International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Data Mining Technique for Temporal Association Mining using SPN-Sigmoid Neural Networks

[1*]Vaishali Sahu, [2]Anubhav Sharma, [3]Anshul Sarawagi

[1,2,3]CS Department, IES College of Technology, Bhopal, India

DOI: https://doi.org/10.26438/ijcse/v7i5.14591465 | Available online at: www.ijcseonline.org

Accepted: 26/Apr/2019, Published: 31/May/2019

***Abstract*—** Data mining is a methodology that takes information as information and yields learning. Such information objects, which are overwhelming not quite the same as or conflicting with the staying set of information, are called exceptions. An anomaly is an informational index which is not quite the same as the rest of the information. In recent research extraction of temporal information that too in specific medical domain came into significance, where the different research performed in this segment. In existing work paper CRF based technique which is conditional random field's model is used. They achieved best f-measure, accuracy and precision parameters while comparing with other approach such as Baseline, CRF+ Lexical is used. The future work remain by the research is developing of semi-supervised scheme for the temporal extraction and also working with un-annotated data text to make it annotating and thus obtaining better precision, recall, accuracy and F- Measure values.

Keywords— Rule Mining, Classification, Data Mining Algorithms, K-Theory.

I. INTRODUCTION

In the recent era, a large amount of raw data is being gathering day by day and storing in databases anywhere across the world, which is mainly collecting from different industry and social media sites. There is a requirement to extract and determine useful data and knowledge from such a data that is being collected. Data mining is an interdisciplinary field of computer science. It is referred to as mining knowledgeable data from large databases. It is the process of performing automated extraction and generating the predictive information from a large database. It is the processes of searching the hidden information from the repositories.

The fields that use Data mining techniques include medical research, marketing, telecommunication, and stock markets, health care and so on. In information retrieval, tf–idf, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

Data mining consists of the different technological methods including machine learning, statistics, database system etc. The aim of the data mining process is to discover knowledge from large databases and transform into a human understandable format. Data mining with knowledge discovery is important parts to the organization due to its decision making strategy.

Classification, clustering and regression are three methods of data mining. In these methods instances are grouped into identified classes. Classification is a popular task in data mining especially in knowledge discovery. It gives an intelligent decision making. Classification is not only studies and examines the existing sample data but also predicts the future behaviour of that information. It maps the data into the predefined class and groups. It is used to predict group membership for data instances.

A semantic TF-IDF based weighting method is proposed in the current paper. The vector is used for redefining semantic weights and thus the similarity of tweets. For a given tweet, T, the tags of the Top N similar tweets are recommended. The classical metrics of data mining is used for evaluating current approach. Semantic similarity and relatedness algorithms are compared and results showed significant improvement than normal TF-IDF weighting schema.

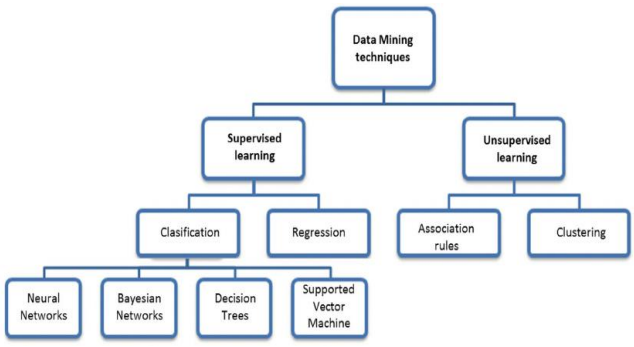

Figure 1: Data Mining Techniques.

Several data mining techniques enable to obtain decision rules (IF-THEN) (Kashani et al., 2011), that can help to understand the accident occurrence and the factors involved on it. These techniques are highly interpretable and could be used to expert learning. These techniques are association rules and decision trees. Therefore, in this project data mining techniques will be used in order to correlate ESM, crashes and TCA. Particularly, decision tree will be applied with the aim of obtaining relations like (IF-THEN) that are easily understandable.

Sentiment analysis [3-4] is a kind of characteristic language handling for following the temperament of people in general about a specific item or point. sentiment analysis, which is likewise called opinion mining, includes in building a framework to gather and look at sentiments about the item made in blog entry, remarks, audits or tweets. There are a few difficulties in opinion Analysis. The first is a feeling word that is thought to be sure in one circumstance might be thinks about negative in another circumstance. A second test is that individuals don't generally express opinions similarly. Most conventional content preparing depends on the way that little differences between two bits of content don't change the importance extremely much [1-10].

