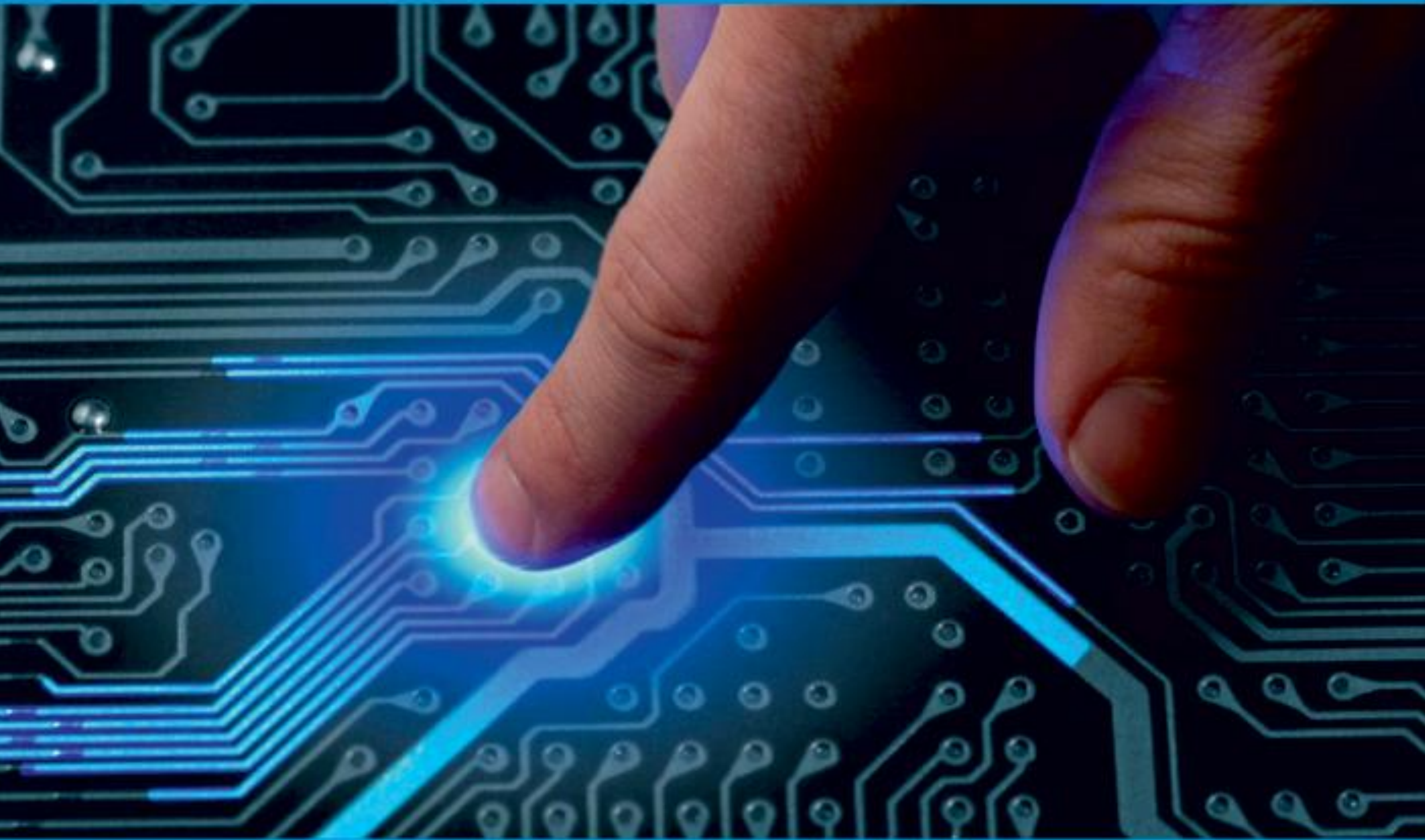
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 4, April 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 4, April 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1004143|

Implementation towards Apriori Versions based on MapReduce for Mining Frequent Patterns on Big Data

Wagh Chetana[1], Gadekar Shital[2], Totala Pranav[3], Gaikwad Prathamesh[4], Prof. R. B. Pawar[5]

UG Student, Dept. of I.T., Dept. of I.T., AVCOE, Ahmednagar, Maharashtra, India[1,2,3,4]

Assistant Professor, Dept. of I.T., AVCOE, Ahmednagar, Maharashtra, India[5]

