У роботі показана необхідність використання спеціалізованого тезауруса для підвищення повноти та точності роботи інтегрованої інформаційно-криміналістичної системи. Розглянуто основні етапи автоматичного формування об'єктно-орієнтованого тезауруса оперативно-розшукової діяльності, що динамічно доповнюється. Запропоновано використання методу компараторної ідентифікації для виділення загальних змістовних ознак ключових слів, що дозволило автоматизувати етап дескріпторізації словникових статей тезауруса

Ключові слова: інформаційно-криміналістична система, об'єктно-орієнтована тезаурус, метод компараторної ідентифікації, автоматична дескріпторізація словника

В работе показана необходимость использования специализированного тезауруса для повышения полноты и точности работы интегрированной информационно-криминалистической системы. Рассмотрены основные этапы автоматического формирования объектно-ориентированного динамически наполняемого тезауруса оперативно-розыскной деятельности. Предложено использование метода компараторной идентификации для выделения общих содержательных признаков ключевых слов, что позволило автоматизировать этап дескрипторизации словарных статей тезауруса

Ключевые слова: информационно-криминалистическая система, объектно-ориентированный тезаурус, метод компараторной идентификации, автоматическая дескрипторизация словаря

УДК 004.822

ИСПОЛЬЗОВАНИЕ МЕТОДА КОМПАРАТОРНОЙ ИДЕНТИФИКАЦИИ ДЛЯ ДИНАМИЧЕСКОГО НАПОЛНЕНИЯ ТЕЗАУРУСА ОПЕРАТИВНО-РОЗЫСКНОЙ ДЕЯТЕЛЬНОСТИ

Н. Ф. Хайрова
Доктор технических наук, доцент, профессор*
E-mail: nina_khajrova@yahoo.com
Д. Ю. Узлов
Соискатель*
E-mail: poputcik@mail.ru
С. В. Петрасова
Аспирант*
E-mail: svetapetrasova@gmail.com
*Кафедра интеллектуальных компьютерных систем
Национальный технический университет
"Харьковский политехнический институт"
ул. Фрунзе, 21, г. Харьков, Украина, 61002

1. Введение

Одной из основных задач информационного поиска в рамках оперативно-розыскной деятельности является удовлетворение потребности оперативных подразделений государственных органов в криминально значимой информации. Один из способов получения такой информации заключается в обращении к различным интегрированным информационно-криминалистическим системам или информационно-поисковым системам общего доступа.

Для того чтобы при работе с неструктурированными или слабо структурированными массивами текстовой информации обеспечить работников оперативных подразделений полной и релевантной информацией необходимо приблизить информационно-поисковый язык (ИПЯ) запроса к естественному языку. Одним из путей решения данной задачи является использование ИПЯ дескрипторного типа. Под языками дескрипторного типа, или дескрипторными ИПЯ чаще всего понимают систему описания документов и запросов, осуществляемых с помощью дескрипторов информационно-поискового тезауруса. Использование дескрипторов позволяет поисковой системе реагировать не на формальное совпадение ключевых слов, а на совпадение понятий или смыслов. Понятия при этом задается как совокупность всех способов обозначения его в текстах, т.е. в виде множества синонимов.

Таким образом, для использования языка дескрипторного типа при поиске криминально значимой информации необходимо ввести в интегрированную информационно-криминалистическую систему специальный тезаурус, дескрипторизация словарных статей в котором позволит показать какие ключевые слова и выражения означают одно и то же или близкое понятие. Информационно-поисковый тезаурус (ИПТ) представляет собой словарь терминов заданной области знания, в котором путем ссылок между терминами зафиксированы смысловые связи понятий, отражающих взаимодействие объектов и явлений действитель-

ности. Иногда говорят, что тезаурус отражает онтологию предметной области [1].

Методика разработки ИПТ основывается на государственном стандарте – ДСТУ 4032-2001 (ISO 2788: 1986) – одноязычный тезаурус, межгосударственном стандарте – ГОСТ 7.25-2001 – тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления в международном стандарте ISO 12620:2009 – Terminology and other language and content resources – Specification of data categories and management of a Data Category Registry for language resources [2]. В соответствии с методикой данных стандартов разработка ИПТ основывается на определении семантики отдельного термина или словосочетания и осуществляется в три этапа:

1) определение тематического профиля тезауруса;

2) сбор лексики и формирование словаря ключевых слов;

3) формирование классов эквивалентности и выделение дескрипторов.