II. RELATED WORK

Hao Yan, Bo Zhang In this paper an algorithm to determine short text using semantic knowledge is discussed. Here two modes of detection which is offline and online mode is provided by the author. The given processes first take the input from the user and then process it first by text segmentation process. The segmentation process creates the different segment of values. Further term building using the segmented value and tag generation from the value is performed. Then based on term understanding maximum clique is determined. Single chain and pair is detected so that data strength can be taken for processing. Weight detection is performed over the large data understanding and thus value output is generated. MaxCMC and CMaxC both the algorithms were used for computation. Twitter dataset is used for processing and further computation cost, precision is computed for the analysis purpose. A high precision is shown for the computation with data analysis pair wise and chain model [21].

A. B. M. Rezbaul Islam This paper work towards the TF-IDF approach which work with similarity measure algorithm with dataset. This approach work with similarity recommendation approach. Data text determination, computation of relation in between the algorithm given words such as #frd and #friend can be computed is solved in this paper. The algorithm computes with high accuracy, precision and better recall over previous IDF approach. A similarity measure score is computed and weight determination to solve the given issue. This paper lacks in processing with large number of data and noise removal entity [22].

The writers SongnianLi, Suzana Dragicevic, et al. in [23] made survey on different geospatial hypothesis and techniques used to deal with geospatial huge information. Given some uncommon properties, creators considered that standard information taking controlling philosophies and systems are missing and the accompanying spaces were perceived as in necessity for promote headway and examination in the control. This fuses the headways in counts to oversee constant investigation and to help progressing flooding information, and in addition enhancing new spatial ordering strategies. The change of hypothetical and methodological approaches to manage exchange of huge information from illustrative and parallel research and applications to ones that examines agreeable and illustrative associations.

There are heaps of information mining examines the world over. Understudies Mood acknowledgment [23] was proposed by Christos N. Moridis et. al. for online self-evaluation test. Exponential rationale and recipes were utilized in this respects. The sources of info were understudy's past answers and slide bar status. The exponential rationale factors were an all out number of inquiries for the online self-appraisal test, understudy's objective, and slide bar esteem. Suitable criticisms are recorded dependent on current status of mind-sets of the understudies. Understudy's manual determination of their temperament utilizing slide bar with no computerization is the confinement of the framework.

In An improved Apriori Algorithm for Association Rules of Mining [24] the essential ideas of affiliation rule mining and the traditional Apriori calculation is talked about. The plan to improve the calculation is likewise talked about. The new calculation is made that chips away at the accompanying procedure, right off the bat, separate each gained information as per discretization of information things and tally the information while filter the database, furthermore, prune the procured thing sets. After investigation, the improved calculation lessens the framework assets involved and improves the effectiveness and quality.

An Improved Apriori Algorithm [25] called APRIORI-IMPROVE is proposed dependent on the constraints of Apriori. APRIORI-IMPROVE calculation presents enhancements on 2-things age, exchanges pressure and uses hash structure to produce L2, utilizes an effective.

An Improved Apriori-based Algorithm for Association Rules Mining [26] explains that in view of the fast development in overall data, effectiveness of affiliation rules mining (ARM) has been worried for quite a long while. In this paper, in light of the first Apriori calculation, an improved calculation IAA

is proposed. IAA receives another check based technique to prune hopeful itemsets and utilizations age record to diminish absolute information examine sum. Trials show that our calculation beats the first Apriori and some other existing ARM techniques... In this paper, an improved Apriori-based calculation IAA is proposed. Through pruning competitor itemsets by another check based strategy and diminishing the mount of sweep information by hopeful age record, this calculation can lessen the repetitive activity while producing incessant itemsets and affiliation governs in the database. Approved by the trials, the improvement is eminent. This work is a piece of our Distributed Network Behavior Analysis System, however we have considered C-R issue in our calculation, for explicit dataset, more work is as yet required. We additionally need further research to actualize this calculation in our conveyed framework.

Streamlining of Association Rule Mining through Genetic Algorithm [27] clarifies the Strong principle age is a significant region of information mining. In this paper creators plan a novel technique for age of solid guideline. In which a general Apriori calculation is utilized to produce the guidelines after that creators utilize the advancement systems. Hereditary calculation is a standout amongst the most ideal approaches to improve the principles. Toward this path for the advancement of the standard set they structure another wellness work that utilizes the idea of regulated adapting then the GA will probably create the more grounded guideline set.