ABSTRACT: All the proposed models are based on the well-known Apriori algorithm and the MapReduce framework. The proposed algorithms are divided into three main groups. Two algorithms Apriori MapReduce (AprioriMR) and iterative AprioriMR (IAprioriMR) are properly designed to extract patterns in large datasets. These algorithms extract any existing item-set in data regardless their frequency. Pruning the search space by means of the antimonotone property. Two additional algorithms space pruning AprioriMR (SPAprioriMR) and top AprioriMR (TopAprioriMR) are proposed with the aim of discovering any frequent pattern available in data. Maximal frequent patterns. A last algorithm maximal AprioriMR (MaxAprioriMR) is also proposed for mining condensed representations of frequent patterns, i.e., frequent patterns with no frequent supersets.

KEYWORDS: Frequent Itemset, Apriori ,Apriori MapReduce ,Iterative AprioriMR, Space pruning AprrioriMR

I. INTRODUCTION

Pattern mining is one of the most important tasks to extract meaningful and useful information from raw data. This task aims to extract item-sets that represent any type of homogeneity and regularity in data.MapReduce is an emerging paradigm that has become very popular for intensive computing. Pruning the search space by means of the antimonotone property. Two additional algorithms [space pruning AprioriMR (SPAprioriMR) and top AprioriMR (TopAprioriMR)] are proposed with the aim of discoveringany frequent pattern available in data.To live in the Big Data Era implies data being gathered everywhere, at every moment, from different devices and, most of thetime, in an almost imperceptible way. Taking advantage of such information is essential for many organizations as well asgovernments in decision-making to improve our daily life (Kraska, 2013). Data analytics systems are therefore boomingthanks to their capacity to extract hidden, effective, and usable knowledge from large collections of data. Though many different tasks come under the umbrella of data analysis or data mining, frequent itemset mining (FIM) is, from the very outset, anessential task due to its ability to extract frequently occurring events, patterns, or items (symbols or values) in data(Aggarwal & Han, 2014).In the process of transforming raw data into significant and meaningful information for making sense of the data, the keyelement is the pattern (a singleton or set of items) which represents any type of homogeneity and regularity, and it is thereforeconsidered as a good descriptor of intrinsic and important properties of the data (Han & Kamber, 2000). Numerous FIM algorithms have been proposed since the first approach was described at the beginning of the 1990s (Agrawal, Imielinski, &Swami, 1993). In that approach, a levelwise breadth first search methodology was responsible for producing candidateitemsets whose frequency counting was performed by reading the dataset multiple times (one for each size of candidateitemsets). Later algorithms such as FP-Growth (Han, Pei, & Yin, 2000) and ECLAT (Zaki, 2000), on the contrary, were based.

II. RELATED WORK

In this paper, a hybrid version of Apriori and MapReduce for the fast and efficient execution is shown. The Apriori algorithm deployed on the MapReduce platform with suitable frequent key values. The hybrid approach is executed on the dataset and provides more accurate result. Experimental results show that the algorithm scales up linearly with respect to dataset sizes.

In this paper, Mining class association rules (CARs) with the item set constraint is concerned with the discovery of rules, which contain a set of specific items in the rule antecedent and a class label in the rule consequent. This task is commonly encountered in mining medical data. For example, when classifying which section of the population is at high risk for the HIV infection, epidemiologists often concentrate on rules which include demographic information

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 4, April 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1004143|

such as gender, age, and marital status in the rule antecedent, and HIV-Positive in the rule consequent. There are two naive strategies to solve this problem, namely pre-processing and post-processing. The post-processing methods have to generate and consider a huge number of candidate CARs while the performance of the pre-processing methods depend on the number of records filtered out. Therefore, such approaches are time consuming. This study proposes an efficient method for mining CARs with the itemset constraint based on a lattice structure and the difference between two sets of object identifiers (diffset)[1]

