Improved Frequent Pattern Mining for Educational Data by using Mapreduce Approach in Hadoop

Than Htike Aung
Faculty of Information Science
University of Computer Studies, Yangon

Nang Saing Moon Kham
Faculty of Information Science
University of Computer Studies, Yangon

Soe Soe Mon
Faculty of Computer Systems and Technologies
University of Computer Studies, Hinthada

ABSTRACT

In this paper, we describe the formatting guidelines for IJCA Journal Submission.

In the education area of Myanmar, computers, mobile and internet have become important tools for high school students. To enable the quality and the flexibility of the education, verities of education programs and methods are greatly included but with different manners. In this paper, the field of large educational data and how big educational data can be analysis to provided quality improvement in education.

For the frequent pattern mining and exploitation of educational data, proposed system present improved data mining techniques and popular applied hadoop mapreduce for large data manipulation such as parallel processing data analysis such as learning, academic and visual analytics, providing examples of how these techniques and methods could be used.

The proposed system has been started pay attention to the teacher assessment application of data and data analytics to handle large data generated in the educational sector. These data stored in Hadoop file system, then discover frequent pattern by using mapreduce support apriori, eclat and prefix tree methods. These approached is effective and scalable for large data instead use of traditional standard data mining tools.

Keywords

Hadoop, Mapreduce, Eclat, Apriori, Prefix

1. INTRODUCTION

Since recent developing country have raised to produce the enormous amounts of data and storage, the discovery of knowledge large data analysis has become more critical for educational research. The modern-day cloud technology aid teacher best. When used precisely and appropriately information can provide an instructor with the wisdom and evidence required to aid educational politics while upgrade classroom experiences for every student.

Explosion of the Internet, social media and impact of technological innovations, train a basic part of youthful lives and future race. Mobile devices used in Myanmar alone was over $30 million in 2019.

Various education standards are developing electronic learning and information technology and communications subjects among the pre-school to higher level school. Consequently, electronic information data is increasing every year. To manage and maintenance storages required on every educational information to provide enhance educational purpose decision making. Educational data applied in data mining method evaluate quality of education system find out useful knowledge from large data set has getting to improve management and decision. Through the actual application of teacher assessment data of university of computer studies, Yangon in Myanmar, data amount grows and continues to increase. How to know student and teacher requirements in universities classroom and how to improve the quality of universities status for effective management is interest of administrator.

Data Mining is an essential for data analysis and finding knowledge for frequent pattern. Discovering frequent pattern was used the previous time use with sample data mining program and algorithms. Their memory capacity and processing power of computers is limit to both software program and hardware. Unfortunately, this nature is limited at future tense of analysis for large data. Large data to frequent pattern mining application is required for educational data mining. Processing to large data not only need large memory capacity and fast process power but also required parallel or batch processing frame work.

Now a day, hadoop and mapreduce is a open source processing framework suitable for academic work such as universities used for education. This proposed system is used Hadoop processing frame work. Yearly produced teacher assessment data can be enlarged and support that big data sets. Tradition data mining algorithm is modified to translate mapreduce algorithms run on hadoop. Hadoop is distributed framework and it can be stored vast data amount in distributed environment clusters. Mapreduce works mapper and reducer two important tasks. Mapper maps input key/value pairs to a set of intermediate key/value pairs. Reducer reduces a set of intermediate values which share a key to a smaller set of values. MapReduce uses parallel computing approach and HDFS is fault tolerant system. Hadoop distributed file system is exploited to find out frequent itemsets.

2. RELATED WORK

Distributed data is a process which obtained building of statistical data model to mining massive data ,L.Jure work the issue2.

Knowledge discovery process produced many steps cleaning the data to presentation of knowledge which work by Gorunescu 3.

The FIM problem is considered as the root of the pattern mining field, which encompasses multiple tasks that aim at extracting itemsets on multiple forms and for various purposes.

The FIM issue is express as field of frequent pattern mining, which includes various activities aimed at collecting itemsets in various ways and for specific purpose (Aggarwal & Han, 2014) 5.

Simple method FIM is the extraction of itemsets. The anti-monotone FIM methods have been rejected.

The problem of frequent pattern mining in different ways of methods survey as rare pattern mining in dynamic or static data that frequent or infrequent (Koh & Ravana, 2016) 14.

