Research Article Impact Factor: 0.621 ISSN: 2319-507X
LK Gautam, IJPRET, 2014; Volume 2 (9): 133-138 IJPRET

INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

NETWORK TRAFFIC ANALYSIS USING CLUSTERING

PROF. L. K. GAUTAM[1], PROF. V. P. VAIDYA[2], PROF. RASHMI N. GADBAIL[3]

1. Assistant Professor, Sipna's COET, Department of Computer Science And Engineering, Maharashtra, India.
2. Assistant Professor, Sipna's COET, Department of Computer Science And Engineering, Maharashtra, India.
3. Assistant Professor, IBSS COE Department of Computer Science And Engineering, Maharashtra, India.

Accepted Date: 27/02/2014 ; Published Date: 01/05/2014

Abstract: This paper exploits a newish self-organizing clustering method to analyze network data. The method is based on cooperative behavior of ants. Collective intelligence of an ant colony rests on interactions of individual ants and the environment. Hence no central control is needed and the whole process is performed unsupervised by simple agents. The data used in the analysis was recorded from a border gateway of a campus network. The theory of clustering ants is introduced before the clustering process takes place. The target of the research is outlined and preprocessing of the data is described in detail. Several clustering procedures are performed and the results are visualized. Daily variation in network traffic is discovered and the method is found to be working. A way to detect long range changes is presented and some changes are observed in the data. Future work and lacks of the clustering method are discussed.

Keywords: Ant Colony, Clustering, Network Traffic Analysis, Long Range Dependencies, Netflow, Peer-to-Peer

Corresponding Author: PROF. L. K. GAUTAM

Access Online On:

www.ijpret.com

How to Cite This Article:

LK Gautam, IJPRET, 2014; Volume 2 (9): 133-138

PAPER-QR CODE

INTRODUCTION

There has been a continuously growing need for research and work on network management and traffic classification since the early 1990's when Internet broke through into people's everyday life. In the recent years, Quality of Service (QoS) has become an important issue, because new real-time applications like Internet calls and net meetings have become more and more popular. Thus network traffic management, characterization and classification has been an issue for many years and a lot of work related to it can be found. An early work on flow profiling is done in [2]. Profiling network applications has been done in [8] and network data has been classified in [5] and [7]. The aim of this paper is to cluster network data, perceive the known daily variation in the amount of traffic and after that develop a method for detecting long range changes in networks. Ant clustering was chosen to be the clustering method, because it does not require a priori knowledge of the number of clusters formed. Another purpose of the research is to test the feasibility of ant clustering, how it works on a real world problem. According to our knowledge, this is the first time when ant clustering has been used in network traffic analysis. The next chapter introduces briefly the basics of the clustering ants. Chapter 3 describes data collection and in chapter 4 the data is given a closer look and the analysis is described. Development of the clustering method was inspired by cemetery organization and larval sorting phenomena detected on some ant species [1]. Deneubourg et al. [3] developed a basic model which was redeveloped by Lumer and Faieta [9]. In this work, the model developed by Lumer and Faieta is used. Many proposals to improve this model have been made, e.g. [4], [10], [11], [12], but the model used here is rather easy to understand and satisfies the needs of this paper

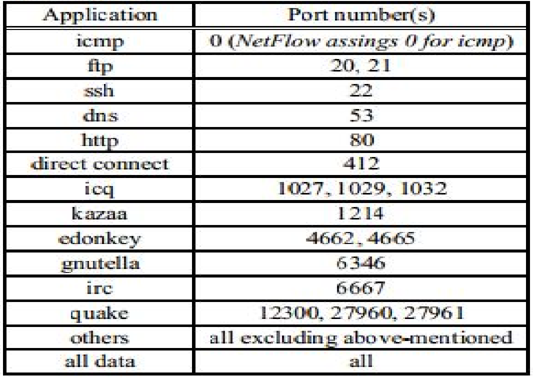
DATA COLLECTION

The data used in this work is real world network data recorded from a border gateway of a campus network. This kind of data enables us to get more realistic knowledge about the Internet users and applications than data collected from local area or test network, which usually is the case. Technology used for collecting data was NetFlow, developed by Cisco [14]. NetFlow enables flow data collection from routers. In the data produced by NetFlow, a flow is a unidirectional traffic stream of packets having common source and destination, protocol, type of service and input interface [13]. Even though flow data produced by NetFlow is unidirectional, the data used in the analysis is bidirectional, because it contains both incoming and outgoing data. Using flows instead of packets causes some information losses, but collecting data on packet level would make the analysis impossible because of the massive amount of information.

