Improving Computer Network Performance by Using Knowledge Discovery

Mohammed Alhamdan[1], Aziza Saad Asem [2] and Mohammed Hashem [3]

**[1] Computer Science, Mansoura University,
Mansura, ZIP 35516, Egypt**

**[2] Computer Science, Mansoura University,
Mansura, ZIP 35516, Egypt**

**[3] Computer Science, Ain Shams University,
Cairo,ZIP 11566, Egypt**

Abstract

During the past decade, the computer world has been changing a dramatic change in many aspects. In the past, personal computer was known as a desktop machine that functioned alone. All the resources, data, and computing power were within the same machine. However, in the network world today, more and more computers are connected together through the media to share data and resources or parallel the computation. The necessity to evaluate the performance of these computer networks has been recognized for a long time. Performance measurement has been significantly important to computer networks designers, administrators, and analysts to justify the impact of a new design or change compared to existing system. On the other hand, this information is also important to the users or customers to evaluate different systems from different vendors to determine whether their needs can be fulfilled, and to measure the performance of the running network Overall network performance is integration between two categories network-related like packet size, throughput, delay time and workstation-related category like CPU utilization, memory utilization, resource response time, resource utilization. By recording all the events triggered in the network workstations using network monitoring tools we have a very large number of records that describe in details the network characteristics. By analyzing and mining this large amount of data we will get valuable information and extract a hidden predictive knowledge that will guide network administrators how to adjust different network parameters and adjust hardware and software resources attributes.

Keywords: *Computer Network, Network Performance, Data Mining, Data Analysis, Log File Preprocessing.*

1. Introduction

1.1. The Initiative of Network Performance Evaluation

A computer network, or even the Internet, is composed of dozens to thousands or even millions of computers connected together. The goal is to allow information and valuable resources to be shared among computers located at different sites. As the usage of the Internet exploded in mid 90s, more and more communication software applications and distributed applications (which take advantage of parallel processing to achieve high throughput and performance) will be developed and deployed over the Internet or corporate networks. Therefore, the importance of network performance measurements was proved by the rapid deployment of these network software applications in this decade [3].

As Liu mentioned [2], “Network system performance is the performance of a computer system where networking plays a significant (if not the dominant) role, hence it is a continuum of computer system performance.” Computer system performance studies have already been around for a long time. In traditional computer performance evaluation, the subject is a single computer comprised of hardware and software components. Most studies concerns are concentrated on the comparison of performance between two or more computer system and optimization of performance of a single system. On the other hand, the network system is a product of the combination of the computer technology with the networking technology, whereby independent computers are interconnected through a network. With this approach, independent computers are able to share resources including CPU cycles, data, applications, and services among others [5].

For a network system, also known as distributed system, the scope of performance evaluation is considerably extended and complicated by the existence of the network component. Not only is the performance of individual systems, but the interaction among the computers as well as the network which interconnects them also plays a significant role.

1.2. 2. The Elements of Network Performance

Much work has been devoted to the attempt to define network performance exactly. It is not the intention of this paper to bore you with numerous equations that describe theoretical network philosophy about how packets traverse networks. Network performance is a complex issue, with lots of independent variables that affect how clients access resources across a network. However, most of the elements involved in the performance of networks can be boiled down to a few simple network principles that can be measured, monitored, and controlled by the network administrator with simple—often free—software.

Most network performance tools use a combination of separate elements to measure network performance like availability, response time, memory utilization, CPU utilization, network delay, network throughput, network bandwidth capacity, packet size, resource utilization [6].

In this paper we introduce a new methodology to improve the network overall performance using large data sets techniques by recording and gathering all the events that took place on each workstation on the network on a log files -using network monitoring tools that is resident on the workstation- and the event delay time, packet size, CPU utilization and memory utilization and then find correlations among these attributes, Association rules, classify the data according to specific criteria.
In the past these log files were headache to the system administrators because it increase in size rapidly and should be handled periodically, sometimes they disable this property or just backup it as a managerial process or delete it periodically to free the storage media or at the most they benefit from it as individual cases not as bulk of data to analyze or mine it to get a new knowledge that will help in predict the future or improve the performance of the current time. Therefore we tried to benefit from this large amount of data stored in the network by enhancement it and prepare it to be able to analyzed and even mined.