Starting from it, the pattern mining concept was extended with new ideas (Aggarwal & Han, 2014), considering small variations in some cases, like the one considering an item within a pattern (either frequent or infrequent) as interesting not only if it is present (postive patterns) but also if it is absent (negative patterns) in data (Savasere, Omiecinski, & Navathe, 1998;

H. Wang, Zhang, & Chen, 2008). Of course, an item and its opposite form (positive or negative) always produce a zero support value since any record cannot satisfy both at the same time.

Sergio A. Alvarez describe statistical method, such as Chi-squared method for datasets and association rule compute correlation between datasets 6 .

The mapreduce programming model is become parallel and clustering the input data. Intermediate key value combining, large clustering, large amount of memory fit and load balancing work present by J.Dean and; S.G. jeff 15 .

Frequent pattern mining model used compress and storing data structure for large data based which not made candidate generation. J.Han , J. Pei and Y. Yin work Mining Frequent Patterns with- out Candidate Generation 10 .

Agarwal et al. uses apriori algorithm for generating frequent itemsets and association rules. Level wise search. Frequency of an itemsets are counted by scanning the database D and then candidate k+1 itemsets are generated from frequent k-itemset by applying support and threshold condition.

Zaki et al. uses eclat algorithm which requires vertical database D needs to be stored in main memory. DistEclat developed by Zaki and Gouda store diffset(vertical data representation); diffset is difference between candidate itemset of size k and prefix frequent itemsets of size k-1; support value is based on diffset gaining performance growth than Eclat; it is not efficient when the database is sparse.

3. EDUCATIONAL DATA COLLECTION

The collection program proposed system designed is based on cloud database and mobile android application. The android teacher assessment apk installed on not below android 7 version. Teachers, Students and Administer must register before use of this application. Students register verification used QR code scanner with their student identification card. Teacher and Administer register verification used with preregister by administration staff records from their relevant Universities. Teacher assessment survey questions can used three languages as Myanmar Unicode, Myanmar Zawgyi, and English fonts. Thus, teacher assessment users can easily understand in survey questions. Moreover, their answers are accurately confident to their understandable questionnarie. The Figure 1. show mobile devices uses teacher assessment apk and cloud system.

Mobile phone application is used internet connection for data collection. Data collection mobile app installed on smart phone that provided teacher assessment survey system. The teacher assessment survey app is used google cloud platform Figure 2.

Fig 1: Cloud and Teacher Assessment Survey System Mobile Application

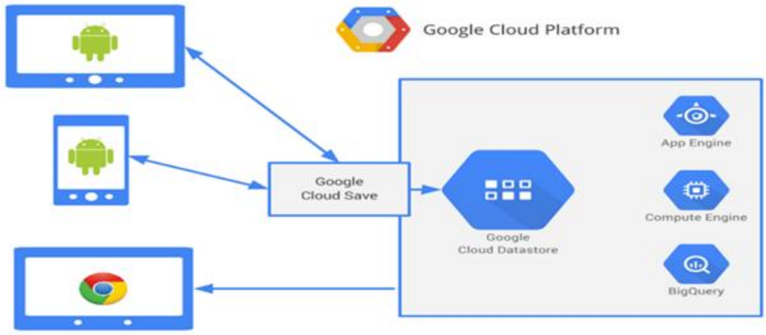
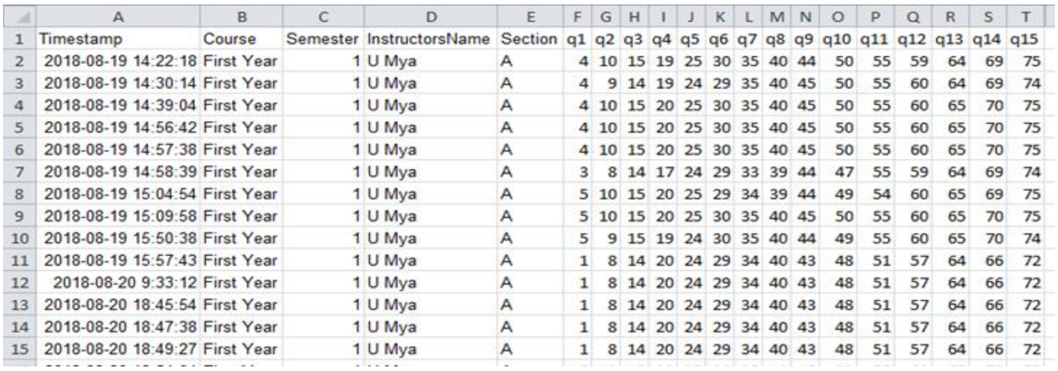


