**SYSTEM ANALYSIS**

**EXISTING SYSTEM**

There are two distinct categories of cyber-attack detection methods, namely signature based and anomaly based. The machine learning techniques are used in both of them. Recently machine learning-based algorithms have been used for developing signatures that will efficiently identify both the code and behaviour of the malicious code.

The Network-based Signature Generation (NSG) [6], Length-based Signature Generation (LSEG) [7] and F-Sign [8] are the examples of algorithms designed for automated and fast extraction of signatures of polymorphic worms.

The LESG algorithm targets those worms that use buffer overflow attack to infect victims, whereas the F-Sign extracts the signature on a basis of the code of a worm (such signature can be used to detect and stop the worm from spreading). In literature there are also algorithms such as SA(Semantic Aware [9]) that are designed to generate the signatures of malicious software on a basis of the network traffic they generate. Such solutions can even properly identify malicious behaviour when the traffic is noise-like [9].

The anomaly based methods for cyber-attack detection typically build a model that describes normal and abnormal behaviour of network traffic. Commonly, such methods use three types of algorithms taken from machine learning theory, namely unsupervised, semi-supervised and supervised.

For unsupervised learning often clustering approaches are used that usually adapt algorithms like *k*-means, fuzzy *c*-means, QT and SVM [10–12]. The clustered network traffic established using the mentioned approaches commonly requires the decision whenever given cluster should be indicated as malicious or not. Pure unsupervised algorithms use a majority rule telling that only the biggest clusters are considered normal. That means that network events that happen frequently have no symptoms of the attack. In practice, it is a human role to tell which cluster should be considered as an abnormal one.

The supervised machine learning techniques require at least one learning phase to establish the traffic model. The learning is typically off-line and is conducted on the specially prepared (cleaned) traffic traces.

**PROPOSED SYSTEM**

The proposed method adapts machine learning paradigm. During the learning phase the labelled data is required to establish the model parameters of the normal application behaviour.

We propose to use a graph-based approach to build a set of regular expressions that model the normal HTTP requests sent by client to the web application.

In the proposed approach, the segmentation components *S* are the regular expressions further explained in Section 3.3. In other words, our goal is to group the similar HTTP requests and represent them with a single pattern. In fact, the algorithm is not only limited to the HTTP protocol and can be easily adapted to other kinds of textual data, like different kinds of log files generated by the

Application or databases.