2. RELATED WORKS

Researches in network performance domain have toke place long time ago. For example, performance of wireless network domain discussed by Zhang [8]. Achuthsankar Nair, and J. Jayasudha augmented improving performance of world wide web by adaptive web traffic reduction [9]. T. Rokicki et al discussed the problem of improving network performance by reducing network contention in source-based cows with a low path-computation overhead [10]. Sivanantharasa P. discussed the Performance Evaluation of information and communications technology infrastructure for smart distribution network applications [11]. Chuck C., et al, introduced high performance network monitoring with a SQL interface [12]. If a computer network system consists of a server and two subsystems connected in parallel, in this way the whole system will be down only when both the subsystems are down or server fails. By supplementary variable technique it is concluded that Steady state transition probability is around 99.55%, system will degrade slowly. This study also describes that how MTTF (Mean Time to Fail) will affect with respect to failure [13]. S, Deepankar el discussed network performance from network reliability point of view [14]. Kaur, R., discussed and evaluated “Optimized Link State Routing Protocol” OLSR routing measurement performance analysis based on different simulation parameters. [15]. Partha S., et al Discussed Reliability as one of the most important factors for assessing the performance of the network [16]. Braner et al augmented that, packets should pass through the most reliable path. This paper presents a technique for selecting the most reliable path for communication between node pairs of a computer network. Reliability has been calculated based on the bandwidth utilization by the nodes. Bandwidth utilization has been calculated using throughput, latency-time. In the main time Stefan Worm discussed monitoring of large-scale cluster [17].

3. EXPERIMENTAL NETWORK DESCRIPTION

In our experiment applied on organization contain about 112 workstations organized as a domain under operating system Microsoft Windows server and Microsoft windows workstation on clients, the network was in class C, according to the hardware resources there are 92 printers distributed in different departments. *connected weather directly to workstations or to the network through IP address, the printers are different types ,models and capacities.*

Regarding to scanners there are number of scanners connected to the workstations. Some of workstations have fax modem cards connected to it enabling the workstation to send and receive faxes through network fax utilities installed in workstations, the switches in each departments are from the same type and speed, there are a firewall installed behind the gateway.

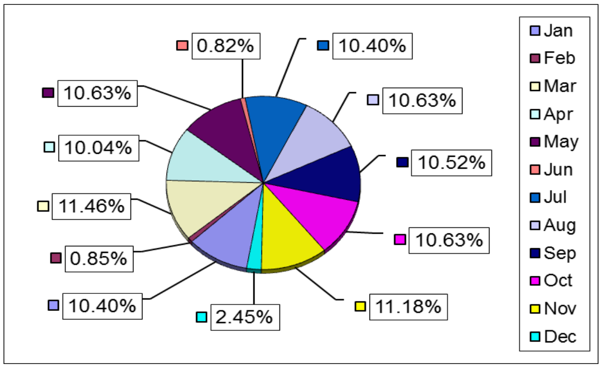

Fig1.Network events per month

According to software there are large number of software programs installed on workstations like operating system, tools and applications.

The University is divided to 11 departments, in many cases a workstation prints out of the department that it's physically located to achieve managerial purposes. The time and the date of the network daily adjusted through Microsoft home page time setting.

Data gathered historically from the beginning of year 2006 till October 2012, the data is approximately 4,232,000 records.