ANALYSIS

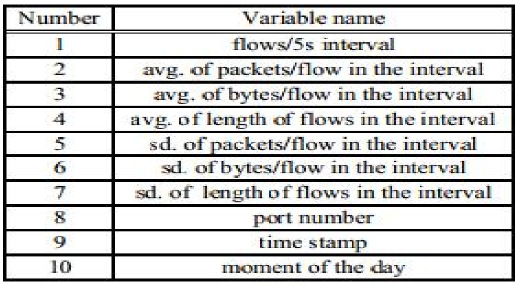
The analysis done for this work consisted of preprocessing the data, running several clustering procedures, visualizing the clustering outcome and interpreting the results. All data pre- and post processing was done with Matlab®. The clustering was performed with a small self-made program called antclust written in C++ for this purpose. The performance of ant lust appeared to be hundreds of times faster than the performance of a corresponding Matlab®-function, which was first used to cluster the data. Data Description and Processing Each datarow, a flow, contains 37 variables. Most of the variables are categorical, e.g. source and destination IP-addresses, source and destination port numbers, input and output interface indices, protocols and type of service. The only continuous variables are packet and byte count and variables related to time stampThe amount of data was huge and had to be reduced. To reduce the data, four five-minute files were chosen from each 19 days included in the analysis. To detect the daily variation the files were chosen at different moments of the day. The moments were just before 6,12,1824 o'clock. There were still millions of lines of data so it was split into smaller sets by applications. These applications are presented in Table I. The original data together and all the other applications not included were also considered as datasets. Now five-second averages were evaluated for each of these datasets. This reduced the data considerably and provided us with fourth continuous variable, the amount of flows in the interval. Usually averaging loses the self-similar nature of the traffic, but here the time interval was so short that the typical burstiness of self-similar traffic could still be seen. Table II represents the final format of the data used in clustering.

TABLE I. Datasets used in clustering. Data was divided by port numbers, e.g. a flow was considered to be an HTTP flow if either one of the port numbers was 80 [15].



Application	Port number(s)
icmp	0 (*NetFlow assings 0 for icmp*)
ftp	20, 21
ssh	22
dns	53
http	80
direct connect	412
icq	1027, 1029, 1032
kazaa	1214
edonkey	4662, 4665
gnutella	6346
irc	6667
quake	12300, 27960, 27961
others	all excluding above-mentioned
all data	all

TABLE II. The final format of the data. The first seven variables were used in evaluating the euclidean distance between datapoints.



Number	Variable name
1	flows/5s interval
2	avg. of packets/flow in the interval
3	avg. of bytes/flow in the interval
4	avg. of length of flows in the interval
5	sd. of packets/flow in the interval
6	sd. of bytes/flow in the interval
7	sd. of length of flows in the interval
8	port number
9	time stamp
10	moment of the day

Due to the chosen clustering method, the data was scaled inside a hypercube where minimum value for an attribute was 0 and maximum value 1. Formally, if ū = [u1,...,ui,...,un] is an n-dimensional data point, after scaling every ui [0,1]. Because of the self-similar characteristic of network traffic, logarithmic scaling had to be used. With simpler scaling, e.g. only dividing each attribute with the maximum value of the attribute in the dataset, almost all of the data would have been scaled close to zero. Moreover, there had not been enough variance to distinguish small values from medium values. However after logarithmic scaling each attribute was divided with the maximum value of the attribute in the dataset. With this kind of logmax scaling every data point was inside the hypercube. There was also enough variance to distinguish originally small values from originally medium values.

∈ Ant clustering needs some parameter values to be adjusted for good performance. The values α=0.5, k1=0.1 and k2=0.15 were adopted from an earlier work, where their use is also explained more carefully [1]. The value for dmax=0.6 is based on experience gained during the research done for this work. Iterations were limited with tmax to 500000, because it seemed to be a sufficient time to form clear clusters, although with bigger datasets tmax could have been increased. Also the size of the grid and the number of ants had to be adjusted. Some rules of thumb were sought to produce comparable circumstances for each clusteri procedure. The following formulas were finally used and they seemed to perform well with bigger datasets. width of grid amount of data

8

1 = ∗ number of ants amount of data

100

7 = ∗

Original values were saved before clustering. This omitted the need for rescaling the data before analysing the results. Clustering produces two columns of coordinates so the original data was attached to the coordinates and scaled values were simply ditched. The clusters based on the coordinate colums were formed with Matlab®. Clusters with less than 10 objects were ignored.