Fig 2: Google Cloud Platform

Table 1. Teacher Assessment Questionnaire

Attributes	Description		Ordinal
q1	The course content, course arrangement and assignments support course objectives.	Strongly Disagree=1, Disagree=2, Neutral=3, Agree=4, Strongly Agree=5	q1
q2	The course materials that the instructor prepared are good enough to grasp the concept of the course chapters.	Strongly Disagree=6, Disagree=7, Neutral=8, Agree=9, Strongly Agree=10	q2
----	-----	-------	----
q15	Overall, this course has been efficient to advance my learning.	Strongly Disagree=71, Disagree=72, Neutral=73, Agree=74, Strongly Agree=75	q15



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Timestamp	Course	Semester	InstructorsName	Section	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	q11	q12	q13	q14	q15
2	2018-08-19 14:22:18	First Year	1	U Mya	A	4	10	15	19	25	30	35	40	44	50	55	59	64	69	75
3	2018-08-19 14:30:14	First Year	1	U Mya	A	4	9	14	19	24	29	35	40	45	50	55	60	64	69	74
4	2018-08-19 14:39:04	First Year	1	U Mya	A	4	10	15	20	25	30	35	40	45	50	55	60	65	70	75
5	2018-08-19 14:56:42	First Year	1	U Mya	A	4	10	15	20	25	30	35	40	45	50	55	60	65	70	75
6	2018-08-19 14:57:38	First Year	1	U Mya	A	4	10	15	20	25	30	35	40	45	50	55	60	65	70	75
7	2018-08-19 14:58:39	First Year	1	U Mya	A	3	8	14	17	24	29	33	39	44	47	55	59	64	69	74
8	2018-08-19 15:04:54	First Year	1	U Mya	A	5	10	15	20	25	29	34	39	44	49	54	60	65	69	75
9	2018-08-19 15:09:58	First Year	1	U Mya	A	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75
10	2018-08-19 15:50:38	First Year	1	U Mya	A	5	9	15	19	24	30	35	40	44	49	55	60	65	70	74
11	2018-08-19 15:57:43	First Year	1	U Mya	A	1	8	14	20	24	29	34	40	43	48	51	57	64	66	72
12	2018-08-20 9:33:12	First Year	1	U Mya	A	1	8	14	20	24	29	34	40	43	48	51	57	64	66	72
13	2018-08-20 18:45:54	First Year	1	U Mya	A	1	8	14	20	24	29	34	40	43	48	51	57	64	66	72
14	2018-08-20 18:47:38	First Year	1	U Mya	A	1	8	14	20	24	29	34	40	43	48	51	57	64	66	72
15	2018-08-20 18:49:27	First Year	1	U Mya	A	1	8	14	20	24	29	34	40	43	48	51	57	64	66	72

Fig 3: Teacher Assessment Survey Database

Table 2. Teacher Assessment Survey Data

Dataset	Number of Transaction	Number of Different Items	Average Transaction Width
Tass (Teacher Assessment Survey Data)	1000011	75	15

4. PROPOSED SYSTEM ARCHITECTURE

System design is implemented in Figure 4. Firstly, input section consists of four parameters with datasets, number of maps and support count via the command line. AprioriMapReduce phase processed from input data set from the first stage.Then the next stage generate prefix tree and check prefix. If it is no prefix then go to Generate Prefix Tree stage repeat again. Check Prefix condition is yes then next to prefix tree process. EclatMapReduce process partition the data and generate frequent pattern the next stage. Finally output the frequent itemsets.