TABLE.1 Network events per month

Month	No. of Events	Percentage
Jan	440000	10.40%
Feb	36000	0.85%
Mar	485000	11.46%
Apr	425000	10.04%
May	450000	10.63%
Jun	34500	0.82%
Jul	440000	10.40%
Aug	450000	10.63%
Sep	445000	10.52%
Oct	450000	10.63%
Nov	473000	11.18%
Dec	103500	2.45%

4. Data Modeling Approaches

After data is preprocessed and ready to be handled it's useful to get some data visualization tools to understand the data feature.

For each event recorded in the log file there are availability status that describes weather the resource called was available or not in the requested time. Figure 1 show that the availability with status "yes" versus time series. The graph clarify that the maximum availability found in the network at 10 O'clock at the morning (after 1 hour of working) and less availability found at 4 O'clock at the end of the day this is due to at 10 O'clock all the resources are adjusted on line but at 16 O'clock the some resources turned off not in a correct order that means a printer may be switched off while a print job is sent to it from another computer.

For each event triggered there are two values generated by the monitoring tool "the packet size" and "throughput" of the event as figure 2 shows the department name and the average of network throughput of each department.

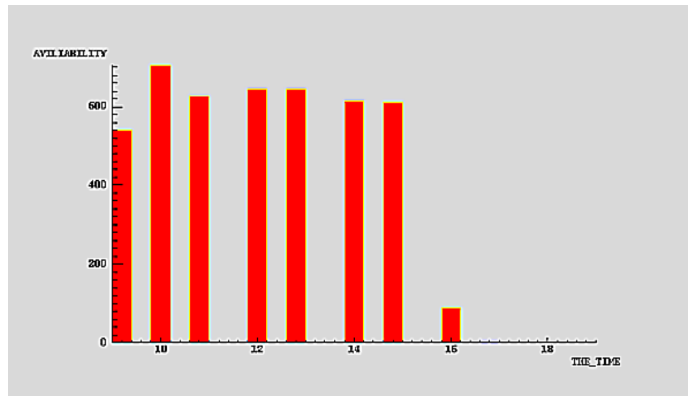

Fig2.Availability Vs. The time Series

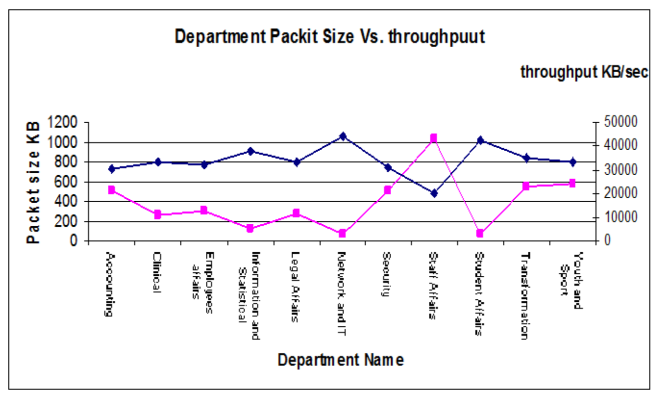

Fig3.Throughput vs. Packet size in each department

By going in depth in data of each department as shown from table no.1 the reason of the throughput is packet size that in department Network & IT the packet size is 64 kb and the throughput was 44000 KB per Sec, on the other hand in department Staff Affairs the average of packet size was 1024 and the packet size was 20200 MB per Sec with the same type of switch.

The table 2 shows that the more the packet size increase the more throughput decrease and vice versa witch direct the network administrator to force the OS to decrease the packet size as much possible.

TABLE.2 The relation between throughputs Vs. Packet sizes

Department Name	Average Throughput	Average Packet Size
Accounting	30620	512
Clinical	33524	265
Employees affairs	32052	300
Information and Statistical	38000	128
Legal Affairs	33240	280
Network and IT	44000	64
Security	29550	512
Staff Affairs	20200	1028
Student Affairs	42521	75
Transformation	35220	550
Youth and Sport	33524	580
Totals	33859.18	390.3636
Average	33859.1	390.4
Minimum	20200	64
Maximum	44000	1028

1. CARMA Algorithm

The continuous association rule mining algorithm (Carma) is an alternative to Apriori that reduces I/O costs, time, and space requirements, Hidber .C [18]. It uses only two data passes and delivers results for much lower support levels than Apriori. In addition, it allows changes in the support level during execution.

