Effective Utilization of Processor in a Distributed Environment For PC Clusters Analysis

N.Kavitha [1] *S.Karthikeyan* [2]

Abstract - Distributed Data mining is expected to perform partial analysis of data at clients and then to send the outcome, as results to the server where it is sometimes required to be aggregated to the global result .The primary issues to be considered for Distributed Data mining are Scalability, privacy of data and autonomy of data. These issues can be easily handled when we go for intelligent software agents for Distributed Data mining, because of its inherent features of being autonomous, capable of adaptive and deliberative reasoning. So by using software intelligent agents the average idle.

Key words: Distributed Data Mining, Software Intelligent Agent, and Scalability

1. Introduction

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data for the, extraction of interesting knowledge that could help in decision-making. Distributed data mining refers to the mining refers to the mining of distributed data sets .The data sets are stored in local data bases, hosted by local computers, which are connected through a computer network. Data mining takes place at a local level and at a global level where local data mining results are combined to gain global findings. In some applications, data are inherently distributed, but it is necessary to gain global insights from the distributed data sets. Distributed Data mining has emerged as a means for identifying patterns and trends from large quantities of data in a quick manner. The implementation of data mining in distributed computing has become crucial for ensuring system scalability and interactivity as data continues to grow inexorably in size and complexity time per node is kept low. To gain good understanding of the utilisation of the CPU time of identified workstations time to time, the software agent for the distributed environment was implemented. Also, this intelligent agent is capable of giving alert information to the server, when the CPU usage time of any identified workstations exceeds threshold value. Usage of intelligent agent approach to distributed data mining was applied and the expected performance was obtained successfully.

2. Methodology

PROPOSED SYSTEM MODEL

The Proposed System was aimed to develop the effective utilisation of the CPU time in workstations time to time; the software agent for the distributed environment was implemented. Also, this intelligent agent is capable of giving alert information to the server, when the CPU usage time of any identified workstations exceeds threshold value. The Apriori algorithm was implemented to mine the transaction data sets.

[1]Research Scholar, Department of CS, Karpagam University, Coimbatore. Email: nkavee1@yahoo.co.in

[2]Assistant Professor, Department of Information Technology, College of Applied Sciences, Sohar, Sulatanate of Oman.Email: skaarthi@gmail.com

Intelligent agent (IA): is an autonomous entity which observes and acts upon an environment (i.e. it is an agent) and directs its activity towards achieving goals (i.e. it is rational).Intelligent agents may also learn or use knowledge to achieve their goals. They may be very simple or very complex: a reflex machine such as a thermostat is an intelligent agent, as is a human being, as is a community of human beings working together towards a goal.

Primarily, the server interacts with the agent to know about the status of the clients in terms of idle time, periodically. Now, agent contacts client to know about its idle time. This in turn makes the client to response with its current idle time. This will be given as response to the server. Based on this information, now the server will start splitting the large data sets among various clients. After performing association rule mining on this given data set, clients send the processed results to the agent. Now, the agent collects the processed result and submits to the server to display in the required format.

3. Implementation Of Apriori Algorithm In A Single Machine And Analyze The Results

In this step, the implementation of Apriori algorithm for association rule mining in a single machine was considered and the results were analyzed. The apriori algorithm was implemented using JDK 1.4 and executed successfully to generate association rules for the given identified inputs in a single machine.

Different numbers of records were given as input and the results were analyzed and the same was presented in the following Table 1. The processing time required for mining 1, 00,000 records with single system is 1.1 secs.

Table : 1

No Of Records	Duration(In Secs)
	Single System
10	0.03
20	0.06
30	0.08
40	0.11
60	0.19
75	0.27
85	0.58
100	1.1

4. Implementation Of Apriori Algorithm In The Number Of Workstations Using Intelligent Agents And Analyze The Results

Apriori Algorithm in N (N=6) identified workstations were considered. This involves the calculation of CPU idle time in the workstations. Distribution of dataset among these workstations based on the calculated CPU idle time using intelligent agents were considered and the results were analyzed. The distribution of data sets among the identified workstations and the agent program that monitor the CPU idle has been implemented using JDK 1.4.Proposed approach consists of two logical components.

