УДК 343.982.4:004

ЗАЦЕРКЛЯНИЙ М.М., д.т.н., проф. ХНУВС, м. Харків,
УЗЛОВ Д.Ю., начальник відділу УБОЗ ГУ МВС України в Харківській області, м. Харків

ЛІНГВІСТИЧНИЙ ПРОЦЕСОР ДЛЯ ПОШУКУ ТА ОПРАЦЮВАННЯ КРИМІНАЛЬНО ЗНАЧИМОЇ ІНФОРМАЦІЇ В НЕСТРУКТУРОВАНИХ МАСИВАХ

В статті запропонована модель лінгвістичного процесора для пошуку та опрацювання кримінально значимої інформації з неструктурованих текстових масивів з використанням статистичних методів виділення коллокацій та латентно-семантичного аналіза.

Ключові слова: лінгвістичний процесор, коллокація, латентно-семантичний аналіз, кримінально значима інформація, модель.

Постановка проблеми. Інтенсивне наповнення інформаційного простору і наявності в ньому інформації, що має значення для оперативно-службової діяльності правоохоронних органів ставить ряд проблем, пов'язаних із можливістю автоматичного вилучення кримінально значимої інформації. Однією з важливих задач є організація автоматичного чи напівавтоматичного вибору фрагментів текстів, що відповідають контрольному переліку інформації аналітичних підрозділів правоохоронних органів [1]. Пошук нової інформації в нових джерелах тільки одна із задач необхідних для оперативно-службової діяльності. Іншою важливою задачею є ретроспективний пошук латентних закономірностей у неструктурованих масивах, подібних добовим відомостям подій у правоохоронних органах, тобто задача полягає в пошуку подій схожих за якимись параметрами (місце вчинення, вид, механізм, учасники, тощо), зареєстрованим у добових відомостях. Розв’язування цієї задачі дозволить зменшити кількість нерозкритих злочинів, сприяти прогнозуванню і регулюванню криміногенної обстановки.

Аналіз літератури. Для пошуку необхідної інформації користувач звертається до інформаційно-пошуковій системі (ІПС) у вигляді запиту. Формулювання запиту є відповідальним і важко формалізованим етапом. Крім того, проблема полягає в тому, що потреба правоохоронців знаходиться в досить вузькоспеціалізованій області зі спеціальною термінологією і формулювання запиту природною мовою часто призводить до великої захараченості результату нерелевантними даними у зв’язку, зокрема, з полісемією природної мови [2]. Використання спеціальної термінології досить звужує коло пошуку. Ця проблема обумовлена тим, що розробники ІПС не прагнуть до розвитку мов

запитів. Їхні зусилля спрямовані на облік інформації, яка міститься в запиті, а також виявлення переваг і очікувань "масового користувача" [3]. Для користувачів зі специфічними інформаційними потребами необхідно усунути дисбаланс між універсальністю пошукової машини (ПМ) і специфічністю інформаційних потреб на етапі формулювання запитів. З цією метою необхідно модифікувати запит у спеціалізованій метапошуковій машині (ММ). Традиційно модифікація запиту здійснюється за допомогою семантичних словників – тезаурусів. Тезауруси не застосовуються в універсальних повнотекстових ПМ у зв'язку зі складністю побудови тезауруса, який відповідав би тематичним розмаїттям інформації, що індексується універсальною ПМ. Проте, можна вважати, що саме тезауруси сформовані з метаданих фактографічних систем баз даних правоохоронних органів будуть ефективними при побудові відповідної ММ. Більш ефективним буде використання словника стійких словосполучень кримінальної значимості – колокацій.

В зв'язку з цим виникають дві задачі:

– формування тезаурусів і колокацій, які мають кримінальну значимість;

– виявлення окремих слів і колокацій, які мають кримінальну значимість, у текстових масивах.

Метою статті є розгляд методів статистичного аналізу текстів щодо виявлення кримінально значимих слів та словосполучень, можливість складання відповідних тезаурусів та словників колокацій, модифікація методу LSA під застосування колокацій та розробка функціональної моделі лінгвістичного процесору для автоматичної обробки текстів щодо виявлення кримінально значимої інформації.

1. Функціональна модель лінгвістичного процесора. На рис. приведена функціональна схема запропонованого лінгвістичного процесора для пошуку та опрацювання кримінально значимої інформації в неструктурованих масивах. Його призначення полягає у розв'язуванні щойно сформульованих задач.

