УДК 004.82 + 004.89 + 004.91 + 004.93’1 + 004.62

МОДЕЛИ И АЛГОРИТМЫ МЕТОДА ИНТЕГРАЦИИ ИСТОЧНИКОВ В КОНТЕКСТЕ ПРОБЛЕМЫ АНАЛИЗА ТРАФИКОВ ТЕЛЕФОННЫХ СЕТЕЙ

Савельев О.О., Шевченко А.И.
Институт информатики и искусственного интеллекта
Донецкого национального технического университета, Украина

Рассматривается задача извлечения и трансформации данных из трафиков телефонных сетей. Результаты анализа трафика как документа позволяют использовать существующие подходы обработки документов. Проведен синтез моделей представления и содержания, а также алгоритма преобразования документа трафика.

Введение

Трафики телефонных сетей – статистическая информация о действиях абонентов мобильной и стационарной телефонии, накопленная операторами связи. Трафики находят активное применение в работе оперативно-технических подразделений правоохранительных структур в ходе осуществления следственных мероприятий. Для повышения эффективности решения таких задач в работе [2] была предложена концепция архитектуры СППР при анализе трафиков. В качестве основы подобной СППР в работе [3] было предложено использовать реляционное хранилище. К сожалению, не существует единого утвержденного формата предоставления трафиков операторами связи, который можно было бы использовать для загрузки данных из трафиков в хранилище.

Целью синтеза моделей и алгоритмов метода интеграции источников является увеличение степени автоматизации в процессах извлечения, трансформации и ввода полезной информации из трафиков. Необходимо решить следующие ***задачи***: анализ трафика как документа и возможности применения существующих подходов обработки документов, синтез моделей представления и содержания документа трафика, синтез алгоритмов анализа и преобразования документа трафика.

Актуальность поставленных задач состоит в том, что на настоящий момент не существует готовых инструментальных средств анализа документов различной структуры, которые можно использовать для трансформации трафиков в единый формат. По данной причине аналитику приходится вручную обрабатывать каждый трафик. Предлагаемые модели и алгоритмы могут стать основой нового инструментального средства, позволяющего значительно сократить степень ручного труда в ETL процессах СППР при анализе трафиков.

1 Анализ трафика как документа и возможности применения существующих подходов обработки документов

Операторы связи используют промышленные СУБД, высокоуровневая интеграция на уровне ядра с которыми по ряду технических и правовых причин не возможна. Следовательно, оправдано применение низкоуровневого обмена информацией – в виде документов, где отчеты одних систем являются исходными данными других.

Как показано в классификации [3] трафики могут: быть получены от разных операторов, содержать данные разных типов мониторинга в различных форматах, и находится на различных носителях. Любой документ трафика может быть приведен к некоторой форме в виде простого текстового документа. Данная предобработка возможна в контексте трафиков, так как эти документы не содержат графических изображений, что значительно упрощает формализацию трафика как документа и его дальнейшую обработку. Содержимое документа трафика может быть представлено в виде таблицы, таблицы с блоками текста, блоков текста, просто текста. Поэтому необходимо рассматривать все множество методов обработки текстовых документов.

Классификацию существующих подходов в области анализа документов можно провести по следующему ряду критериев: характер применяемых методов, тип используемых моделей документа, принцип алгоритмов обработки [1].

По характеру применяемых методов подходы можно разделить на три класса – синтаксические методы, использующие аппарат формальных грамматик, методы, основанные на механизме правил, и гибридные методы.

Первый класс методов рассматривает документ как последовательность компонент грамматики. Второй класс рассматривает документ как представление некоторой модели, компоненты которой отличаются наборами правил. Используется два типа моделей: для физической и логической структуры документа. Гибридные методы представляют собой объединение синтаксических и методов, основанных на правилах.

По принципу алгоритмов обработки, подходы можно классифицировать на использующие стратегию “сверху-вниз”, стратегию “снизу-вверх” и гибридные.

2 Синтез моделей представления и содержания документа трафика

Принимая во внимание преимущества и недостатки рассмотренных подходов обработки документов, а также особенности документа трафика было решено акцентировать внимание на подходе, обладающем следующими характеристиками: метод – использующий механизм правил, типы моделей – иерархические древоподобные модели физической и логической структуры документа, алгоритм – использующий стратегию “снизу-вверх”.

