2. Семакин И., Залогова Л., Русаков С., Шестакова Л. Информатика: уч. по базовому курсу.
3. Угринович Н. Информатика и информационные технологии. Учебное пособие для общеобразовательных учреждений.
4. Шафрин Ю.А. Информационные технологии
5. Иванова Г.С. Технология программирования: Учебник для вузов.
http://ru.wikipedia org/wiki.

УДК 004.89

ИСПОЛЬЗОВАНИЕ ИНТЕЛЛЕКТУАЛЬНЫХ МЕТОДОВ ИДЕНТИФИКАЦИИ ИНФОРМАЦИИ В ИНТЕГРИРОВАННЫХ ИНФОРМАЦИОННО-КРИМИНАЛИСТИЧЕСКИХ СИСТЕМАХ

Хайрова Н. Ф., Узлов Д. Ю., Борисова Н. В.

Национальный технический университет «Харьковский политехнический институт», Харьков,
nina_khajrova@yahoo.com, poputcik@mail.ru, n_borisova2004@yahoo.com

Сегодня в информационных хранилищах собраны терабайты неструктурированной текстовой информации. И хотя, в процессе своей деятельности правоохранительные органы используют, в основном, внутренний ресурс – информацию, которая вырабатывается в процессе оперативно-служебной деятельности, открытые источники служат внешним дополнительным ресурсом. Для внутреннего ресурса характерно наличие больших массивов данных, представленных в файлах различных форматов (неструктурированные и слабоструктурированные текстовые данные, вырабатываемые в процессе административной, оперативно - розыскной, следственной, аналитической и иной деятельности). Внешний ресурс представлен массивами файлов, аналогичных внутренним, а также веб-сайтами, социальными сетями, электронной почтой и RSS-рассылками.

Чтобы получать криминально значимую информацию из подобных неструктурированных текстовых массивов и проводить ее анализ, необходимо иметь специальный инструментарий, в основе которого должна находиться определенная технология. Целью такого инструментария является поиск неструктурированных источников, содержащих криминально значимую информацию по заранее определенным формальным признакам и составление методов и моделей извлечения релевантной информации с учетом особенностей предметной области.

Потребность правоохранительных органов, в частности ОВД, в информации весьма разнообразна и, определяясь тактической и стратегической необходимостью решаемой задачи, очень часто не является четко криминально выраженной [1]. То есть, ее криминальная окрашенность и предметная направленность в некотором слабоструктурированном массиве текстовой информации будет определяться динамически в процессе проводимого аналитического поиска.

Актуальная криминально значимая информация, зачастую не имеющая причинно-следственных связей с событием преступления, но имеющая потенциальное криминалистическое значение, не позволяет при ее поиске использовать предварительно разработанный тезаурус заранее известной предметной области, а также использовать для ее идентификации только ключевые слова, которые описывают преступные деяния и часто, являясь своего рода индикативным признаком, имеют свою специфику. Поэтому
для решения задачи обеспечения работника правоохранительных органов полной и релевантной информацией необходимо разработать модели, методы и алгоритмы, осуществляющие моделирование процессов интеллектуальной обработки слабоструктурированных текстовых информационных элементов, реализующих функции понимания и систематизации.

В качестве математического аппарата задачи моделирования интеллектуальной деятельности по пониманию, идентификации и систематизации криминально значимой информации в слабоструктурированных и неструктурированных текстовых массивах, используемого для описания дискретных, детерминированных и конечных объектов интегрированной информационно криминалистической системы используем алгебру конечных предикатов (АКП) и предикатных операций [2].

Вводим универсум элементов U, включающий все возможные документы, поступающие аналитику-криминалисту на обработку (справки, сводки, выписки, отчеты, описания портретов, протоколы, газетные и Интернет-публикации, электронные ресурсы и т.д.), а также понятия и объекты анализа рассматриваемой предметной области (ключевые слова, метаданные, авторов и УДК документов), элементы специализированных словарей и тезаурусов.

Из элементов универсума, в соответствии с конкретной задачей обработки информации, образуем подмножества M_{1i} , M_{2i} , …, M_m, на декартовых произведениях которых $M_{1i}\times M_{2i}\times \ldots\times M_m$ определяются предикаты P_i, характеризующие работу модели.

Базисным для алгебры конечных предикатов является предикат узнавания предмета *a* по переменной *x*, равный единице в том случае, если *x* равен *a*, и нулю в противном случае [2]:

$$x_i^a = \begin{cases} 1, \text{если } x_i = a \\ 0, \text{если } x_i \neq a \end{cases}, \ (1 \le i \le n),$$

где *a* — это любой элемент универсума.

В предлагаемой модели динамической идентификации актуальной криминально значимой информации в слабоструктурированных текстовых массивах вводятся предметные переменные, определяющие отношение исследуемых текстов к существующим криминалистическим учетам: ключевые слова— *l*, значения УДК — *u*, определяющие тему документа, и источник криминально значимой текстовой информации (автор сведений, зафиксированных в документе, электронный адрес корреспондента, доменный или IP адрес, а также адрес RSS-фида) — *a*. Данные предметные переменные отражают суть документа, назначение и взаимосвязь его составляющих, то есть объективно представляют извлекаемую из документа актуальную информацию.

В модели используется также базовое для наших рассмотрений, понятие криминалистического учета: *b*, под которым понимается область знаний, образовавшаяся в сфере мышления эксперта при углубленном анализе значимых сведений о субъектах и объектах преступлений и связанных с ними событий. Область знаний формируется в сфере мышления и имеет внеязыковую природу. Но поскольку мысль не может существовать вне слова, под криминалистическим учетом мы подразумеваем фразу или словосочетание, называющее или определяющие объект, место, время преступления, предмет посягательства, признаки способа совершения преступлений и т.д., сформированные в виде определенно версии или предположения.

В рассмотренном примере из 12 документов, поступивших следователю ОВД на обработку, область значений предметных переменных: $L=\{l^i\}$, $1\le i\le 14$, $U=\{u^i\}$, $1\le i\le 5$, $A=\{a^i\}$, $1\le i\le 4$ и $B=\{b^i\}$, $1\le i\le 16$. Можно построить парадигматическую таблицу, отображающую связь между криминалистическими учетами b^i и предметными переменными *l*, *u* и *a*(табл. 1).

Таблица 1. Фрагмент парадигматической таблицы связей предметных переменных

источник информации	a^1 = автор сведений-1	a^4	a^1	a^1	a^1	a^2	a^1	a^1	a^1	a^2	…	a^4
значение УДК	u^1=343.71 – Присвоение имущества	u^2	u^1	u^1	u^3	u^4	u^3	u^3	u^3	u^4	…	u^2
ключевые слова	l^{12}= сбыт	l^{12}	l^{13}	l^{14}	l^1	l^1	l^2	l^3	l^4	l^4	…	l^{11}
криминалис-тический учет	b^1=xxx	b^2	b^3	b^4	b^5	b^6	b^7	b^8	b^9	b^{10}	…	b^{16}

Используя данную таблицу можно выразить отношения между предметными переменными, объективно характеризующими информацию, содержащуюся в текстовых документах, поступающих на обработку аналитиком, и имеющимися криминалистическими учетами.

Затем выполняя операцию почленной дизъюнкции возможно большего числа родственных равенств [3], формируем функцию перехода от криминалистических учетов к областям текущих дела, которыми занимается следователь или иное процессуальное лицо, **s**. Введение почленной дизъюнкции с использованием родственных равенств обусловлено необходимостью получения области знаний текущих дел, находящихся в производстве, того или иного следственного или аналитического отдела. Такие области могут включать больше чем одно исчисляемое ограниченное количество криминалистических учетов.

В результате не сложных преобразований [4] получен предикат локализации области знаний текущих дел, находящихся в производстве следователя или иного процессуального лица, который описывает связь областей знаний конкретных дел и предметных переменных, объективно определяющих извлекаемую из документов актуальную информацию:

$$P(a, l, u, s)= s^1a^1u^1\,(l^{12}\vee l^{13}\vee l^{14}) \vee s^1a^2u^5\,(l^7\vee l^8) \vee\ s^2a^4u^2$$
$$(l^{12}\vee l^9\vee l^{10}\vee\vee l^{11}) \vee s^3\,a^1\,u^3(l^1\ \ l^2\vee l^3\vee l^4) \vee s^4a^2u^4(l^1\vee l^4\vee l^5).$$

Данный предикат можно наглядно изобразить в виде логической сети (рис.1), которая является графическим представлением результата бинарной декомпозиции многоместного предиката. Каждому полюсу логической сети ставится в соответствие своя предметная переменная модели. С каждым полюсом связывается область изменения атрибута этого полюса. Любой полюс логической сети в определенный момент времени несет некое знание о значении своего атрибута.

