ISSN 2347 - 3983
Volume 8. No. 10, October 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter1168102020.pdf
https://doi.org/10.30534/ijeter/2020/1168102020

DATA UNCERTAINTY AND DATA QUALITY PROCESSING WITH APACHE SPARK

P. Madhuri[1], Dr. K. Srinivas Rao[2], Dr. S.K. Yadav[3]
[1]Research Scholar, Shri Jagdish Prasad Jhabarmal Tibrewala University, Jhunjhunu, Rajasthan.
[2]Professor, Principal, MLR Institute of Technology, Telangana.
[3]Director Research, Shri Jagdish Prasad Jhabarmal Tibrewala University, Jhunjhunu, Rajasthan.

ABSTRACT

Spatiotemporal analysis refers to the processing of spatiotemporal data and discovering knowledge, patterns from it. A practical example is identifying different city's histories and earthquakes. This data is collecting from the source where it is being referred with countries their regions, population, speaking habits from time to time, evolutions from time to time in the cities and also the developments happened constitutes space and time, and this data are often highly noisy and sparse consequently extract useful knowledge from such noisy and sparse data is a challenging task. Another challenging problem is integrating multiple-model data from three different sectors like countries, states, cities and their socio-economic development. There are different factors involved in ST data such as location, time, and text, these heterogeneous factors are highly coupled to reflect people's activities in a collective way, yet they have totally different modes, sizes, and allocations. How to effectively integrate those different data types for knowledge acquisition remains another challenge with the reference data structure is still unsolved. Furthermore, still, its time-consuming process to store a huge volume of spatiotemporal data which are rapidly accumulated and processing queries on the vast amount of spatiotemporal data is highly difficult in terms of space or time complexity. This paper discusses about the different sources around us to navigate and work along with the spatio data and explore more into the data set. The methods used in the analysis are EDA (Exploratory Data Analysis), Data deficiency, Data uncertainty so on. The architecture is been used in order to make a perfect their representation of the models. Lastly a model is also proposed as to work with the dataset and work more on the model with a proper consistent method.

Key words : Data sparcity, spatio temporal, Apache Spark and data uncertainty.

I. INTRODUCTION

In preprocess step how to take care of missing values or data values and also noisy data so as to overcome data sparcity. Many researchers examined spatiotemporal data mining, however they realized one difficulty that spatio temporal data sparsity i.e. spatiotemporal data is collecting in four corners of selected place, ST data in many cases are exceptionally noisy and lean hence extract useful knowledge from such rigid and sparse data is hard; yet another hard problem is integrating multiple-model data. These facets are highly combined to signify people's tasks in a collective fashion, yet they will have completely different manners, sizes, and allocations. The way to effortlessly incorporate those different data types for comprehension acquisition remains a second challenge with benchmark block structure remains disgusting. More over still its time-consuming procedure to put away a massive level of spatio temporal data that are rapidly accumulated and processing inquiries on huge number of ST data is exceptionally troublesome concerning time or space complexity. To tackle those problems, a broad group of approaches are suggested i.e. we urge to use intelligent classification which could improve spatio temporal data sparsity for effective classification, and Impalement Novel way of tackling huge volume of continuing unstructured data collection for query processing, and finally we must design effective multi-dimensional data-integration frame for rapid evaluation and aggregation of all spatio temporal data. Because of this our job will research in Business of Spatio Temporal data mining to enhancing efficacy and efficacy of data mining jobs such as clustering, prediction, anomaly detection, and pattern mining whilst working on spatio-temporal information.

Data collection Sources

Spatio temporal data are accumulated from several resources for traffic surveillance analysis field. Probably one among most usual type is Videos cameras out of point, produced by GPS-equipped vehicles. Additionally, other forms of trajectories probably originate in smart phones, on line check in data, geo tagged messages or websites in societal networks, RFID readers, etc. Thus, moving items might be human beings, creatures and vehicles etc.

