УДК 004.048

Об одном методе сокращения размерности анализируемых признаков сетевых трафиков, используемых для мониторинга компьютерных сетей

Р.Г. ШЫХАЛИЕВ, канд. техн. наук

Институт информационных технологий НАНА, г. Баку
E-mail: ramiz@science.az

В статье предложен метод сокращения размерности признакового пространства сетевого трафика, который используется для сетевого мониторинга компьютерных сетей (КС). Этот метод основывается на применении алгоритмов построения ассоциативных правил. Предлагаемый подход позволяет систематизировать собранные мониторинговые данные, что существенно сокращает время, затрачиваемое администраторами КС на анализ сетевого трафика и принятие обоснованного решения по управлению КС. Кроме того, нахождение в мониторинговых данных ассоциативных правил позволяет обнаружить происходящие в КС аномалии (например, аномалии в поведении пользователей КС).

Ключевые слова: *компьютерные сети, сетевой трафик, сетевой мониторинг, мониторинговые данные, признаки сетевого трафика, ассоциативные правила, обобщенные ассоциативные правила.*

Возрастающие сложность и масштаб сегодняшних компьютерных сетей (КС), а также количества используемых в них сетевых сервисов и приложений, механизмов защиты, аппаратного и программного обеспечения приводят к появлению в сети большого объема трафика с информацией различного рода. Вместе с тем, из-за использования в КС таких приложений, как P2P, VoIP, IPTV и т.п., изменяются характер и объем сетевого трафика и расширяется диапазон неисправностей сетевых устройств, а также повышается вероятность их появления. В результате усложняются задачи диагностирования состояния и управления КС, а также обеспечения приемлемого уровня QoS (Quality of Service). При этом для администраторов КС важным становится возможность постоянного получения необходимой информации о состоянии сетевых устройств, сервисов, приложений и о действиях пользователей в сети. На основании этой информации администраторы КС могут принять оперативное решение по управлению сетью. Поэтому состояние сегодняшних КС должно систематически контролироваться.

На сегодняшний день для сбора необходимой информации о состоянии КС в основном используются системы сетевого мониторинга (ССМ). К задачам ССМ относятся: сбор значений характеристик каналов передачи и коммутирующего оборудования КС; выявление в КС аномальных ситуаций и узких мест; прогнозирование последствии изменений в топологии КС; мониторинг поведения пользователей КС и т.д. Однако проведение сетевого мониторинга в современных КС с помощью традиционных ССМ становится очень сложной задачей, так как для анализа результатов мониторинга от администраторов сетей требуются большой опыт и знания. В основном это связано с разнородно-

стью структуры и большим объемом сетевого трафика. Поэтому при сетевом мониторинге КС самыми важными задачами являются быстрый анализ большего объема сетевого трафика и выделение наиболее значимых данных для гибкого и оперативного реагирования на изменяющиеся в КС условия. Решение этой задачи может быть достигнуто путем сокращения размерности мониторинговых данных. Для этого предлагается метод сокращения признакового пространства сетевого трафика, используемый для сетевого мониторинга КС, т.е. выделения из множества признаков (параметров) сетевого трафика наиболее значимых. При этом сокращение размерности большого объема мониторинговых данных может быть достигнуто с использованием методов интеллектуального анализа данных (ИАД) и заключается в описании анализируемого исходного набора признаков сетевого трафика КС новым набором наиболее значимых признаков гораздо меньшего размера.

Использование методов ИАД для анализа мониторинговых данных объясняется их структурной разнородностью, сложностью получения аналитической информации из сетевого трафика значительно большего объема, а также большим числом одновременно работающих в КС пользователей, серверов, сетевого оборудования и приложений, необходимостью постоянного контроля функционирования КС и принятия обоснованных решений по управлению сетями.

Другим мотивом применения методов ИАД для сокращения размерности признакового пространства сетевого трафика, используемым для сетевого мониторинга КС, является сокращение временных и вычислительных затрат (например, оперативной памяти, дискового пространства, процессорного времени и т.д.), употребляемых для обработки и хранения мониторинговых данных без потери полезной информации.

Для решения задачи сокращения размерности признакового пространства сетевого трафика, используемого для сетевого мониторинга КС, предлагается метод, который основан на применении алгоритмов нахождения ассоциативных правил. Основная идея подхода заключается в применении алгоритмов поиска ассоциативных правил для выявления часто встречающейся устойчивой комбинации признаков сетевого трафика КС, используемых для мониторинга.