Client- There may be in N-number of clients connected in intranet. In contrast to the model proposed by, here the client receives data sets and mining operation to be performed as parameters from the server agent. Local agent running in the client machine takes care performing specified operation and storing the final results. The client side agent sends resource utilization information periodically to the server machine. Additionally, it is also having the Responsibility of alerting the server, if it is overloaded or can't execute the task because of various reasons such as taking appropriate actions in the unwanted situation like overloaded message from the client agent, network failure etc.,

Server - In this proposed model, the server machine stores the large database or Data warehouse where the millions of records are stored and intelligent agent running in this machine. The System flowchart for the proposed work is illustrated in the Fig1.and the frame work for the proposed system is illustrated in Fig 2.

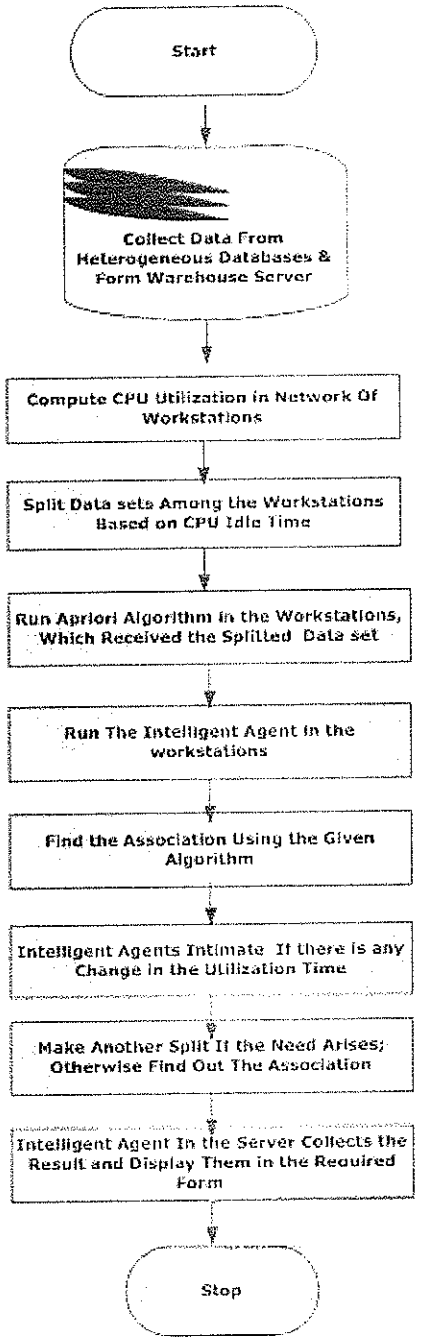

Figure 1: System Flow Diagram for the Proposed Work.

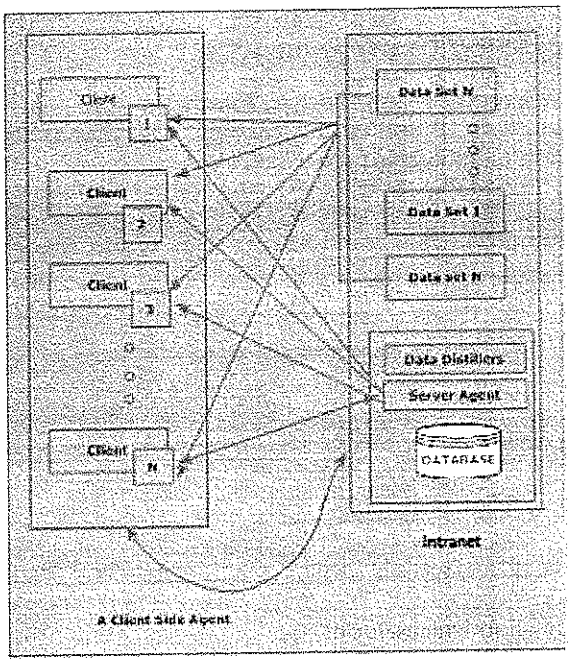

Figure 2 : Framework for the proposed work

5. Apriori Algorithm – Implementation

The implementation of Apriori algorithm basically consists of the following stepsJoin Step: Ck is generated by joining Lk-1 with itself.Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset.