2. Анализ литературных данных и постановка проблемы

При автоматизации технологии построения тезауруса этапы обработки базовых текстов, из которых извлекаются ключевые слова, остаются аналогичными ручной интеллектуальной обработке [3]. При автоматизации процесса формирования многоязыкового тезауруса криминально-значимой информации на этапе формирования словаря ключевых слов необходимо выделить из разноязычных текстов, обрабатываемых в процессе оперативно-розыскной деятельности, информационные термины, с определенной степенью достоверности, отражающие объективную сторону преступления [4–6].

Особенностью такого выделения с точки зрения информации является определения совокупности лингвистических переменных естественного языка – специальных терминов, которые характеризуют действие и его результат [7–9]. Наличие подобных терминов в текстовом корпусе позволяет с определенной вероятностью соотнести корпус к одной из заранее определенных категорий множества противоправных деяний [10].

При этом для выделения общих содержательных признаков ключевых слов на этапе формирования классов эквивалентности и выделения дескрипторов необходимо использовать модели и методы теории интеллекта, позволяющие моделировать функцию интеллекта по пониманию и классификации смысла лингвистических элементов [11, 12].

3. Цель и задачи исследования

Целью данного исследования является разработка методики автоматического изменения тезауруса криминально-значимой информации, включающей два основных этапа: наполнение словаря ключевых слов предметной области и дескрипторизация статей данного словаря. Для достижения поставленной цели необходимо решить задачи формализации этапа формирования классов семантической (смысловой) эквивалентности и выделения дескрипторов словарных статей.

Предлагается рассмотреть решение данных задач, базирующееся на использовании методов теории интеллекта, которые позволяют моделировать интеллектуальные функции понимания, идентификации смысла и смысловой классификации лингвистических единиц.

4. Технология автоматического наполнения тезауруса криминально-значимой информации

На втором этапе стандартной методики формирования тезауруса осуществляется наполнение словаря ключевых слов предметной области. Для автоматического создания тезауруса криминально значимой информации на данном этапе реализуется разработанный лингвистический процессор, решающий задачи определения языка документа, выделения лексем, анализа графемного оформления текста, контекстного анализа коллокаций и статистического определения ключевых слов и словосочетаний [13]. В результате поэтапной лингвистической обработки всех текстовых документов коллекции формируется лексикон криминально-значимых информационных понятий анализируемой коллекции (рис. 1).

Третий этап формирования тезауруса – этап дескрипторизации ключевых слов, предполагает разделение выделенных информационных понятий по дескрипторным словарным статьям. На этом этапе дескрипторизации должна быть устранена неоднозначность в виде омонимии ключевых слов и осуществлена группировка полученных ключевых слов и коллокаций словаря в классы эквивалентности.

При такой смысловой классификации все ключевые термины должны быть разделены на взаимно исключающие классы, в каждом из которых термины имеют один или более общих содержательных признаков.

Этими содержательными признаками в тезаурусе терминов оперативно-розыскной деятельности являются отношения некоторого смыслового признака синонимичности, гипонимии и/или гиперонимии.

Рис. 1. Структурная схема лингвистического процессора расширения объектно-ориентированного тезауруса оперативно-розыскной деятельности

Два слова (или словосочетания) будут определены как синонимы, если в одном из своих лексических значений они равнозначны в некоторых или во всех контекстах. Т. е. мы будем говорить об одинаковости именно пропозиционного (propositional) значения [14] и синонимической связи именно между смыслами слов, а не между словами.

Мы будем говорить, что одно значение лексемы является гипонимом (hyponym) по отношению к одному значению другой лексемы, если первое значение является подчиненным, более специфическим или подклассом второго. И обратно, мы говорим, первое значение является гиперонимом (hyperonym) второго, если первое значение является надклассом или родительским классом другого.

Таким образом, отношения гипонимии, гиперомии и синонимии между значениями слов показывают наличие общих содержательных (семантических) признаков между одним или несколькими лексическими значениями данных слов. Наличие одного из таких признаков семантической эквивалентности позволяют отнести ключевые слова к общей дескрипторной статье тезауруса [15].

5. Использование метода компараторной идентификации для факторизации пространств концептов

Для дескрипторизации ключевых слов по взаимоисключающим классам семантической эквивалентности используется метод компараторной идентификации [11].