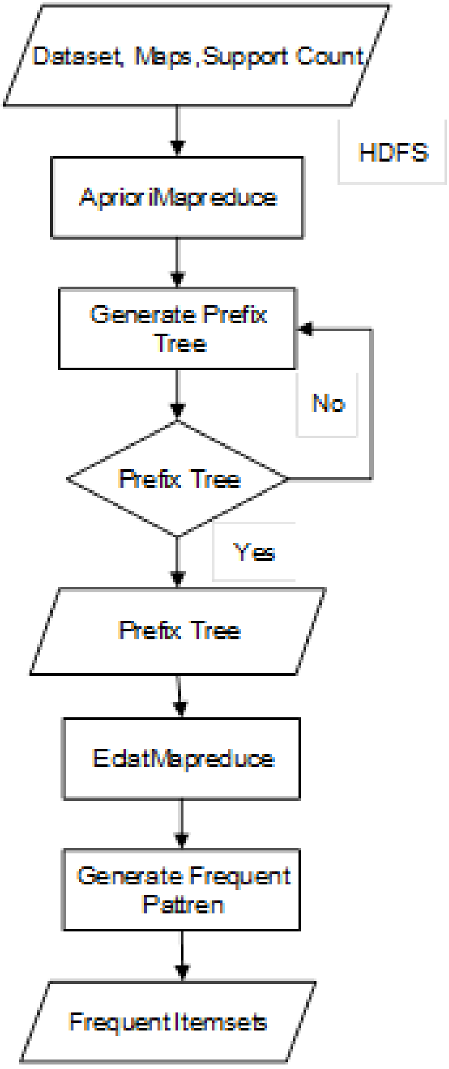


Fig : 4 Effects of selecting different switching under dynamic condition

1) The proposed system contains input dataset, number of maps and support count.
2) AprioriMapReduce process generated k-Frequent itemsets by using input large transaction list. Mappers received <key,value> pairs of transaction list. Reducer combine all local frequencies. Reducer redistributed global frequencies of item/itemsets to all mappers as candidate for the next time. Then repeated k-times process to get k-frequent itemsets. The candidates <key,value> pairs are partition into across the mappers can solved memory problem.
3) PrefixTree process generated k+1 Frequent itemsets (local frequent superset). Reducer combine all local frequencies to global frequencies list. Then redistributed to complete prefix groups of itemsets to all mappers.
4) EclatMapReduce mining part utilizes the diffsets to mine the prefix groups as a conditional database for frequent itemsets.

5. RESULTS AND ANALYSIS

For experiments machines are going to be used Intel® Core ™ i5-3230M CPU@2.60GHz processing units and 16.00GB RAM with Oracle BigData Lite 4.1 and Oracle VirtualBox VM 6.1.2.

Datasets used from Teacher Assessment Survey Dataset repository in University of Computer Studies, Yangon, Myanmar in order to compare results with existing systems such as Eclat Prefix Tree (ET) and Apriori Prefix Tree Eclat (ATE) 6,7. The following datasets used instead of educational data.

Table 3. Experimental used Datasets

Dataset	Number of Transaction	Number of Different Items	Average Transaction Width
dataset-1	1000011	75	15
dataset-2	100011	75	15

Table 4. Elapse Time (sec.) of mapreduce phases on dataset-1 (support count = 0.2) and (number of maps=1 ,2,3,4,5)

No of Mappers	Map=1	Map=2	Map=3	Map=4	Map=5
Support Count (%)	0.2	0.2	0.2	0.2	0.2
ET method (Time in sec)	79	80	80	81	81
ATEproposed method (Time in sec)	74	74	75	76	76

