A network diagnostics method based on pattern recognition algorithms

Olizarovich E.V. [1)], Rodchenko V.G. [2)],
1) Olizarovich Evgeny, Yanka Kupala State University of Grodno;
230005, Grodno, Belye Rocy st., 45-31; e.olizarovich@grsu.by.
2) Rodchenko Vadim, Yanka Kupala State University of Grodno;
230005, Grodno, Malyshchinskaya st., 23-21; rovar@mail.ru.

Abstract: *This report deals with the problem of designing and building of a computer network diagnostic system. Diagnostic content problems are reviewed, as well as ways of their solution based on mathematical and computer modelling methods. A traffic analysis-based diagnostic method is suggested for process statuses in a computer network. The method is based on algorithms of the mathematical pattern recognition theory. To build a diagnostic system, a multi-level model building and verification arrangement is suggested. Methods and techniques for data collection and preliminary analysis are considered.*

Keywords: simulation models, network diagnostics, pattern recognition.

1. INTRODUCTION

There exist a number of methods and tools for study and analysis of computer network (CN) performance. However, the problem of building efficient diagnostic systems is still pressing, due to the fact that this type of technical systems undergoes constant modification. Accordingly, constant renovation of diagnostic tools is an objective necessity.

Changes do not only involve upgrading of CN components and technical parameters. Just as constantly modified are CN building and application principles. To build a CN is not usually a target, but a technological way to support certain production processes. Accordingly, features and, especially, consumer-oriented characteristics of networks differ significantly. Every CN may be regarded as a unique system which all too often requires specific diagnostic tools.

Constant changes take place in both CN building principles and their technical bases. Transmission medium undergoes changes, and so do physical and logical topologies, switching methods, statistical characteristics of traffic, etc. Hence, the following causes for the need to change diagnostic tools: a) incompatibility with new technical and technological solutions, b) insufficient precision in solving new diagnostic tasks.

The first problem objectively requires replacement of equipment. This is successfully settled by the producers of instrumental control tools.

Problems of the second type are characteristic of technical systems which are computer-based. Such systems are a priori characterised by universality and adaptivity to practical application in solving most diverse tasks. Thus, for example, within the same CN certain tasks may require to determine the use of servers, or the degree of activity of various users, etc., with technical structure and possibilities to measure the network parameters being unchanged.

Thus, there exist tasks which require studies of a number of various aspects of the system's functioning, on the basis of one data set. Similar problems occur in many of the applied spheres of human activities, and they are successfully solved through methods of applied statistical analysis [1,2]. Methods of exploration for mineral resources may serve as examples, along with medical diagnostics, processing of sociological research findings, and many others.

The main advantage of the use of mathematical and computer modelling during studies of complicated systems is a unique opportunity to study the processes which cannot be measured directly due to their physical or other specificity. An additional advantage is a significant extension of the measuring tools' abilities and reduction of costs through extended equipment life.

Regardless of a particular sphere of realisation, operation of any CN causes a number of complicated tasks, such as calculation of use factor for channels and servers, registration of network resource use and user activity, classification of emergency conditions, detection of abnormal events, etc. In this relation, it becomes relevant to find a solution to the problem of development of a CN diagnostic system which would allow essential changes in the composition of the tasks currently being solved, without changing technical measuring tools.

This paper deals with description of a method which would allow to extend the abilities of instrumental CN diagnostic tools, based on computer processing with the use of applied statistics methods.

2. PROBLEM DESCRIPTION

The problem area of computer network diagnostics covers a wide range of tasks. These include technical, infrastructural, and informational parameters of a network. Solution of each of these tasks calls for specific approaches, methods, and tools.

In spite of a number of possible technical solutions, the information component of a CN is based, as a rule, on a limited set of standard data transfer protocols. As regards modern computer networks, most widespread are the Ethernet family – for lower levels of the ISO/OSI reference model, TCP/IP stack – for high-level protocols.

It is essential that in organisation of CN parameter measuring subsystems an important role is played by the model of available equipment and software. Depending on the manufacturer, equipment class, and production year, abilities to obtain particular data may differ significantly. Examples may include availability and degree of support of network control protocols (SNMP, RMON) and their libraries, availability of switch port mirroring functions, data presentation format. These and other features are realised by manufacturers in a varying

number of ways.

Basic CN diagnostic tools are greatly oriented at certain proprietary solutions. This means that maintenance of a heterogeneous CN will require either additional inefficient expenses for procurement of special equipment, or extension of qualified staff. Another disadvantage is the basic impossibility to obtain the kind of information regarding a CN, which is not provided by the producer.