На множестве лексикона терминов, объективно описывающих состав преступления, $T=\{t_1,t_2,...,t_n\}$ и множестве рассматриваемых в документах оперативно-розыскной деятельности концептов $\theta=\{\rho_1,\rho_2,...,\rho_m\}$ вводим функцию понимания термина $\rho=f(t)$, где ρ — концепт (или понятие) термина. Под концептом, в данном контексте, мы будем понимать информацию, которую термин t несет о возможных денотатах τ [16], т. е. совокупность суждений о каком-либо объекте, предмете, орудии, средстве и т. п. преступления, выражающим его сущность.

Понимая ключевой термин t, выражаемый определенными лексическими единицами, аналитик или эксперт соотносит его с определенным концептом (смыслом) ρ.

Функция понимания описывает процесс установления экспертом тождества между термином и концептом, знаком которого он является. Если эксперт рассмотрел все множество терминов лексикона коллекции документов, функция f отобразит множество терминов лексикона на множество всех значений функции, т.е. совокупность всех концептов, порождаемых терминами из множества θ. Причем множество θ значительно меньше множества T, т. к. разнообразие концептов значительно меньше разнообразия знаков этих концептов.

Ключевые слова, отнесенные к родственным понятиям или концептам, имеющим общие элементы смысла, мы будем понимать, как эквивалентные в одном из своих семантических значений. Критерием включения слов или словосочетаний в класс эквивалентности является семантическая значимость этих КС при поиске документов в определенном контексте.

На практике такая значимость проявляется в следующем: если при документальном поиске один термин может быть заменен другим термином так, что при любом запросе результат выдачи документов будет такой же, как и до замены, то такие два КС объявляются эквивалентными в определенном контексте поиска и включаются в один класс эквивалентности.

Таким образом, ключевые слова, входящие в класс эквивалентности, соответствуют близким по смыслу концептам. Денотаты таких концептов, как показывают исследования [17], рассматриваются в одном связном тексте, который на уровне семантики характеризуется единой тематичностью.

Анализируя содержание связного текста d из рассматриваемого множества документов оперативно-розыскной деятельности и понимая его, аналитик, обычно, формирует в своем сознании некий инсайтный смысл ω, являющийся основным значением текста [18]. Смысл документа однозначно определяется породившим его текстом. Понимание аналитиком текста документа обозначает компонент его мышления, психологическое состояние, определяющее верное восприятие или интерпретацию данного документа, т.е. установление связи раскрываемых новых свойств объекта познания с уже известными.

Функцию $\omega=g(d)$ зависимости смысла связного текста от определяющей его знаковой смысловой единицы назовем функцией понимания связного текста [19]. Функция g отображает множество текстов D на множество рассматриваемых в них смыслов $\Re$.

При организации семантического поиска термины t_1 и t_2 считаются эквивалентными в определенном поисковом контексте, если они соответствуют понятиям $\rho_1=f(t_1)$ и $\rho_2=f(t_2)$, денотаты которых τ_1 и τ_2 рассматриваются в одном тексте: $\tau_1\in d$ и $\tau_2\in d$. Концептуально-смысловой предикат $\varepsilon=Q(\rho,\omega)$ отражает соответствие $(\varepsilon=1)$ и несоответствие $(\varepsilon=0)$ денотата τ рассматриваемого концепта ρ смыслу документа ω.

Согласно [15] значение предиката Q можно установить с помощью объективно определяемого контекстно-дескрипторного предиката:

$$Q(\rho,\omega)=Q(f(t),g(d))=L(t,d). \quad (1)$$

Таким образом, удается перейти от субъективного восприятия концептов, денотатов и смыслов текстов к объективному отношению между контекстом и термином, соответствующим $L(t,d)=1$ или не соответствующим $L(t,d)=0$ данному тексту. Согласно [10, 15] данный предикат является предикатом эквивалентности, который факторизует пространство ключевых терминов лексикона, однозначно разбивая его на классы эквивалентности (рис. 2).

Легко показать, что данный предикат является предикатом эквивалентности, который можно использовать для объективного определения соответствия двух любых ключевых терминов одной дескрипторной статье. Действительно, если $G_2(t_y,t_w)=1$, то $L(t_y,d)=L(t_w,d)$ для любого документа коллекции, выданного в результате запроса. В результате чего ключевые термины t_y и t_w для любого документа кол-

лекции дадут одинаковый результат выдачи: документ будет одинаково либо выдан, либо не выдан по результатам запроса с данными ключевыми словами. Таким образом, объективно ключевые понятия t_y и t_w будут относиться к одному дескриптору тезауруса.

