Работа выполнена при финансовой поддержке РФФИ (гранты № 18-41-220004 и № 18-08-01475).

СПИСОК ЛИТЕРАТУРЫ

1. LAMMPS Molecular Dynamics Simulator. URL: http://lammps.sandia.gov (дата обращения 06.10.2018).

2. Stukowski A. Visualization and analysis of atomistic simulation data with OVITO–the Open Visualization Tool // Modelling and Simulation in Materials Science and Engineering. 2009. Vol. 18. No. 1. Article ID 015012.

3. Zope R. R., Mishin Y. Interatomic potentials for atomistic simulations of the Ti-Al system // Physical Review B. 2003. Vol. 68. No. 2. Article ID 024102.

4. Jordan V.I., Shmakov I.A. The study of microstructure and propagation of the combustion wave of SHS in nanodimensional multilayer systems of Ni-Al with using molecular-dynamic simulation // IOP Conference Series: Journal of Physics: Conf. Series. 2018. Vol. 1134. Article ID 012023.

5. Ackland G. J., Jones A. P. Applications of local crystal structure measures in experiment and simulation // Physical Review B. 2006. Vol. 73. No. 5. Article ID 054104.

УДК 004.048

Р. Г. Шыхалиев
e-mail: ramiz@science.az

Институт Информационных Технологий НАНА, Баку, Азербайджан

ОБ ОДНОМ МЕТОДЕ ОБНАРУЖЕНИЯ АНОМАЛИЙ В КОМПЬЮТЕРНЫХ СЕТЯХ НА ОСНОВЕ АНАЛИЗА ЛОГ-ФАЙЛОВ

В статье предлагается метод обнаружения аномалий в компьютерных сетях (КС). Предложенный метод основывается на анализе лог-файлов. Обычно данные в лог-файлах являются неструктурированными, а форматы данных лог-файлов различных серверов бывают разными. Предлагается метод анализа лог-файлов на основе алгоритма извлечения последовательных шаблонов. Классификация этих шаблонов позволить обнаружить аномалии независимо от типов форматов лог-файлов.

Введение. Любые инциденты компьютерных сетей (КС), включая отказ в обслуживании и снижение качества обслуживания, могут привести к остановке работы сетевых приложений. Своевременное обнаружение аномалий позволит выявить и устранить эти проблемы.

Анализ лог-файлов может быть использован в системах мониторинга КС для прогнозирования и обнаружения аномалий. В большинстве случаев, лог-файлы являются единственным источником данных по обнаружению и

определению аномалий в КС. Поэтому обнаружение аномалий на основе анализа лог-файлов является очень важной областью исследований.

Обычно лог-файлы являются текстовыми ASCII файлами и данные в лог-файлы поступают динамично и стохастически. Вместе с тем, данные в лог-файлах не сортируются по определенной структуре, то есть они являются неструктурированными данными. Форматы данных лог-файлов различных серверов бывают разными. Поэтому для их анализа может потребоваться большие вычислительные ресурсы, например, даже для обнаружения и устранения незначительного сетевого инцидента может потребоваться анализ большого объема лог-файлов.

Анализ лог-файлов в основном, включает в себе фильтрацию, изменение формата и обобщение данных [1]. При этом, подходы и цели анализа лог-файлов могут быть различными.

В лог-файлах современных КС объем структурированных и неструктурированных данных постоянно увеличивается, так как постоянно увеличивается количество сетевых приложений, сервисов, устройств и пользователей. Вместе с этим, появляются новые виды сетевых приложений, сервисов и устройств, в результате чего усложняется процесс определения аномалий и угроз в КС на основе анализа лог-файлов.

Основная проблема при анализе лог-файлов заключается в том, что непосредственный анализ лог-файлов человеком (администратором) является очень сложной, практически невозможной задачей [2]. Для решения этой проблемы используются различные приложения (известные как анализаторы лог-файлов), которые могут анализировать лог-файлы конкретных серверов и создать различные типы отчетов. Такие приложения могут иметь различные варианты конфигурации, но при этом могут создать только определенные типы отчетов, основанные на заранее определенных встроенных запросах.

Целью данной статьи является разработка универсального метода анализа лог-файлов, независимого от типов их форматов. Для этого предлагается использовать методы добычи данных (англ., data mining), а именно алгоритм извлечения последовательных шаблонов (англ., mining sequential patterns algorithm).

Метод анализа лог-файлов. Для представления процесса обнаружения аномалий на основе анализа лог-файлов предлагается следующая структура, которая состоит из трех этапов: сбора лог-файлов, извлечения из лог-файлов последовательных шаблонов и классификация извлеченных последовательных шаблонов, то есть обнаружения аномалий (рис.).

