

A Methodology of Human Behavioral Pattern Mining Approach to Big Data Analytics

Dr.P.Srimanchari¹, Dr.G.Anandharaj²

*1Assistant Professor and Head, Department of Computer Applications,
Erode Arts and Science College(Autonomous), Erode – 638001*

*2Associate Professor and Head, Department of Computer Science,
Adhiparasakthi College of Arts and Science (Autonomous), Kalavai, Vellore – 632506
Corresponding Author: Dr.P.Srimanchari*

Abstract: This work introduces a set of scalable algorithms to identify patterns of human daily behaviors. These patterns are extracted from multivariate temporal data that have been collected from smartphones. We have exploited sensors that are available on these devices, and have identified frequent behavioral patterns with a temporal granularity, which has been inspired by the way individuals segment time into events. We briefly introduce the basics of related research topics, review state-of-the-art approaches, and present some preliminary thoughts on future research directions. This paper proposes a fusion of three different data models like relational, semantically, and big data based data and metadata involving their issues and enhanced capabilities.

Keywords: Multivariate temporal data, Big-data, real-time analytics.

Date of Submission: 11-11-2018

Date of acceptance: 23-11-2018

I. INTRODUCTION

The computing and networking capabilities of mobile and wearable devices, makes them appropriate tools for obtaining and collecting information about user activities (mobile sensing). This has led to a significant expansion of opportunities to study human behavior ranging from public transport navigation [1] to well-being [2]. Moreover, the advent of mobile and wearable devices enables researchers to unobtrusively identify human behavior to an extent that was not previously possible. Nevertheless, there is still a lack of wide acceptance of mobile sensing applications in real-world settings [3]. There are different reasons for this mismatch between capability and acceptance. First, the limitation of resources and a lack of accuracy in the collected contextual data, especially is a challenge with regard to the battery life [4]. Furthermore, the small size of sensors that are dealing with radio frequency, i.e., Bluetooth, WiFi and GPS, affects the quality of their data [5] (the smaller the device, the less accurate the data).

Big data in real time have diverse and autonomous representations bringing highly unstructured and unrelated data based relationships in producing results which are getting complex and faulty. The heterogeneous data features represent different representations for data. Decrease the effect of heterogeneous and complex data; there can be computationally introduced at localized systems considering they are having better computational power. There can be a way of transforming data into a common data fusion. As a result, the common forms of data in consequence to data fusion will be highly compatible for data linkage and relativity indexing for getting better analytical outcomes. Major of data is stored either using relational, semantical or big data formats. Relational data is stored in the form records containing a collection of singleton cells representing fields supported by its data structure and constraints for an entity.

II. RELATED WORKS

The relational data model was first invented with the term "relational database" by E. F. Codd from IBM in 1970. Whereas, Codd had defined relational in his paper titled "A Relational Model of Data for Large Shared Data Banks" in which he had introduced 12 rules for implementing relational data model also known as Codd's rules. These rules were completely taken but up to a minimum and necessary level in defining a table as a relation and operators used to manipulate this data form. Whereas, a language was introduced for querying by Chamberlin and Boyce in 1974 from IBM. It was first named a SEQUEL (Structured English Query Language) which was made standard in ANSI X3H2 committee with SQL (Structured Query Language) in 1986 [9]. In 1976 a designing model to view relational data with the entity-relational model by Peter Chan. In 1990's third generation database system manifesto was introduced by Stonebraker in 1990 which in 1996 became ORDBMS

(Object Relational Databases Management System) [10]. Further history of RDB is concerned with the managementsystem of the relational model. The evolution of RDB data and querying model linkingthem together according to the timeline at the side to showtheir arrival according to the history using year and author details. Now in next section evaluation of XML is beingrepresented [13], [14].

