Discovery of the causes of high latency network using data mining techniques: A case study

Leandro D. Pulgatti[1]

[1]Universidde Federal do Paraná (UFPR)

Curitiba – PR – Brasil

ldpulgatti@inf.ufpr.br

Abstract. *Today, computer networks are essential not only for businesses but for the whole society, especially with the advent of increasingly powerful devices that are connected and need to communicate larger amounts of data to perform their functionalities. Latency, available bandwidth and packet loss ratio, are generally essential metrics used for monitor the network status. Monitor and control these resources tend to become a chore, but these components are critical to keep scalability, reliability and high availability in the networks. The main idea of this project is to use the various tools of data mining techniques for analyzing error logs of a system Management Network to find which technique is the most suitable to locate patterns or trends on this logs that can explain or indicate what is causing the high latency in the network.*

1. Introduction

Today, computer networks are essential not just for business but for all of society, especially with the advent of increasingly powerful devices that need to be connected and to communicate larger amounts of data and perform their functions. The network management has become complex due to the almost exponential growth that has occurred in recent years as well as their heterogeneity, in this scenario the administration and management of computer networks tend to become increasingly pressured to maintain quality of the service and customer satisfaction. Along with this scenario is important to note that the current tools for network management cannot cover all problems that can be found. In this scenario can be seen that a high latency has a stronger effect on user satisfaction and usability of software that depend on the transfer of data to its full operation, however identify a cause, or even a set of causes that are causing the latency in the network tends to be complex task. In this first version of the paper an overview of what is latency and how to discover will be presented. Then, we will explain the techniques of data mining and an overview of the Weka implementation, demonstrating how it intends to use these resources to the discovery of the causes of latency. Presenting at the end what the next steps and work being carried out to completion of the final paper.

2. Latency

Latency can be defined as a delay in the communication process. Network latency is an expression of how much time it takes for a packet of data to get from one designated point to another. In some environments, latency is measured by sending a packet that is returned to the sender; the round-trip time is considered the latency. The contributors to network latency include: Propagation: The time it takes for a packet to travel between one place and another. Transmission: The medium itself introduces some delay. The size of the packet introduces delay in a round trip since a larger packet will take longer to receive and return than a short one. Router and other processing: Each gateway node takes time to examine and possibly change the header in a packet. Other computer and storage delays: Within networks at each end of the journey, a packet may be subject to storage and hard disk access delays at intermediate devices such as switches and bridges. In TCP / IP connections, latency can also directly affect the overall performance of the communication.

2.1. Measurement and detection

In most cases the ping command of the operating system can be used to measure round-trip latency. Ping performs no packet processing, he just sends a response back when it receives a packet, so it is a first rough measure of latency. Ping cannot make accurate measurements, and differs from the actual communication protocols, such as TCP. Also routers and ISP can apply different policies to different traffic shaping protocols. However a typical packet is forwarded over many links via many passages, each of which will not begin transmitting the packet until it has been completely received. In a network of this type, the minimum latency is the sum of the minimum latency of each connection, the transmission delay over each link. In practice, this minimal latency is augmented by queuing and processing delays. Latency can be caused by many environmental factors, such as Ethernet problems, congestion or network applications, network devices, long distance, packet loss, or even poorly written applications.

The network logs, including syslogs, contain important information that help to diagnose any abnormalities or fluctuations in the network. The syslog messages are often used to enable operators to verify the root cause of problems and decide what should be done to recover the failure, or return the network to its state of better performance. However, despite the large amount of information available in the records of the network, these tend not to be fully utilized in production due to the fact of unstructured texts possessing logs are generated by specific rules of each supplier. Since a large network is composed of elements from multiple vendors, record formats are highly diversified. Table 1 shows an example of the syslogs

Table 1: An Example of Syslogs

ID:	Timestamp:	Event Severity:	Att1	Att2	Message
##:	Nov 13 00:06:23:	ERR:	bridge:	!brdgursrv:	queue is full. discarding a message.
##:	Nov 13 10:15:00:	WARN:	INTR:	ether2atm:	Ethernet Slot 2L/1 Lock-Up!!
##:	Nov 13 10:15:00:	WARN:	INTR:	ether2atm:	Ethernet Slot 2L/2 Lock-Up!!
##:	Nov 13 10:15:00:	WARN:	INTR:	ether2atm:	Ethernet Slot 2L/3 Lock-Up!!