In this paper, proposes the Apriori Algorithm based frequent trajectory pattern mining algorithm to efficiently and effectively handle the trajectory database transaction. Prior to that the trajectory dataset is extracted from a text file and is imported to a Oracle database after doing the initial data cleaning process. Initial frequency count is done in Oracle database using its programming feature. Then the data is written in the operating system then further processing is done to find the frequent trajectory pattern. Advantage of this method is later iterations are much faster than the initial iterations of the algorithm. The results obtained by this method are more accurate and reliable. This algorithm uses large coordinate set property. Each iteration in this algorithm can be parallelized so that execution time can be reduced. More over this algorithm is easy to implement. Disadvantage of this method are, it uses a generate, prune and test approach generates candidate coordinate sets (1-coordinate, 2- coordinate, 3-coordinate,…), to check the generated sequence of coordinates are already generated or not, and tests if they are frequent by scanning the database and counting their support each time. Generation of candidate coordinate sets is expensive (in both space and time). Since generation and pruning steps are in memory resident, it needs more RAM. Another disadvantage is it needs n+1 database scans, n is the length of the coordinates in the longest pattern.[2]

In this paper most existing algorithms mine frequent patterns from traditional transaction databases that contain precise data. In these databases, users definitely know whether an item (or an event) is present in, or is absent from, a transaction in the databases. However, there are many real-life situations in which one needs to deal with uncertain data. In such data users are uncertain about the presence or absence of some items or events. For example, a physician may highly suspect (but cannot guarantee) that a patient suffers from a specific disease. The uncertainty of such suspicion can be expressed in terms of existential probability. Since there are many real-life situations in which data are uncertain, efficient algorithms for mining uncertain data are in demand. Two algorithms have been proposed for mining frequent patterns from uncertain data. The previous two algorithms follow the horizontal data representation. In this paper we studied the problem of mining frequent itemsets from existential uncertain data using the Tidset vertical data representation. We introduced the U-Eclat algorithm, which is a modified version of the Eclat algorithm, to work on such datasets. A performance study is conducted to highlight the efficiency of the proposed algorithm also a comparative study between the proposed algorithm and the well known algorithm UF-growth is conducted and showed that the proposed algorithm outperforms the UF-growth.[3]

In this paper, we have proposed new efficient pattern mining algorithms to work in big data. All the proposed models are based on the well-known Apriori algorithm. This algorithm has been also proposed for mixing condensed representations of frequent patterns. Pruning the search space by means of anti-monotone property. Two additional algorithms have been proposed with the aim of discovering any frequent pattern available in data. In Future, We will use the Top – K Ranking Algorithm to find the top k frequent patterns from the given dataset. Ranking functions are evaluated by a variety of means; one of the simplest is determining the precision of the first k top-ranked results for some fixed k; Frequently, computation of ranking functions can be simplified by taking advantage of the observation that only the relative order of scores matters, not their absolute value; hence terms or factors that are independent of the features may be removed, and terms or factors that are independent of the feature may be precomputed and stored with the dataset.[4].

III. PROPOSED SYSTEM

A. *Methodology:*

We propose new efficient pattern mining algorithms to work in big data. All of them rely on the MapReduce framework and the Hadoop open-source implementation.Two of these algorithms (AprioriMR and IAprioriMR) enable any existing pattern to be discovered.Two additional algorithms (SPAprioriMR and TopAprioriMR) use a pruning strategy for mining frequent patterns.Finally, an algorithm for mining MaxAprioriMR is also proposed.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 4, April 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1004143|