Carma deals with items and itemsets that make up transactions. Items are flag-type conditions that indicate the presence or absence of a particular thing in a specific transaction. An itemset is a group of items which may or may not tend to co-occur within transactions. Carma proceeds in two stages: It identifies frequent itemsets in the data, and Then it generates rules from the lattice of frequent itemsets, more in Hidber .C [18].

From analyzing the data gathered from the log files we can detect the relations between computers & printer's as a Web describing the strong connections and week connections Fig no. 3 and 4, Table 3.

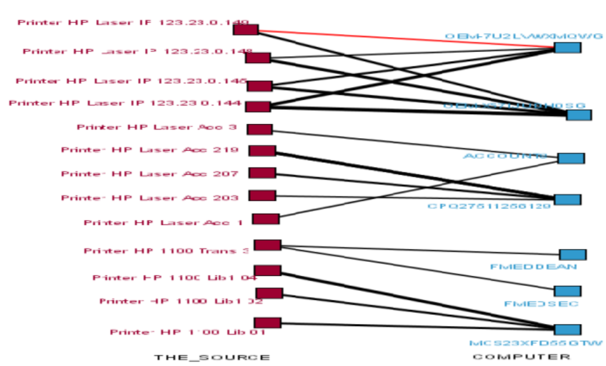

Fig4. Connection among computers & Printers

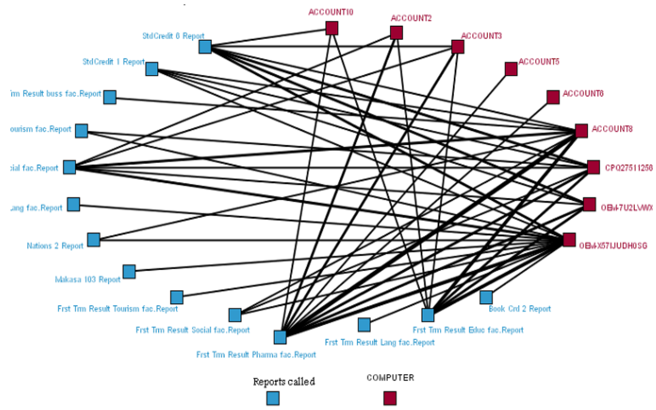

Fig5. Connections between reports called and computers

TABLE.3 Strong connections among computers & printers

Strong Links

Links	Field 1	Field 2
1,985	COMPUTER = "OEM-X57IJUDH0SG"	THE_SOURCE = "Printer HP Laser IP 123.23.0.144"
1,355	COMPUTER = "MOS23XFD55GTW"	THE_SOURCE = "Printer HP 1100 Lib1 04"
1,282	COMPUTER = "OEM-X57IJUDH0SG"	THE_SOURCE = "Printer HP Laser IP 123.23.0.145"
1,215	COMPUTER = "CPQ27511256129"	THE_SOURCE = "Printer HP Laser Acc 219"
1,039	COMPUTER = "OEM-X57IJUDH0SG"	THE_SOURCE = "Printer HP Laser IP 123.23.0.148"
1,015	COMPUTER = "OEM-7U2LVWXMQ...	THE_SOURCE = "Printer HP Laser IP 123.23.0.144"
779	COMPUTER = "MOS23XFD55GTW"	THE_SOURCE = "Printer HP 1100 Lib1 02"
752	COMPUTER = "CPQ27511256129"	THE_SOURCE = "Printer HP Laser Acc 207"
724	COMPUTER = "MOS23XFD55GTW"	THE_SOURCE = "Printer HP 1100 Lib 01"
621	COMPUTER = "OEM-7U2LVWXMQ...	THE_SOURCE = "Printer HP Laser IP 123.23.0.145"
587	COMPUTER = "OEM-X57IJUDH0SG"	THE_SOURCE = "Printer HP Laser IP 123.23.0.149"
453	COMPUTER = "OEM-7U2LVWXMQ...	THE_SOURCE = "Printer HP Laser IP 123.23.0.148"
411	COMPUTER = "CPQ27511256129"	THE_SOURCE = "Printer HP Laser Acc 203"
383	COMPUTER = "ACCOUNT8"	THE_SOURCE = "Printer HP Laser Acc 3"
381	COMPUTER = "ACCOUNT8"	THE_SOURCE = "Printer HP Laser Acc 1"
276	COMPUTER = "FMEDSEC"	THE_SOURCE = "Printer HP 1100 Trans 3"
270	COMPUTER = "FMEDDEAN"	THE_SOURCE = "Printer HP 1100 Trans 3"
246	COMPUTER = "OEM-7U2LVWXMQ...	THE_SOURCE = "Printer HP Laser IP 123.23.0.149"