2.2. Data mining

In the Knowledge Discovery in Databases (KDD) discipline, data mining is one of the processes used, and is the computational process of discovering patterns in large data sets involving several methods including the use of artificial intelligence, machine learning, statistics, and database systems (Figure 1). The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. On the context of this paper, data mining is the process of extracting useful information from server logs e.g. network logs and syslogs, specifically seeking to find patterns or trends that can lead to finding or direct the search of the main causes that may be causing a high latency in network.

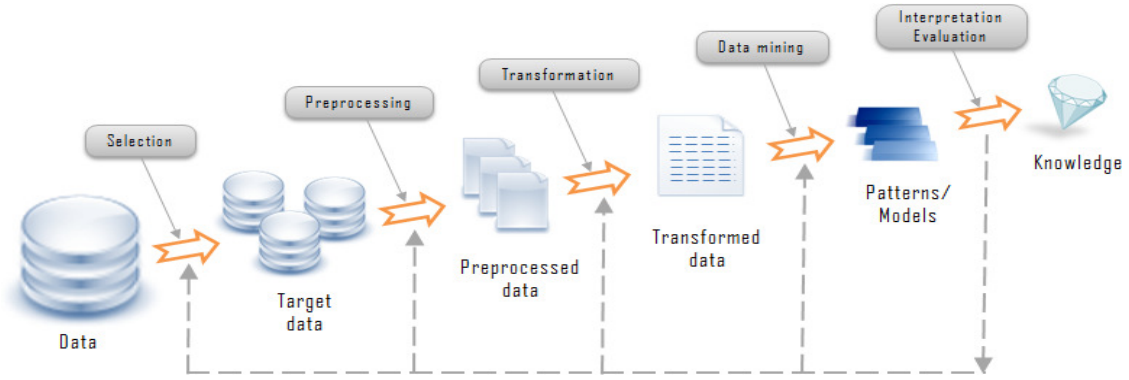

Figure 1. Main steps of a typical KDD process

There are several major data mining techniques have been developing and using in data mining, the main and most utilized are :

2.2.1. **Association**

In association, a pattern is discovered based on a relationship between items in the same transaction. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together. Based on historical sale data, retailers might find out that customers always buy crisps when they buy beers, and therefore they can put beers and crisps next to each other to save time for customer and increase sales.

2.2.2. **Classification**

Classification is a technique based on machine learning and is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics.

2.2.3. **Clustering**

Clustering is a data mining technique that makes meaningful or useful cluster of objects which have similar characteristics using automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes.

2.3. Weka

Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. The capacity of Weka to support several data mining tasks, specifically, data preprocessing, clustering (Figure 2), classification, regression, visualization, and feature selection turns this tool in a natural choice for the type of task proposed in this paper.