Data types

We describe four common Types of ST information there's many different st data types this you may encounter in numerous real estate software. They differ from manner times and space is alternative st subjects. As first two Data Types and also raster info, where observations Trajectory info, In Used from custom of information collection and representation, which cause different forms of STDM problem formulations. As a Result of this, it is Vital to determine Shape of ST information on market at a Particular program to build most genuinely effective use of STDM methods forms:

Point mention information of those s Field have now already been collected in cells that are predetermined within a s grid Trajectories) list observation of different events and things, and next Event information, which comprises of are 2 data types (purpose routine and rasters) capture information regarding continuous.

II. A conceptual 3-tier spatiotemporal knowledge model

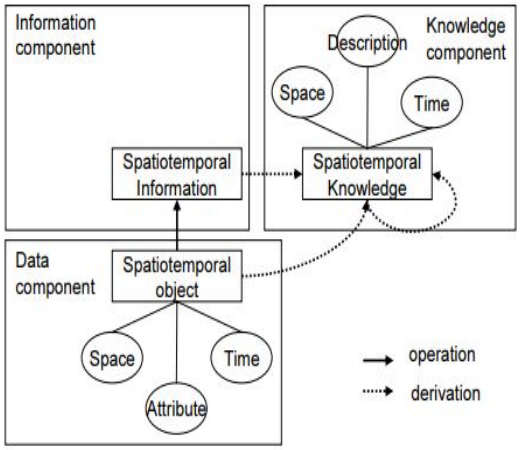


Figure 1: 3-tier model of Spatiotemporal Knowledge

Conceptual consciousness model for symbolizing spatio temporal happenings really reflects detected spatio temporal data, refined spatio-temporal info, as well as discovered spatio-temporal knowledge. According to Fig. inch, info component in this version is made up of all of this spatio temporal data which is present from spatio temporal world class. Info component concerns purposeful and selected data derived by data component using a generalization or conversion surgery so as to extract significance. Last, information component reflects knowledge items which can be of use to people. Awareness might be directly triggered by collection of spatio temporal items or indirectly caused by spatio temporal info or perhaps even a priori knowledge.

Table 1 Content of spatiotemporal observation on geographic object

No.	Geographic object	Time item	Observation Frequency	Geographic event
1	Transportation	Location, speed	Minute	Traffic jam, clear road, slow speed
2	Urban climate	Temperature, humidity, PM2.5	Hour	Wind, rain, thunder, light
3	Logistics	Transition information	Day	Trans post, sign up
4	Energy consumption	amount	Hour, month	Peak, troughs
5	Estate	Owners, scale	Year	Transfer, allot
6	Community	Population	Season	Move in, move out, rent
7	Mobile phone	Location	Minute	On duty, work, off duty

Process of spatiotemporal data mining

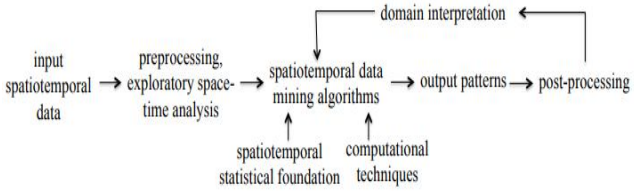


Figure 2: process of spatiotemporal data mining

Specifically, layer of spatiotemporal data mining techniques contains five components listed as follows:

Pre-processing

At pre-processing phase, spatio temporal data usually are washed, segmented, calibrated, and sampled for agents, or inferred from indeterminate temporal data. In department, we presented five shared surgeries in pre-processing phase.

Cleaning

Spatio temporal data mining methods try to find questionable moving items or to catch features of many strange moving objects. As a result of ambiguity of all RFID data, i.e., there is no geographical place given multiple subscribers discovered a thing; cleanup rectal data intends to shed hopeless locations or data harnessing specific limitations, e.g., max rate, space limits.

Segmentation

In most application scenarios, a spatio-temporal is invisibly to temporal data, every one that can be known as a segment, a partition or even a framework. Generating rectal is ordinary as it evolves to inherent structures in spatio temporal data, e.g., a course with numerous road sections together with moment interval. A partition-and-summarization process that endeavours to build human-readable outline of spatio temporal data additionally divides into several walls according to activities of moving items. Spatio temporal data are set into eyeglasses as a way to effectively store sample points of a moving thing that are coordinated by time periods, leveraging advanced column-oriented storage procedure.