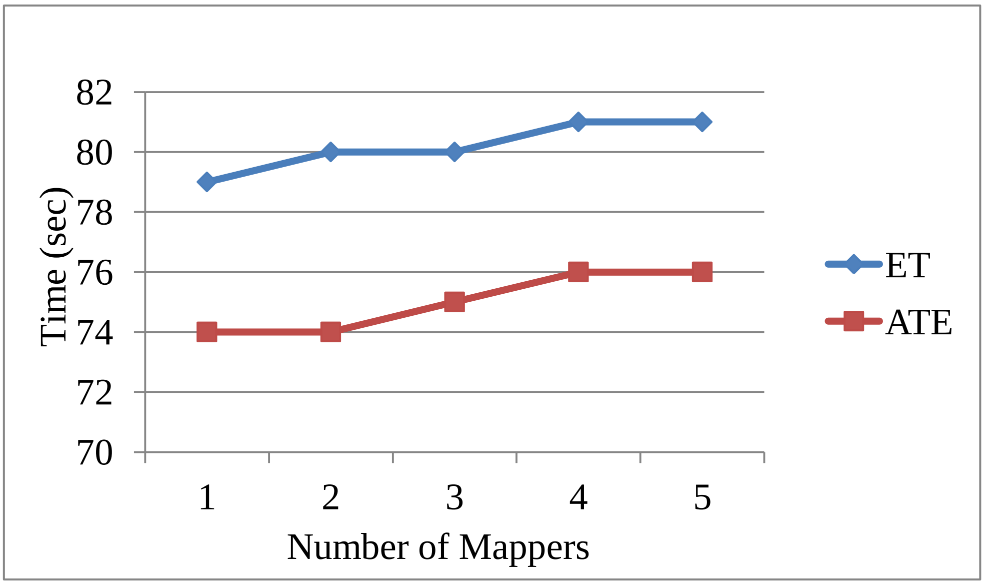


Fig :5 Execution time with varying number of Mappers for dataset-1

Table 5.Elapse Time (sec.) of mapreduce phases on dataset-1 (min_sup = 0.1,0.15,0.2,0.25,0.3) and (number of maps=2)

No of Mappers	Map=2	Map=2	Map=2	Map=2	Map=2
Support Count (%)	0.1	0.15	0.2	0.25	0.3
ET method (Time in sec)	81	81	80	74	74
ATEproposed method (Time in sec)	74	74	74	71	71

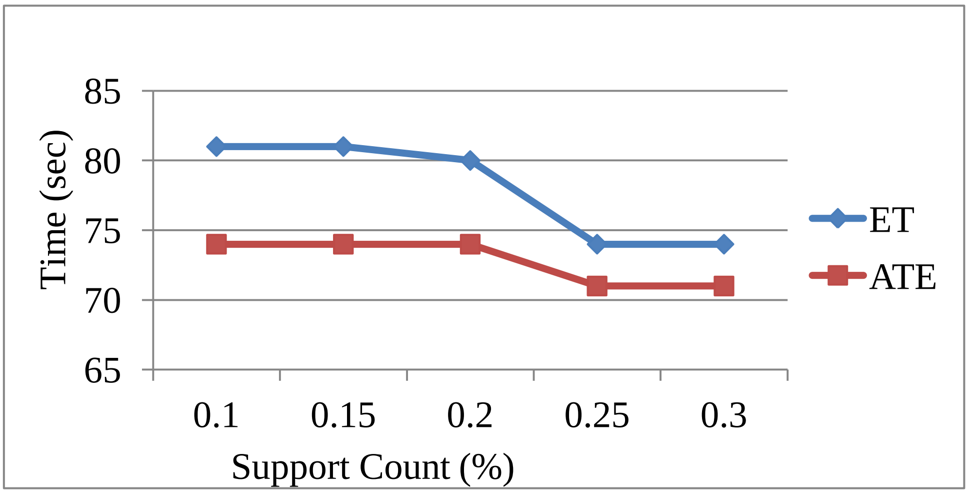


Fig : 6 Execution time with varying support count for dataset-1

Table 6. Elapse Time (sec.) of mapreduce phases on dataset -2 (support count = 0.2) and (number of maps=1 ,2,3,4,5)

No of Mappers	Map=1	Map=2	Map=3	Map=4	Map=5
Support Count (%)	0.2	0.2	0.2	0.2	0.2
ET method (Time in sec)	212	219	224	225	229
ATEproposed method (Time in sec)	209	211	216	218	222