B. *System Architecture*

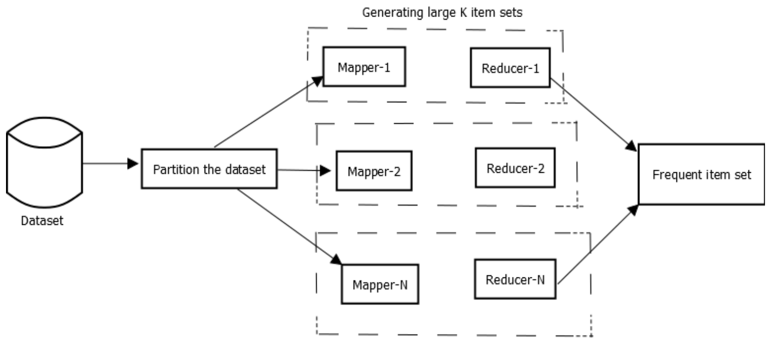


FIG1: System Architecture

C. *Flowchart:*

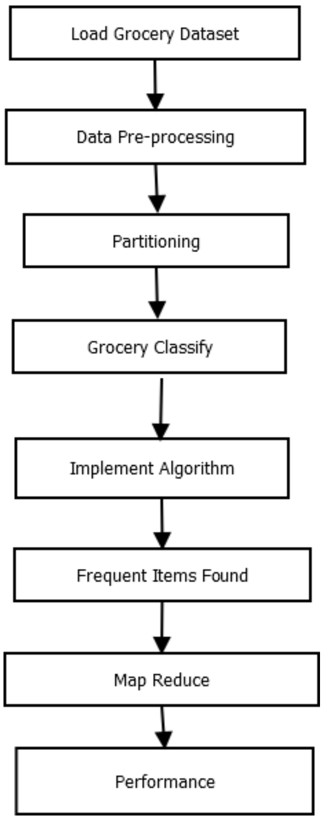


Fig2: Flowchart

D. *Modules:*

1) *Pattern Mining*

The term pattern is defined as a set of items that represents any type of homogeneity and regularity in data, denoting intrinsic and important properties of data

It is noteworthy the support of a pattern is monotonic i.e., none of the super-patterns of an infrequent pattern can be frequent

2) *Map Reduce*

Map Reduce is a recent paradigm of parallel computing. It allows to write parallel algorithms in a simple way, where the applications are composed of two main phases defined by the programmer: 1) map and 2) reduce. In the map phase, each map per processes a subset of input data and produces key-value (k, v) pairs.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 4, April 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1004143|

E. *Algorithms:*

Algorithm 1 Original Apriori Algorithm

Input: $\mathcal{T}$ // set of transactions
Output: L // list of patterns found in data
  $L = \emptyset$
  for all $t \in \mathcal{T}$ **do**
    for $(s = 1; s \leq |t|; s{+}{+})$ **do**
      $C = \{\forall P : P = \{i_j, \ldots, i_n\} \wedge P \subseteq t \wedge |P| = s\}$
      // candidate item-sets in t
      $\forall P \in C$, then $support(P) = 1$
      if $C \cap L \neq \emptyset$ **then**
        $\forall P \in L : P \in C$, then $support(P) + +$
      end if
      $L = L \cup \{C \setminus L\}$ // include new patterns in L
    end for
  end for
  return L

Algorithm 2 AprioriMR Algorithm

begin procedure AprioriMapper(t_l)
  for $(s = 1; s \leq |t_l|; s{+}{+})$ **do**
    $C = \{\forall P : P = \{i_j, \ldots, i_n\} \wedge P \subseteq t_l \wedge |P| = s\}$
    // candidate item-sets in t_l
    $\forall P \in C$, then $supp(P) = 1$ // support is initialized
    for all $P \in C$ **do**
      emit $\langle P, supp(P)_l \rangle$ // emit the $\langle k, v \rangle$ pair
    end for
  end for

IV. RESULTS

Fig. 3. IaprioriMR

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |
|| Volume 10, Issue 4, April 2022 ||
| DOI: 10.15680/IJIRCCE.2022.1004143|