An Improved Apriori Algorithm Based on Pruning Optimization and Transaction Reduction [28] explains the fundamental thoughts and the weaknesses of Apriori calculation, ponders the present significant improvement procedures of it. The improved Apriori calculation dependent on pruning advancement and exchange decrease is proposed. As indicated by the exhibition examination in the reproduction analyze, the quantity of regular thing sets is substantially less and the running time is essentially decreased just as the presentation is upgraded then at long last the calculation is improved.

III. PROBLEM DEFINITION

Today the World Wide Web is popular and interactive medium to distribute information. The web is huge, diverse, dynamic and unstructured nature of web data, web data research encountered lot of challenges for web mining. Information user could encounter following challenges when interacting with web.

Working with the short text and finding its proper meaning and usage is one of the important task objectives for the work.

1. Finding Relevant Information: Individuals either peruse or utilize the inquiry administration when they need to discover explicit data on the web. The present pursuit devices have issues like low exactness which is because of insignificance of a considerable lot of the list items. This outcomes in a trouble in finding the significant data. Another issue is low review which is because of powerlessness to record all the data accessible on the web.

2. Making New Knowledge out of the Information Available on the Web: This issue is fundamentally sub issue of the above issue. Above issue is inquiry activated procedure (recovery arranged) however this issue is information activated procedure that presumes that as of now has gathering of web information and concentrate possibly helpful learning out of it.

3. Personalization of Information: At the point when individuals associate with the web they contrast in the substance and introductions they like.

4. Finding out About Consumers or Individual Users: This issue is about what the client does and need. Inside this issue there are sub issue, for example, modifying the data to the planned purchasers or even to customize it to singular client, issue identified with web architecture and the board and promoting.

5. Finding or Analyzing the Large Data: Huge Amount of the information is unfit to screen and improve as indicated by the client necessity, so here the prerequisite is to locate the most ideal approach to break down it effectively.

IV. PROPOSED WORK

Here we briefly describe a technique to discovered frequent item pattern and hybrid approach with make use of Line-Up Approach in proposed system.

Symantec Mining: A Web mining from the slither is done first, we are separating the data from the online on the comparable kind of article and their accessibility in semantic way, the information is been extricated and use to make Entropy.

Synaptic Mining: In this calculation, the examples are sorted by the length executed on grid model. Examples will frame a cross section dependent on the example length and example recurrence. Also, utilizing this grid, visit designs are sought profundity first.

Grid Construction: The fundamental component of the cross section is a particle for example single page. Every iota or page represents length-1 prefix comparability class. Starting from base components the recurrence of upper components

with length n can be determined by utilizing two n-1 length examples having a place with a similar class.

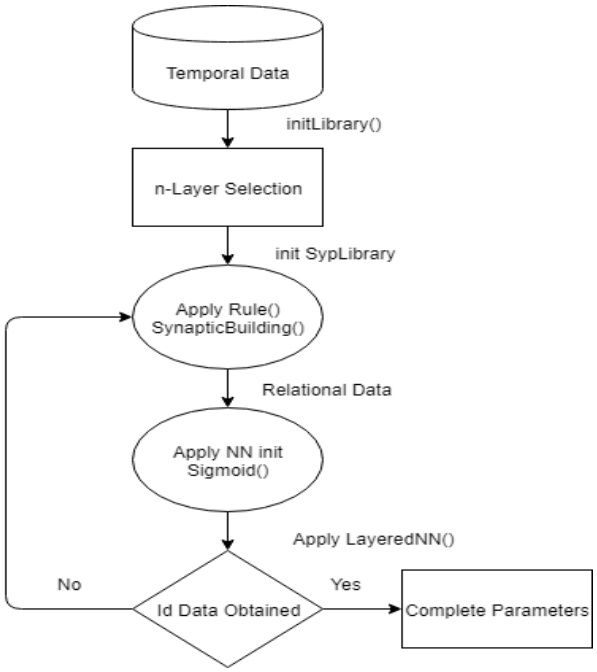

Figure 2: Flowchart of proposed algorithm.

We are applying now the imagine and SPN-Sigmoid NN system where the things can be substitute in different dataset and the outcome saw from the different semantic information and client can enhance as indicated by the Visualize and perception required.

Pseudo Code:

Input: temporal data TDi-n, n layer Symlib.

Output: temporal analysis, relational data, computation time, efficiency analysis.