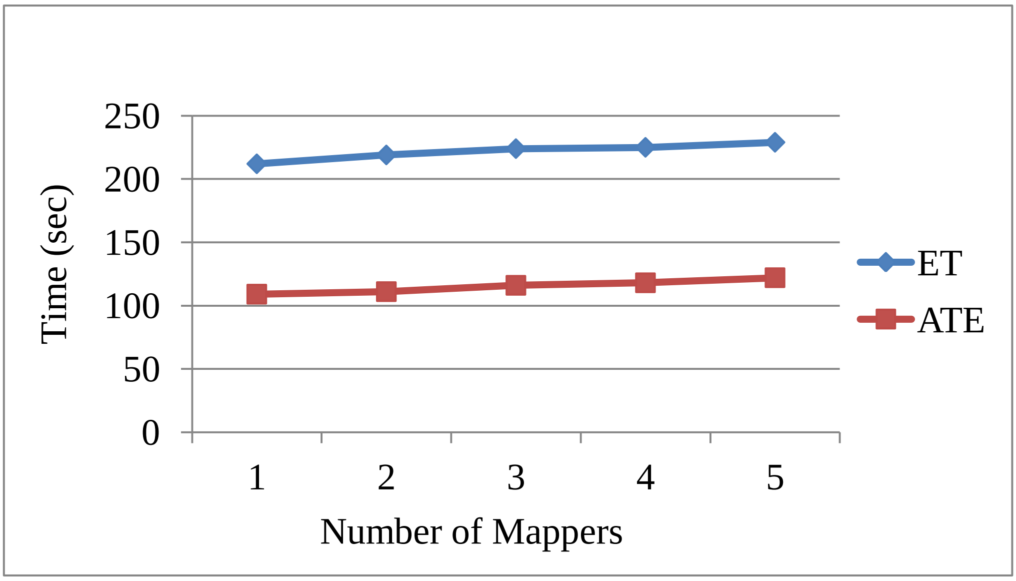


Fig : 7 Execution time with varying number of Mappers for dataset-2

Table 7. Elapse Time (sec.) of mapreduce phases on dataset -2 (min_sup = 0.1,0.15,0.2,0.25,0.3) and (number of maps=2)

No of Mappers	Map=2	Map=2	Map=2	Map=2	Map=2
Support Count (%)	0.1	0.15	0.2	0.25	0.3
ET method (Time in sec)	222	222	219	216	216
ATEproposed method (Time in sec)	114	114	111	100	100

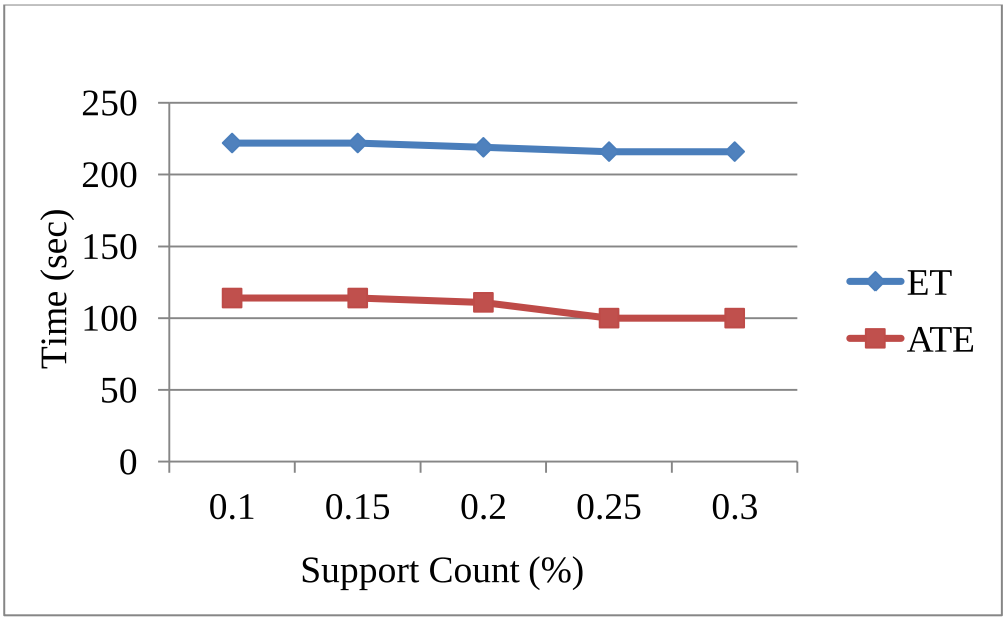


Fig : 8 Execution time with varying support count for dataset-2

Table 8. Elapse Time (sec) comparison between two methods on dataset -1 and dataset-2 (support count=0.3, number of maps=2)

Dataset	in dataset-1	dataset-2
No of Mappers	**Map=2**	**Map=2**
Support Count (%)	0.3	0.3

Dataset	in dataset-1	dataset-2
ET method (Time in sec)	74	216
AETproposed method (Time in sec)	71	100

