ABOUT ONE DIMENSION REDUCTION METHOD OF ANALYZING FEATURES OF NETWORK TRAFFICS USED FOR COMPUTER NETWORKS MONITORING

Ramiz Shikhaliyev

Institute of Information Technology of ANAS, Baku, Azerbaijan
ramiz@science.az

Dimensions reduction method of feature space of network traffic, used for network monitoring of computer networks (CN) is being reviewed. Method is based on application of construction algorithms of association rules. Proposed approach allows systemizing of collected monitoring data, with significantly reduces the time spent by the administrators of CN for analysis of network traffic and making a justified decision on CN administration.

Currently, network monitoring systems (NMS) are used for collection of necessary information on condition of CN. Following are included in tasks of NMS: collection of values of characteristics of transmitting channels and switching equipment of CN; detection of abnormal situations, as well as bottlenecks in CN; forecasting of aftermaths of changes in topology of CN; monitoring of users' actions etc. However, conduction of network monitoring in modern CN with the assistance of traditional NMS is becoming a very difficult task, because significant experience and knowledge are required from network administrators for analysis of monitoring results. Mainly, it is connected to diversity of the structure and large volume of network traffic. Thus, during the network monitoring of CN, most important tasks are rapid analysis of significantly larger volume of network traffic and extraction of the most important data, in order to flexibly and operatively react to changing conditions of CN. Solution of this task can be achieved by reducing the dimensions of monitoring data. For this reason, we propose the dimensions reduction method of feature space of network traffic, used for network monitoring of CN, i.e. extraction of the most important indications out of the majority of indications (parameters). Thereby, reduction of dimensions with a large volume of monitoring data can be achieved with the use of intellectual data analysis (IDA) methods and consists of description of analyzed initial set of features of network traffic of CN and their replacement with a new set of more valuable features with significantly less volume.

Using the methods of IDA for analysis of monitoring data is explained by diversity of structure of monitoring data, complexity of obtainment of analytical information from the network traffic with a significantly larger volume, as well as the great number of simultaneously working users, servers, network equipment and applications in CN, necessity of continuous control of functioning of CN and making substantial decisions on network administration.

Another motive for application of IDA methods for reduction of the dimensions of feature space of network traffic used for network monitoring of CN, is reduction of temporary and computational costs (for example, random access memory, disc space, processor time etc) consumed for processing and storage of monitoring data without losing useful information.

Mainly network traffic of CN consists of the traffic of clients, servers and applications [1]. Client engines start the generation of traffic from the moment they are switched on, and do not stop generating traffic until they are switched off physically. In their turn, sources of client traffic can be as following: traffic connected to protocols, search of different objects in the network etc. Different types of existing servers and applications in CN generate a very large volume of traffic.

Usually, network traffic of CN is characterized with a variety of features that are used for network monitoring of CN. DNS-request, DHCP-request, DHCP-reply, WINS-traffics, number of packages, volume and speed of incoming and outgoing traffics. IP-address of sender and receiver, MAC-address of hosts, types of used protocols (for example HTTP, FTP, SMTP) and applications, time etc can be applied as such features. On this basis, diversity of freatures descriing the nework traffic of CN, can reach several hundreds. Usually these and other features of network traffic of CN are collected in log-files and/or data bases.

Reduction task of monitoring data dimension consists of following: Let us suppose that $X = \{x_1, x_2, ..., x_n\}$ - is a set of features characterizing network traffic of CN. Let us suppose that $T = \{t_1, t_2, ..., t_m\}$ - is the network traffic of CN, which consists of m traffics of CN subjects (for example, users, servers, applications), where each t_i th traffic consists of a set of features, contained in a set of X, i.e. $T \subseteq X$. It is required to find association rules for detection of commonly occurring robust combinations of CN network traffic features used for network monitoring, which will allow the dimension reduction of features describing the network traffic of CN.

For solution of set task, we propose a method which is based on application of detection algorithms of association rules. The main idea of this approach consists of application of search algorithms of association rules for detection of commonly occurring robust combinations of features of CN network traffic used for monitoring. Search of association rules is one of the effective methods of IDA, based on which concealed connections between objects of certain subject fields are detected and presented [2, 3].