Medium Links
Weak Links

As cleared from the Web graph there are strong connections, medium connections and week connections. According to the strong connections between the computer and the printer means that this computer has large no. of accesses to this printer, it may be useful to reduce the distance between these computers and printers –put this in the same department- or within the same switch to reduce the traffic on the network lines or try to convert some print jobs to batch printing instead of online printing.

We may perform this task with all the resources to re arrange the locations of the network again. By blotting the average of CPU utilization over months we found that the maximum CPU utilization found in December, January, June and July, which is : Actually the largest no. of events recorded in these months because all the department of the university is working in a full power to accept students for graduations.

Using analytical methods:

The most printers used were "Printer HP Laser 4V Acc 7".

The fewer printers give Error result was "Printer HP Laser

SEC". The maximum software used "Oracle 8i". The most web page Browsed first "WWW.MSN.COM".
No. of events triggered distribution during the day's working hours in Fig 5

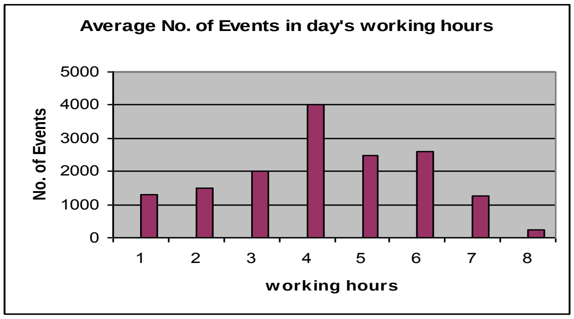

Fig6. Average no. of events distributed in day's working hours

As seen from figure 5 the distribution of events is highly distributed from the third hour of work (11 AM) to the sixth hour of work (2 PM), that gives us guide line for resource load redistribution like loading of print job tasks:

- The maximum computers triggered no. of events was "CPQ27511256129".
- By tracing the main oracle program we found that the most report called is "Frst Trm Result Educ fac.Report" and from most departments so we may guide the database administrator to distribute this report locally on workstations that call it with the fixed data of the report and just request the changes of the data each time to reduce the traffic on network lines.
- Plotting CPU utilization versus memory utilization for workstations and network throughput versus day's working hours will guide us to manage the network efficiently, for example:-
  - Allow internet browsing only in peak times and allow downloading in rest of times.
  - Prevent program copying in peak times and allow it at the rest of time.
  - Convert some of print jobs to batch printing instead of online printing if the nature of the print job allows this action.
  - Upgrade the bottle necks hardware components like switches, routers bridges to increase the overall performance.

By Using such proposed algorithm to extract association rules which is finding interesting relationships among attributes in a given data set [7] table no. 4 illustrates –as a sample-the associations between workstations and software. We found the following results:-

- For the Computer named "ACCOUNT10" has 59388 instances with support 3.61% of all the data and confidence with the "MS Excel " S.W. 65% which means that 65 % from the occurrences of this computer has request to MS Excel .

