О МЕТОДЕ СОЗДАНИЯ ПРОФИЛЯ ДЛЯ ВЕБ-ПОЛЬЗОВАТЕЛЕЙ

ABOUT THE METHOD OF CREATING A PROFILE FOR WEB USERS

R. Alguliev
Y. Imamverdiyev
B. Nabiyev

Summary. There are some tools for securing computer networks and optimizing processes. Considering this, to determine the behavior profile of traffic on the network, a special tool has been developed. To determine the behavior profile, the K-means clustering method was applied. The reason for choosing the K-means algorithm is that this method is very fast and simple for solving the clustering problem.

As a result of the application of the clustering model, certain clusters were formed. Clusters, in the main, form social networks, video resources and scientific and practical resources. The result is obtained for 20 clusters using the bigml.com resource. Most of all, the cluster under consideration consists of scientific and practical resources. The 2nd cluster in turn, these are social networks. The third cluster consists of calls to video resources. Appeal to other clusters is much less.

Keywords: network traffic, clustering, behavior profile, abnormal traffic, communication channels, centroids, video resources.

Алгулиев Расим Магамед оглы
Институт информационных технологий при НАНА; rasim@science.az
Имамвердиев Ядигяр Насиб оглы
Институт информационных технологий при НАНА; yadigar@iit.science.az
Набиев Бабек Расим оглы
Институт информационных технологий при НАНА; babek@iit.science.az

Аннотация. Существует множество средств для обеспечения безопасности компьютерных сетей и оптимизации процессов. Учитывая это, для определения профиля поведения трафика в сети, разработан специальный подход. Для определения профиля поведения применён метод кластеризации К-средних. Причиной выбора алгоритма К-средних является то, что для решения задачи кластеризации этот метод является очень быстрым и простым.

В результате применения модели кластеризации были сформированы определённые кластеры. Кластеры, в основном, формируют социальные сети, видео-ресурсы и научно-практические ресурсы. Результат получен для 20 кластеров с помощью bigml.com ресурса. Наиболее часто обращаемый кластер состоит из научно-практических ресурсов. 2-й по порядку обращаемый кластер — это социальные сети. Третий кластер состоит из обращений к видео ресурсам. Обращение к другим кластерам значительно меньше.

Ключевые слова: сетевой трафик, кластеризация, профиль поведения, аномальный трафик, каналов связи, центроиды, видео-ресурсы.

Введение

В стремительно глобализирующемся мире ускоренное получение любого ресурса или информации с помощью Интернета стало очень легко и доступно. Это очень позитивная и необходимая ситуация в условиях информационного общества. Но, как мы знаем, не вся генерируемая информация, является необходимой и полезной. Это, создаёт излишнюю нагрузку на компьютерную сеть, что в свою очередь, снижает доступность каналов связи. Это событие, является одним из тех событий, с которыми рано или поздно могут столкнуться корпоративные сети, неадаптированные к правилу поведения.

Согласно отчёту фирмы Symantec, представленному в 2014 году [1], число предотвратимых нападений на веб-ресурсы в течение одного дня составляет 586700. Принимая это во внимание, для того, чтобы пользователи сети могли избежать столкновений с угрозами, эффективно использоват корпоративные ресурсы, с ограниченными возможностями и для повышения пропускной способности информационных каналов, предлагается формирование профиля поведения в трафике сети (в дальнейшем профиль поведения) на основе метода кластеризации сетевого трафика. Анализируя данные, полученные с помощью сетевого мониторинга трафика на основе оценки кластеризации, могут быть получены кластеры поведения определенного трафика, и реализация этого процесса осуществляется через алгоритм кластеризации. К-средних.