Let us suppose that $X = \{x_1, x_2, ..., x_n\}$ - is a set of features characterizing network traffic of CN. Let us suppose that $T = \{t_1, t_2, ..., t_m\}$ - is the network traffic of CN, which consists of m traffics of CN subjects (for example, users, servers, applications), where each t_i th traffic consists of a set of features, contained in a set of X, i.e. $T \subseteq X$. It is supposed that, network traffic T contains a set of features A, contained in the set X, if $A \subseteq X$. Then, implication $A \Rightarrow B$ is the association rule of features of CN network traffic, whereas $A \subset X$, $B \subset X$ and $A \cap B = \varnothing$. In CN network traffic, $A \Rightarrow B$ rule is performed with c certainty, if $c\%$ of subject traffics of CN contained in T contains a set of A features, as well as a set of B features. Herein, certainty of the rule is calculated based on below formula:

$$c(A \Rightarrow B) = \frac{s(A \cup B)}{s(A)}.$$

Rule $A \Rightarrow B$ is supported by s, if $s\%$ of subject traffics of CN, contained in T network traffic of CN, contains $A \cup B$, where $s(A \Rightarrow B) = s(A \cup B)$.

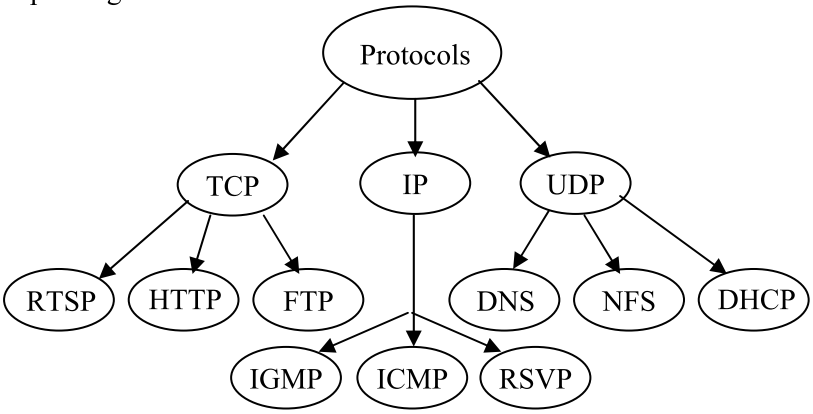
Currently, for detection of association rules of data with a large volume, APriori algorithm is broadly used, and its primary virtue is flexibility [4]. Herein, it is possible to assign both *min_sup* and *min_conf* rules, which allows to receive a set of different groups of rules. However, generation of a large number of association rules creates a serious problem for their analysis. For that reason, detection of associations in data monitoring is insufficient for using only APriori algorithm. Accordingly, it is proposed to generalize “similar” rules. i.e. determine the generalized association rules, which contain received rules that include predecessor features of features contained in CN subject traffics. As a result, we can detect association rules not only with separate features of CN subject traffics, but also between CN subject traffics.

Upon detection of generalized association rules, taxonomy (hierarchy) of features of CN network traffic is an important element. The meaning of taxonomy here, is a forest of directed trees, leaves of which are features of CN network traffic, and their internal nodes – are their groups. An example of hierarchy of some protocols and groups of protocols, which are the features of CN network traffic are depicted on Picture 1. In rules achieved as a result of such taxonomy, there can be elements situated on different levels of taxonomy both antecedently and subsequently. For example, “if HTTP-protocol is present in user traffic, then presence of DNS-protocol is probable”.

Introduction of additional information about grouping of features of CN network traffic in hierarchic form can provide following advantages:

1. Association rules not only between separate features of CN network traffic, but also among different feature groups can be detected.
2. In some cases, separate features of CN network traffic can have very little support, but the support value of the entire group containing this feature may exceed the *min_sup* threshold.

3. Introduction of information about grouping of features of CN network traffic can e used for pruning of un-informative rules.



Pic.1. Hierarchy of some protocols and groups of protocols.

Therefore, $A \Rightarrow B$ implication is called generalized association rule, where $A \subset X$, $B \subset X$ and $A \cap B = \varnothing$; and none of the elements contained in set B is a predecessor of any element in set A. Support and authenticity are calculated in the same manner as in association rules case.

For detection of generalized association rules, it is efficient to use specialized algorithm [5], which is more effective than standard Apriori algorithm.

References

1. Эд Уилсон. Мониторинг и анализ сетей. Методы выявления неисправностей. Изд. «Лори», 2002, 350 с.
2. A. Savasere, E. Omiecinski, S. Navathe. An Efficient Algorithm for Mining Association Rules in Large Databases, *Proc. 21st Int'l Conf. of Very Large Data Bases*, 1995.
3. J. Han, M. Kamber– Data mining: concepts and techniques, – 2000. – C. 279–310.
4. R.Agrawal, R.Srikant. Fast Algorithms for Mining Association Rules in Large Databases, Proc. Conf. Very Large Databases, 1994, pp. 487–499.
5. R.Srikant, R.Agrawal. Mining Generalized Association Rules, Proc. Conf. Very Large Databases, 1995, pp. 407–419.