PSEUDO CODE

Ck: Candidate itemset of size k

Lk : frequent itemset of size k

L1 = {frequent items};

for (*k* = 1; *Lk* !=∅ *k*++) **do begin**

Ck+1 = candidates generated from *Lk*;

for each transaction *t* in database do

increment the count of all candidates in *Ck+1* that are contained in *t* *Lk+1* = candidates in *Ck+1* with min_support

end

return ∪*k Lk*;

GENERATION OF CANDIDATE SET

Suppose the items in Lk-1 are listed in an order

Step 1: self-joining Lk-1

insert into Ck

select p.item1, p.item2, …, p.itemk-1, q.itemk-1 from Lk-1 p, Lk-1 q

where p.item1=q.item1, …, p.itemk-2=q.itemk-2, p.itemk-1 < q.itemk-1

Step 2: pruning

for all item sets c in Ck do

for all (k-1)-subsets s of c do

if (s is not in Lk-1) then delete c from Ck

6. Results And Discussion

This research work uses database from the Departmental Store of Nilgris. It consists of 1, 00,664 records with 15 different fields. The primary information in this item code module includes, Date of item purchased, item quantity, item weight, item cost. With this information, we are aiming at the following association and predictions.

- Number of items for each transaction.
- Regularity of Items sold
- Selection of items by the customer and the transaction for the particular items in terms of the days and months and year.

The original database consists of additional fields and information. It is essential to remove these fields and make transformations to improve the accurate results in the mining process and to improve the efficiency of the whole system.

The unnecessary information like Leave Type, Created User, Crated Date, Modified User, Modified Date and Batch were deleted from original database which are not related to the generation of association rules of the mining process.

All the entries marked with item code for each transaction, of the transformed database.

All the entries in AT_DAY, which represents the day in which the transaction was marked, were assigned a unique number in the transformed database to speed up the mining process.

AT_DATE, which represents the date in which the transaction was entered of the original database, is transformed to two fields AT_MONTHS, AT_DAYS contains Month Information and the Number of Days (calculated from the Difference between first record and existing date entry).

The Table 1 shows the original database considered and the Table 2 reflects the transformations performed on the original table.

Table 2 – Structure of Original Database:

I_NAME	Item Name For Each Transaction
I_CODE	Purchased Items Code For Each Transaction
T_ID	Unique Value For Each Transaction: Transaction Id
TD_DATE	Corresponding Day Of The Item Purchased
I_QTY	Purchased Item Quantity
I_COST	Purchased: Item Cost

Table 3 – Structure of Transformed Database

I_CODE	Purchased Items Code For Each Transaction
I_QTY	Purchased Item Quantity
I_COST	Purchased: Item Cost

Data Transformation and Cleaning

In this part of work involves forming the data warehouse and perform any transformations and cleaning needed in our collected data. The transformed and cleaned data can be useful to be mined using Apriori algorithm. Here, about 1, 00,664 records were transformed and cleaned.

3 TEST CASES

Number of records = 1, 00,000

Number of nodes = 4

Expected Performance

The expected results are tabulated in the following Table.