From this knowledge we can discover witch Workstations call which S.W. and when if we add the Date or time to the rule set. *We can extract other interesting knowledge from these data if we change the attributes we mine*.

TABLE.4 correlation between Computer name & Software

Consequent	Antecedent	Instances	Support	Confidence
Computer name	SW			
ACCOUNT10	MS Excel XP	59388	3.61%	65.00%
ACCOUNT11	Internet Explorer 6	59192	3.60%	63.00%
ACCOUNT12	MS Access	57428	3.49%	62.00%
MOS23XFD55GTW	MS Assembler	59192	3.60%	61.00%
OEM-7U2LVWXMQWG	MS Excel XP	59388	3.61%	60.00%
OEM-X57IJUDH0SG	Ms Power Point XP	33320	2.03%	62.00%
PC1	MS Visual Basic 6	33908	2.06%	63.00%
PC10	MS Visual C++	32732	1.99%	67.00%
PC11	MS Visual Fox Pro 3	54096	3.29%	65.00%
PC12	MS Word XP	56252	3.42%	64.00%
PC13	Oracle 8i	313600	19.00%	85.00%
PC14	Paint Pro	94080	5.72%	71.00%
SAFE-KDE8K6SCE8	PcTools 4 XP	101920	6.20%	8.00%
SAFE-KDE8K6SCE9	Photo Shop 7	9212	0.56%	6.00%
SYSTEMMANAGER	Power Point XP	98392	5.98%	6%
WEB10	MS Excel XP	255584	15.54%	75.00%
WEB11	Sound Recorder	49000	2.98%	15.00%
WEB12	Win AMP 3	6468	0.39%	60.00%
WEB13	Win RAR 2.6	44100	2.68%	5%
WEB14	Win Zip 8	167776	10.20%	65%

Network Throughput Classification

Classification is one form of data analysis which can be used to extract models describing important data classes or to predict future data trends. More specifically, classification process consists of two steps. Firstly, model construction; describing a set of predetermined classes. Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute. The set of tuples used for model construction called training set. The model is represented as classification rules, decision trees, or mathematical formulae. Secondly, model usage; for classifying future or unknown objects. Estimation of the accuracy of the model can be performed by. Firstly, the known label of test sample is compared with the classified result from the model. Secondly, Accuracy rate is the percentage of test set samples that are correctly classified by the model.

2. a. C4.5 Algorithm

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan[19]. C4.5 is an extension of Quinlan's earlier ID3 (Iterative Dichotomiser 3) algorithm-ID3 operates only on numeric data . The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set S=s1,s2,… of already classified samples. Each sample Si=x1,x2,… is a vector where x1,x2,… represent attributes or features of the sample. The training data is augmented with a vector C= c1,c2,… where c1,c2,… represent the class to which each sample belongs.

At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The general algorithm for building decision trees is [20]:

- Check for base cases
- For each attribute a
- Find the normalized information gain from splitting on a
- Let a best be the attribute with the highest normalized information gain
- Create a decision node that splits on a_best
- Recurse on the sublists obtained by splitting on a_best, and add those nodes as children of node

By implementing such classification algorithms we can classify the data and predict the new entry characteristics [7], in the following model we got random sample of the data about 28980 records, we classify the average throughput of the network according to packet size, the result shown in Fig 6