Постановка задачи

В основном сетевой трафик в КС состоит из трафиков клиентов, серверов и приложений [1]. Клиентские машины начинают генерацию трафика с того момента, как только они включаются, и не прекращают создавать трафик, пока не будут выключены физически. В свою очередь источниками клиентского трафика могут быть: трафик, связанный с протоколами, трафик, связанный с поиском различных объектов в сети и т.п. Кроме этого, имеющиеся в КС серверы и приложения различных видов порождают очень большой объем трафика.

Обычно сетевой трафик КС характеризуется множеством признаков, которые используются для сетевого мониторинга КС. В качестве таких признаков могут применяться DNS-запросы, DHCP-запросы, DHCP-ответы, WINS-трафики, числа пакетов, объем и скорость входящего и выходящего трафиков, IP-адрес отправителя и получателя, MAC-адрес хостов, виды используемых протоколов (например, HTTP, FTP, SMTP и т.д.) и приложений, время и т.д. Исходя из этого, размерность признаков, описывающих сетевой трафик КС, может достигать нескольких сотен. Обычно эти, а также другие признаки сетевого трафика КС накапливаются в log-файлах и/или базах данных.

Задача сокращения размерности мониторинговых данных состоит в следующем. Пусть $X = \{x_1, x_2,..., x_n\}$ — множество признаков, которое характеризует сетевой трафик КС. Пусть $T = \{t_1, t_2,..., t_m\}$ — сетевой трафик КС, который состоит из трафиков m субъектов КС (например, пользователей, серверов, приложений), где каждый t_i-ый трафик содержит набор признаков, входящих во множество X, т.е. $T \subseteq X$. Требуется найти ассоциативные правила для выявления часто встреча-

ющейся устойчивой комбинации признаков сетевого трафика КС, используемых для сетевого мониторинга, который позволит снизить размерность признаков, описывающих сетевой трафик КС.

Решение задачи

Поиск ассоциативных правил, на основе которых выявляются и представляются скрытые связи между объектами некоторой предметной области [3], является одним из эффективных методов ИАД. Они определяются следующим образом.

Пусть дано $I = \{i_1, i_{i_2}, ..., i_n\}$ — множество n элементов, которые формируют отдельную запись набора данных $T = \{t_1, t_{i_2}, ..., t_m\}$, состоящую из m записей. Ассоциативным правилом называется импликация вида $A \Rightarrow B$, где A, B — непересекающиеся наборы элементов множества I, $A \subset I$, $B \subset I$, $A \cap B = \varnothing$.

При этом каждое правило оценивается с помощью двух показателей: поддержки (support) и достоверности (confidence). Если $A \subseteq t$, то запись t содержит набор элементов A. Количество записей, которые содержат набор элементов A, называется поддержкой набора элементов A и обозначается sup A. В терминах теории вероятности поддержка sup A представляет собой вероятность наступления события $A \subseteq t$.

Поддержкой ассоциативного правила $A \Rightarrow B$ обозначается sup $A \Rightarrow B$ и называется вероятность наступления события $A \cup B \subseteq t$, $\sup A \Rightarrow B = \sup A \cup B$.

Достоверностью ассоциативного правила $A \Rightarrow B$ обозначается conf $A \Rightarrow B$ и называется условная вероятность наступления события $B \subseteq t$ при условии, что наступило событие $A \subseteq t$, $A \subseteq t$, $\mathrm{conf}\, A \Rightarrow B = \sup A \cup B / \sup A$.

Если задано пороговое значение поддержки и достоверности, соответственно min_sup и min_conf, то задача поиска ассоциативных правил в наборе данных T сводится к решению двух подзадач. Первая задача заключается в поиске всех наборов элементов $A \in I$, для которых sup $A \geq$ min_sup. При этом для заданного порогового значения поддержки min_sup набор элементов A называется часто встречающимся набором. Исходя из этого, первый этап поиска ассоциативных правил сводится к поиску часто встречающейся устойчивой комбинации элементов. Вторая задача заключается в том, что для каждого найденного на первом этапе набора элементов A генерируются ассоциативные правила вида $A' \Rightarrow A \setminus A'$, причем $A' \subset A$ и $\mathrm{conf}\, A' \Rightarrow A \setminus A' \geq$ min_conf.

Значения для параметров min_sup и min_conf выбираются таким образом, чтобы ограничить количество найденных правил. Если поддержка имеет большое значение, то алгоритмы будут находить настолько очевидные правила, что нет никакого смысла проводить такой анализ. А если поддержка имеет низкое значение, то алгоритмы будут находить огромное количество правил. Однако слишком низкое значение поддержки ведет к генерации статистически необоснованных правил. Поэтому на начальных этапах обработки данных в ряде случаев сложно задать значения параметров min_sup и min_conf.

В работе [4] вместо поиска всех часто встречающихся наборов элементов и соответствующих ассоциативных правил предлагается найти так называемые замкнутые, часто встречающиеся наборы элементов и на их основе построить множество ассоциативных правил. Часто встречающиеся наборы элементов A, называются замкнутыми, если не существует набора элементов A', такого, что $A' \supset A$ и $\sup A' = \sup A$.

Алгоритмы нахождения ассоциативных правил хорошо зарекомендовали себя при обработке больших объемов данных [2, 5], поэтому именно эти алгоритмы применяются при выявлении значимых признаков. При этом значимыми считаются те признаки, которые используются в построенных ассоциативных правилах.

Исходя из вышесказанного, для анализа собранных мониторинговых данных КС находим ассоциативные правила. На основании ассоциативных правил может быть определена часто встречаемая устойчивая комбинация признаков сетевого трафика, которая

определяется значением поддержки правила. Для этого выделяется сетевой трафик за требуемый временной период. При этом ассоциативные правила используются для выявления корреляций между признаками сетевого трафика КС.

Пусть $X = \{x_1, x_2, ..., x_n\}$ — множество признаков, которое характеризует сетевой трафик КС. Пусть $T = \{t_1, t_{i_2}, ..., t_m\}$ — сетевой трафик КС, который состоит из трафиков m субъектов КС (например, пользователей, серверов, приложений), где каждый t_i-ый трафик содержит набор признаков, входящих во множество X, т.е. $T \subseteq X$. Предполагается, что сетевой трафик T содержит множество признаков A, входящих во множество X, если $A \subseteq X$. Тогда ассоциативным правилом признаков сетевого трафика КС является импликация $A \Rightarrow B$, где $A \subset X$, $B \subset X$ и $A \cap B = \varnothing$. В сетевом трафике КС правило $A \Rightarrow B$ выполняется с достоверностью c, если $c\%$ трафиков субъектов КС, входящих в T, содержит множество признаков A, а также множество признаков B. При этом достоверность правила рассчитывается по формуле:

$$c(A \Rightarrow B) = \frac{s(A \cup B)}{s(A)}.$$

Правило $A \Rightarrow B$ имеет поддержку s, если $s\%$ трафиков субъектов КС, входящих в сетевой трафик T КС, содержит $A \cup B$, причем $s(A \Rightarrow B) = s(A \cup B)$.

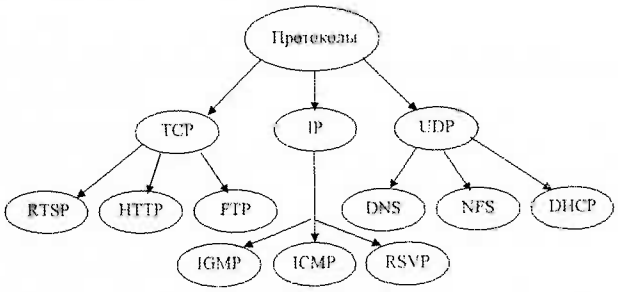
Сегодня для нахождения в больших объемах данных ассоциативных правил наиболее широко используется алгоритм APriori, основным достоинством которого является его гибкость [6]. При этом имеется возможность задавать и min_sup и min_conf правила, что позволяет получать множество разных групп правил. Однако генерация большого количества ассоциативных правил создает серьезную проблему для их анализа. Поэтому для нахождения ассоциаций в мониторинговых данных недостаточно использования только алгоритма APriori. Исходя из этого, предлагается обобщить «подобные» правила, т.е. определить обобщенные ассоциативные правила, в которых получаемые правила включают признаки, являющиеся предками признаков, входящих в трафики субъектов КС. В результате можно выявить ассоциативные правила не только между отдельными признаками сетевого трафика КС, но и между трафиками субъектов КС.