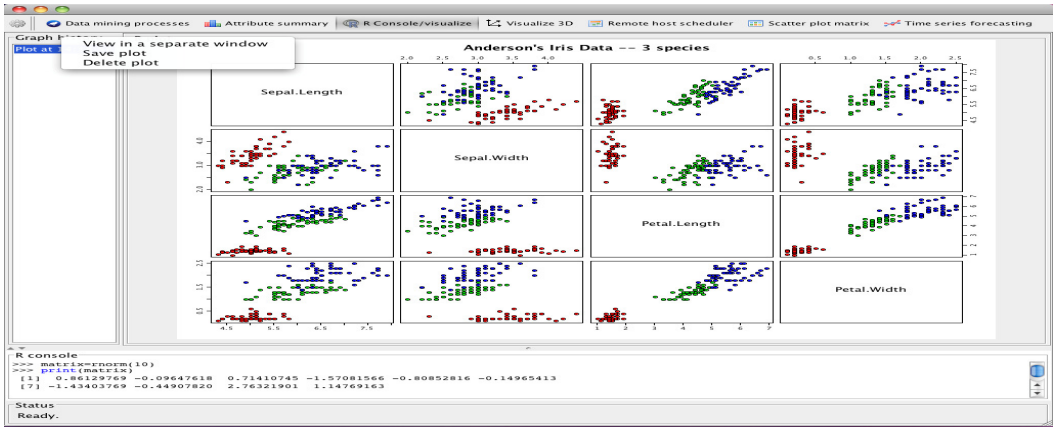

Figure 2. Clustering screen on the Weka software.

2.3.1. DataSets

All of Weka's techniques are predicated on the assumption that the data is available as a single flat file wherein each data point is described by a fixed number of attributes. These files are called Datasets, and are an ASCII text file that describes a list of instances sharing a set of attributes. This files have two distinct sections. The first section is the Header information, which is followed the Data information.

```
% 1. Title: Iris Plants Database
% 2. Sources:
%      (a) Creator: R.A. Fisher
%      (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%      (c) Date: July, 1988
@RELATION iris

@ATTRIBUTE sepallength  NUMERIC
@ATTRIBUTE sepalwidth   NUMERIC
@ATTRIBUTE petallength  NUMERIC
@ATTRIBUTE petalwidth   NUMERIC
@ATTRIBUTE class       {Iris-setosa,Iris-versicolor,Iris-virginica}
```

The Data of the file looks like the following:

```
@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
```

3. Conclusion and Next steps

Network latency is one of the parameters that most impact the perception of users in relation the quality of services offered. Despite being common and have multiple causes are not always properly diagnosed because more in-depth analysis in the system log tend to be lengthy and complex. The use of data mining techniques is proposed to fill this space to simplify and streamline the process of finding out what are the main causes of this situation. A major challenge is to know which of the various techniques which exist in data mining is the most suitable for the type of data generated and displayed by these logs, and the process of converting these into a format that can be used by the tool (pre-processing) can be considered a non-trivial step of the process. In the full paper will be proposed a comparison of the different methods to identify the most suitable for the task. Table 2 shows a schedule of upcoming activities planned.

Table 2. Schedule of next activities

Schedule

Main Task	Dates
Research of new papers	Oct, 27th to Nov, 27th
Transform syslogs in .aff format	Oct, 27th to Nov, 3
Run the analysis in Weka tool	Nov, 3 to Nov, 27th
Compare the results and white a report	Nov, 3 to Nov, 27th

References

J. Dean and S. Ghemawat, “MapReduce: simplified data processing on large clusters,” in Symposium on Opearting Systems Design & Implementation (OSDI). USENIX Association, 2004.

M. A. Kolosovskiy and E. N. Kryuchkova. Network congestion control using netflow. arXiv preprint arXiv:0911.4202, 2009.

T. Telkamp. Traffic characteristics and network planning. In Proc. Internet Statistics and Metrics Analysis Workshop, 2002.

Bouckaert, Remco R.; Frank, Eibe; Hall, Mark A.; Holmes, Geoffrey; Pfahringer, Bernhard; Reutemann, Peter; Witten, Ian H. (2010). "WEKA Experiences with a Java open-source project". Journal of Machine Learning Research 11: 2533–2541.

Ian H. Witten; Eibe Frank; Mark A. Hall (2011). "Data Mining: Practical machine learning tools and techniques, 3rd Edition". Morgan Kaufmann, San Francisco. Retrieved 2011-01-19.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

KDD Cup 1999 dataset, converted to ARFF format

http://tunedit.org/repo/KDD_Cup/KDDCup99.arff - Last accessed 24/11/2014

Directory contains measurements of the latencies between a set of DNS servers

http://pdos.csail.mit.edu/p2psim/kingdata/ - Last accessed 24/11/2014

TANENBAUM, A. S. Redes de Computadores. 4a edição. Editora Campus. 2003.

Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases"