MATHEMATICAL MODELING OF COMPUTER SYSTEM USERS' BEHAVIOR

Tojiev N. Sh
Samarkand Institute of Economics and Service, Teacher

The large-scale use of computer technology in almost all spheres of human activity has brought more and more attention to the user himself. Knowledge of what actions he performs (or should perform) can be applied in different areas, for example, in security systems, to create a personalized environment for users and so on. Therefore, the task of building models of user behavior of computer systems is relevant. In this paper, we propose a comprehensive user model consisting of interactive and session parts, which in both models to detect deviations from the usual take into account, respectively, dynamic and statistical properties of user behavior for expected user behavior neural networks are used. Thus, the interactive model is based on predicting user commands based on previous commands. Since the choice of neural network architecture is a non-trivial task, it is important to know how much its current behavior depends on the pre-history. In the case of the session-based model, there is a problem with the sample size that is used to train the neural network. The fact is that when the training set size is small, the neural network tends to memorize images locally, which is undesirable. In the session model, the input to the neural network is the data collected during the session as a whole. Accordingly, the size of the training set is directly determined by the number of sessions during which the user's activity was monitored. However, even over a long period of time, this data will not be sufficient for qualitative training of the neural network. Therefore, it is very important to provide a more representative sample of data in the session model for qualitative training of the neural network.

Comprehensive neural network model of computer systems user, the comprehensive user model proposed in the paper takes into account both dynamic (interactive part) and statistical (session part) properties of user behavior. The developed model is based on a feed forward neural network which consists of input, output and one or more hidden layers of neurons. The interactive model is used to detect abnormal activities during user's activity. For each user of the computer system, a neural network is constructed and trained so as to predict the next command based on previous commands. Based on the number of commands that were correctly predicted by the neural network, it is concluded whether the user's current behavior matches the previously built model.

It should be taken into account that users tend to change their behavior over time. Therefore, in order to ensure adaptation to their behavior, the neural network should be periodically retrained. The session model is designed to identify uncharacteristic user activity for the session as a whole and for this purpose a statistical data set is used. This information, in turn, is used to build and train a neural network that determines how much user activity matches the previously built model. When describing the data for modeling the behavior of computer system users, we used real data that were collected in the local network of the Space Research Institute of NASU-NSAU for two to three months.

In this network, workstations operate under Windows 98, XP, 2000 operating systems (OS). Since these operating systems do not fully provide the necessary information about user activity, a special software application was developed. A separate audit file is created for each user session (its name unambiguously defines the name of the user account and the date of its work), which stores information in the following format: command start time|command identifier|command name|start or end flag.

The following is an example of such an audit file: 11:28:50|620|WINWORD.EXE|STARTED

1 l: 30:58|2276|Far.exe|STARTED.

11:39:17|730|WINWORD.EXE|STARTED 11:39:17|620|WINWORD.EXE|FINISHED.

It should be noted that the command identifier is assigned by the OS and is unique (it is different for commands with the same name, and it also changes from session to session). Therefore, when encoding commands, it is important to ensure that the same values correspond to the same commands. To this end, for the interactive part of the complex model, an alphabet of commands (i.e., the set of commands entered by the user over a specified period of time) was constructed for each user based on the information collected. Each command was then assigned a corresponding decimal number, which was subsequently used in converting the audit files into sequences of commands. As a result, the following data set was available for each user { } i j N j N i i j N i i s =1 =1 "{ } i j N j N i i s =1 =1, (5) where i j s is the decimal number of the entered jth command of the i-th session;

N - number of sessions; i N - total number of commands in i -th session.

The analysis of the plotted curves shows that the autocorrelation functions decrease as the number of lags increases. At the same time, the extrema are observed at values of lags from 1 to 8. Thus, this number of commands should be used when predicting user commands. At the same time it is necessary to take into account the following: the use of too many commands will lead to the fact that during this period of time it will be impossible to carry out the prediction of commands, which will reduce the ability to detect abnormal user activity.

In general, the functioning of a neural network significantly depends on the quality of the training sample. The point is that when the training set size is small, the neural network tends to hard memorization of images, which leads to a decrease in its ability to generalize. Thus, when building the interactive user model in our case, the problem with representative data sampling did not arise, because even for a short period of user time, a sufficient number of (representative) images can be collected to train the neural network. (For example, given that the user inputs an average of 80 to 150 commands per session, then over ten sessions the Training sample could number up to 1000 images.) In the case of the session model, the size of the training set is directly determined by the number of sessions during which the user's work was observed. Thus, for three months the number of such sessions can reach 100, which in our case was insufficient for qualitative training of the neural network and optimization of its architecture.

Two approaches can be used to solve this problem. The first one is to significantly increase the time interval within which the user's behavior is observed (say, up to 8-10 months). However, in this case, there is a high probability that the user's behavior will change during this period of time and thus the training set will contain contradictory images.

The second approach is statistical modeling of data based on the available sample. It will be used in this paper. Since the session model uses information about the number of commands entered per session, computer number, duration and start time of the session to train the neural network, this particular data set was modeled for each user. For this purpose, the consistency of the empirical distribution with a set of theoretical distributions (normal, log-normal, uniform, etc.) was checked. As a criterion of agreement we used the criterion of 2% The different distributions were analyzed, but the best value of 2 x was obtained for the log-normal distribution, which provided 97% correspondence of the hypothesis to the real data

Computer number. When analyzing the values of the parameter it is necessary to take into account the following: the user, as a rule, has his main place at the computer and very rarely works at others. Therefore, there is always a value that has the highest probability of 0.8...0.95. Events related to the user's work at other workstations can be considered equally probable.

Session duration. Analysis of the values of this parameter shows that they are distributed equally likely. The value of the criterion 2 x provides 95% compliance of the hypothesis with the real data.

Session start time. The best value of the 2 x criterion for this parameter was obtained for a normal distribution, which provided 90% compliance of the hypothesis with real data. Based on the theoretical dependencies obtained, a program was developed to simulate the user's work per session. The data generated with its help made it possible to optimize the architecture of the neural network and improve the quality of its functioning.

Thus, in this paper, mathematical modeling of computer system users' behavior was conducted. For this purpose, the complex model proposed in the paper was used. The dynamics of user behavior during a session was studied. Since determining the optimal neural network architecture is a non-trivial task, in the interactive model it is important to know how many commands should be used during training to predict the next one. Based on the autocorrelation curves constructed, it was found that up to eight commands should be considered for this purpose. Statistical modeling of the data used to train the neural network in the session-based model was also performed. Empirical distributions were approximated by theoretical ones and the necessary data set was generated based on them, which made it possible to optimize the neural network architecture and improve its functioning.

Literature

1. Kussul N., Sokolov A. Adaptive detection of anomalies in the behavior of computer system users using Markov chains of variable 340 order. 4.2 Anomaly detection methods and experimental results // Problems of Management and Informatics. - 2003. - № 4. - C. 83 - 88.
2. Manavoglu E., Pavlov D., Lee Giles S... Probabilistic User Behavior Models // Proc. of the 3rd IEEE International Conf. on Data Mining (ICDM 2003). - Melbourne, Florida (USA). -2003.-P . 203-210.
3. Davison B. D., Hirsh H. Probabilistic Online Action Prediction // Working Notes of the AAAI Spring Symposium on Intelligent Environments. -1998. - P. 148 -154.
4. Skakun C.B., Kussul H.H. Neural network model of computer system users // Cybernetics and Computer Science. - 2004. - Vyp. 143. - C. 55 - 68.