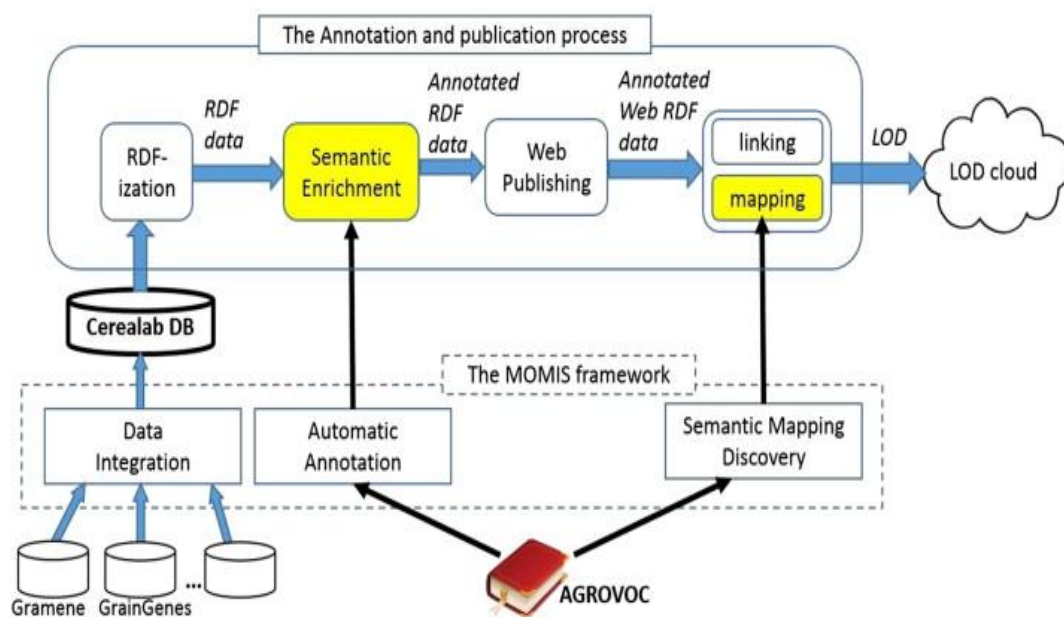


Figure 2.1 The annotation and publication process

Inventory Forecasting Use Case: Inventory forecasting is one of the major tasks in manufacturing and includes raw material inventory forecasting. Accurate and reliable inventory prediction can guarantee a smooth production process [16]. Here we assume that the forecasting target is the inventory of a PCB (Printed Circuit Board) and the address of the training data is inventoryOfPCB.csv. Car Evaluation Use Case: When manufacturing enterprises are going to launch a new product, it needs to match the public's aesthetic. We can use a rule extraction model to extract rules regarding the public's aesthetic that were hidden in previous user feedback data on various car models. Here we take a Car Evaluation Data Set from the UCI machine learning repository as training data [17]; the address is carEvaluation.csv and the target is carAcceptability. Tool Condition Monitoring Use Case: Manufacturing systems are becoming more complex and are subject to failures that adversely impact their reliability, availability, safety, and maintainability [18]. For example, in the high-speed milling process, a worn milling tool might irreversibly damage a workpiece [19]. In such a case, real-time monitoring of the condition of the tools can help the operator avoid catastrophic events. Here we take a Steel Plates Faults Data Set as training data [20]; the address is steelPlatesFaults.csv and the target attribute is Faults.

3. Event Model Description

As the interface for the whole system, GMDL is one of its most essential parts. It directly determines what tasks GMDA can perform and thus is crucial to the whole system, therefore it requires perfect design. We attempt to make it concise and easy to understand for inexperienced users.

3.1 GMDL for use cases

To fully describe a task, we must specify at least three key points:

The goal of task;

The dataset;

Target attribute (if the task is prediction).

Each point is independent, and to make GMDA easier for inexperienced users, we initially propose a basic formal structure for the GMDL language, where COMMAND is the description of the task that GMDA can understand.

That is:

**COMMAND -> SETTING COMMAND | SETTING;
SETTING -> PROPERTY = "VALUE".**

3.2 Definition of language GMDL

COMMAND includes multiple independent SETTINGS, which are independent because they are assignments for different PROPERTIES.