Completion

Because of account of transmission and storage, spatio temporal data are largely collected at comparatively low emission speeds, just providing partial observations of actual paths. All these spatio temporal statistics are called cloudy rectal data.

Calibration

Routes that signify different approximation of initial avenues using different sampling strategies as well as different sampling speeds are equivalent. Heterogeneity includes a poor effect in dimension of course similarity, e.g., it's tough to compare two avenues based from distinct sampling methods by right utilizing cognitive proximity established on measures like Euclidean space. Focus on altering such heterogeneous trajectories into people together with unified sampling plans.

III. Research Gap

Over spatial-temporal information to be able to look after data sparsity massive spatiotemporal data with Apache Star K framework. Designing multiple Spatio-temporal data mining is classified into three different types. Data Classification dimensional data version for effective investigation over Spatio-temporal data. The scope of the Investigation work effectively handles Streaming and so on.

Many investigators analyzed spatiotemporal data mining, nevertheless by now know one language which the spatiotemporal data sparsity difficulty i.e spatio-temporal data is collecting at four different corners of selected place, Spatio Temporal data oftentimes are excessively noisy and sparse consequently extract useful knowledge from such stiff and lean data is hard; nonetheless still another tricky problem is incorporating multiple-model data You may detect 3 different elements a part of s t data: location, period, and text. All these factors are tremendously joined to signify people's activities in a collective manner, yet they'll have different ways, sizes, and allocations. How to incorporate those distinct data types for understanding acquisition remains another challenge with the grade cube arrangement is still disgusting. Over still its time intensive procedure to put a gigantic degree of STD which is immediately accumulated and calculating queries over the massive number of Spatio Temporal data is exceptionally problematic regarding space or time complexity. To handle those issues, a wide set of approaches have been indicated we advocate to use intelligent classification calculations that could improve Spatio Temporal data sparsity for effective classification, along with also Impalement Novel method of tackling huge degree of continuing unstructured data collection to query processing, last but not least we must generate effective multidimensional data integration framework for fast analysis and aggregation of most spatio temporal data. As a result of our occupation will probably explore over the specialization of spatio temporal data mining into improving the efficiency and efficiency of data mining tasks such as clustering, prediction, anomaly detection, and pattern mining whilst working with spatiotemporal data.

IV. RESULTS AND DISCUSSION

In industry of spatiotemporal data mining, approaches in many cases are worked on large temporal data bases and therefore surgeries are complicated, time-consuming and expensive. S t sampling methods intention to decrease a big rectal database taking only agent types of initial rectal database. Undoubtedly, subset of samples must exude freedom patterns hidden in first rectal database.

To overcome issues related with Data uncertainty and Data quality. Here we are trying to the get the data to the apache spark based hadoop platform, for which we are using the hive as programming model in order to work along with the data and we are trying to impose our methods on sample data to perform our first objective. So, such to work we improvising the dataset as 311 a data mining approved data set which has been engaged from the UCI machine leaning repository for machine learning methods.