RESULTS

Results are best presented as figures of cluster centers. Before examining the results it must be remembered that the data is highly compressed and consists of five-second averages. Hence these results must be taken with a grain and consider them only as suggestive. Daily variation can now be seen as movement of the cluster centers. The clearest variations among the applications included are in HTTP (Figure 1) and DNS (Figure 2). HTTP is in fact the main source for the daily variation of all network traffic, because web is mainly browsed in the daytime and the amount of HTTP traffic is big enough to affect the nature of the whole traffic. Most of the wide area traffic is web-related. Web connections are usually preceded by DNS lookups so it is likely that DNS behaviour is closely linked to HTTP [6]. This can be clearly seen in the similarity of figures 1 and 2. The best regressors seem to be flows and bytes. Packets and bytes are highly correlated so there is no significant difference which one to use when plotting the cluster centers.

HTTP creates roughly two times more flows than DNS. One reason for this is that web browsers use multiple flows to download a page, e.g. pictures may be downloaded as separate flows. Thus the loading of a web page causes more HTTP flows than DNS flows. importantly, because the data was collected on a border gateway of a campus network, no DNS request is seen in this data if the information was found on a server inside the network. Days were compared to other days by clustering all data of a single day and plotting the cluster centers together. This is illustrated in Figure 3. There is a clear group of days that have more flows than others. These days are 22nd, 23rd, 24th, 25th, 26th and 29th of April 2002. The 27th and 28th day were Saturday and Sunday and can be found in the left group. It is a known fact that there is less traffic during the weekends.

There can be several reasons for this growth, e.g. a new sub network may have been attached to the network or a new popular web or file sharing server has been founded inside the network. Although in this case the rise in flows is not due to increased web traffic, because there was no similar phenomenon in HTTP traffic. Long range changes in peer-to-peer traffic are interesting, because these applications seem to take over more and more of available bandwidth. There can be seen an obvious rise in Gnutella traffic during April 2002, this is illustrated in Figure 4. On the right side of the figure are all the days after 13th day. The reason for such an obvious rise in flows could be due to a new server inside the network. It is now obvious that network traffic growth can be detected by observing the movement of cluster centers. This information can be used as a decision support when deciding about new investments on hardware etc. It would also be possible to notice if e.g. peer-to-peer applications begin to occupy too much bandwidth and their use should be limited. Another reason is DNS caching. Browsers may save some recently visited pages and more

CONCLUSION

In this paper flow data was analyzed using clustering ants. The basic theory was briefly introduced, the problem at hand was described and the data was studied. Clustering processes were performed and the results were interpreted. The data used in all of the analyses was real world network data which was produced by daily users of a campus implies, every data point does not necessarily belong to any cluster, adjustable parameters require attention and the effectiveness of the algorithm is directly related to CPU power available. Some work to improve these flaws can be found in [1], [4], [10], [11], [12]. No clear advantages in ant clustering compared to traditional methods have been found so far. The results presented in this paper would have been achieved with a better known method like substractive clustering or self-organizing map. Those yet undiscovered benefits of ant clustering may underlie in its distributed nature. Whether to take advantage of this trait with a multiprocessor system or distinct workstations over Internet is left as future work. Another issue left as future work is that the method for perceiving long range changes could also be used as a fault detector. To do this every five minute interval should be included. This way every five minutes could be compared to each other. For example the jams that occur in DNS might not last longer than five minutes so it is reasonable to compare a lot shorter periods than days to detect the possible crashing of a server.

REFFERENCE

1. Bilgili MS, Ahmetdemir and Bestamin Ozkaya, Quality and Quantity of Leachate in Aerobic Pilot-Scale Landfills., Environmental Management, 2008; 35(2): 189-196.

2. Florida Center for Solid and Hazardous Waste Management (FCSHWM), Analysis of Florida MSW landfill leachate quality. University of Central Florida, Florida. 1998.

3. Hossain MS, Pennethsa KK and Hoyos L, Permeability of Municipal Solid Waste in Bioreaktor Landfill with Degradation, GeitechGeolEng, 2009; 27: 43-51

4. Jaramillo. J, Guidelines for the Design, Construction and Operation of Manual Sanitary Landfills, Universidad de Antioquia, Colombia. 2003.

5. Powerie, W, and Beaven RP, Hydraulic Properties of household waste and implication." Proceedings Institution of Civil Engineers Geotechnical Engineering. 1999; 137: 235-247