ПОСТРОЕНИЕ МОДЕЛИ ИДЕНТИФИКАЦИИ И СИСТЕМАТИЗАЦИИ КРИМИНАЛЬНО ЗНАЧИМОЙ ИНФОРМАЦИИ В ТЕКСТОВЫХ РЕПОЗИТОРИЯХ

УДК 004.89

УЗЛОВ Дмитрий Юрьевич

соискатель кафедры интеллектуальных компьютерных систем Национального технического университета «Харьковский политехнический институт».

Научные интересы: интеллектуальные системы, идентификация знаний, криминально значимая информация, автоматизированная обработка текстовой информации, искусственный интеллект.

ХАЙРОВА Нина Феликсовна

доктор технических наук, доцент, профессор кафедры интеллектуальных компьютерных систем Национального технического университета «Харьковский политехнический институт».

Научные интересы: лингвистические технологии идентификации знаний в слабоструктурированной текстовой информации, Text Mining, Opinion Mining, Web Mining, Natural language processing, когнитивная лингвистика, функции интеллекта, искусственный интеллект.

ВВЕДЕНИЕ

При прогнозировании преступлений, выявлении признаков скрытых преступлений, установлении зависимости между личными качествами преступников и выбором места совершения преступления, а так же другой аналитической следственно-розыскной деятельности, следователю (или иному процессуальному лицу) необходимо проработать огромное количество электронных текстовых документов, вычленив из них криминально значимую информацию. Этими электронными текстами могут быть как документы, имеющие электронную форму: объяснительные/служебные записки, отчеты, словесные портреты фигурантов, протоколы и т. п., накопившиеся в результате расследования или расследований, так и электронные коллекции интернет публикаций, RSS - рассылок и социальных сетей.

Все подобные электронные документы представлены в виде слабоструктурированной текстовой информации, под которой понимается текстовый электронный документ, имеющий высокую степень вариативности контента, меняющегося в зависимости от конкретной ситуации. В целом, эти документы представляют доступный репозиторий криминалистических знаний. В таком репозитории качество криминально значимой информации определяется содержимым, способствующим поиску доказательств и закономерностей, присущих именно криминалистическим аспектам преступной деятельности. Другими словами, криминалистическая характеристика преступления, как средства оптимизации расследования, должна представлять собой совокупность информации, имеющей не квалифицирующее или процедурное и предупредительное, а именно поисково-познавательное значение [1].

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.

ОБЩАЯ ПОСТАНОВКА ЗАДАЧИ

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 [2]. То есть, ее криминальная окрашенность, и предметная направленность в некотором репозитории слабоструктурированной текстовой информации будет определяться динамически в процессе проводимого аналитического поиска.

В общем случае информационные процессы, связанные с расследованием состава преступления и получением криминально значимой информации, а так же криминально значимых данных и фактов из массивов электронных текстовых документов и электронных ресурсов, представляются следующей схемой (рис. 1):

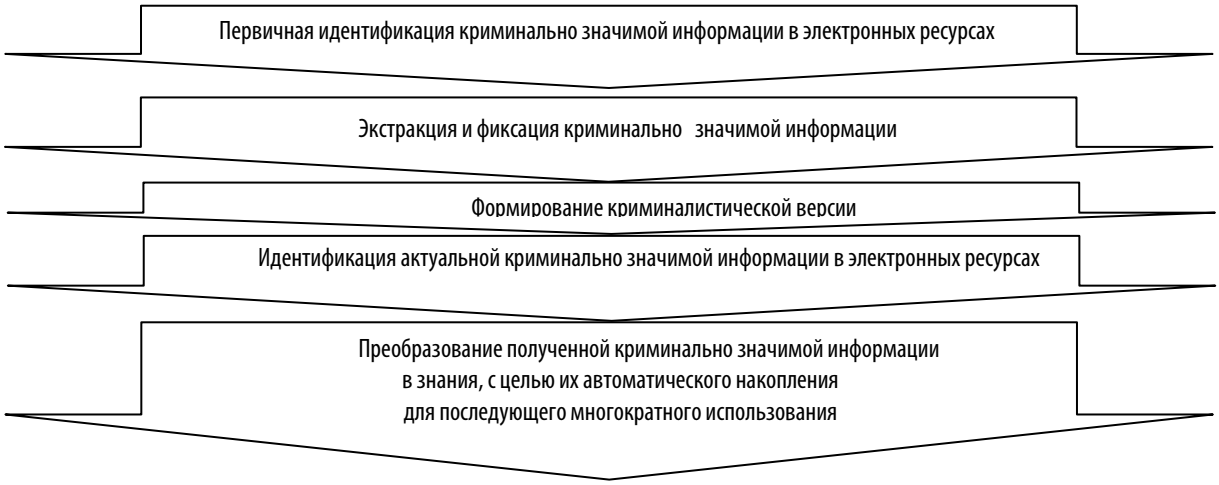

Рисунок 1 – Общая схема информационных процессов аналитической следственно-розыскной работы.