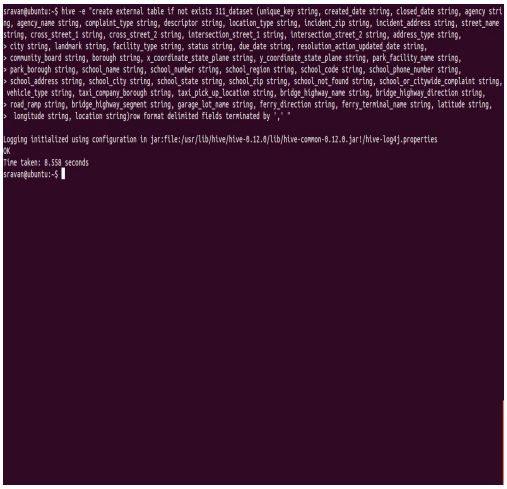


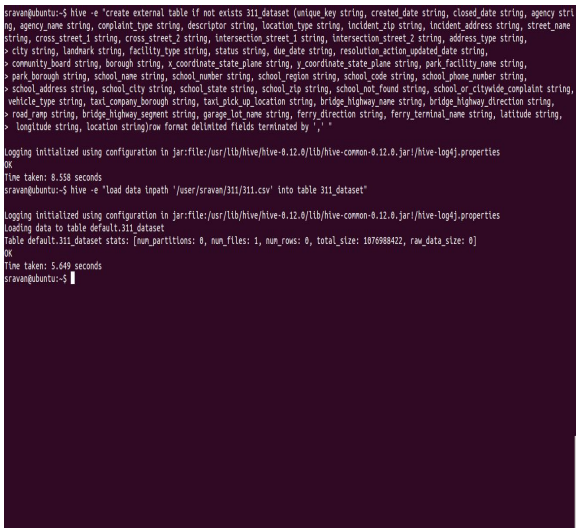
Table created successfully

Successfully created table

Importing data from hdfs to hive warehouse:

To Load Data:

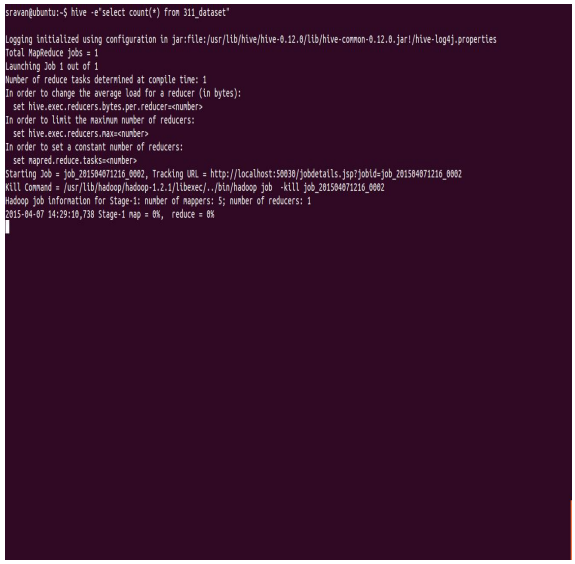
hive -e "load data inpath '/user/sravan/spatio-data /spatio-data .csv' overwrite into table spatio-data "



Data loaded to hive warehouse and table.

spatio-data Table data verification

Query : hive –e “ select *from spatio-data _dataset limit 5”



data from spatio-data _dataset

data from spatio-data dataset

Table 2 Performance of memory consumption using dataset 1

Minimum Bounding Rectangle (cm)	Memory consumption(kb)	
	Threshold value of Fτ =2.5	Threshold value of Fτ =3.5
1	440	540
2	420	520
3	380	490
4	360	470
5	335	450
6	315	426
7	300	398
8	287	370
9	275	349

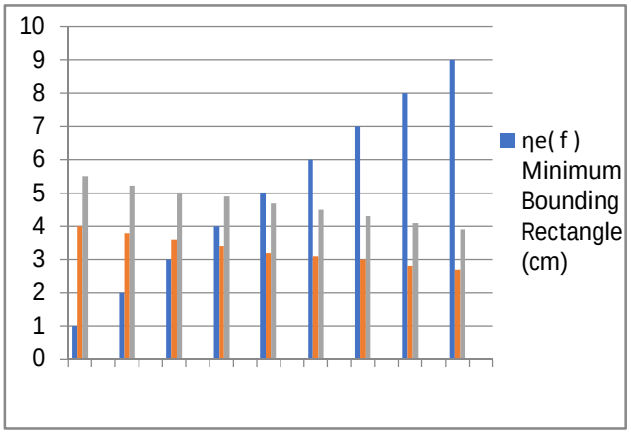


Figure 3 Performance of memory consumption using dataset 1

From the experimental investigation, the threshold values of FR are mended as 2.5 and 3.5 along with also the parameters

like computational memory and time use for distinct MBR state are reported. The time for brink worth of FR 2.5 is significantly less than the brink worth of FR 3.5. Likewise, the memory intake because of brink worthiness of FR 2.5 is significantly less than the brink worth of FR 3.5. From operation investigation utilizing data-set inch it's discovered the consequences with threshold worthiness of FR 2.5 are far better compared to people FR 3.5. Hence FR 2.5 is traditionally popularly employed like a threshold price to its relative investigation utilizing data-set inch that can be presented from the next portion.