Представим документ трафика как текстовый документ, состоящий из строк и символов, в следующей форме:

$$D = \{ s_0, s_1, ..., s_i, ..., s_{n^D} \}, \quad (1)$$

где s_i – i-я строка документа; n^D – количество строк в документе. А строку будем рассматривать как

$$s_i = \{c_0, c_1, ..., c_j, ..., c_{n^{S_i}} \}, \quad (2)$$

где c_j – j-й символ строки; n^{S_i} – количество символов в конкретной строке.

Тогда документ (1) можно рассматривать как множество символов, или как модель представления документа (3).

$$D = \left\{ c_{i,j} \mid \begin{cases} i \in [0, n^d] \\ j \in [0, n^{s_i}] \end{cases} \right\} = \langle M_V \rangle. \quad (3)$$

Введем понятие слова – некоторой последовательности символов (4).

$$w = c_{i,j}, ... c_{i+p,j+q}, \quad i \in [0, n^d), j \in [0, n^{S_i}), p \in [i, n^d), q \in (j, n^{s_p}). \quad (4)$$

Предположим, что слова, могут быть объединены во множество по некоторому критерию, тогда пусть некоторая последовательность слов в исходном документе составляет строку (5).

$$s' = w_i, .. w_j, \quad i \in [0, j), j \in \left(i, \sum_{k=0}^{n^D} n^{s_i} \right). \quad (5)$$

А последовательность строк, может быть объединена во множество, которое назовем блоком (6).

$$b = s'_i, ..., s'_j. \quad (6)$$

Пусть слово, строка и блок обобщенно могут рассматриваться как элемент документа (7).

$$e = \langle t, p, E \rangle, \quad (7)$$

где t – тип сущности элемента (слово, строка, блок); p – родительский элемент в иерархии; E – множество дочерних элементов.

Тогда все множество элементов, выстроенное в дерево, будет составлять документ, а само дерево элементов будет отражать модель физической структуры (содержания) документа (8).

$$D = \langle E, e = \langle t = e, p = null, E = E' \rangle, E' \subseteq E \rangle = \langle M_D \rangle. \quad (8)$$

3 Синтез алгоритмов анализа и преобразования документа трафика

Рассмотрим алгоритм преобразования документа из модели представления в модель физической структуры. Данный алгоритм призван решать задачу выделения физических элементов документа из его представления и построения их иерархии. Для формализации процесса преобразования необходимо ввести понятия правил, в соответствии с которыми происходит формирование элементов физической модели.

Пусть r – регулярное выражение, которое может быть использовано для поиска слов в исходном представлении документа. Тогда объединив несколько регулярных выражений, предназначенных для поиска слов, соответствующих всем вариантам представления некоторого текстового шаблона, получим множество, которое назовем текстуальным (текстовым) правилом (9).

$$t = \langle X_t \rangle, \qquad X_t = \{ r_0^t, ... r^t_{n^{X_t}} \} . \tag{9}$$

Аналогично введем понятие относительного правила (определяющего связи между элементами) (10).

$$h = \langle t, H \rangle , \tag{10}$$

где t – тип правила: относительное либо текстуальное; $H = \{h_0, ..., h_{n^H}\}$ – множество дочерних правил.

Исходными данными для алгоритма являются:

M_V – модель представления исходного документа;

$X = \{r_0, ... r_{n^X}\}$ – универсальное множество регулярных выражений в контексте предметной области;

$T = \{t_0, ... t_{n^T}\}$ – множество текстуальных правил;

$H = \{h_0, ..., h_{n^H}\}$ – множество относительных правил.

Выходными данными алгоритма являются узлы дерева физической структуры, заполненные контентом и организованные в дерево.

Блок схема алгоритма представлена на рисунке 1.

На первом шаге осуществляется анализ модели представления всеми регулярными выражениями, результат запоминается в контейнере (11).

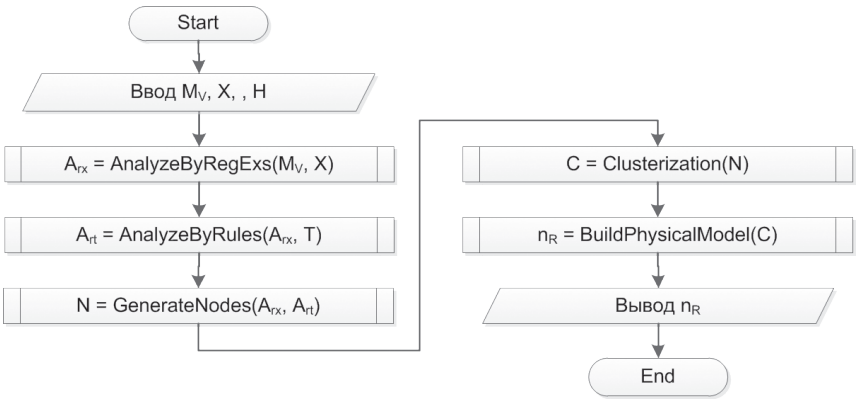

Рисунок 1. Блок-схема алгоритма преобразования документа из модели представления в модель физической структуры

$$A^{rx} = \langle W^{rx}, X^{rx}, R^{rx} \rangle , \tag{11}$$

где $W^{rx} = \left\{ w_0, ..., w_{n^{W^{rx}}} \right\}$ – множество всех обнаруженных слов;

X^{rx} – множество регулярных выражений, по которым были обнаружены слова; R^{rx} – матрица отношений между словами и регулярными выражениями.

На втором шаге производится анализ полученных слов при помощи текстуальных правил – используя регулярные выражения в качестве критерия, производится классификация, результат

представляется в виде (12):

$$A^{rt} = \left\langle E^{rt}, T^{rt}, R^{rt} \right\rangle \tag{12}$$

где E^{rt} – множество элементов, построенных на основе слов; T^{rt} – множество текстуальных правил, каждому из которых соответствует минимум одно слово; R^{rt} – матрица отношений между элементами и правилами.

На третьем шаге происходит генерация множества N начальных узлов (13) – листьев дерева физической модели содержания.

$$N = \{n_0, ..., n_{n^N}\}, \qquad n^N = \left| E^{RT} \right|, \tag{13}$$

где узел определяется как (14).

$$n = \langle e, H, D \rangle \tag{14}$$

где e – элемент, на основе которого построен узел; H – множество правил, соответствующих элементу; D – расстояния до других узлов (15).

$$D = \{ d_0, ..., d_{n^N} \} = \{..., \langle d_p, d_x, d_r \rangle_i, ...\} \tag{15}$$

где d_p – расстояние по критерию близости позиций элементов (слов) исходном представлении; d_x – расстояние по критерию близости регулярных выражений; d_r – расстояние по критерию близости правил.

На четвертом шаге производится многоэтапная кластеризация. В первую очередь производится иерархическая кластеризация узлов, полученных на третьем шаге: сначала приоритет отдается d_p, что позволяет, последовательно устанавливая различные граничные величины объединения узлов в кластеры, выделить кластеры соответствующие строкам, а затем блокам. Кластеризация по d_x и d_r позволяет решать конфликты при исправлении ошибок, опечаток, шума. Все полученные кластеры заносятся во множество кластеров C.

На пятом шаге производится построение дерева модели физической структуры (содержания) документа M_D, используя элементы, хранящиеся в узлах кластеров в соответствии с относительными правилами.

Выводы

В ходе исследования был проведен анализ трафика как документа, осуществлен синтез моделей представления и физической структуры документа трафика и алгоритма преобразования документа трафика из модели представления в модель физической структуры. Следует отметить, что разработанные модели и алгоритмы могут применяться ко всему множеству текстовых документов.

На основе полученных результатов можно сформулировать дальнейшие задачи исследования: синтез модели логической структуры документа трафика, синтез алгоритма преобразования документа трафика из модели физической структуры в модель логической структуры.

Литература

[1] Mao S. Document Structure Analysis Algorithms: a Literature Survey / Mao S., Rosenfeld A., Kanungo T. // Proc. SPIE Electronic Imaging. – 2003. P. 197-207.

[2] Савельев О.О. О концепции создания информационной системы интеллектуального анализа данных телекоммуникационных компаний в рамках разработки интеллектуальной системы поддержки принятия решений / О.О. Савельев // Искусственный интеллект. – 2010. – № 3

[3] Савельев О.О. Особенности разработки подсистемы хранения информации для системы поддержки принятия решений в области анализа телекоммуникационных данных / О.О. Савельев, А.И. Шевченко // Материалы VI международной научно-технической конференции студентов, аспирантов и молодых ученых «Информатика и компьютерные технологии». – Донецк: ДонНТУ. – 2010. С. 343-349.