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.

ЦЕЛЬ ИССЛЕДОВАНИЯ

В исследовании для идентификации актуальной криминально значимой информации в репозитория текстовых документов неограниченных динамически меняющихся предметных областей предлагается использовать методы теории интеллекта, позволяющие моделировать интеллектуальное понимание и идентификацию смысла [3]. В работе используются наработки и подходы компьютерной лингвистики, искусственного интеллекта, когнитивной семантики и инженерии знаний, связанные с «пониманием» компьютером естественных языков, базирующиеся на семантических моделях представления знаний, и использующие символьные и логические сети.

ОПИСАНИЕ МАТЕМАТИЧЕСКОЙ МОДЕЛИ

В качестве математического аппарата задачи моделирования интеллектуальной деятельности по пониманию, идентификации и систематизации криминально значимой информации в слабоструктурированных и неструктурированных текстовых репозиториях, используемого для описания дискретных, детерминированных и конечных объектов интегрированной информационно криминалистической системы используем алгебру конечных предикатов (АКП) и предикатных операций [4].

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.

В предлагаемой модели динамической идентификации актуальной криминально значимой информации в слабоструктурированных текстовых репозиториях 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 — *a*. Данные предметные переменные отражают суть документа, назначение и взаимосвязь его составляющих, то есть объективно представляют извлекаемую из документа актуальную информацию.

Значения соответствующие предметным переменным представлены множествами *L*, *U* и *A*. В рассмотренном примере из 12 документов, поступивших следователю ОВД на обработку, множество ключевых слов и словосочетаний, определенное статистико-позиционными методами на этапах предлингвистического и лингвистического анализов, $L=\{l^i\}$, $1 \le i \le 14$: l^1 = депозитные операции; l^2= похитить; l^3= кража; l^4= частная собственность; l^5= имущество; l^6= сговор лиц; l^7= ущерб; l^8= обман; l^9= растрата; l^{10}= присвоение; l^{11}= уничтожение; l^{12}= сбыт; l^{13}= электрические сети; l^{14}= приборы учета.

Иерархическая классификация UDK представлена множеством значений, прямо или косвенно связанных с реальными или потенциальными противоправными действиями. В рассматриваемом примере это: $U=\{u^i\}$, $1 \le i \le 5$, где u^1=343.71 – Кража. Ограбление. Присвоение имущества. Пиратство. Вымогательство. Шантаж; u^2= 347.73 – Коммерческое право. Финансовое право; u^3= 343.72 – Мошенничество. Обман, u^4 =343.77 – Злостное повреждение имущества; u^5= 343.63 – Преступления против чести. Клевета.

Источник криминально значимой информации может представлять собой автора документа, автора сведений, зарегистрированных в документе, документ из определенного зарегистрированного дела, электронный адрес корреспондента электронного письма, доменный или IP адрес сайта, получения информации, адрес RSS-фида и т д. В нашем примере множество источников информации $A=\{a^i\}$, $1 \le i \le 4$, где a^1= автор сведений-1; a^2=автор сведений-2; a^3=электронное письмо адресата-1; r^4= документ из архивного уголовного дела № 000 000.

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. Введем достаточно четко очерченное множество криминалистических учетов $B=\{b^i\}$, $1 \le i \le 16$.

Можно построить парадигматическую таблицу, отображающую связь между криминалистическими учетами b^i и предметными переменными *l*, *u* и *a* (табл. 1).

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

$$\begin{gathered} a^1u^1l^{12}= b^1 ; a^4u^2l^{12}= b^2; a^1u^1l^{13}= b^3 ; a^1u^1l^{14=} b^4 ; a^1 u^3 l^1= b^5; \\ a^2u^4l^1= b^6; a^1u^3l^2= b^7; a^1u^3l^3= b^8; a^1u^3l^4= b^9; a^2u^4l^4= b^{10}; a^2u^4l^5= b^{11}; \\ a^4u^2l^{11}= b^{12};a^2u^5l^7= b^{13}; a^2u^5l^8= b^{14}; a^4u^2l^{10}= b^{15}; a^4u^2l^9= b^{16}. \end{gathered} \quad (1)$$

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. Родственными равенствами называются равенства, которые после выполнения над ними операции почленной дизъюнкции приводят к равенствам с левой частью в виде логического произведения, каждый сомножитель которого зависит от одной предметной переменной [4]. 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.

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}

$$a^1u^1(l^{12}\vee l^{13}\vee l^{14}) = b^1\vee b^3\vee b^4;\ a^4u^2(l^{12}\vee l^9\vee l^{10}\vee l^{11}) = b^2\vee b^{16}\vee b^{15}\vee b^{12};$$
$$a^2u^5(l^7\vee l^8) = b^{13}\vee b^{14};\ a^1u^3(l^1\vee l^2\vee l^3\vee l^4) = b^5\vee b^7\vee b^8\vee b^9;\ a^2u^4(l^1\vee l^4\vee l^5) = b^6\vee b^{10}\vee b^{11}. \quad (2)$$

Осуществляя переход от криминалистических учетов *b* к конкретным делам, находящимся в производстве *s*, получаем:

$$b^1\vee b^3\vee b^4\vee b^{13}\vee b^{14} = s^1;\ b^2\vee b^{16}\vee b^{15}\vee b^{12} = s^2;$$
$$b^5\vee b^7\vee b^8\vee b^9 = s^3,\ b^6\vee b^{10}\vee b^{11} = s^4. \quad (3)$$

Переопределяя описание дел, находящихся в производстве, предметные переменные характеризующие информацию, извлекаемые из документов репозитория, получим следующие зависимости:

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

Таким образом, получен предикат локализации области знаний текущих дел, находящихся в производстве следователя или иного процессуального лица P(a, l, u, s), который описывает связь областей знаний конкретных дел и предметных переменных, объективно определяющих извлекаемую из документов актуальную информацию:

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

ПОСТРОЕНИЕ ЛОГИЧЕСКОЙ СЕТИ МОДЕЛИ

Логическая сеть определяется как устройство, предназначенное для решения уравнения алгебры конечных предикатов [5] Для построения логической сети модели идентификации актуальных криминалистических знаний необходимо преобразовать предикат *P(a, l, u, s)* (5) в бинарную систему:

$$P_a(a,s) = s^1(a^1\vee a^2)\vee s^2a^4\vee s^3a^1\vee s^4a^2;$$
$$P_l(l,s) = (l^{12}\vee l^{13}\vee l^{14}\vee l^7\vee l^8)s^1\vee(l^{12}\vee l^9\vee l^{10}\vee l^{11})s^2\vee (l^1\vee l^2\vee l^3\vee l^4\vee l^8)s^3\vee(l^1\vee l^4\vee l^5\vee l^6)s^4;$$
$$P_u(u,s) = s^1(u^1\vee u^5)\vee s^2u^2\vee s^3u^3\vee s^4u^4. \quad (6)$$

Где P_a представляет собой бинарный предикат, определяющий отношения переменной s и предметной переменной a, бинарный предикат P_l определяет отношения переменной s и предметной переменной l, а бинарный предикат P_u определяет отношения пере-

менной s и предметной переменной u. Данный предикат можно наглядно изобразить в виде логической сети (рис. 1), которая является графическим представлением результата бинарной декомпозиции многоместного предиката.

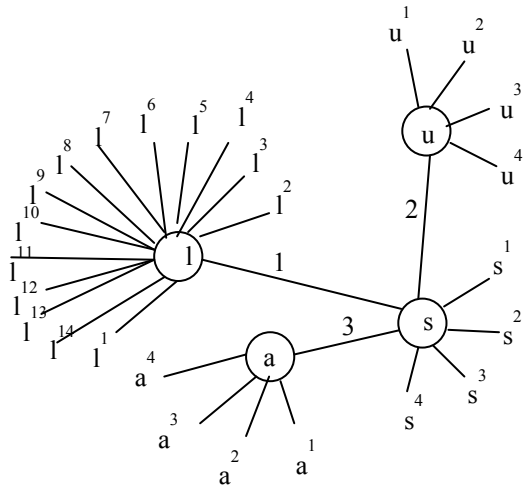

Рисунок 2 – Логическая сеть локальной области знаний текущих криминальных дел.

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.

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.

ВЫВОД

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 дел, которым занимается следователь или иное процессуальное лицо. Дополнительное использование в модели логической сети позволяет осуществить также аппаратную реализацию модели.

ЛИТЕРАТУРА:

1. Knjaz'kov A. S. O kriterijakh znachimosti kriminalisticheskojj kharakteristiki prestuplenija / Vestn. Tom. gos. un-ta, 2007. –№304. – S. 122–128.
2. Christopher Westphal. Data Mining for Intelligence, Fraud, & Criminal Detection. Advanced Analytic & Information Sharing Technologies. 2009. CRC Press. – 426 p.
3. Khajjrova N. Model' izvlechenija znanijj iz nestrukturirovannykh dokumentov korporativnojj informacionnojj sistemy./ N. Khajjrova, N. Sharonova. // Applicable Information Models. ITHEA. — Varna, Bulgaria, 2011. — C. 131—139.
4. Bondarenko M. F. Teorija intellekta: uchebnik/ Bondarenko M. F., Shabanov-Kushnarenko Ju. P. Khar'kov: Komp. SMIT, 2007. — 576 s.
5. Shabanov-Kushnarenko S. Ju. Reshenie bulevykh uravnenijj s pomoshh'ju logicheskikh setejj / S. Ju. Shabanov-Kushnarenko, L. G. Sitnik, D. V. Bilenko, K. V. Silivejjstrov //ASU i pribory avtomatiki. — Kharkiv: KhNURE, 2008. — № 142. — S. 23-28.