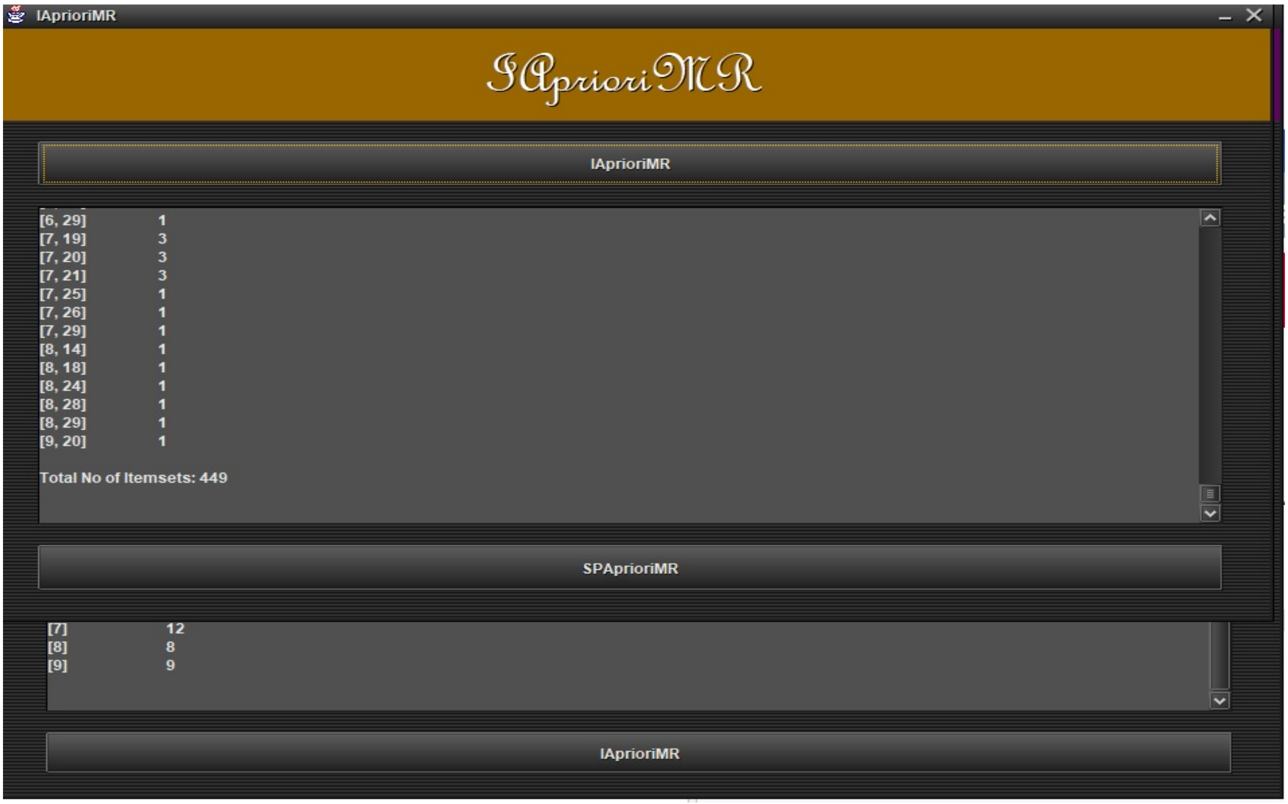


Fig.4. AprioriMR

Fig.5. TopaprioriMR

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 4, April 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1004143|

Fig.6. MaxAprioriMR

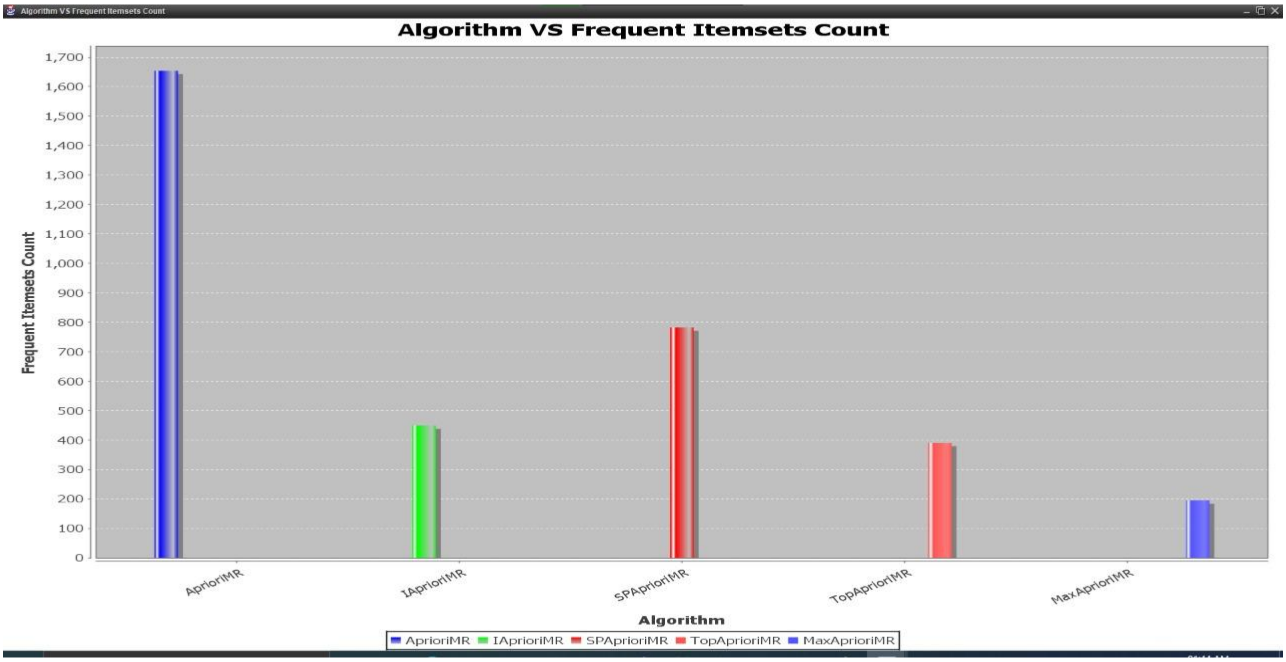


Fig.7. Analysis Of all Algorithms

V. CONCLUSION AND FUTURE WORK

In this project, we have proposed new efficient pattern mining algorithms to work in big data. All the proposed models are based on the well-known Apriori algorithm and the MapReduce framework.The proposed algorithms are divided into three main groups.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 4, April 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1004143|