Каждой ветви логической сети ставится в соответствие свое бинарное отношение модели, которое называется отношением этой ветви. Каждая ветвь соединяет два полюса,

отвечающие тем предметным переменным, которые связываются отношением, соответствующим данной ветви (см. 1-3 на рисунке 1)

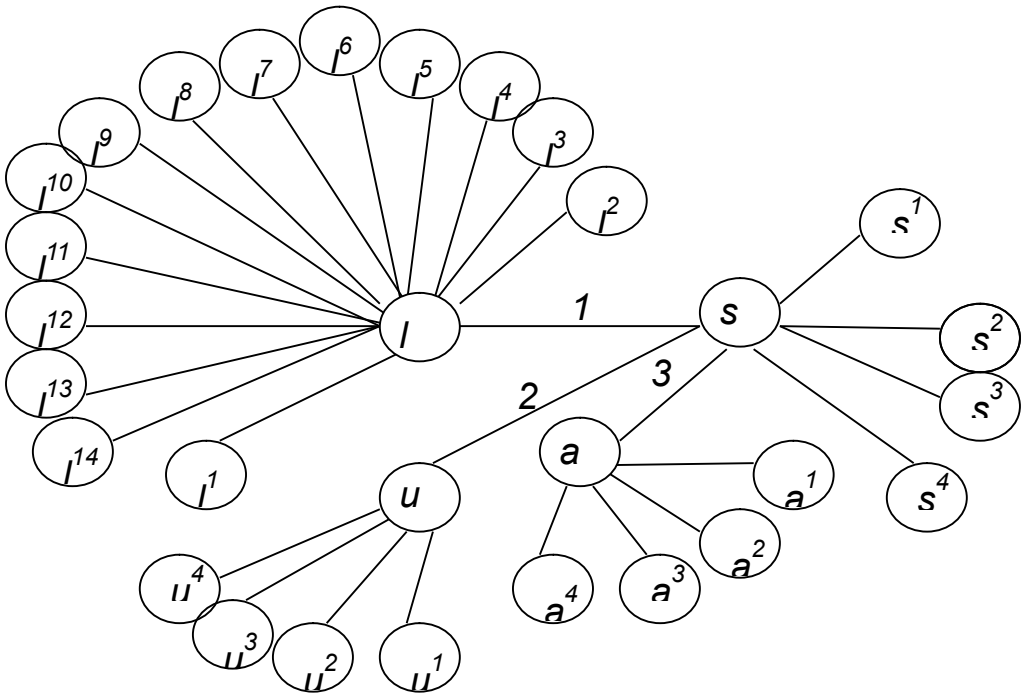

Рис. 1. Логическая сеть локальной области знаний текущих дел.

Логическую сеть можно содержательно понимать как графическое представление условного устройства, предназначенного для решения уравнений алгебры предикатов, предварительно преобразованных в бинарную систему. Логическая сеть задается парой ⟨X, R⟩, где X — конечное непустое множество унарных уравнений АКП, описывающих вершины логической сети, через предметные переменные модели и области их определения, а R — представляет собой конечное непустое множество бинарных уравнений АКП, описывающих ветви сети, через отношения между предметными переменными модели.

Логическая сеть работает в итеративном режиме, каждая итерация представляет такт. Исходные данные поступают в соответствующие полюса сети, результат решения задачи также содержится в полюсах сети после остановки ее работы, которая происходит, когда на очередном такте состояние сети повторяется [5]. На момент тактовой остановки сети можно определить значения неизвестного атрибута некоторой вершины или область значений данного атрибута.

Таким образом, предложенный подход к моделированию интеллектуальной деятельности по пониманию и систематизации поступающих на обработку в криминалистическую информационную систему текстовых массивов, позволяет осуществлять систематизацию потоков полнотекстовой электронной информации по областям текущих дела, которым занимается следователь или иное процессуальное лицо. Дополнительное использование в модели логической сети позволяет осуществить также аппаратную реализацию модели.

Список литературы

1. Christopher Westphal. Data Mining for Intelligence, Fraud, & Criminal Detection. Advanced Analytic & Information Sharing Technologies. 2009. CRCPress. – 426 p.
2. Бондаренко М. Ф. Теория интеллекта: учебник/ Бондаренко М. Ф., Шабанов-Кушнаренко Ю. П. Харьков: Комп. СМИТ, 2007. — 576 с.
3. Dieter Jungnickel. Graph, Networks and Algorithms. Algorithms and Computation in mathematics. Volume5. – Springer BerlinHeidelbergNew York, 2008. — 650 p.
4. Хайрова Н. Модель извлечения знаний из неструктурированных документов корпоративной информационной системы./ Н. Хайрова, Н. Шаронова. // Applicable Information Models. ITHEA. — Varna, Bulgaria, 2011. — С. 131—139.
5. Шабанов-Кушнаренко С. Ю. Решение булевых уравнений с помощью логических сетей / С. Ю. Шабанов-Кушнаренко, Л. Г. Ситник, Д. В. Биленко, К. В. Силивейстров //АСУ и приборы автоматики. — Харків: ХНУРЕ, 2008. — № 142. — С. 23-28.