Анализ подобных исследований

Одним из ключевых элементов управления сетью являются идентификация сетевого трафика и категоризация. В качестве примера можно привести приоритезацию потока формирования трафика, транспортной политики и диагностику мониторинга. Во всем мире с помощью IP сетей передается и принимается огромное количество информации. Специалисты держат под контролем весь этот процесс и благодаря чему, выявляются и ликвидируются угрозы. Функции и параметры, включая заголовки пакета IP, позволяют получить большую информацию о сети и пользователях. Кроме того, результаты анализа заголовков

Таблица 1. Описание переменных кластеризации

Индекс	Объяснение переменных
1	Штамп времени
2	Время процесса
3	IP адрес
4	Результирующие коды
5	Объем контента
6	Метод запроса
7	URL
8	Код иерархии
9	IP отвечающего
10	Содержание

IP пакетов могут быть использованы для управления сетью и оптимизации, устранения угрозы и создания новых услуг. В [2], используя заголовки IP-пакетов, предлагается способ многоуровневой кластеризации в расширенной форме, объясняющий течение процесса в сети и профиль поведения пользователя. Кроме того, необходимо сказать, что проведенный процесс анализа используя заголовок IP-пакетов, обеспечивает неприкосновенность личной информации пользователей. Сетевой трафик или журналы файлов, собранные из трафика сети могут быть использованы для обнаружения аномалий и угроз. Для этого процесса используются различные методы и средства. Например, в [3], используя алгоритм кластеризации K-средних, предложен метод обнаружения аномалий в потоке трафика. Немаркированные данные сетевого трафика разделяются на два кластера, т.е. на нормальный и аномальный. В основе обнаружения аномалий в данных нового мониторинга лежит использование центра тяжести для выбора эффективного расстояния в определенных кластерах. Самоорганизующийся без центрального управления и без процесса контроля метод кластеризации является одним из самых новых подходов. Для этого, в [4] используется, основанный на взаимосвязи, метод поведения муравьев. Преимущество данного метода заключается в том, что нет необходимости в первичных данных и предварительного определения количества кластеров. Каждый из виртуальных муравьёв в отдельности и самостоятельно, исследуя сеть, выполняет процесс кластеризации. Но, поскольку этот метод является новым, коэффициент точности выполненного процесса вызывает сомнение.

В трафике сети подход «Machine learning» широко используется для определения аномальных потоков, основываясь на их уникальных статисти-ческих характеристиках. По сравнению с традиционной кластеризацией, нечёткая кластеризация является более гибкой, а для обнаружения вторжений и естественной обработки данных более целесообразной [5].

Многие методы кластеризации для обнаружения вторжений предусматривают разделение трафика на нормальный и аномальный. Методы кластеризации применяются для обнаружения разницы и схожих особенностей сессии трафика и для классификации каждого из них разделением на соответствующие группы [6]. Эти группы представляют присвоенные им знаки. В дальнейшем эти знаки используются для прогнозирования типов входящих сетевых трафиков.

Быстрая и точная идентификация сетевого трафика является одной из самых важных задач функции управления — QoS, мониторинга безопасности сети и т.д. Однако, в последнее время, количество узлов, использующих P2P увеличилось, и они, используя различные порты, скрываются под различными устройствами,

Таблица 2. Пример данных, собранных прокси-сервером Squid

Штамп времени UNIX	Время процесса (мсек)	IP адрес	Результирующие коды	Объем контента (байт)	Метод запроса	URL	Код иерар-хии	IP отвечающего	Содержание
1444780867.298	39	10.100.80.51	TCP_ MISS/200	10946	GET	http://pagead2. googlesyndication. com	HIER_ DIRECT	216.58.208.98	application/x-shockwave-flash
1444795608.042	3598	10.100.80.23	TCP_ MISS/301	567	POST	http://v.icecentury. com/	HIER_ DIRECT	54.169.165.185	text/html
1444795738.177	222	10.100.80.14	TCP_ MISS/304	318	GET	http://code. createjs.com	HIER_ DIRECT	23.77.228.124	application/x-javascript
1444799392.183	38	10.100.80.61	TCP_ MISS/200	345	HEAD	http://ds.download. windowsupdate. com	HIER_ DIREC	188.43.72.35	application/ octet-stream

необходимыми потоками сообщений или кодированными потоками сообщений, генерируя ненужные информационные потоки. В этом случае использование, считаемы классическими «port mapping» или «payload analysis» подходов, не эффективно. Альтернативным подходом является классификация сетевого TCP трафика исследованием поведения трафика внутри нескольких первичных пакетов. Это в будущем, кластеризируя всю информацию, позволяет облегчить процесс идентификации.

Лог-файлы обращений в интернет

Данные собраны в сетевой среде AzScienceNet состоящей из более чем 5000 адресов и эта сеть, также разделяется на несколько маленьких подсетей. С целью обеспечения ненарушения конфиденциальности пользователей, AzScienceNet основана на пользова-тельской политике и дополнительно ограничены данные о личности пользователей. Эти данные состоят из 10 переменных [7], приведённых в таблице 1.

Приведенные в таблице 1 десять переменных можно объяснить следующим образом:

1. Штамп времени. В целом, в области информационных технологий — символ или последовательность кодированной информации для регистрации даты появления, ликвидации, отправки или приема любого типа информации.
2. Время процесса. Регистрирует время процесса проведенное в кэше. То есть промежуток времени между началом и концом передачи пакетов HTTP.
3. IP адрес. Здесь регистрируются адреса обращений за информацией и к ресурсам.
4. Результирующие коды собирают информацию об ответе, отказе на запросы и т.д.
5. Объем контента важно для определения объема общего трафика с регистрацией объема контентов всех отправляемых и принимаемых пакетов.
6. Метод запроса, как правило, пишутся заглавными буквами, состоят из коротких GET, HEAD и т.д. английских слов. На основе этих методов определяется для чего был отправлен запрос от пользователя веб ресурса.
7. URL (Uniform Resource Locator) регистрирует имена доменов первого уровня и ссылки обращающихся пользователей сети.
8. Код иерархии предоставляет информацию о форме обработки запросов. Например, запрос был отправлен на прямую или через партнерский сервер и т.д.
9. IP отвечающего — IP адрес отвечающего на запросы
10. Содержание находится в заголовке HTTP ответа и показывает тип содержимого в объекте.

Данные приведённые в таблице 2 собраны с помощью прокси-сервера Squid. Прокси-сервер Squid

Таблица 3. Матрица информативных признаков пользователи / категории

	Кат1/ Объем (Гб)	Кат2/ Объем (Гб)	Кат3/ Объем (Мб)	Кат1/ Время (мин)	Кат2/ Время (мин)	Кат3/ Время (мин)	Кат1/ Запрос (количество)	Кат2/ Запрос (количество)	Кат3/ Запрос (количество)
Полз. 1	12	6	800	126	98	22	5355	4742	1586
Полз. 2	14	4,8	350	148	71	18	10163	3102	1475
Полз. 3	3,1	2,7	787	78	38	28	608	1554	3217