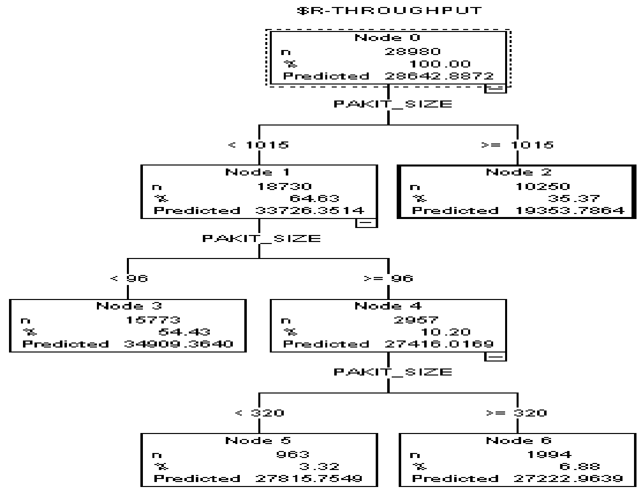

Fig7.Classification of sample data according to packet size to predict throughput

2.b. C4.5 Improvement

C4.5 made a number of improvements to ID3. Some of these are [20]:

- Handling both continuous and discrete attributes: In order to handle continuous attributes, C4.5 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it.
- Handling training data with missing attribute values: C4.5 allows attribute values to be marked as ? for missing. Missing attribute values are simply not used in gain and entropy calculations.
- Pruning trees after creation: C4.5 goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes.

A C5.0 works by splitting the sample based on the field that provides the maximum information gain. Each subsample defined by the first split is then split again, usually based on a different field, and the process repeats until the subsamples cannot be split any further. Finally, the lowest-level splits are reexamined, and those that do not contribute significantly to the value of the model are removed or pruned.

C5.0 offers a number of improvements on C4.5. Some of these are [21]:

- Accuracy: C5.0 uses Boosting improves the trees and gives them more accuracy. Boosting is a technique for generating and combining multiple classifiers to improve predictive accuracy

- Speed: C5.0 is much faster. For instance, if C4.5 required nine hours to find the ruleset, the C5.0 completed the same task in 73 seconds.
- Memory: C5.0 commonly uses an order of magnitude less memory than C4.5 during ruleset construction. For example [21], C4.5 needs more than 3GB (the job would not complete on earlier 32-bit systems), but C5.0 requires less than 200MB.

Figure 7a and 7b shows the accuracy rate and tree size respectively in C4.5 and C5.0. it is shown that C5.0 has more accuracy rate over the different datasets and also less generated tree size.

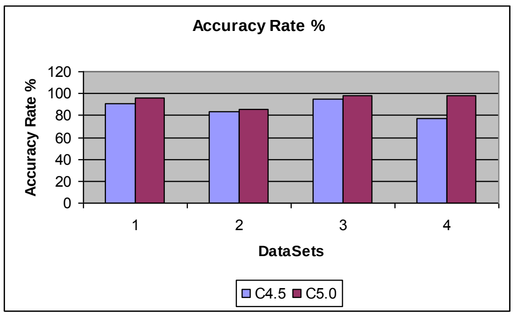

Fig8.7b. C4.5 and C5.0 tree size

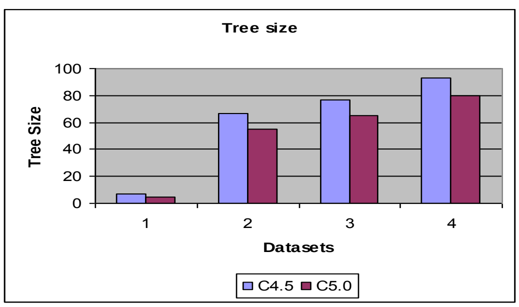

Fig9.7a. C4.5 and C5.0 accuracy rate

4. Conclusions

The data recorded in network log files whether by monitoring tools or by operating system itself has valuable knowledge, these knowledge may be mined to discover interesting patterns will be useful in improve network performance like:-

- Decide highly resources usage and decide whether the load is suitable for it or not and redistribute the recourse loads.
- Decide which S.W. to upgrade and pay for a new license or remove it from network due to no. of usage and results.
- Decide the machines need to be upgraded by monitoring CPU utilization or memory utilization not due to user requirements.
- We can determine which computer calls which S.W., when and how much.
- Decide the workstations that need to be memory upgraded by monitoring memory utilization graph.
- Decide the periods the network delay increases to decide which applications to stop this time or limit it in specific machines like internet browsing or downloading.
- Monitoring the network hardware that enables IT department to decide the best hardware to buy and facilitate the ability to redesign network.
- Decide the maintenance contract conditions for the resources such as printers and others depends on the resource utilization.
- Determine the future needs of accessories like papers, ink and replace Spare parts.
- Decide training plans due to different Hardware and Software results in different departments.
- Increase the availability of help desk staff according to physical needs.