Вхідний текст із пошукової машини надходить на модуль індексації, де відбувається його розкладання на елементарні слова з урахуванням знаків пунктуації. Далі, розкладений текст, перетворюється в модулі "Транслятор" у текст з метаданими, які використовуються в тезаурусі, при цьому, в модулі коригування помилок відбувається пошук слів із друкарськими помилками і здійснюється їх коригування з використанням тезауруса. Перетворений текст опрацьовується в модулі "Тематичний фільтр", де на основі метаданих (із використанням онтологічної моделі)

визначається передбачувана тематична область тексту (визначається семантика) і підключається для проведення порівняння методом LSA з відповідною колекцією бази опрацьованих текстів "Text". Одночасно відбувається статистичний аналіз за частотою і мірою взаємозв'язку слів у модулі статистичного аналізатора з метою виявлення колокацій в нових текстах і поповнення відповідного словника, а також виділення додаткових метаданих у вигляді певної статистики, прив'язки їх до тексту і переміщення в базу даних опрацьованих текстів.

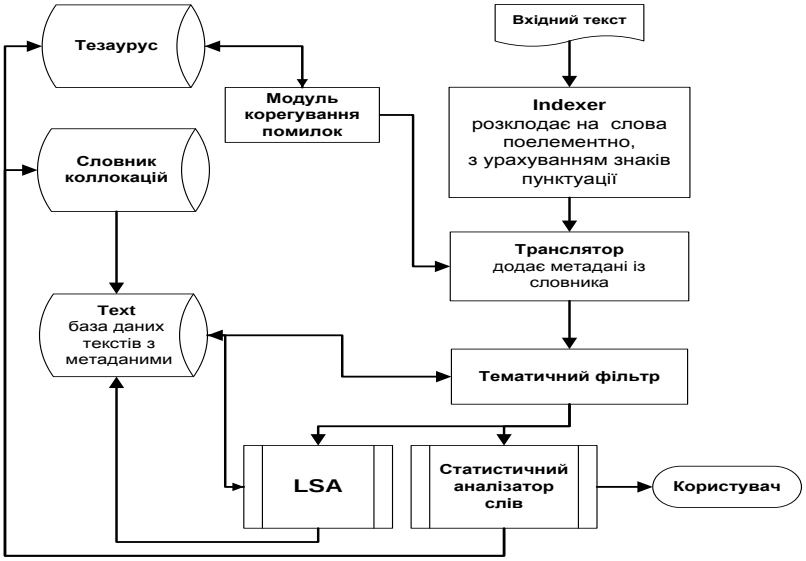

Рис. Функціональна модель лінгвістичного процесора

2. Метод латентно-семантичного аналізу ЛСА. Латентно-семантичний аналіз, або індексування, (LSA/LSI) – це метод вилучення "прихованих" контекстно-залежних значень термів зі структури семантичних взаємозв'язків між ними шляхом статистичного опрацювання великих наборів текстових даних.

Вибір цього методу для виділення кримінально значимої інформації в неструктурованих текстових масивах ґрунтується на тому, що:

– в якості початкових даних LSA використовує частоту використання слів в уривках тексту, а не частоту спільного використання слів;

– метод збирає дані не про спільне використовування слів, а про використану множину слів у великому текстовому масиві.

Зміст методу ЛСА полягає у порівнянні множини всіх контекстів, у яких вживаються потрібні слова чи групи слів із подальшим ранжуванням за мірою близькості слів або груп слів, тобто розв'язуються задачі класифікації, кластеризації, складання тезаурусів і ранжування виділеної інформації.

При розв'язуванні цих задач потрібно виконати розбиття текстових масивів на систему підмножин, відмічених змістовними описувачами (автоматичну кластеризацію) на основі метаданих фактографічних баз даних. Для цього використовується матриця, яка описує вживання слів або сталих словосполучень у текстах. Стовпчики матриці відповідають документам, а рядки – словам (сполученням слів), що зустрічаються в документах. Елементом матриці в даному випадку є кількість використання конкретного слова (сполучення слів) в цьому тексті.

Аналіз текстових масивів природною мовою пов'язаний із розв'язуванням деяких проблем:

– синонімії, коли одне поняття описується різними наборами слів;

– полісемії, коли одне слово або одна комбінація слів описують різні поняття;

– невизначеність, коли слово (сполучення слів) містить розмиту якісну оцінку поняття.

Останню проблему варто віднести до fuzzy logic (нечітка логіка) [Л. Заде]. Дві інші проблеми можна розв'язати шляхом пониження рангу матриці, тобто шляхом апроксимації її матрицею меншого рангу. Пониження рангу випливає зі зменшення розміру матриці, пов'язане зі словами (сполученнями слів), які мають близькі значення. У разі полісемії слова, коли одне з його значень відповідає змістовим елементам тексту, його елемент об'єднується з матричними елементами слів з таким самим значенням. В іншому випадку – відповідний елемент буде відкинутим. Зниження рангу необхідно також для фільтрації шуму – випадкового попадання слів, що не відповідають змісту тексту.