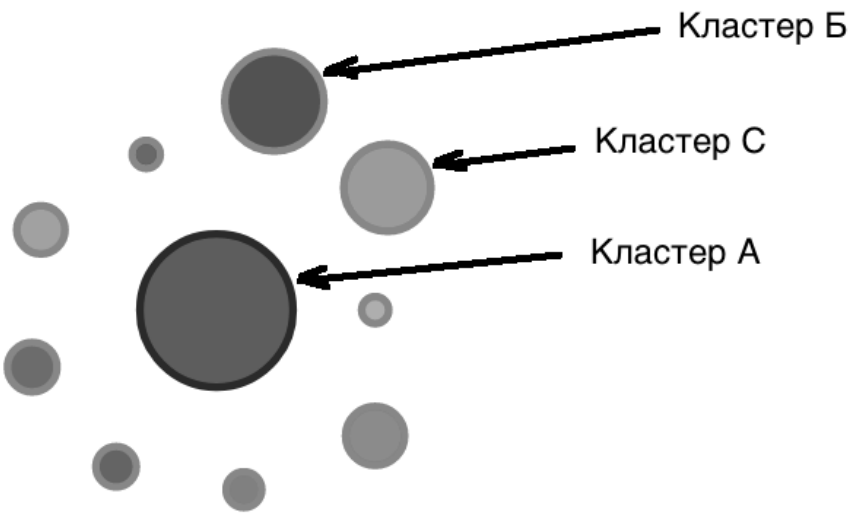

Рис. 1. Результаты применения кластеризации

[8] используется для реализации процесса накопления и управления лог-файлов сетевого трафика. Прокси-сервер Squid является программным обеспечением, с открытым кодом и его использование целесообразно в крупных сетях, где суточное число пользователей превышает 2000. Преимущество прокси-сервера Squid в том что, он является кэшируемым прокси-сервером, а в этом случае обращаемые ресурсы накапливаются в кэше и при повторном обращении процесс обработки завершается более ускоренно. Это в свою очередь положительно влияет на доступность сети. Лог-файлы, с помощью прокси-сервера Squid, накапливаются на специальной базе данных и используются в процессе анализа.

Очистка информации в лог-файле

Лог-файлы, накапливаемые с помощью прокси-сервера Squid, создают широкие возможности для интерпретации. Это в свою очередь создаёт условие для использования лог-файлов для различных целей. Пример данных, накопленных прокси-сервером Squid приведен в таблице 2. Однако, в рамках данной статьи нет необходимости конкретного рассмотрения всех 10-ти переменных представленных прокси-сервером Squid. При подходе со стороны информационной безопасности для идентификации профиля пользователя нет необходимости рассмотрения содержания обращения, ip назначения, http иерархического кода, способа опроса и кодов результата. Поэтому, во время анализа лог-файлов для облегчения и ускорения процесса обработки необходимо учитывать указанные переменные.

Идентификация профиля пользователя

Когда мы говорим о профиле идентификации, имеем ввиду вектор интересов и тематические выборы построенные на основе обращаемых веб ресурсов. Сбор те-

матических профилей пользователей создаёт матрицу. В этой матрице на каждой строке указывается пользователь, а в каждом столбике показаны признаки. В зависимости от частоты обращения ресурсов входящих в категории поведения пользователей и объёма входящего трафика, вычисляется значимость признаков. Для повышения качества модели проводится процесс нормализации свойств в интервале [0;1].

После завершения процесса проектирования признаков, для построения модели выбираются более информативные и достоверные признаки. Это уменьшает объем обрабатываемой информации, создаёт условие для предотвращения повторения процесса обучения, а также, в целом, повышает качество модели. В рассматриваемом случае ресурсы группируются согласно тематической категории. Понятно, что ресурсы, которые относятся к одной тематической категории, могут быть размещены в различных источниках.

Первым этапом решения проблемы Data mining является проектирование признаков (feature engineering). Это является ответственным и трудоёмким этапом и наряду с этим, непосредственно, влияет на результаты процесса. В рассматриваемом случае объектами являются пользователи сети, а в качестве признаков рассматриваются веб ресурсы, к которым обращаются пользователи. В результате полученного изображения признаков, формируется тематический профиль пользователей и получается матрица пользователь/категории, состоящая из информативных признаков. Полученная матрица имеет большие размеры (таблица 3), но по форме соответствует разреженной матрице (sparse matrix).

Алгоритм K-средних

Мы будем использовать алгоритм К-средних для кластеризации трафика сети. Причиной является то, что для решения задачи кластеризации алгоритм К-средних оказывается очень быстрым и простым. Если $X = \{x_1, \ldots, x_n\}$, то множество данных состоит из n сессий трафика. x_i представляет собой каждую трафик-сессию в d — мерной Евклидовой среде. $x_i = (f_1, \ldots, f_d)$, когда i трафик-сессия имеет значения $f_1, \ldots, f_d$, d-значение свойств. Это является основной целью разделения трафик-сессии по кластерам. Во время этого процесса ставится условие, что бы расстояние между n данными и соответствующими центроидами К кластеров было минимально. У каждого кластера имеется центр μ_k известный как центроид, и он может считаться представителем этой группы.

Таким образом, nxd матрица данных является входом алгоритма K-средних, K — количество кластеров, а центроиды являются первичными данными:

1. Сначала необходимо определить K точки, представляющие центроидные группы.
2. Для расчета Евклидова расстояния между каждым данным и самым близким центроидом используется уравнение:

$$dist(x, y) = \left(\sum_{i=1}^{n} (x_i - y_i)^2\right)^{\frac{1}{2}}$$

3. После определения всех точек, позиции К сентроидов заново вычисляются и это означает, что середина всех точек определенной группы μ_k должна также заново вычисляться.