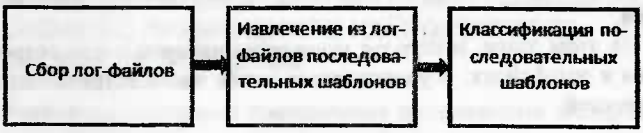

Рис. Структура процесса обнаружения аномалий на основе анализа лог-файлов

Первый этап анализа заключается в сборе лог-файлов. В КС обычно генерируются лог-файлы, в которых записываются состояния сети. Записи в лог-файлах имеют временные отметки, то есть записи в лог-файлах имеют определенные временные атрибуты (час, день, неделя, месяц, год и т.д.), что является необходимым для обнаружения аномалий.

Второй этап анализа заключается в том, что из лог-файлов извлекаются последовательные шаблоны. А в третьем этапе осуществляется классификация извлеченных шаблонов.

Задачей данной статьи является исследование второго этапа анализа лог-файлов, а именно извлечения из лог-файлов так называемых максимальных последовательностей шаблонов, для которых предлагается использовать алгоритм PrefixSpan [3]. Максимальные последовательности – это последовательности, которые имеют наибольшую поддержку пользователей сети. Так как, в КС трафики в основном инициируются пользователями сетей и несмотря на то, что их поведения отличаются, в их трафиках можно обнаружить общие тенденции.

Алгоритм извлечения последовательных максимальных шаблонов из лог-файлов заключается в следующем:

1. Лог-строки сортируются по идентификатору пользователей (например, по IP-адресу пользователей), которые представляют трафики пользователей. В свою очередь трафики пользователей сортируются по времени (для определения краткосрочных тенденций) или по дате (для определения долгосрочных тенденций). Таким образом, из лог-строк получаются пользовательские последовательности.

2. Осуществляется поиск набора всех часто встречающихся элементов, полученных в первом этапе пользовательских последовательностей. Потом, наборы часто встречающихся элементов обозначаются буквами (например, HTTP – a, TCP – b, SMTP – c, FTP – d и т.д.) и превращаются в альтернативное представление, что позволяет упростить алгоритмическое решение задачи.

3. Определяется, какие из часто встречающихся последовательностей относятся к пользовательским последовательностям. Для этого, последовательности пользовательских трафиков заменяются набором их часто встречающихся элементов. Если в пользовательском трафике нет никакого набора часто встречающихся элементов, то этот трафик вообще не рассматривается. Если в пользовательской последовательности нет никакого набора часто встречающихся элементов, то этот трафик последовательностей тоже не рассматривается. После преобразования, каждая пользовательская последовательность представляется в виде множества наборов часто встречающихся элементов.

4. На этом этапе, используя множества наборов часто встречающихся элементов в лог-файлах, осуществляется поиск часто встречающихся последовательностей.

5. Наконец на последнем этапе, среди часто встречающихся в лог-файлах последовательностей осуществляется поиск максимальных последовательностей.

Результаты этого этапа используются в качестве входных данных для классификатора. В качестве классификатора могут быть использованы различные классификаторы – искусственные нейронные сети, методы машинного обучения (англ., data mining) и т.д.

Заключение. Своевременное обнаружение аномалий является ключевым элементом в управлении инцидентами в масштабе КС. Для обнаружения аномалий была предложена структура анализа лог-файлов, состоящая из трех этапов: сбор лог-файлов; извлечение последовательных шаблонов и классификации шаблонов. В данной статье был исследован второй этап анализа лог-файлов, для которого был предложен алгоритм извлечения последовательных шаблонов. На наш взгляд, данный подход позволит анализировать лог-файлы вне зависимости от типов их форматов, что сделает этот метод универсальным.

СПИСОК ЛИТЕРАТУРЫ

1. Alspaugh S., Chen B., Lin J., et al. Analyzing Log Analysis: An Empirical Study of User Log Mining. Proceedings of the 28th USENIX conference on Large Installation System Administration, 2014, pp. 53-68.

2. Alspaugh, S., et al. Better logging to improve interactive data analysis tools. In KDD Workshop on Interactive Data Exploration and Analytics 2014, pp. 19-25.

3. Pei J., Han J., Mortazavi-Asl B., and etc., Mining sequential patterns by pattern-growth: The prefixspan approach // IEEE Transactions on Knowledge and Data Engineering. 2004. Vol. 16, no. 11. Pp. 1424-1440.

4. Alspaugh, S., et al. Better logging to improve interactive data analysis tools. In KDD Workshop on Interactive Data Exploration and Analytics 2014. Pp. 19-25.

УДК 004.932

Д. А. Юдин, В. В. Прахов
e-mail: ydin.da@bstu.ru, betalik97@yandex.ru

Белгородский государственный технологический университет им. В. Г. Шухова, Белгород

ПРОГРАММНЫЙ КОМПЛЕКС АВТОМАТИЗИРОВАННОЙ РАЗМЕТКИ ИЗОБРАЖЕНИЙ С ПРИМЕНЕНИЕМ НЕЙРОСЕТЕВОГО ДЕТЕКТИРОВАНИЯ ОБЪЕКТОВ

В статье рассмотрены современные программные инструменты для разметки изображений при решении задачи детектирования объектов.