Begin-

Init Param()

{

TDi-n();

Layer Selection n;

}

Initialize SynapticLibrary();

Foreach (TDi-n)

{

RuleApply(SynLib);

Apply Relational Building;

Finding Relational Data;

}

Apply nn;

Sigmoid init();

Apply Layered NN;

Return Temporal Relation;

Return Data;

}

Computation Time;'

Return parameters();

Exit;

}

END;

Applying Proper Vocabulary Mined Data: We are applying now the imagine and ANN system where the things can be substitute in different dataset and the outcome saw from the different semantic information and client can enhance as indicated by the Visualize and perception required.

V. EXPERIMENTAL SETUP & RESULT ANALYSIS

All experiment execution is performed on i3 Machine, Windows 10 Operating System with 750 GB HDD and 4 GB of RAM. Apache server foundation is used for running the application and WAMP server is used for data storage. A processing is performed with two different web service dataset.

Result Analysis

As per the experiment performed on different dataset and recommendation is generated. Below are the result observed in statically manner and graphical manner. Result analysis are performed in computational aspect and then confusion matrix performance aspect which is discussed below.

Confusion matrix aspect:

A confusion matrix generation and then computing of resources utilization parameter is given below.

In the confusion matrix aspect some variable parameter such as true positive, false positive, true negative and false

negative computation is performed. These parameter taken as input and further accuracy, precision, recall were computed.

Table 1: Comparative Analysis between Existing And Proposed Algorithm.

Dataset	Algorithm System	Accuracy %	Precision %	Recall%	MAE
WSDREAM-Dataset 1	Location & Region based approach	70.4	86.5	80.10	60.49
	SPN-Sigmoid NN	82.90	88.43	82	66.23
WSDREAM-Dataset 2	Location & Region based approach	83.3	78.23	87.68	64.78
	SPN-Sigmoid NN	84.70	81	83.90	66.0

As per the statistical result in table 1, further a comparison is made individually using graph by which a proper monitoring and observation can be made.

Table 2: Comparative Analysis Computation Parameter.

Dataset	Algorithm System	Computation time
WSDREAM- Dataset 1	Location & Region based approach	220
	SPN-Sigmoid NN	164
WSDREAM- Dataset 2	Location & Region based approach	298
	SPN-Sigmoid NN	203

In the table 2 above, other computation related to the computation time is given.

Graphical Result Analysis

An analysis of result graphically is discussed which help in understanding the observe parameter and their graphical monitoring.

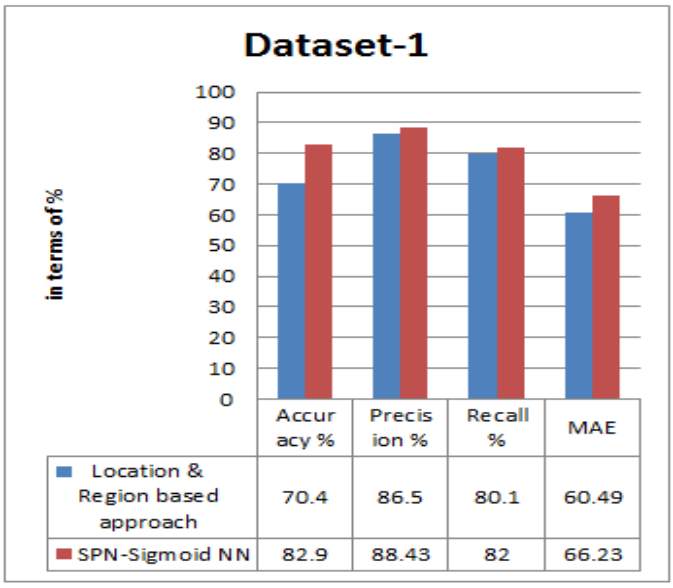

Figure 3: Result Analysis with the Dataset-1.

In the figure 3 above a graphical analysis of computational parameter, with the dataset-1 is computed.

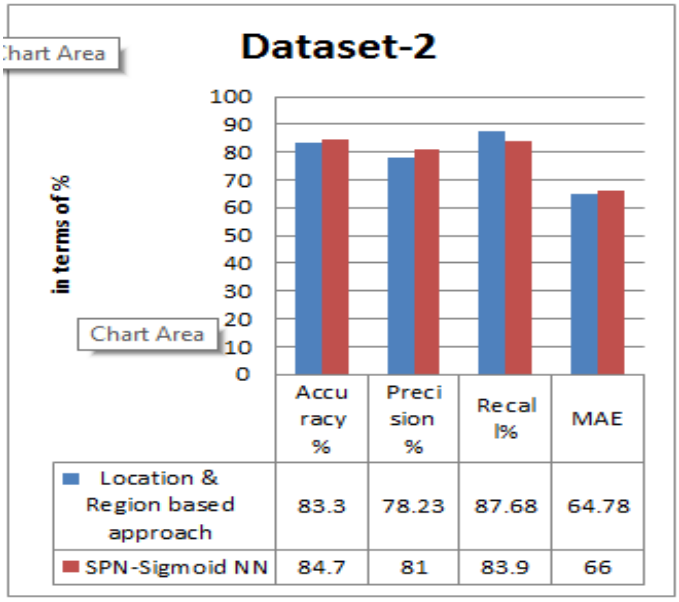

Figure 4: Dataset 2 Result Analysis.

In the figure 4 above a graphical analysis of computational parameter, with the dataset-2 is computed.

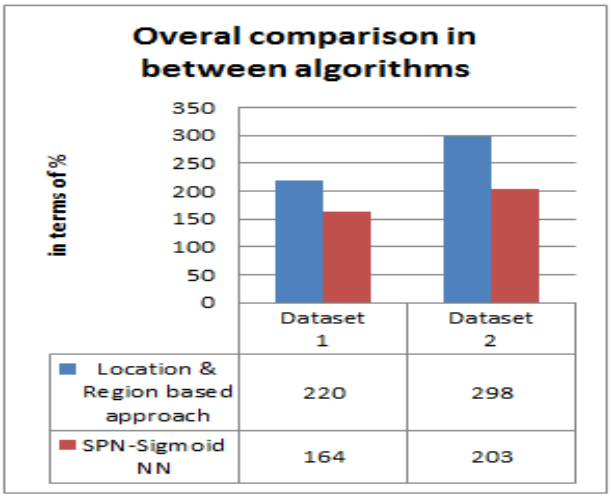

Figure 5: An Overall Comparison Analysis with Both the Dataset.

In the figure 5 above, an overall analysis over dataset-1, dataset-2 result monitoring is shown in the figure.

VI. CONCLUSION & FUTURE WORK

First data mining and web mining categories have been discussed. Semantic –Synaptic mining also has been analysed along their advantages and Entropy having their usage. An algorithm has been proposed for Mining and Visualizing of mining in web usage mining which is efficient than traditional Tabular scheme. The first part of algorithm, i.e. backwards can. Firstly scan the web log database and obtain the longest candidate level length. After that count the occurrence of each candidate. Each candidate count satisfy the minimum threshold value and then obtain the maximum forward reference from candidate count length. The new approach requires minimum repeated database scan for mining in web usage mining. It will reduce the time and space execution. Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Web usage mining consists of three phases, namely pre-processing, pattern discovery, and pattern analysis. One of the algorithms which is very simple to use and easy to implement is the SPN-Sigmoid NN algorithm. In this paper a new technique is proposed to discover the web usage patterns of websites from the server log files with the foundation of clustering and improved SPN-Sigmoid NN algorithm. The effective algorithm will be proposed with the improvements as well as the implementation of SPN-Sigmoid NN Algorithm.

The forthcoming step in the research work shall be to design the improved version of the SPN-Sigmoid NN Algorithm that shall be implemented on the synaptic mining and spatial data information.

Acknowledgment

I am very thankful to my guide who has helped me throughout my work.

REFERENCES

[1] M. Nikhil Kumar*, K.V.S. Koushik, K. John Sundar, Data Mining and Machine Learning Techniques for Cyber Security Intrusion Detection, 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307.

[2] Deepashri.K.S, Survey on Techniques of Data Mining and its Applications, International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-6, Issue-2) 2017.

[3] Aarti Sharma et al, Application of Data Mining – A Survey Paper, International Journal of Computer Science and Information advancements', Vol. 5 (2), 2014.

[4] Smita, Priti and Sharma, Use of Data Mining in Various Field: A Survey Paper IOSR Journal of Computer Engineering, 8727Volume 16, Issue 3, Ver. V (May-Jun. 2014)

[5] Brijesh Kumar Baradwaj, Saurabh Pal Mining Educational Data to Analyze Students Performance (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011

[6] J. Han and M. Kamber. Data Mining, Concepts and Techniques, Morgan Kaufmann, 2000.

[7] Nikita Jain, Vishal Srivastava DATA MINING TECHNIQUES: A SURVEY PAPER IJRET: International Journal of Research in Engineering and Technology, Volume: 02 Issue: 11 | Nov-2013.

