

**РОЗРОБКА ІНФОРМАЦІЙНО-ЛІНГВІСТИЧНОГО ЗАБЕЗПЕЧЕННЯ  
СИСТЕМИ ПОШУКУ  
КРИМІНАЛЬНО ЗНАЧУЩОЇ ІНФОРМАЦІЇ  
НА ОСНОВІ АНТИЦІПАЦІЙНОГО АЛГОРИТМУ**

**Узлов Д.Ю., Хайрова Н.Ф.**

*Національний технічний університет  
«Харківський політехнічний інститут», м. Харків*

В роботі представлено підхід до побудови тезауруса для інформаційної системи пошуку кримінально значущої інформації в неструктурованих або слабкоструктурованих текстових масивах.

Основна ідея створення тезауруса для інформаційного пошуку: ми наперед знаємо що шукаємо на основі власних суджень заснованих на розпізнаних і зрозумілих раніше колекціях документів.

Пошукова система повинна мати антиціпаційну програму, що базується на ймовірно-статистичних характеристиках розпізнаних і зрозумілих раніше колекціях документів. Саме схема передбачення буде основою структури пошукового тезауруса.

У випадку з кримінально значущою інформацією антиціпаційна програма являє собою алгоритм розбору злочину за складовими, виділення кваліфікуючих ознак, осіб, об'єктів і складання по суті онтологічної моделі конкретного злочину.

Основними індикативними ознаками наявності кримінально значущої інформації в тексті, є слова, що означають злочинні діяння, що описують об'єктивну сторону складу злочину. Інформаційно-лінгвістичне визначення об'єктивної сторони злочину містить сукупність лінгвістичних змінних природної мови і спеціальних термінів, що характеризують дію і її результат, наявність яких в текстовому корпусі дозволяє, з певною ймовірністю, віднести текст корпусу до однієї з задалегідь визначених категорій множини протиправних діянь.

Суть пропонованого підходу полягає в описі кожного терміна або колокацій тезауруса вектором статистичних даних про зустрічальність, вагу, унікальність та інше 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 мультифункції приналежності кожного терміна або колокацій, кінцевому масиву класів (статті КК).