3.3 Annotation for Big Data

Real-time data collection is found mostly in the form of sensor data collected through physical or biological resources. In the current era of information analytics Internet of Things (IoT) is playing the main role in managing, controlling and monitoring of the resources even at remote locations. With the involvement of social medium and mobile communication data is increasing rapidly. At the end of big data, Hadoop is playing a key role through its platform in data collection, computational clustering of distributed units, and dramatic fast analytics. However, still, it lacks in real-time boosted analytics for a localized fast outcome. For that to work data fusion is proposed at the level of localized or short area cluster of units to have highly interactive.

3.4 Understanding Challenges

The challenges involved in the methodology for real-time data fusion for localized big data's analytics concerns with data updates. Other issues involve one data model support and limitation to other data model during the process of data fusion. Data collected in traditional data storage representing relation database where data is placed separately from metadata. The new generation data formats like, JSON and RDF are more data and hierarchy oriented.

III. COMMUNICATION PROTOCOLS

In order to implement our algorithms for the problem described above, first the data format should be converted

from heterogeneous data to machine-processable data, i.e., the raw data needs to be converted to the previously described entity format. As previously stated, the data has been collected from heterogeneous sources. Some sensors have multiple values, for instance WiFi has BSSID, SSID and Capabilities (WPA, PSK, etc.). Nevertheless, for each sensor our model chooses only one value. In particular, each sensor (attribute) A, requires a single data point (value) D. Therefore, "BSSID" has been chosen for WiFi and Bluetooth, the pseudonymized phone number for SMS and Calls, "process name" for Application and tilting, in-vehicle, on-bicycle, walking, still, and unknown for the activity sensors (UbiqLog uses Google play services for activity recognition and therefore there is no raw accelerometer data inside the dataset). A similar approach has also been used for the Device Analyzer dataset, which we do not report here to preserve space. During the second step, we propose an algorithm that identifies the movement (based on location changes) state, which will be used to enrich the semantics of the data within the notion of the location. In third step, we need to convert the timestamp to a time similar to the human perception of time. Afterward, in the fourth step we describe the behavior similarity and FBP detection algorithms.