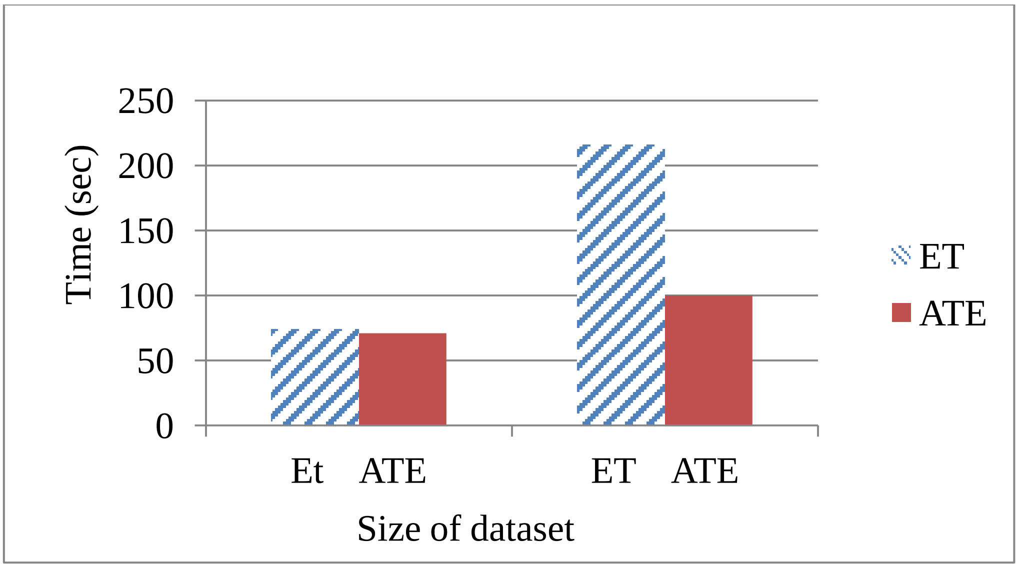


Fig : 9 Execution time with varying support count (0.3) and number of maps (2)

Experiments are performed on dataset-1 and dataset-2 dataset and execution time required for generating number of frequent itemset is compared based on number of mappers and Minimum Support.

Results shown that ATE is faster than ET on both dataset-1 and dataset-2. ET algorithm is not working on large datasets. ET is not scalable enough and faces memory problems as the dataset size increases.

Experiments performed on dataset-1 in order to compare execution time with different Minimum Support and number of mappers on ET and ATE.

Figure 8 and Figure 9 shows timing and size of dataset execution comparison for various methods on dataset-1 which shows that ATE has faster performance over ET algorithm.

Execution time decreases as Minimum Support value increases which shows effect of Minimum Support on execution time. Execution time increases as number of mappers increases as communication cost between mappers and reducers increases.

Results have been shown that ATE algorithm works on

Large Data. Experiments are performed on dataset-2. ET algorithm faced memory problem with dataset-2.

Results of ATE are compared with ET algorithm which is scalable. Table IV. and Table V. shows execution time taken for ET and ATE algorithm on dataset-2 dataset with variable Minimum Support and No. of Mappers.

Table VI. shows execution time taken for ET and ATE algorithm on dataset-1 and dataset-2 datasets with Number of mappers is 2 and Minimum Support is 0.3 for the experiments. Figure 22. shows that ATE algorithm has better performance over ET algorithm. ATE algorithm is scalable then ET algorithm.

6. CONCLUSION

In the future frequent item sets finding will be used real educational survey data on MapReduce framework. Apriori Prefix Tree Eclat finds frequent itemsets having large itemset size. This algorithm can be resolved memory problem. Hadoop MapReduce platform can be used extensively for mining large amount of educational dataset.

7. REFERENCES

[1] B. Singh, R. Singh,N. Kushwaha, O. P. Vyas, "An Efficient Approach for Discovering Closed Frequent Patterns in High Dimensional Data Sets",Advanced Computing, Networking and Informatics- Volume 1 pp 519-528

[2] LESKOVEC, Jure; RAJARAMAN, Anand; ULLMAN, Jeffrey D. Mining of Massive Datasets, 2nd Ed. Cambridge University Press, 2014. ISBN 978-1107077232.

[3] GORUNESCU, Florin. Data Mining - Concepts, Models and Techniques.

[4] Springer, 2011. Intelligent Systems Reference Library. ISBN 978-3- 642-19720-8. Available from DOI: 10.1007/978-3-642-19721-5.