- No pruning strategy. Two algorithms (AprioriMR and IAprioriMR) for mining any existing pattern in data have been proposed.
- Pruning the search space by means of anti-monotone property. Two additional algorithms (SPAprioriMR and TopAprioriMR) have been proposed with the aim of discovering any frequent pattern available in data.
- Maximal frequent patterns. A last algorithm (MaxAprioriMR) has been also proposed for mining condensed representations of frequent patterns.

REFERENCES

[1] Sharma A., Tripathi K. (2021) Hybrid Version of Apriori Using MapReduce. In: Marriwala N., Tripathi C.C., Kumar D., Jain S. (eds) Mobile Radio Communications and 5G Networks. Lecture Notes in Networks and Systems, vol 140. Springer, Singapore.

[2] Wang C, Zheng X (2019) Application of improved time series Apriori algorithm by frequent itemsets in association rule data mining based on temporal constraint. EvolIntell 1–11.

[3] Zaki FA, Zulkurnain NF (2019) Frequent itemset mining in high dimensional data: a review. Computational science technology. Springer, Singapore, pp 325–334

[4] Xie DF, Wang MH, Zhao XM (2019) A spatiotemporal Apriori approach to capture dynamic associations of regional traffic congestion. IEEE Access

[5] Luna, J. M., Padillo, F., Pechenizkiy, M., & Ventura, S. (2018). Apriori versions based on MapReduce for mining frequent patterns on big data.IEEE Transactions on Cybernetics, 48(10), 2851–2865.

[6] Padillo, F., Luna, J. M., Herrera, F., & Ventura, S. (2018). Mining association rules on Big Data through Map Reduce genetic programming. Integrated Computer-Aided Engineering, 25(1), 31–48. https://doi.org/10.3233/ICA-170555

[7] Luna, J. M., Ondra, M., Fardoun, H. M., & Ventura, S. (2018). Optimization of quality measures in association rule mining: An empirical study.International Journal of Computational Intelligence Systems, 12, 59–78.

[8] Noaman, A. Y., Luna, J. M., Ragab, A. H. M., & Ventura, S. (2016). Recommending degree studies according to students' attitudes in high schoolby means of subgroup discovery. International Journal of Computational Intelligence Systems, 9(6), 1101–1117.

[9] J. M. Luna, "Pattern mining: Current status and emerging topics," Progr. Artif. Intell., vol. 5, no. 3, pp. 165–170, 2016.

[10] Luna, J. M., Cano, A., Pechenizkiy, M., \& Ventura, S. (2016). Speeding-up association rule mining with inverted index compression. IEEE Transactions on Cybernetics, 46(12), 3059–3072.

[11] C. C. Aggarwal and J. Han, Frequent Pattern Mining, 1st ed. Cham, Switzerland: Springer, 2014.

[12] Luna, J. M., Romero, J. R., Romero, C., & Ventura, S. (2014). Reducing gaps in quantitative association rules: A genetic programming freeparameter algorithm. Integrated Computer-Aided Engineering, 21(4), 321–337.

[13] J. M. Luna, J. R. Romero, C. Romero, and S. Ventura, "On the use of genetic programming for mining comprehensible rules in subgroup discovery," IEEE Trans. Cybern., vol. 44, no. 12, pp. 2329–2341, Dec. 2014. [Online]. Available: http://dx.doi.org/10.1109/TCYB.2014.2306819

[14] Moens, S., Aksehirli, E., & Goethals, B. (2013). Frequent itemset mining for big data. In Proceedings of the 2013 IEEE international conference onbig data (pp. 111–118). Santa Clara, CA.
Luna, J. M., Romero, J. R., & Ventura, S. (2012). Design and behavior study of a grammar-guided genetic programming algorithm for mining association rules. Knowledge and Information Systems, 32(1), 53–76.

[15] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: Current status and future directions," Data Min. Knowl. Disc., vol. 15, no. 1, pp. 55–86, 2007.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.165

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING