УДК 81’374.26:343.9

М.М. Зацеркляний[1], Д.Ю.Узлов[2]

[1] *Харківський національний університет внутрішніх справ, Харків*
[2] *Управління інформаційно-аналітичного забезпечення ГУ МВС України в Харківській області, Харків*

ОБ’ЄКТНО- ОРІЄНТОВАНИЙ ТЕЗАУРУС І СЛОВНИК КОЛОКАЦІЙ ДЛЯ БАЗИ ЗНАНЬ КРИМІНАЛІСТИЧНИХ ІНФОРМАЦІЙНИХ СИСТЕМ

У статті на основі існуючих інформаційно-пошукових тезаурусів та методів ймовірно-статистичного розбору корпусів текстів запропоновано побудова тезаурусу на основі антиципаційної програми, тобто алгоритму передбачення того, що повинно бути знайдено. Алгоритм базується на ймовірнісно-статистичних характеристиках розпізнаних раніше колекціях документів.

Ключові слова: *інформаційно-пошукові тезауруси, словники колокацій, статистичні ознаки, антиципаційна програма.*

Вступ

Завданням інформаційного пошуку в рамках оперативно-розшукової діяльності є задоволення потреби в кримінально значимій інформації. Одним із способів одержання цієї інформації є звернення до різних криміналістичних інформаційно-пошукових систем (КІПС). Звернення здійснюється за допомогою формальної мови запитів і вимагає наявності знання про принципи роботи КІПС та відповідного досвіду. Без досвіду і знання принципів роботи КІПС, при відсутності уявлень про структуру вмісту цих систем сформулювати ефективний запит досить складно. Неправильна побудова запиту призводить до нерелевантних результатів. Ситуація ускладнюється, коли необхідно знайти інформацію в неструктурованих чи слабкоструктурованих масивах, наприклад у мережі Internet. В таких випадках, як правило, пошук здійснюється з використанням різних пошукових машин (Google, Yandex, Rambler, Meta та ін.). В даному випадку формулювання запиту для одержання релевантної відповіді у вузько спеціалізованій області стає процесом нетривіальним. Це обумовлено відсутністю розвинених мов запитів у всіх пошукових машинах для мережі Internet, їх спрямованості на переваги і очікування «масового користувача». Класичним підходом до вирішення даної проблеми є використання тезаурусу вузької предметної області (у нашому випадку кримінологічної) як основи для незалежної програми-асистента, призначеної для формування запитів до пошукових машин [1].

Використання асистента дозволяє зробити процес формулювання запитів більш осмисленим і ефективним. Тезаурус є необхідним елементом будь-якої пошукової чи довідкової системи і може використовуватися як при побудові запиту, так і при розв’язуванні задач класифікації, кластеризації, ідентифікації та порівняння різних масивів даних. Досвід роботи з пошуковими запитами показує, що при пошуку кримінально значимої інформації значно підвищити релевантність пошуку можна при використанні в пошуковому запиті термінів та колокацій кримінальної спрямованості, у тому числі з використанням сленгових виразів. Тому видається важливим визначення принципів, методології та структури побудови кримінально значимого тезауруса.

Аналіз літератури. Інформаційно-пошукові тезауруси, як правило, створюються для конкретних предметних областей. Їх побудова базується на таких сутностях як «поняття» і «термін», під якими розуміється слово чи словосполучення, номінуються поняття певної галузі знань або діяльності. Саме таке розуміння терміна є підставою розглядати інформаційно-пошукові тезауруси як вид онтологічних ресурсів [2].

Класичний підхід до складання тезауруса є складання дескрипторів і аскрипторів, які дають смислове (семантичне) навантаження термінів предметної області. Даний підхід закріплений у міждержавному стандарті ГОСТ 7.25-2001 (тезаурус інформаційно-пошуковий одномовний), і міжнародному ISO 12620: 1999 Computer applications in terminology - Data categories. Складання таких тезаурусів ґрунтується на виділенні семантики окремого терміна чи словосполучення.

Інформаційно-пошукові тезауруси, створювані в тому вигляді, як це закріплено міжнародними та національними стандартами, призначені для використання їх у ручному режимі індексування. За своєю суттю такий тезаурус є штучною мовою опису, побудованою на основі природної мови. Тому є значна дистанція між лексичним складом документів предметної області та словниковим складом інформаційно-пошукового тезауруса в цій предметній області. Саме тому традиційні інформаційно-пошукові теза-

уруси, розроблені для ручного індексування, складно використовувати при автоматичному індексуванні документів, застосовувати в інших застосуваннях інформаційного пошуку, хоча такі тезауруси містять в собі багато корисної інформації про предметну область.