При нахождении обобщенных ассоциативных правил важным элементом является таксономия (иерархия) признаков сетевого трафика КС. Под таксономией здесь понимается лес направленных деревьев, листьями которых являются признаки сетевого трафика КС, а внутренними узлами — их группы. Пример иерархии некоторых протоков и групп протоколов, которые являются признаками сетевого трафика КС, приведен на рисунке. В получаемых на основании такой таксономии правилах, как в предпосылке, гак и впоследствии, могут присутствовать элементы, находящиеся на разных уровнях таксономии. Например, «если в трафике пользователя присутствует HTTP-протокол, го скорее всего будет присутствовать и DNS-протокол».

Введение дополнительной информации о группировке признаков сетевого трафика КС в виде иерархии может дать следующие преимущества:

1) могут быть выявлены ассоциативные правила не только между отдельными признаками сетевого трафика КС, но и между различными группами признаков;

2) в некоторых случаях отдельные признаки сетевого трафика КС могут иметь очень маленькую поддержку, однако значе-



Иерархия некоторых протоколов и групп протоколов

ние поддержки всей группы, в которую входит этот признак, может быть больше порога min_sup;

3) введение информации о группировке признаков сетевого трафика КС может использоваться для отсечения неинформативных правил.

Таким образом обобщенным ассоциативным правилом называется импликация $A \Rightarrow B$, где $A \subset X$, $B \subset X$ и $A \cap B = \varnothing$ и ни один из элементов, входящих в набор B, не является предком ни одного элемента, входящего в A. Поддержка и достоверность подсчитываются так же, как и в случае ассоциативных правил.

Для нахождения обобщенных ассоциативных правил целесообразно использовать специализированный алгоритм [7], который является более эффективным, чем стандартный алгоритм APriori.

Заключение

Предложен метод сокращения размерности признакового пространства сетевого трафика, который используется для сетевого мониторинга КС. Этот метод основывается на применении алгоритмов построения ассоциативных правил. При этом было предложено обобщить «подобные» правила, позволяющие значительно снизить объем мониторинговых данных, что дает возможность намного облегчить интерпретацию и визуализацию.

Предлагаемый подход позволяет систематизировать собранные мониторинговые данные, что существенно сокращает время, затрачиваемое администраторами КС на анализ сетевого трафика и принятие обоснованного решения по управлению КС. Кроме того, нахождение в мониторинговых данных ассоциативных правил позволяет обнаружить происходящие в КС аномалии (например, аномалии в сетевом трафике, поведении пользователей КС).

СПИСОК ЛИТЕРАТУРЫ

1. **Уилсон Э.** Мониторинг и анализ сетей. Методы выявления неисправностей. М.: Лори. 2002. 350 с.
2. **Savasere A., Omiecinski E., Navathe S.** An Efficient Algorithm for Mining Association Rules in Large Databases//Proc. of the 21st VLDB Conference, Zurich, Swizerland. 1995. P. 432—444.
3. **Han J., Kamber M.** Data mining: concepts and techniques//Second Edition, Morgan Kaufmann Publishers, San Francisco. 2006. P. 279—310.
4. **Pasquier N., Bastide Y., Taouil R., Lakhal L.**//Efficient Mining of Association Rules Using Closed Itemset Lattices, Information Systems. 1999. 24(1). P. 25—46.
5. **Agrawal R., Imielinski T., Swami A.N.** Mining Association Rules between Sets of Items In Large Databases//Proc. ACM SIGMOD Conf., 1993. P. 207—216.
6. **Agrawal R., Srikant R.** Fast Algorithms for Mining Association Rules in Large Databases//Proc. Conf. Very Large Databases, 1994. P, 487—499.
7. **Srikant R., Agrawal R.** Mining Generalized Association Rules//Proc. Conf. Very Large Databases, 1995. P. 407—419.

ООО «Наука и технологии»

Учредитель журнала ООО «Наука и технологии»
Журнал зарегистрирован в Комитете Российской Федерации по печати.
Свидетельство о регистрации № 018873 от 27 мая 1999 г.

Редактор *Морозова И. М.*

Оригинал-макет и электронная версия изготовлены в ООО «СиД»
Сдано в набор 12.04.2011. Подписано в печать 18.05.2011.
Формат 60 × 88 1/8. Усл.-печ. л. 5,88. Уч.-изд. л. 5,76. Печать цифровая. Тираж 120 экз.

Отпечатано в ООО «СиД»