2- и 3-й пункты должны повторяться до тех пор, пока не из менится позиция центроидов.

Выбор количества кластеров

В этом разделе до применения алгоритм К-средних, будет показано, как выбирается количество кластеров. Первым измеряется внутрикластерное расстояние, определяющее расстояние между точкой и центроидом. После этого определяется усредненное значение всех этих расстояний:

$$intra = \frac{1}{N}\sum_{i=1}^{K}\sum_{x \epsilon C_i} \|x - x_i\|^2$$

где, N — количество сессий (точек), K — количество кластеров, а z_i является центроидом кластера C_i**.** Далее, необходимо измерить межкластерное расстояние и при этом необходимо учитывать, что чем больше это расстояние, тем лучше. Для этого используется приведенная ниже формула:

$$inter = min(\|z_i - z_j\|^2),\ i = 1, 2, \ldots, K - 1;$$

$$j = i + 1, \ldots, K$$

Для определения количества K кластеров в алгоритме К-средних необходимо использовать следующую формулы:

$$validity = \frac{intra}{inter}$$

Результаты экспериментов

В результате применения модели кластеризации были сформированы определённые кластеры. Кластеры, в основном, формируют социальные сети, видео-ресур-

сы и научно-практические ресурсы (рис. 1). Результат показанный на рис. 1 получен для 20 кластеров с помощью

bigml.com ресурса [9]. Больше всех обращаемый кластер А состоит из научно-практических ресурсов. 2-й по порядку обращаемый кластер Б, это социальные сети. Кластер С состоит из обращений к видео-ресурсам. Обращение к другим кластерам значительно меньше. Это связано с тем, что пользователи основную часть необходимой информации получают от социальных сетей и видео-ресурсов.

Заключение

Данная статья посвящена проблеме определения профилей пользователей AzScienceNet на основе кластеризации. Для этого выбрана самая высокоскоростная и простая модель кластеризации на основе К-средних. В результате проведённых исследований были обеспечены: целесообразное распределение сетевых ресурсов, оптимизация сетевого трафика, определение источников аномальной активности и обеспечение своевременной ликвидации угроз.

ЛИТЕРАТУРА

1. http://www.itu.int/en/ITUD/Cybersecurity/Documents/Symantec_annual_internet_threat_report_ITU2014.pdf;
2. Kumpulainen P., Hätönen K., Knuuti O., Alapaholuoma T., Internet traffic clustering using packet header information / Joint International IMEKO TC1+ TC7+ TC13 Symposium, Jena, Germany, 2011, pp. 13–20;
3. Gerhard M., Sa L., Georg C., Traffic Anomaly Detection Using K-Means Clustering / In Proceedings of performance, reliability and dependability evaluation of communication networks and distributed systems, 4GI/ITG-Workshop MMBnet, Hamburg, Germany, 2007, pp. 25–33;
4. Ekola T., Laurikkala M., Lehto T., Koivisto H., Network traffic analysis using clustering ants / Proceedings. World Automation Congress, v. 17, Seville, Spain 2004, pp. 275–280;
5. Duo Liu, Chung-Horng Lung, LambadanЇs I., Seddigh N. Network traffic anomaly detection using clustering techniques and performance comparison / Proceedings the 26th Annual IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), Canada, 2013, pp.1–4;
6. Shokri, R., Oroumchian F., Yazdani N., CluSID: a clustering scheme for intrusion detection improved by information theory / Proceedings of the 7th IEEE Malaysia International Conference on Communications and IEEE International Conference in Networks, Kuala Lumpur, Malasia, 2005, pp.553–558;
7. http://wiki.squid-cache.org/SquidFaq/SquidLogs;
8. http://www.squid-cache.org/Intro/why.html**;**
9. http://www.bigml.com**.**

Национальная академия наук Азербайджана