[5] AGGARWAL, Charu C.; HAN, Jiawei (eds.). Frequent Pattern Mining.Springer, 2014. ISBN 978-3-319-07820-5. Available from DOI: 10 . 1007/978-3-319-07821-2.

[6] ALVAREZ, Sergio A. Chi-squared computation for association rules: preliminary result. In: Technical Report BC-CS-2003-01. 2003. Avail- able also from: http : / / www . cs . bc . edu / ~alvarez / ChiSquare / chi2tr.pdf.

[7] AGRAWAL, Rakesh; SRIKANT, Ramakrishnan. Fast Algorithms for Mining Association Rules in Large Databases. In: VLDB’94, Proceed- ings of 20th International Conference on Very Large Data Bases, Septem- ber 12-15, 1994, Santiago de Chile, Chile. 1994, pp. 487–499. Available also from: http://www.vldb.org/conf/1994/P487.PDF.

[8] AGRAWAL, Rakesh; IMIELINSKI, Tomasz; SWAMI, Arun N. Mining Association Rules between Sets of Items in Large Databases. In: Pro- ceedings of the 1993 ACM SIGMOD International Conference on Man- agement of Data, Washington, D.C., May 26-28, 1993. 1993, pp. 207– 216. Available from DOI: 10.1145/170035.170072.

[9] ZAKI, Mohammed Javeed; GOUDA, Karam. Fast vertical mining us- ing diffsets. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, 2003. 2003, pp. 326–335. Available from DOI: 10.1145/956750.956788.

[10] AGARWAL, Ramesh C.; AGGARWAL, Charu C.; PRASAD, V. V. V.Depth first generation of long patterns. In: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, Boston, MA, USA, August 20-23, 2000. 2000, pp. 108–118. Available from DOI: 10.1145/347090.347114.

[11] HAN, Jiawei; PEI, Jian; YIN, Yiwen. Mining Frequent Patterns with- out Candidate Generation. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dal- las, Texas, USA. 2000, pp. 1–12. Available from DOI: 10.1145/342009. 335372.

[12] ZAKI, Mohammed Javeed; PARTHASARATHY, Srinivasan; OGI- HARA, Mitsunori; LI, Wei. New Algorithms for Fast Discovery of Association Rules. In: Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97), Newport Beach, California, USA, August 14-17, 1997. 1997, pp. 283–286. Available also from: http://www.aaai.org/Library/KDD/1997/kdd97-060.php.

[13] ZAKI, Mohammed Javeed. Scalable Algorithms for Association Min- ing. IEEE Trans. Knowl. Data Eng. 2000, vol. 12, no. 3, pp. 372–390. Available from DOI: 10.1109/69.846291.

[14] LIN, Goh Chun; DESMOND, Koh Eng Tat; HTOON, Naing Tayza; THUAT, NV. A Fresh Graduate’s Guide to Software Development Tools and Technologies. Chapter-6: Scalability, School of Computing, National University of Singapore. 2012.

[15] DEAN, Jeffrey; GHEMAWAT, Sanjay. MapReduce: Simplified Data Processing on Large Clusters. In: 6th Symposium on Operating Sys- tem Design and Implementation (OSDI 2004), San Francisco, California, USA, December 6-8, 2004. 2004, pp. 137–150. Available also from: http://www.usenix.org/events/osdi04/tech/dean.html.

[16] HURWITZ, Judith; NUGENT, Alan; HALPER, Dr. Fern; KAUFMANN,Marcia. Big Data for Dummies. John Wiley & Sons, Inc., 2013. ISBN 978-1-118-50422-2.

[17] SAVASERE, Ashok; OMIECINSKI, Edward; NAVATHE, Shamkant B. An Efficient Algorithm for Mining Association Rules in Large Databases. In: VLDB’95, Proceedings of 21th International Conference on Very Large Data Bases, September 11-15, 1995, Zurich, Switzerland. 1995, pp. 432–444. Available also from: http : / / www . vldb . org / conf/1995/P432.PDF.

[18] A. Tabarcea, V. Hautamäki, P. Fränti,”AD-HOC GEOREFERENCING OF WEB-PAGES USING STREET-NAME PREFIX TREES”, 6th International Conference on Web Information Systems and Technologies,April-2010, DOI: 10.1007/978-3-642-22810-0_19 ·