[8] Prof. Dr. Wolfgang Karl Hardle, Time Series Data Mining Methods: A Review, Berlin, March 25, 2015.

[9] Pradnya P. Sondwale, Overview of Predictive and Descriptive Data Mining Techniques IJARCSSE, Volume 5, Issue 4, April 2015

[10] Data Mining: Concepts and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management System's) third Edition.

[11] Sonali Agarwal, Data Mining in Education: Data Classification and Decision Tree Approach, International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 2, No. 2, April 2012.

[12] G. N. Pandey Data Classification and Decision Tree Approach, International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 2, No. 2, April 2012.

[13] A. Merceron and K. Yacef, Educational Data Mining: a Case Study, In C. Looi; G. McCalla; B. Bredeweg; J. Breuker, article director, Proceedings of the twelfth overall Conference on Artificial Intelligence in Education AIED, pp. **467–474. Amsterdam, IOS Press, 2005.**

[14] Chanchal Yadav, Algorithm and approaches to manage handle immense Data-A Survey, IJCSN International Journal of Computer Science and Network, Vol 2, Issue 3, 2013 ISSN (Online) : 2277-5420.

[15] Mehmet Koyuturk, Ananth Grama, and Naren Ramakrishna, Compression, Clustering, and Pattern Discovery in uncommonly High-Dimensional Discrete-Attribute Data Sets, IEEE Transactions on Knowledge and Data Engineering, April 2005, Vol. 17, No. 4.

[16] Trupti A. Kumbhare Prof. Santosh V. Chobe, An Overview of Association Rule Mining Algorithms, Trupti A. Kumbhare et al,/(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 927-930.

[17] Jyoti Arora1, Nidhi Bhalla2, Sanjeev Rao3, A REVIEW ON ASSOCIATION RULE MINING ALGORITHMS, ISSN ONLINE (2320-9801).
[18] Soumadip Ghosh1, Association Rule Mining Algorithms and Genetic Algorithm: A Comparative Study, 2012 Third International **Conference on Emerging Applications of Information Technology (EAIT).**
[19] Zebang Chen, Takehiro Yamamoto, Katsumi Tanaka, Query Suggestion for Struggling Search by Struggling Flow Graph, IEEE 2016.
[20] Huiping Peng, Discovery of Interesting Association Rules Based on Web Usage Mining **2010** International Conference.
[21] Hao Yan, Bo Zhang, Yibo Zhang, Fang Liu, Zhenming Lei Web use mining reliant on WAN customers' practices **2010** International Conference.
[22] A. B. M. Rezbaul Islam; Tae-Sun Chung, An Improved Frequent Pattern Tree Based Association Rule Mining Technique,IEEE 2011.
[23] Songnian Li, SuzanaDragicevic, FrancesAnton Castro, Monika Sester, Stephan Winter, ArzuColtekin, Christopher Pettit, Geospatial gigantic data managing theory and systems: A review and research challenges, Volume2 | Issue2 || March-April-2017 | www.ijsrcseit.com 97 ISPRS Journal of Photogrammetry and Remote Sensing, pp. 119-133, Volume 115, May 2016.
[24] WEI Yong-qing, YANG Ren-hua, LIU Pei-yu, An Improved Apriori Algorithm for Association Rules of Mining IEEE(2009).
[25] Rui Chang, Zhiyi Liu, An Improved Apriori Algorithm, 2011 International Conference on Electronics and Optoelectronics **(ICEOE 2011).**
[26] Huan Wu, Zhigang Lu, Lin Pan, Rongsheng Xu, Wenbao Jiang, An Improved Apriori-based Algorithm for Association Rules Mining, Sixth International Conference on Fuzzy Systems and Knowledge Discovery, **IEEE society, 2009**.
[27] Rupali Haldulakar, Prof. Jitendra Agrawal, Optimization of Association Rule Mining through Genetic Algorithm, International Journal on Computer Science and Engineering (IJCSE), Vol. 3, Issue. 3, Mar 2011.
[28] Zhuang Chen, Shibang CAI, Qiulin Song and Chonglai Zhu, An Improved Apriori Algorithm Based on Pruning Optimization and Transaction Reduction, IEEE 2011.
[29] Stonebraker, M., Çetintemel, U., Zdonik, S.: The 8 necessities of continuous stream taking care of. ACM SIGMOD Record. 34, 42-47 (2005).
[30] Cugola, G., Margara, A.: Processing Flows of Information : From Data Stream to Complex Event Processing. ACM Computing Surveys. Niblett, P.: Event Processing In Action. (2010).

