
mergeTC into TC', re-calculate the RT for TC'
break
end for
end for

IV. Conclusion

Recent advances in technologies such as radio-frequency identification (RFID) have made automatic tracking and tracing possible in a wide range of applications. However, there are still numerous technical difficulties in realizing traceability applications in large-scale, uncertain environments such as the emerging Internet of Things (IoT). In this paper, we have introduced an efficient trajectory model and developed a novel clustering algorithm to cluster RFID trajectories with the capability to recover missing readings. Our algorithm is scalable and efficient, outperforming existing methods such as Time-Focused Clustering (TFC) algorithm and Fuzzy C-Means, as demonstrated by the results from extensive experimental studies. The experimental studies of our clustering algorithm have been conducted using synthetic and offline data. Our Future work includes further performance evaluation with real data from a large-scale supply chain management system, and online clustering and recovering of RFID trajectories.

References

[1] Shawn Jeffery, Minos Garofalakis, and Michael Franklin. Adaptive Cleaning for RFID Data Streams. In Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB'06), Seoul, Korea, September 2006.
[2] H. Ziekow and L. Ivantysynova. A Probabilistic Approach for Cleaning RFID Data. In Proceedings of the 24th International Conference on Data Engineering Workshop (ICDEW'07), Istanbul, Turkey, April 2008.
[3] Christopher Re, Julie Letchner, Magdalena Balazinksa, and Dan ´ Suciu. Event Queries on Correlated Probabilistic Streams. In Proceedings of the 2008 ACM International Conference on Management of Data (SIGMOD'08), Vancouver, Canada, 2008.
[4] Thanh Tran, C. Sutton, R. Cocci, Nie Yanming, Diao Yanlei, and P. Shenoy. Probabilistic Inference over RFID Streams in Mobile Environments. In Proceedings of the 25th International Conference on Data Engineering (ICDE'09), Shanghai, China, April 2009.
[5] Thanh T.L. Tran, Liping Peng, Boduo Li, Yanlei Diao, and Anna Liu. PODS: a New Model and Processing Algorithms for Uncertain Data Streams. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (SIGMOD'10), Indianapolis, Indiana, USA, 2010.

[6] L. Wang, D. Cheung, R. Cheng, S. Lee, and X. Yang. Efficient Mining of Frequent Item sets on Large Uncertain Databases. IEEE Transactions on Knowledge and Data Engineering, PP (99):1, 2011.
[7] Jae-Gil Lee, Jiawei Han, and Kyu-Young Whang. Trajectory Clustering: a Partition-and-Group Framework. In Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data (SIGMOD'07), Beijing, China, 2007.
[8] Hoyoung Jeung, Man Lung Yiu, Xiaofang Zhou, Christian S. Jensen, and Heng Tao Shen. Discovery of Convoys in Trajectory Databases. Proceedings of VLDB Endowment, 1(1):1068–1080, August 2008.
[9] Yanbo Wu, Hong Shen, and Quan Z. Sheng, "A Cloud-friendly RFID Trajectory Clustering Algorithm in Uncertain Environments", IEEE transactions on parallel and distributed systems, Vol. 26, No. 8, pp. 2075-2088, 2014.
[10] Kongfa Hu,Xiangqian Xue, Jianfei Ge, Yan Sun, Ling Chen "Dividing And Clustering Algorithms Of Moving Objects In Rfid Tracing System " Journal of Theoretical and Applied Information Technology E-ISSN: 1817-3195 2012. Vol. 45 No.1.
[11] Ying Shen and Weihua Zhu "Classification of RFID-enabled Trajectory using Pattern Recognition Approach" International Journal of Signal Processing, Image Processing and Pattern RecognitionVol.7, No.2(2014), pp.345-354.
[12] Kongfa HuandLong Li "Mining aNew Movement Pattern in RFID Databaseon Internet of Things"International Journal of Database Theory and ApplicationVol.7, No.2(2014), pp.37-44.
[13] Susanta Satpathy, Lokesh Sharma, Ajaya K. Akasapu, Netreshwari Sharma" Towards Mining Approaches for Trajectory Data"International Journal of Advances in Science and Technology Vol. 2, No.3, 2011.
[14] Akasapu A., Sharma L. K., Ramakrishan G. 2010. Efficient Trajectory Pattern Mining for both Sparse and Dense Dataset. Int. J. of Computer Applications (0975 - 8887) Volume 9– No.5. 45-48.