Table : 4

No Of Records	Duration(In Secs)	
	Single System	N=4
10	0.03	0.03
20	0.06	0.04
30	0.08	0.06
40	0.11	0.085
60	0.19	0.14
75	0.27	0.21
85	0.58	0.38
100	1.1	0.58

Actual Performance

The actual results are tabulated in the following table

Table 5

No Of Records	Duration(In Secs)	
	Single System	N=4
10	0.03	0.03
20	0.06	0.04
30	0.08	0.06
40	0.11	0.085
60	0.19	0.14
75	0.27	0.21
85	0.58	0.38
100	1.1	0.58

Deviation

There is no deviation between the expected result and the actual result. So the status of the test case is said to be success.

7. PERFORMANCE ANALYSIS

The Figure 3 shows the performance comparison of data mining in the single system versus distributed system with 2 workstations. The processing time required for mining 1,00,000 records with single system is 1.1 secs, where as, with the distributed system, took 0.64 secs for the same number of records. This clearly indicates the effectiveness of distributed data mining in terms of maximum utilization of resources.

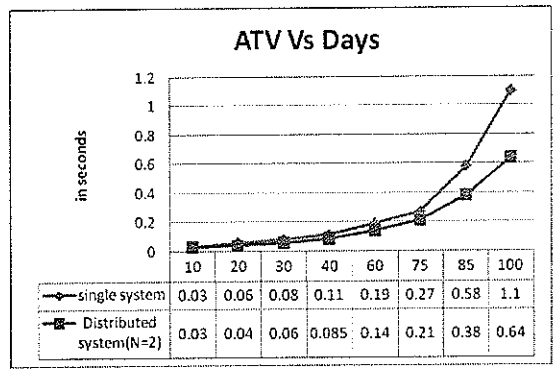

FIGURE: 3 Performance analysis – Single Vs Distributed System. (N=2)

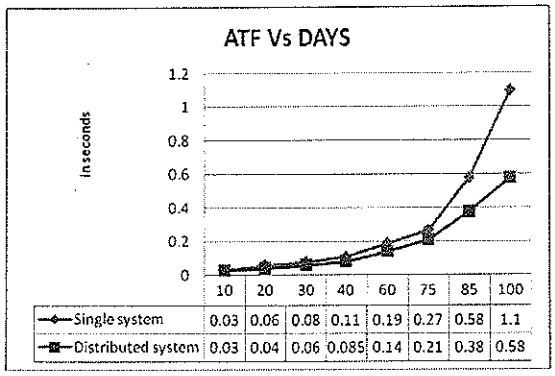

FIGURE: 4 Performance analysis – Single Vs Distributed System (N=4)

The Figure: 4 show the performance comparison of data mining in the single system versus distributed system with 4 workstations. The processing time required for mining 1, 00,000 records with single system is 1.1 secs, where as, with the distributed system, took 0.58 secs for the same number of records. This clearly indicates the effectiveness of distributed data mining in terms of maximum utilization of resources

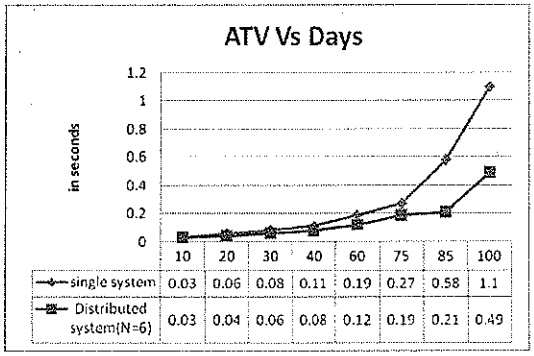

FIGURE: 5 Performance analysis – Single Vs Distributed System (N=6)

The Figure.5 shows the performance comparison of data mining in the single system versus distributed system with 6 workstations. The processing time required for mining 1, 00,000 records with single system is 1.1 secs, where as, with the distributed system, took 0.49 secs for the same number of records. This clearly indicates the effectiveness of distributed data mining in terms of maximum utilization of resources.

By seeing the performance analysis graph, conclusion made that distributed data mining saves the time as well as reduces the average idle time of the CPU in PCclusters.