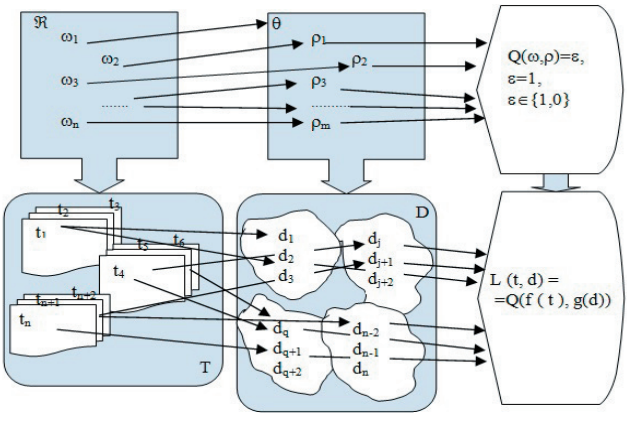

Рис. 2. Логическая схема динамического формирования дескрипторных статей

Можно ввести предикат дескрипторизации G_2, заданный на декартовом квадрате $T*T$:

$$G_2(t_y,t_w)=\forall d\in D(L(t_y,d)\sim L(t_w,d)). \quad (2)$$

То есть предикат G_2 определяет разбиение множества терминов T, выделенных на первом этапе разработки тезауруса, на слои ключевых слов, представляющие собой различные дескрипторные словарные статьи.

Классу всех $\psi_b(t)$ терминов $t\in T$, относящихся к одной дескрипторной статье, включающей ключевое слово или коллокацию $b\in T$ можно показать как:

$$\psi_b(t)=G_2(t,d)=\forall d\in D(L(t,d)\sim L(b,d)). \quad (3)$$

Данная формула явно показывает, что при организации информационного поиска во множестве документов оперативно-розыскной деятельности D ключевые слова и коллокации, относящиеся к классу $\psi_b(t)$, могут быть заменены в определенном контексте термином b с аналогичным результатом выдачи.

6. Апробация результатов исследований

Разработанный метод реализован в виде автоматического тезауруса подсистемы информационного поиска интегральной криминалистической системы «Портал». Проведенное контрольное исследование позволило ключевые слова, автоматически выделенные из массива электронных документов, поступивших следователю ОВД на обработку, факторизовать по четырем криминалистическим учетам: мошенничество; кража частной собственности; растрата и присвоение; финансовые преступления.

На следующем этапе ключевые слова автоматически дескрипторизовались по узким словарным статьям тезауруса: «мошенничество», «обман», «кража», «ограбление», «ущерб», «растрата», «присвоение», «уничтожение», «сговор лиц», «коммерческое право», «финансовое право».

В качестве показателей качества работы системы вычислялись коэффициенты полноты и точности, определяемые по результатам выдачи подсистемы информационного поиска системы «Портал», с использованием динамически разрабатываемого тезауруса. Полученные средние показатели полноты 0,86 и точности 0,92 позволяют использовать полученные решения в практике разработки интегрированных информационно-поисковых систем.

7. Выводы

Рассмотренное в статье использование метода компараторной идентификации для моделирования функции интеллекта по пониманию и выделению общих содержательных признаков в лексических единицах, с определенной степенью достоверности отражающих объективную сторону преступления, позволяет факторизовать пространство концептов криминалистически-значимой информации.

На первом этапе лингвистическим процессором создается словарь ключевых слов массива текстовой информации, используемой в процессе оперативно-розыскной деятельности (сводки, объяснительные/служебные записки, отчеты, газетные и интернет публикации, словесные портреты фигурантов и т. п.).

На втором этапе обработки полученный словарь ключевых слов автоматически разбивается на классы семантической эквивалентности, соответствующие дескрипторным словарным статьям динамически изменяемого тезауруса.

Реализация предложенного метода в подсистеме информационного поиска интегральной криминалистической системы «Портал» позволила подтвердить возможность его практической использования с достаточно высокими показателями полноты и точности результатов выдачи.