However, the above features of CN organisation provide a wide range of opportunities to use computer processing tools. Results of technical and software system operations are registered in electronic logs which contain time-ordered information about events. Each event is presented as a vector which contains information on CN elements in the attribute space of $X = \{x_1,\ldots x_n\}$ characteristics.

Network analyser databases, log files of sniffer programs, of proxy servers, routers, web servers, etc. may be cited as examples of electronic logs. Depending on particular characteristics of the object under study and their particular realisation, various characteristics may become the parameters to be measured. Traditionally, the data set of a log includes: time of the event, quantitative characteristics of the event (statistic data, rate, delay time); informational characteristics (addresses, direction of transmission, priority, data).

Let us call these data primary observation results. Then, a set $X = \{x_1,\ldots,x_n\}$ is an a priori feature vocabulary for building a system of classification of CN statuses against a given diagnostic criterion.

At various stages of a life cycle of a diagnostic system, methods of mathematical pattern recognition theory may be applied with a high degree of efficiency. One of such approaches implies application of similar methods both at the stage of status classification and in the process of building a mathematical model. The concept of building such a recognition system includes two major phases: building of reference standards for CN status, based on analysis of primary data; diagnostics of CN status on the basis of these standards, with the use of mathematical pattern recognition theory apparatus [3].

While building reference standards, five main stages are distinguished: determination of status class sets; formation of a priori feature vocabulary; building of primary classified learning sample; exploratory analysis of feature informativity aimed at building solution space; building of reference standards within solution space.

Thus, in the process of building an observation-based CN diagnostic system two related types of models are to be built: a descriptive mathematical computer network model and an algorithmic CN control model.

3. APPROACH TO SOLUTION

According to the suggested scheme of observation-based diagnostics, one may consider traffic to be the source of data on CN status [4]. For each particular diagnostic task, primary data source is the files and data bases which contain operation logs of various elements of the given CN.

As stated above, in the case when primary data allow direct diagnostics the problem is successfully solved through the use of specialised equipment, e.g., protocol analysers and testers.

Let us consider a status diagnostics method used in situations when classification of a status through direct measurements is impossible. In this case obtaining information on CN status requires enhancement and filtering of primary data. The standard enhancement method is aggregation and calculation of statistical characteristics. Filtering may be realised through recognition methods.

CN possesses a number of unique features which distinguish it from other complicated technical systems. Most important for realisation of the above method are the following ones: firstly, predefined structure of each particular transmission process; secondly, presence in the transmutation medium of not more than one transmitted packet. This makes possible reliable diagnostics of a CN status on the basis of traffic observation, provided that observation methods are statistically independent of the diagnosed processes.

Technologically, diagnostics of ill-defined and obscure processes within a CN should include the following stages:

1) determination of the types of diagnosable CN statuses;
2) building of a set of observable features, i.e., a priori vocabulary;
3) development of algorithms for preliminary transformation of primary data, i.e., enhancement, aggregation, etc.
4) building of primary data array, i.e., classified learning sample;
5) formation of solution space;
6) verification of obtained CN status reference standards.

Primary object for building a diagnostic system is a model of the object under study, i.e., computer network traffic. Within this paper we will use the terms "CN traffic model" and "CN model" as synonyms.

4. MODEL-BUILDING STAGES

The figure below demonstrates main stages of building a model of a CN diagnostic system.

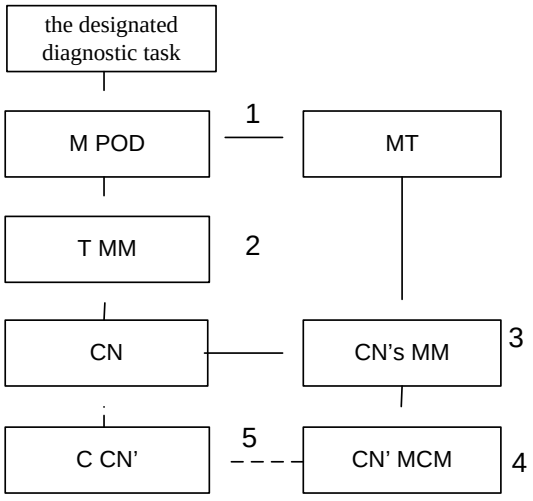

Fig. - Modelling stages of a diagnostic system.

The diagram includes the following stages:

1. Determination of expected methods of obtaining primary observable data (M POD) and corresponding measuring tools (MT). Object study is performed at this

stage, as well as description of its properties and specific features to be taken into account while solving the designated diagnostic task. On the basis of the obtained data, an M POD and MT set is formed through iterative search for solutions. In the case of a specialised CN, a decision may be taken at this stage regarding the need to develop new MT types. For general purpose, as a rule, methods of obtaining primary observable data should be oriented at the use of existing tools.