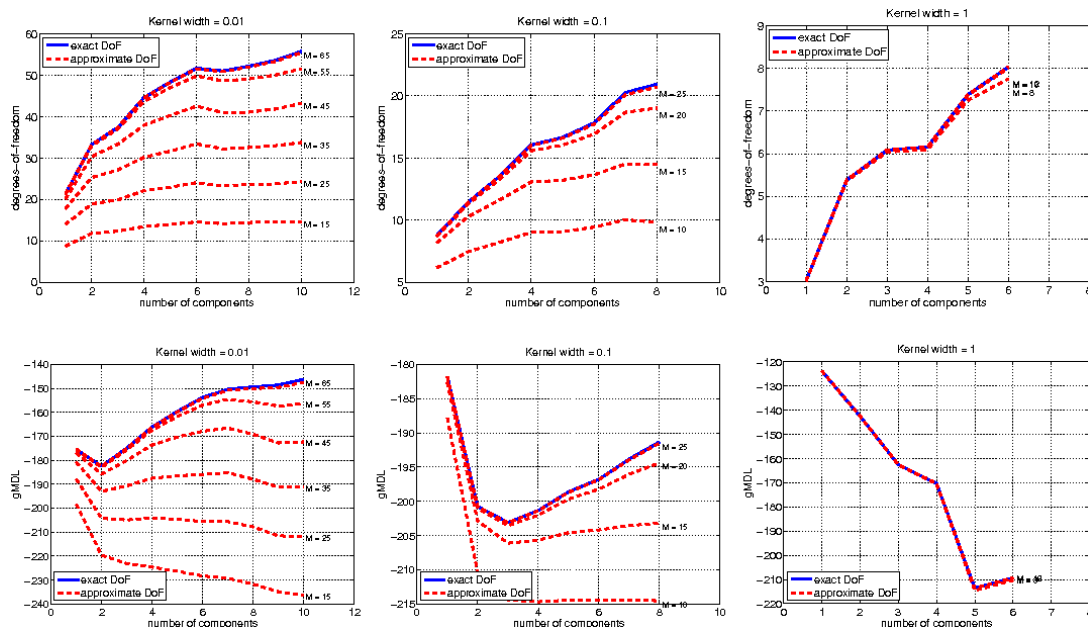
IV. SIMULATION

5.1 Accuracy Analysis -Ground Truth Dataset

In order to evaluate the accuracy and quality of the identified FBPs, we have created a ground truth dataset, which is composed of more than 5,000 identified entities (that participate in FBPs), from five users. It contains randomly identified FBP data that has been labeled by the users as either true or false.

5.1.2 Accuracy of Identified FBPs

After collecting the labels, we carefully examined the accuracy of our algorithms using three temporal segments of the day: 0:00-07:59 (0-8), 08:00-15:59 (8-16) and 16:00-23:59 (16-24) and different TGs. This time based segmentation has been inspired by similar work in mobile data mining, but it is more accurate than the two divisions proposed by Ma et al.



V. CONCLUSION

We have proposed a scalable approach for daily behavioral pattern mining from multiple sensor information. This work has benefited from two real-world datasets and users who use different smartphone brands. We use a novel temporal granularity transformation algorithm that makes changes on timestamps to mirror the human perception of time. Our frequent behavioral pattern detection approach is generic and not dependent on a single source of information; therefore, we have reduced the risk of uncertainty by relying on a combination of information sources to identify frequent behavioral patterns.

REFERENCE

- [1]. P. Y. Vandenbussche, G. A. Atemez, M. Poveda-Villalón, and B. Vatan, "Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web," *Semantic Web*, vol. 8, no. 3, pp. 437_452, 2017.
- [2]. K. Höffner, "Survey on challenges of question answering in the semantic Web," *Semantic Web*, vol. 8, no. 6, pp. 895_920, 2017.
- [3]. M. Nentwig, M. Hartung, A.-C. N. Ngomo, and E. Rahm, "A survey of current link discovery frameworks," *Semantic Web*, vol. 8, no. 3, pp. 419_436, 2017.
- [4]. Gangemi, V. Presutti, D. R. Recupero, A. G. Nuzzolese, F. Draicchio, and M. Mongiovì, "Semantic Web machine reading with FRED," *Semantic Web*, vol. 8, no. 6, pp. 873_893, 2017.
- [5]. R. Goodman, R. P. Mahler, and H. T. Nguyen, *Mathematics of Data Fusion*, vol. 37. Dordrecht, The Netherlands: Springer, 2013.
- [6]. M. Bevilacqua, A. Tsourdos, A. Starr, and I. Durazo-Cardenas, "Data fusion strategy for precise vehicle location for intelligent self-aware maintenance systems," in *Proc. IEEE 6th Int. Conf. Intell. Syst., Modelling, Simulation (ISMS)*, Feb. 2015, pp. 76_81.
- [7]. Hotho, R. Jäschke, and K. Lerman, "Mining social semantics on the social Web," *Semantic Web*, vol. 8, no. 5, pp. 623_624, 2017.
- [8]. D. Calvanese et al., "Ontop: Answering SPARQL queries over relational databases," *Semantic Web*, vol. 8, no. 3, pp. 471_487, 2017.
- [9]. R. Mukherjee, "Interfacing data destinations and visualizations: A history of database literacy," *New Media Soc.*, vol. 16, no. 1, pp. 110_128, 2014.
- [10]. G. Palmer, P. A. Stephens, A. I. Ward, and S. G. Willis, "Nationwide trophic cascades: Changes in avian community structure driven by ungulates," *Sci. Rep.*, vol. 5, Oct. 2015, Art. no. 15601.

Dr.P.Srimanchari. " A Methodology of Human Behavioral Pattern Mining Approach to Big Data Analytics." *IOSR Journal of Engineering (IOSRJEN)*, vol. 08, no. 11, 2018, pp. 78-81.