До інших належать тезауруси типу WordNet, словниковий склад яких є значно більш докладним і близьким до лексики документів. WordNet має досить розвинену структуру родовидових відносин між сінсетами - наборами синонімів, побудованими за ієрархічним принципом. Ця особливість дозволяє використовувати тезауруси даного класу для автоматизації семантичного розбирання тексту на основі морфологічного, синтаксичного аналізу або методів компараторної ідентифікації. Проте даний вид тезауруса також слабо підходить для задач інформаційного пошуку [7, 8].

У даній роботі для побудови тезауруса кримінально значущої інформації пропонується використовувати антиципаційну програму, основна ідея якої зводиться до тези: ми наперед знаємо, що шукаємо, на основі власних суджень, які ґрунтуються на розпізнаних раніше колекціях документів.

1. Побудова тезаурусу на основі антиципаційної програми

Антиципаційна програма (тобто схема передбачення того, що повинно бути знайдено) базується на ймовірнісно-статистичних характеристиках розпізнаних раніше колекціях документів [3]. Саме схема передбачення є основою структури пропонованого тезауруса.

У випадку з кримінально значимою інформацією антиципаційна програма є алгоритмом розбору злочину за складовими, виділення кваліфікуючих ознак, осіб, об'єктів і складання онтологічної моделі конкретного злочину.

Основними індикативними ознаками наявності кримінально значимої інформації в тексті, є слова, що означають злочинні діяння, які описують об'єктивну сторону складу злочину.

Об'єктивна сторона злочину - це один із елементів складу злочину, що включає в себе ознаки, які характеризують зовнішній прояв злочину в реальній дійсності, доступні для спостереження і вивчення. Об'єктивна сторона злочину може також визначатися як «процес суспільно небезпечного і протиправного посягання на охоронювані законом інтереси, що розглядається з його зовнішньої сторони, з точки зору послідовного розвитку тих чи інших подій і явищ, які розпочинаються зі злочинної дії (бездіяльності) суб'єкта і закінчуються настанням злочинного результату». (Кудрявцев В.Н.)

Спробуємо сформулювати це визначення з точки зору інформатики.

Об'єктивна сторона злочину - це сукупність лінгвістичних змінних природної мови і спеціальних термінів, які характеризують дію і її результат, наявність яких у текстовому корпусі дозволяє з певною ймовірністю співвіднести корпус до однієї із заздалегідь визначених категорії множини протиправних діянь.

Суть запропонованого підходу полягає в описі кожного терміна чи колокації тезауруса вектором статистичних даних про частоту появи, вагу, унікальність та ін в колекціях спеціальних текстових корпусів.

Всі тезауруси складаються шляхом виділення термінів (дескрипторів) з колекції тематичних текстів (корпусів), які суб'єктивно виділяються експертами як тематичні в конкретній предметній області.

Формально, маємо необхідність побудови мультифункцій належності кожного терміна або колокації кінцевій множині класів (статті Кримінального Кодексу). Одним із варіантів побудови функції є використання експертних оцінок.

2. Використовувані ознаки слів

Розглянемо ознаки, які можна використовувати для виявлення термінологічності слова, що зустрічається в колекції текстів предметної області.

3.1. Частотність (Freq). Частотність використання слова в колекції.

3.2.Частотность з урахуванням частоти використання в осяжній колекції (Tf*idf). Ця ознака широко використовується в інформаційно-пошукових системах і дозволяє знижувати вагу вживаних слів:

$$Tf*Idf(w) = Tf * \log(n - b),$$

де n - розмір предметної колекції, b - число документів, в яких вживалося слово w з контрастної колекції, Tf - частотність слова в поточній колекції.

За контрастну колекцію для даної ознаки вибрана колекція Зведення пригод.

3.3. Ознака Дивина (Weirdness). Ця ознака враховує пропорційне співвідношення частотності використання слова в робочій текстовій колекції порівняно з контрастною колекцією.

Нехай w - слово. Тоді

$$Weirdness(w) = \frac{Ws}{Ts} Wg,$$

де Ws - частотність слова в колекції предметної області; Ts - сукупна частотність слів у колекції предметної області; Wg - частотність слова в контрастній колекції Зведення пригод; Tg - сукупна частотність слів у контрастній колекції Зведення пригод.

3.4. Ознака C-Value. Ця ознака обґрунтовує рейтинг термінологічності слів з урахуванням частотності словосполучень, до яких входить дане слово. Нехай w - слово. Тоді

$$C-Value(w)=$$
$$=freq(w)-\frac{\sum_{[\text{за всіма b зі множ. Ta}]}freq(b)}{P(Ta)},$$

де Ta - множина всіх словосполучень у колекції, що містять слово w; b – словосполучення, шо містять конкретне слово; P(Ta) - потужність множини Ta.

3.5. Ознака зустрічі слова в термінах тезауруса. Вважаємо, що розробка тезауруса предметної області розпочата і в тезаурус внесена деяка сукупність термінів. Тоді за додаткову ознаку для визначення термінологічності слова можна використовувати ознаку кількості термінологічних словосполучень, у які входить дане слово - ознака FreqByThes.

3. Використовувані ознаки колокацій

4.1. Найбільш частотне словосполучення (Inside). Ця ознака враховує частотність найбільш частотного словосполучення, до складу якого входить дане слово.

Нехай w - слово. Серед усіх словосполучень, що містять слово w, виберемо найбільш частотне. Нехай $F(*w*)_{max}$ - його частота. Тоді

$$Inside(w)=\frac{F(*w*)_{max}}{freq(w)}.$$

Ця ознака перевіряє, чи не вживається дане слово в складі одного і того ж словосполучення. Чим вище значення ознаки, тим нижча ймовірність того, що слово є самостійним значимим елементом предметної області, а, скоріше, є компонентом більш довгого сталого словосполучення.

4.2. Ознаки вживання слова в наборі словосполучень (Sum3, Sum10, Sum50). Ці ознаки перевіряють, наскільки дане слово було продуктивним в утворенні словосполучень предметної області.

Нехай w - слово. Серед усіх словосполучень, що містять слово w, виберемо k найбільш частотних. Нехай Sum - сума їх частотностей. Тоді

$$SumX(w)=\frac{Sum}{X}\frac{1}{freq(w)}.$$

4.3. Критерій MI. MI (коефіцієнт взаємної інформації) порівнює залежні контекстно-пов'язані частоти появи слів із незалежними, тобто з тими, які у тексті з'являлися випадково і обчислюється за формулою:

$$MI=\log_2\frac{f(n,c)\times N}{f(n)\times f(c)},$$

де n - ключове слово; c - колокат; f(n, c) - відносна частота зустрічі ключового слова n у парі з колокатом c; f(n), f(c) - відносні частоти ключового слова n і слова c в тексті; N - загальне число словоформ у тексті.

4.4. Критерій Log-Likelihood - логарифмічна функція правдоподібності або логарифмічна міра подібності:

$$G=2\times$$
$$\times\left(\left(fr_{domain}\times\log\left(\frac{fr_{domain}}{frExp_{domain}}\right)\right)+\left(fr_{general}\times\log\left(\frac{fr_{general}}{frExp_{general}}\right)\right)\right),$$

де $frExp_{domain}$ та $frExp_{general}$ – очікувані частоти в тексті предметної області і в неспеціалізованому тексті відповідно; fr_{domain} та $fr_{general}$ – реально спостережувані частоти в тексті предметної області і в неспеціалізованому тексті відповідно.

Для обчислення очікуваних частот використовуються такі формули:

$$frExp_{domain}=size_{domain}\times R\_fr$$

$$frExp_{general}=size_{general}\times R\_fr,$$

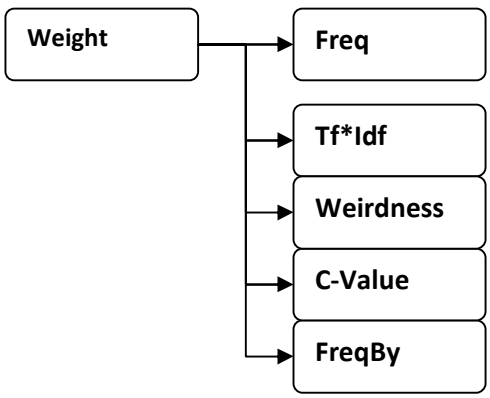
де змінні $R\_fr=\frac{fr_{domain}+fr_{general}}{size_{domain}+size_{general}}$, $size_{domain}$ та $size_{general}$ вказують на розміри відповідних текстів, обчислені за кількістю слів.