2. Selection of the mathematical model type (T MM). According to CN characteristics and task requirements, selection of a method (methods) of building a CN's mathematical model (MM) is performed and requirements to MM are determined. Type of the model and its building method are determined at this stage.

3. Building of a CN's MM on the basis of selected M POD and T MM. At this stage, structure of the model is determined, and its building algorithms. Entry data types are selected, as well as preliminary transformation schemes and algorithms, interfaces, building principles. The obtained model will be regarded as a formalised equivalent of CN for the purposes of the designated task.

4. Building of a CN's mathematical control model (MCM) on the basis of selected M POD and CN's MM. Status classification algorithms and result output methods are determined. The obtained control model should ensure compliance with all the given parameters of a diagnostic system. MCM parameters are determined proceeding from requirements to diagnostic procedure and result presentation format.

5. Verification of model compliance. At the verification stage, the obtained mathematical models are verified for compliance with required characteristics. In particular, model testing is performed under real CN conditions at known functioning requirements. A real CN is transformed into a sequence of states $S = \{s_1,\ldots s_n\}$. If CN's MCM fixes these changes correctly, the model is deemed adequate, and the built models (MM, MCM) are regarded adequate to their corresponding prototypes (CN, C CN) within the designated task.

The resulting system of models should include:

1) a set of classes of states $A=\{A_1,\ldots,A_k\}$, where k is the number of recognisable states;
2) a set of observable data $X=\{x_1,\ldots,x_m\}$, where m is the number of characteristics which can be measured by available tools;
3) a procedure of feature formation $P=\{p_1,\ldots,p_n\}$, where n is the number of features on the basis of which solution space is built;
4) a procedure of final classification.

While building a diagnostic system, a given model should be adapted to the conditions of a particular CN in order to optimise it and reveal its features. The process of adaptation of the diagnostic system includes:

- formation of a class-wise alphabet (a list of diagnosable statuses) specific for the CN under study, and a priori feature vocabulary;
- performance of reference measurements of the network status and formation of a classified learning sample;
- separation of features included into the a priori vocabulary according to the degree of their informativity, with regard to reference divisions in the multi-dimensional feature space;
- formation of a solution space which only contains features most precisely reflecting individual peculiarities of each diagnosable status for a given CN;
- building of reference standards for classes, based on the obtained solution space and the contents of the classified learning sample.

The process of technical state diagnostics includes the following stages:

- monitoring of traffic, parameter changes, calculation of current state vector values;
- diagnosing of the network's technical state;
- analysis, interpretation, and drawing of conclusions.

The use of recognition algorithms allows additional inclusion of the reference verification function into the diagnostic process. Such a mechanism provides for recognition of dynamic processes which are characterised by long duration, as well as presence of hidden events.

The above concept allows building of a relatively easy-to-use software and hardware set. However, the model building and system adjustment processes require participation of network technology experts at the stages of designing and adaptation of a diagnostic system.

5. CONCLUSIONS

The suggested method may be used in designing and building of diagnostic systems for any computer network, regardless of its type, used equipment, and information infrastructure. It is most efficient in high-level tasks related to analysis of a CN as a complicated information system, as well as in other cases when specialised diagnostic methods are not available or inefficient.

The main advantage of the suggested method is its non-dependence on particular technical solutions, available equipment and software. The system built on the basis of the above method may have a comprehensible interface and allows to automate the main stages of computer network control.

6. REFERENCES

[1] Gorelik, A.L. Sovremennoe sostoianie problemy raspoznovania. Nekotorye aspekty. M.: Radio i sviaz, 1985, p. 160.

[2] Zagoruiko N.G. Prikladnye metody analiza dannykh i znaniy. Novosibirsk: Izd-vo Instituta matematiki SO RAN, 1999. p. 268.

[3] A.I. Zhukevich, Ye.V. Olizarovich, V.G. Rodchenko. Metod postroenia etalonov sostoianiy kompiuternoi seti na osnoe primenenia algoritmov teorii raspoznavania obrazov. *Proceedings of the IIIrd International scientific confer.* "*Network computer technologies*", Minsk, BSU, Oct. 17-19, 2007, p 14-17.

[4] Ye.V. Olizarovich, V.G. Rodhenko. Ob odnom metode avtomatizatsii protsessa diagnostiki sostoiania kompiuternoi seti. *Proceedings of the III international confer. "Informatsionnye sistemy i tekhnologii (IST'2006)"*. Minsk, Nov 1-3, 2006. p. 211-214.