V. CONCLUSION

The spatiotemporal analysis is all about dealing the data in a better way like understanding it perfectly. So to make the things perfect the models have been deployed using the apache spark with machine learning models in cleaning the data, working for consistency, data modulation along with space and time factors making data perfection by applying the dimensional reduction over the data set. The methods had been thoroughly applied on the dataset and made good observations on the data by running the deployment model over the data and the resonance time has been estimated over the timing factor. Further, spatio-temporal events have been grouped in line with the event type they belong. A pair of spatio temporal events accumulated over a spot for a time of period creates the spatio temporal database. The underlying process generating the events is regarded as a different occurrence, where space time plane is regarded as filled with entities or objects. Each spatial locale and time period may be inhabited by a couple of items or could be vacant.

REFERENCES

1. Aarya Santhosh; Sumam Mary Idicula (2017), "Home range estimation of migrating organisms considering the spatiotemporal aspects of GPS tracked data", International conference of Electronics, Communication and Aerospace Technology (ICECA), Volume: 1, pp: 162 – 167.
2. Alina Bialkowski et.al (2014), "Large-Scale Analysis of Soccer Matches Using Spatiotemporal Tracking Data", **ISSN:** 2374-8486, IEEE International Conference on Data Mining, pp: 725 – 730.
3. Atluri et al. (2016), "Spatio-Temporal Data Mining: A Survey of Problems and Methods", Vol. 51, No. 4, Article 83.
4. Cheng cheng Jia et.al (2018), "Stacked Denoising Tensor Auto-Encoder for Action Recognition with Spatiotemporal Corruptions", **ISSN:** 1941-0042, IEEE Transactions on Image Processing, Volume: 27, Issue: 4, pp: 1878 – 1887.
5. Damien Dosimont et.al (2014), "A spatiotemporal data aggregation technique for performance analysis of large-scale execution traces", **ISSN:** 1552-5244, IEEE International Conference on Cluster Computing (CLUSTER), pp: 149 – 157.
6. Gatrell et al. (1996)," Spatial point pattern analysis and its application in geographical epidemiology", vol. 21, No. 1 (1996), pp. 256-274
7. Hardisty & Klippel (2010), "Analysing spatio-temporal autocorrelation with LISTA-Viz", Vol. 24, No. 10, October 2010.
8. Jin-Ho Shin et.al (2011), "Spatiotemporal Load-Analysis Model for Electric Power Distribution Facilities Using Consumer Meter-Reading Data", **ISSN:** 0885-8977, IEEE Transactions on Power Delivery, Volume: 26, Issue: 2, pp: 736 – 743
9. Pierre-Antoine Laharotte et.al (2015), "Spatiotemporal Analysis of Bluetooth Data: Application to a Large Urban Network", **ISSN:** 1558-0016, Volume: 16, Issue: 3, pp: 1439 – 1448.
10. Zhongwei Deng; MinheJi (2011), "Spatiotemporal structure of taxi services in Shanghai: Using exploratory spatial data analysis", **ISSN:** 2161-0258, 19th International Conference on Geoinformatics, pp: 1 – 5.
11. Prasadu Peddi (2020), "Public Auditing Mechanism toVerify Data Integrity in Cloud Storage", ISSN 2347 – 3983, Volume 8, No 9, pp: 5220-5225.
12. G. Sekhar Reddy, Dr Ch. Suneetha (2020), "Conceptual Design of Data Warehouse using Hybrid Methodology", ISSN 2278-3091 ,Volume 9, No 3, pp: 2567 – 2573.
13. Nida Kousar G, Dr. Dayanand Lal N (2020), "A Virtual Machine Introspection in Cloud Computing for Intrusion Detection", ISSN 2278-3091 , Volume 9, No 3, pp: 2662– 2666.