Матриця, яка відображає використання слів (сполучень слів) у текстових фрагментах має вигляд:

$$A = \begin{pmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m.n} \end{pmatrix}.$$

Кожний елемент цієї матриці є кількістю використань слова (сполучення слів) *i* у *j*-му фрагменті. Отже, кожний рядок матриці є

вектором, що відповідає слову (сполученню слів) і відображає його використання у кожному документі,

$$t_i^T = [a_{i,1}, \ldots, a_{i,n}],\ i = \overline{1, m}.$$

Аналогічно, кожний стовпчик є вектором, що відповідає певному фрагменту і відображає використання слів (сполучень слів) у цьому фрагменті:

$$d_j = \begin{pmatrix} a_{1,j} \\ \vdots \\ a_{m,j} \end{pmatrix},\ j = \overline{1, n}.$$

На основі роботи [5] одержується сингулярний розклад початкової матриці:

$$A = V \times W \times U^T,$$

або

$$A = \begin{pmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m.n} \end{pmatrix} = \left(\begin{pmatrix} \\ u_1 \\ \\ \end{pmatrix} \cdots \begin{pmatrix} \\ u_l \\ \\ \end{pmatrix} \right) \cdot \begin{pmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_l \end{pmatrix} \cdot \begin{pmatrix} [v_i] \\ \vdots \\ [v_l] \end{pmatrix}.$$

Числа $w_1, \ldots, w_l$ є сингулярними числами, а вектори $u_1, \ldots, u_l$ та $v_1, \ldots, v_l$ правими і лівими сингулярними векторами відповідно.

Якщо з усіх сингулярних значень відібрати k найбільших, то одержуємо апроксимацію початкової матриці матрицею рангу k [4]:

$$A - A_k = \sum_{l=k+1}^{\gamma} w_l v_l u_l^* \Rightarrow \left\| A - A_k \right\|_2 = w_{k+1}.$$

В результаті такої апроксимації у початковій матриці відкидається "надлишкова інформація", а отже, дістається матриця меншого рангу. Вибір найкращого розміру матриці залежить від задачі і встановлюється експериментальним шляхом: з одного боку ранг повинен бути як можна меншим, аби позбавитись "шуму", а з другого – достатньо великим, аби достатньо повно відобразити реально існуючу структуру даних.

3. Коллокації як індикативні ознаки наявності в текстах кримінально значимої інформації. Слова, що описують злочинні діяння мають свою специфіку і часто саме вони є індикативним ознакою, за якою здійснюється відбір документів. Зрозумілими є словосполучення ножове поранення, ознаки насильства, вогнепальне поранення, вибухова речовина, наркотична речовина, угон автомобіля, заволодіння майном,

умисний підпал, крадіжка грошей тощо. Проте є менш звичні, але більш ефективні поєднання слів для пошуку кримінально значимої інформації, наприклад "винт солянка". В цьому прикладі використовуються сленгові назви наркотичних засобів та прекурсорів (допоміжних речовин, що використовуються при виготовлені наркотиків): винт (сленгова назва наркотику) солянка (сленгова назва соляної кислоти, що є прекурсором при виготовлені наркотику "Винт").

Кожне з наведених словосполучень у професійних працівників правоохоронної системи викликає асоціацію з певним видом злочину, а отже, наявність їх у тексті вимагає в крайньому випадку глибокого вивчення цього тексту.

В рамках семантико-синтаксичного підходу стійкі словосполучення (коллокації) розглядаються як семантико-синтаксичні одиниці або лексично визначені елементи граматичних структур. Вони характеризуються семантичною, синтаксичною і дистрибутивною регулярностями [5].

Отже, стійкі словосполучення (коллокації) є потужним засобом у пошуку кримінально значимої інформації. А тому пошукова система їх повинна будувати і віднаходити в неструктурованих чи слабкоструктурованих масивах.

На сьогодні в лінгвістиці існує декілька способів для обчислення міри пов'язаності частин тієї чи іншої коллокацій. В рамках запропонованого лінгвістичного процесора для пошуку та опрацювання кримінально значимої інформації в неструктурованих масивах у модулі "Статистичний аналізатор слів" використовуються статистичні критерії MI, log-likelihood. Кожний із них має як свої переваги, так і свої недоліки.

3.1. Критерій *MI*. Міра *MI* дозволяє виділити стійкі словосполучення, імена власні, а також низькочастотні спеціальні терміни [6]. Слова, у яких *MI* приймає найбільшу величину, менш частотні і мають обмежену сполучуваність.

MI (коефіцієнт взаємної інформації) порівнює залежні контекстно-пов'язані частоти появи слів із незалежними, тобто 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 – коллокат; $f(n, c)$ – відносна частота зустрічальності ключового слова n в парі з коллокатом c; $f(n)$, $f(c)$ – відносні частоти ключового слова n і слова c в тексті; N – загальне число словоформ у тексті.