8. Conclusion

The data mining is an essential activity that is required for any industry that is interested in forecasting the business trends and analysing the behavioural study of customers. The effective and maximal utilisation of the existing resources is the need of the hour for these activities. This paper achieved this objective of effective utilisation of computing resources in efficient manner. Association rules are generated to study and forecast trends in the academic environment by the implementation of Apriori algorithm. Finally the effective utilisation of the CPU is identified.

References

[1] Agent Working Group. *Agent Technology Green Paper*. OMG Document ec/99- 12-02.Version 0.9. 24 December 1999.

[2] R. Agrawal, T. Imielinski, A. Swamy, "Mining association rules between sets of items in large databases". *ACM SIGMOD Conf.* 1993

[3] Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.,.and Verkamo, I. 1996. Fast Discovery of Association Rules. In *Advances in Knowledge Discovery and Data Mining, 3rd* ed., 307–328, AAAI Press

[4] Agrawal, R., and Psaila, G. 1995. Active Data Mining. In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), 3–8. AAAI Press

[5] Apte, C., and Hong, S. J. 1996. Predicting Equity Returns from Securities Data with Minimal Rule Generation. In *Advances in Knowledge Discovery and Data Mining,2nd* ed. , 514–560. AAAI Press

[6] Chan P, Fan W, Prodromidis A, and Stolfo s, "Distributed data mining in credit card fraud detection,", *IEEE Intelligent Systems*, 14(6):67-74, 1999.

[7] Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya, Xiaodong Lin, Michael Y. Zhu ,Purdue University.",Tools for privacy preserving Data mining." In proceedings of the *ACM SIGKDD Explorations*, v.4 n.2, p.28-34, December 2002

[8] Kargupta H,IILkar Hamzoaglu, Brian Stafford., Scalable, "Distributed data mining using an agent based architecture." In the proceedings of *KDD'97.*

[9] Kargupta, H, Park, Hershberger, Johnson ,E., "Collective Data mining: A New perspective toward Distributed Data mining" . *In: Advances in Distributed and Parallel Knowledge Discovery,* AAAI/MIT Press (2000) 131-178

[10] Langley, P., and Simon, H. A. 1995. Applications of Machine Learning and Rule Induction. Communications of the ACM 38:55–64.

[11] Matthias Klusch, Stefano Lodi and Gianulco Moro.,"Agent based Distributed Data mining: The KDEC Scheme." (2003), http://citeseer.ist.psu.edu 2004.

[12] Parthasarathy S, "Towards Network-Aware Data Mining", 15th International Parallel and Distributed Processing Symposium (IPDPS'01) Workshops. *IEEE Computer Society,* p. 30157b

[13] Salvatore Stolfo, Andreas L. Prodromidis, Shelley Tselepis, Wenke Lee, Dave W. Fan, et al. "JAM: Java Agents for Meta-Learning over Distributed Databases." In *International Conference on Knowledge Discovery and Data Mining* -1997

[14] http://portal.acm.org/, 2004.

[15] www.cs.bme.hu/~bodon/en/apriori, 2004.

[16] www.csc.liv.ac.uk/~frans/KDD ,2004

[17] www.agentland.com , 2005

Author's Biography

N. Kavitha Received UG Degree from Bharathiar University and PG Degree from mother Teresa Women's University. I did my Phil Degree at Bharathidasan University. Currently Pursuing Ph.D in Karpagam University. Having 8yrs and 6 months in teaching Experience. My Research Area is Data Mining. I have presented 9 papers in various conferences. Other Areas include Computer Networks.

Karthikeyan S. received the Ph.D. Degree in Computer Science and Engineering from Alagappa University, Karaikudi in 2008. He was working as a Professor and Director in School of Computer Science and Applications, Karpagam University, Coimbatore. At present he is in deputation and working as Assistant Professor in Information Technology, College of Applied Sciences, Sohar, Sulatanate of Oman. He has published more than 14 papers in National/International Journals. His research interests include Cryptography and Network Security