Литература

1. Браславский, П. И. Тезаурус для расширения запросов к машинам поиска Интернета: структура и функции [Электронный ресурс] / П. И. Браславский. – Режим доступа: http://www.dialog-21.ru/Archive/2003/Braslavskij.htm/.
2. "ISO 12620:2009. Terminology and other language and content resources – Specification of data categories and management of a Data Category Registry for language resources".iso.org [Text] / Retrieved 9 November, 2011.
3. Панченко, А. И. Метод автоматического построения семантических отношений между концептами информационно-поискового тезауруса [Текст] / А. И. Панченко // Вестник ВГУ. Серия: Системный анализ и информационные технологии. – 2010. – № 2. – С. 160–168.

4. Mena, J. Investigative Data Mining for Security and Criminal Detection [Text] / J. Mena. – Butterworth Heinemann is an imprint of Elsevier Science, 2003. – 452 p.
5. Nath, S. V. Crime Pattern Detection Using Data Mining [Text] / S. V. Nath // Web Intelligence and Intelligent Agent Technology Workshops. – 2006. – P. 41–44.
6. Phua, C. Resilient Identity Crime Detection [Текст] / C. Phua, K. Smith-Miles, V. C. S. Lee, Ross W. Gayler // IEEE Transactions on Knowledge and Data Engineering. – 2012. – № 24. – P. 533–546.
7. Srivastava, A. N. Text Mining. Classification, Clustering and Applications [Text] / A. N. Srivastava, M. Sahami. – CRC Press. Taylor & Francis Group. London, 2009. – 278 p.
8. Panchenko, A. Serelex: Search and Visualization of Semantically Similar Words [Text] / A. Panchenko, P. Romanov, O. Morozova, H. Naets, A. Romanov, A. Philippovich, C. Fairon // In Proceedings of the 35th European Conference on Information Retrieval. – 2013. – LNCS 7814. – P. 837–840.
9. Трусов, А. В. Модель поиска информации в распределенных информационных системах сети Интернет [Text] / А. В. Трусов, В. А. Трусов // Научно-техническая информация (НТИ). Сер. 2. Информационные процессы и системы. – 2011. – № 8. – С. 29–31.
10. Кудрявцев, В. Н. Объективная сторона преступления [Текст] В. Н. Кудрявцев. – М. : Госюриздат, 1960. – 244 с.
11. Бондаренко, М. Ф. Об алгебре конечных предикатов [Текст] / М. Ф. Бондаренко, Ю. П. Шабанов-Кушнаренко // Бионика интеллекта: науч.-техн. журнал. – 2011. – № 3 (77). – С. 3–13.
12. Дружинин, В. Н. Диагностика общих познавательных способностей [Электронный ресурс] / В. Н. Дружинин // Когнитивное обучение: современное состояние и перспективы. – Режим доступа: http://shp.by.ru/psy/lit/raznoe/00070.shtm (20.11.08).
13. Хайрова, Н. Ф. Использование семантико-ориентированного лингвистического процессора для добывания новых знаний из потока документов корпоративной информационной системы [Текст] / Н. Ф. Хайрова, В. А. Тарловский // Вісник Національного технічного університету «ХПІ». Збірник наукових праць. Тематичний випуск «Системний аналіз, управління та інформаційні технології». – 2010. – № 67. – С. 132–138.
14. Russell, B. Logic and Knowledge [Text] / B. Russell // Essays 1901–1905. London, 1956. – 365 p.
15. Кудинова, Е. А. Концепт и его соотнесение с лексико-семантическим полем [Текст] / Е. А. Кудинова // Филологические науки. Вопросы теории и практики. – Тамбов : Грамота. – 2008. – Ч. 2, № 1 (1). – С. 48–50.
16. Поспелов, Д. А. Введение в прикладную семиотику [Текст] / Д. А. Поспелов, Г. С. Осипов // Новости искусственного интеллекта. – 2002. – № 6. – С. 28–35.
17. Солтон, Дж. Динамические библиотечно-информационные системы [Текст] / Дж. Солтон; пер. с англ. – М. : Мир, 1979. – 557 с.
18. Философия: энциклопедический словарь [Текст] / ред. А. А. Ивина. – М. : Гардарики, 2004. – 1072 с.
19. Хайрова, Н. Модель извлечения знаний из неструктурированных документов корпоративной информационной системы [Текст] / Н. Хайрова, Н. Шаронова. // Applicable Information Models. ITHEA. – Varna, Bulgaria. – 2011. – С. 131–139.