Міра *MI* залежить від розміру тексту: чим більшим є досліджуваний текст, тим вищі в середньому одержувані для нього значення *MI*. Ця властивість відображає міру довіри до даних, одержаних для великих текстів.

MI використовується як засіб ранжування коллокацій всередині одного тексту за мірою їх зв'язності.

При підрахунку міри *MI* порядок слів всередині коллокацій не враховується: ця міра відображає взаємозалежність двох лексем, але не значимість конкретної колокації.

3.2. Критерій Log-Likelihood. Цей критерій відомий як логарифмічна функція правдоподібності або логарифмічна міра подібності. Він показує міру близькості слів у тексті і обчислюється для кожного слова за формулою [7]:

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

де $frExpected_{domain}$ та $frExpected_{general}$ – 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:

$$frExpected_{domain} = size_{domain}\times R\_fr,$$

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

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

Опрацювання даним методом, дозволяє одержати з текстового масиву список слів із їх вагою, яка відповідає ймовірності цих слів бути термінами певної предметної області.

Критерій Log-Likelihood використовується також у модулі "Статистичний аналізатор слів".

В результаті роботи модуля одержуються два словники – тезаурус, що містить кримінально значимі терміни, і словник кримінально значимих коллокацій, що містить характерні для даної предметної області зв'язки слів.

Висновок. В роботі запропоновано та розглянуто функціональну модель лінгвістичного процесору на основі ЛСА та статистичних методів для опрацювання кримінально значимої інформації в текстових масивах.

Список літератури: 1. *Бандурка О.М.* Особливості виділення кримінально значимої інформації в текстових масивах / *О.М. Бандурка, М.М. Зацеркляний, Д.В. Лазарєв, Д.Ю. Узлов.* – Науково-практичний журнал "Наше право". – 2011. – № 2. – Ч. 1. – С. 79-83. **2.** *Андреев А.М.* Лингвистический процессор для информационно-поисковой системы / *А.М. Андреев, Д.В. Березкин, А.В. Брик.* – Режим доступа: http://www.inteltec.ru/publish /articles/textan\art_21br.shtml. **3**. *Браславский П.И.* Тезаурус для расширения запросов к машинам поиска Интернета: структура и функции / *П.И. Браславский* Режим доступа: http://www.dialog-21.ru/Archive/2003/Braslavskij.htm **4.** *Соболев М.С.* Метод латентно-семантического анализа. Магистерская диссертация / *М.С. Соболев.* – М.: МФТИ. – 2007. Режим доступу: http://rykov-dp.narod.ru/Sobolev.pdf. **5.** *Захаров В.П.* Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке / *В.П. Захаров, М.В. Хохлова*. – Режим доступа: http://www.dialog-21.ru/dialog2010/materials/html/22.htm. **6**. *Захаров В.П.* Статистический метод виявления коллокаций / *В.П. Захаров, М.В. Хохлова.* Режим доступа: www.ict.edu.ru/vconf/files/10374.pdf. **7**. *Гельбух А.Ф.* Автоматический поиск и классификация однословных терминов в корпусе предметной области с использованием логарифмической меры сходства с неспециализированным корпусом / *А.Ф. Гельбух, Г.О. Сидоров, Э. Лавин-Вийа.* – Режим доступу: http://www.dialog-21.ru/dialog2010/ materials/html/14.htm. **8.** Introduction to Information Retrieval [Електронний ресурс] *Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze* – Режим доступа: http://nlp.stanford.edu/IR-book/information-retrieval-book.html.

УДК 343.982.4:004

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

В статье предложена модель лингвистического процессора для поиска и выделения криминально значимой информации из неструктурированных текстовых массивов с использованием статистических методов выделения коллокаций и латентно-семантического анализа. Ил.: 1. Библиогр.: 8 назв.

Ключевые слова: лингвистический процесор, криминально значимая информация, статистические методы, коллокации, латентно-семантический анализ.

UDC 343.982.4:004

Linguistic processor for searching and processing criminal information important in unstructured array / Zatserklyany N.M., Uzlov D.Y. // Herald of the National Technical University "KhPI". Subject issue: Information Science and Modelling. – Kharkov: NTU "KhPI". – 2011. – № 36. – P. 87 – 94.

In the article the model of linguistic processor is offered for a search and selection criminally of meaningful information from the unstructured text arrays with the use of statistical methods of selection of collotions and latently-semantic analysis. Figs.: 1. Refs.: 8 titles.

Key words: linguistic processor, criminally relevant information, statistical methods, collocation, latent-semantic analysis.

Надійшла до редакції 30.06.2011