4. Структура тезауруса

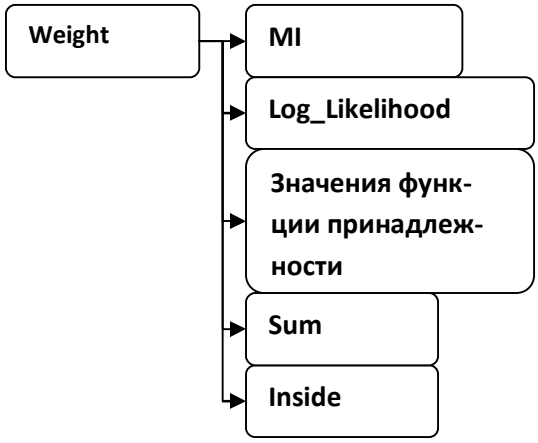
За основу структури тезауруса використана структура запропонована Бреславським [1]. Модернізація цієї структури проводиться на основі розширення числової характеристики (weight) набором (вектором) ймовірнісно-статистичних характеристик для термінів та колокацій.

Важливими числовими характеристиками для термінів та колокацій є значення функцій належності до класу вид злочину, що ґрунтується на Кримінальному Кодексі.

Для термінів:



Для колокацій:



Висновки

На основі існуючих інформаційно-пошукових тезаурусів та методів ймовірно-статистичного розбору корпусів текстів запропоновано побудова тезаурусу на основі антиципаційної програми, тобто алгоритму передбачення того, що повинно бути знайдено.

Алгоритм базується на ймовірнісно-статистичних характеристиках розпізнаних раніше колекціях документів.

Вказаний підхід до побудови тезауруса і словника колокацій дає можливість його використання в лінгвістичному процесорі [5] для виділення кримінально-значимої інформації з неструктурованих або слабоструктурованих текстових масивів.

Список літератури

1. Браславский П.И. Тезаурус для расширения запросов к машинам поиска Интернета: структура и функции [Электронный ресурс] / Браславский П.И.. Режим доступа: http://www.dialog-21.ru/Archive/2003/Braslavskij.htm

2. Лукашевич Н.В. Тезаурусы в задачах информационного поиска / Н.В. Лукашевич. – М.: Издательство Московского университета, 2011. - 512 с.

3. Лукашевич Н.В. Использование методов машинного обучения для извлечения слов-терминов / Н.В. Лукашевич, Ю.М.. Логачев. – М.: Издательство Московского университета, 2009. - 242 с.

4. Горелов И.Н. Разговор с компьютером: Психолингвистический аспект проблемы. С послесловием Д.А. Поспелова / И.Н. Горелов. – М.: Наука. Гл. ред. физ.-мат. лит.,1987. – 256 с.

5. Особливості виділення кримінально значимої інформації в текстових масивах / О.М. Бандурка, М.М. Зацеркляний, Д.В. Лазарєв, Д.Ю. Узлов // Наше право : науково-практичний журнал. – 2011. – №2, ч.1. – С.79-83.

6. Зацеркляний Н.М. Лингвистический процессор для поиска и обработки криминально значимой информации в неструктурированных массивах / Н.М. Зацеркляний, Д.Ю. Узлов // Вестник НТУ "ХПИ". Тематический выпуск: Информатика и моделирование. – Х.: НТУ "ХПИ". – 2011. – № 36. – С. 87 – 94.

7. Feldman R. The text mining handbook / Ronen Feldman, James Sanger. – Published in the United States of America by Cambridge University Press, New York, 2007. – 423 p.

8. Data mining for intelligence, fraud & criminal detection : advanced analytics & information sharing technologies / Christopher Westphal. CRC Press, 2009. – 411 p.

Надійшла до редколегії 22.01.2013

Рецензент: д-р техн. наук, проф. О.А. Серков, Національний технічний університет «ХПІ», Харків.

ОБЪЕКТНО-ОРИЕНТИРОВАННЫЙ ТЕЗАУРУС И СЛОВАРЬ КОЛЛОКАЦИЙ ДЛЯ БАЗЫ ЗНАНИЙ КРИМИНАЛИСТИЧЕСКИХ ИНФОРМАЦИОННЫХ СИСТЕМ

Н.М. Зацеркляный, Д.Ю.Узлов

В статье предлагается модель построения тезауруса и словаря коллокаций для использования в базах знаний криминалистических информационных систем с использованием вероятностно-статистических признаков слов и коллокаций.

Ключевые слова: *информационно-поисковые тезаурусы, словарь коллокаций, статистические признаки, антиципационная программа.*

OBJECT-ORIENTED THESAURUS AND COLLOCATIONS DICTIONARY FORKNOWLEDGE BASE FORENSIC INFORMATION SYSTEMS

M.M. Zacerklyaniy D.Y.Uzlov

The paper propose a model of building a thesaurus and a dictionary of collocations for use in the knowledge bases of forensic information systems using probabilistic-statistical characteristics of words and collocations.

Keywords: *information retrieval thesaurus, dictionary collocations, statistical features, antitsipatsionnaya program.*