As a future work we tend to invent new classification techniques for enhancing the computer network performance and conduct empirical comparisons to such proposed classification algorithms. Such technique may enhance the performance evaluation of the computer network system with new performance measures.

References

[1] Svobodova L. "Computer Performance Measurement and Evaluation Methods: Analysis and Applications", Elsevier Science Publishers, 2001.

[2]Liu, Mei-Ling. "A Look at Network Performance Benchmarks", Department of Computer Science, Cal Poly State University, San Luis Obsipo, CA. August 1997.

[3] Addison Wesley – "Managing a Microsoft Windows Server 2003 Environment " Microsoft Press 2003.

[4] Fayyad U.M, Piatesky- Shapiro G.,Smyth P. and Uthururusamy R.,"From data mining to knowledge discovery: An overview",Advanced in Data mining and knowledge discovery,MIT press.

[5] Jeremy Moskowitz. "Windows 2003 Active Directory Administration Essentials "Microsoft press 2003.

[6] John M.waolfer. "Network.performance.toolkit Using open source testing",John Wiley AndSons 2002.

[7] Jiawei Han, Micheline Kamber, “Data Mining Concepts and Techniques” Morgan Kaufman, 2001.

[8]. Zhang, Q. et al, ’Network planning in wireless ad hoc networks’, IEEE Explore, 2005,Volume 23 issue 1.

[9]. Achuthsankar N. and Jayasudha J., “Improving Performance of World Wide Web by Adaptive Web Traffic Reduction”, World Academy of Science, Engineering and Technology 17 2006.

[10]. T. Rokicki et al, Improving Network Performance by Reducing Network Contention in Source-Based COWs with a Low Path-Computation Overhead, IEEE, 2001.

[11]. Sivanantharasa P., Performance evaluation of information and communications technology infrastructure for smart distribution network applications, Master thesis, Brunel University, 2012.

[12]. Chuck C., et al, High performance network monitoring with an SQL interface, ACM 2002, June.

[13]. Xiaohua, H., A Data Preprocessing Algorithm for Data mining Application 2003.

[14]. Deepankar, S., Ruchi, G., Amit, R., Performance Measures of A Computer Network System, JATIT, 2009.

[15]. Rupinder K., Performance Evaluation of QoS Routing in Computer Network, International Journal of Computational Engineering Research (ijceronline.com) Vol. 2 Issue.5 Issn 2250-3005(online) September| 2012 Page 1480.

[16]. Partha S., et ell, A Framework for Selecting the Most Reliable Path in a Computer Network using Particle Swarm Optimization (PSO) based on Fuzzy Logic, International Journal of Computer Applications, Volume 45– No.8, May 2012.

[17]. Steven, A. Hofmeyr, K., Immunity by Design: An Artificial Immune System VLDB Journal Vol. 3, No. 4, September 2007.

[18]. Hidber, C., “Online Association Rule Mining”, In Proceedings of the ACM SIGMOD International Conference on Management of Data, New York: ACMPress, pp.145–156, 1999.

[19]. Quinlan, J. R.” C4.5: Programs for Machine Learning”, Morgan Kaufmann Publishers, 1993.

[20]. Kotsiantis S.B., “Supervised Machine Learning: A Review of Classification Techniques”, Informatica 31,pp.249-268, 2007.

[21]. http://www.rulequest.com