ДОДАТОК А
ГРАФІЧНИЙ МАТЕРІАЛ

Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет _Комп'ютерних наук
Кафедра Інформаційних управляючих систем

КВАЛІФІКАЦІЙНА РОБОТА

ГРАФІЧНИЙ МАТЕРІАЛ

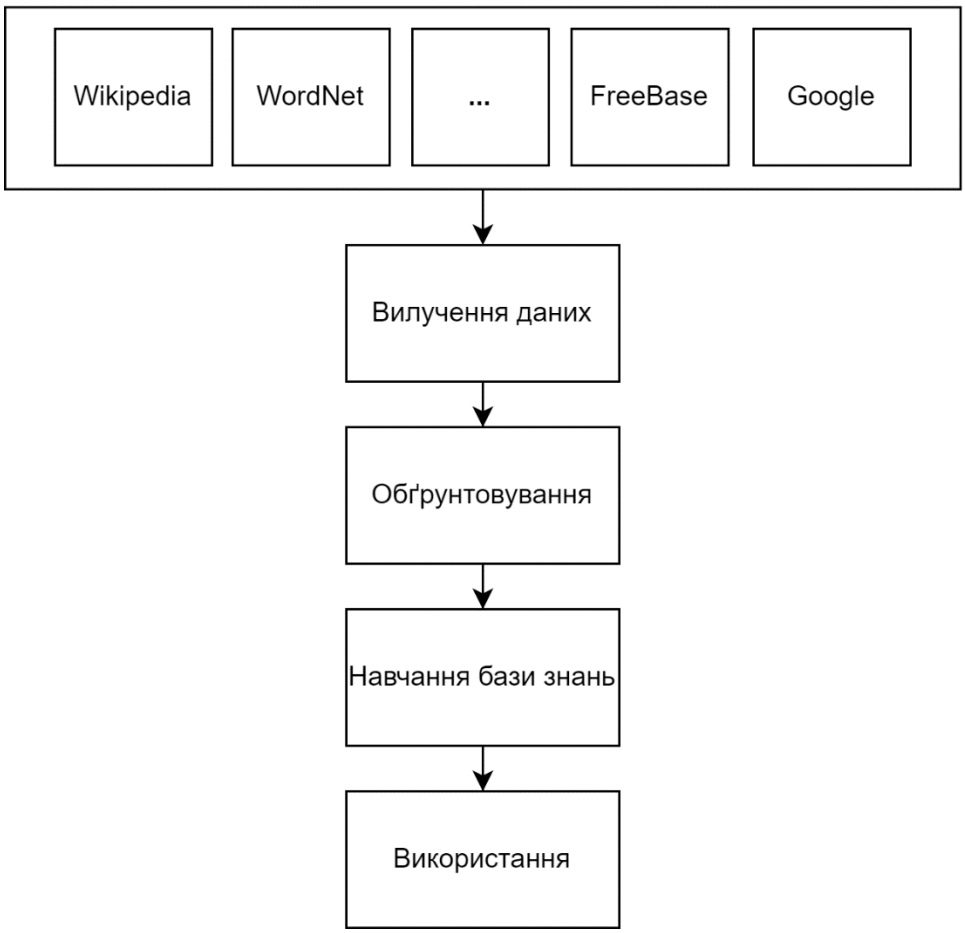
Дослідження методів автоматизованої побудови баз знань в інформаційно-довідкових системах

2021р.

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність	Робота присвячена вирішенню задачі удосконалення методу автоматизованої побудови баз знань в інформаційно-довідковій системі на основі автоматизованого вилучення правил із вхідних документів. Актуальність даної задачі є наслідком невідповідності можливостей існуючих методів автоматизованої побудови баз знань для інформаційно-довідкових систем, які використовують додаткові зовнішні бази типових правил, та практичними потребами процесу автоматизованої побудови баз знань безпосередньо на основі аналізу документів.
Об'єкт дослідження	Процес автоматизованої побудови баз знань для інформаційної довідкової системи.
Предмет дослідження	Методи автоматизованої побудови бази знань для інформаційної довідкової системи.
Мета роботи	Дослідження методів автоматизованої побудови баз знань для інформаційно-довідкових систем для підвищення їх ефективності на основі виявлення правил нових типів з тексту довідкових документів
Наукова новизна	Удосконалено метод автоматизованої побудови зважених правил для інформаційно-довідкової системи на основі використання триплетів «об’єкт, відношення, суб’єкт» шляхом виділення залежностей між суб’єктами, для яких співпадає об’єкт та може співпадати відношення між об’єктом та суб’єктом. Ваги правил визначаються частотою виявлення однакових правил в тексті.
Практичні результати	Вдосконалений метод автоматизованої побудови бази знань для інформаційної довідкової системи, алгоритм та програмний модуль побудови баз знань
Задачі дослідження	– дослідження методів автоматизованої побудови баз знань; – удосконалення методу автоматизованої побудови баз знань в інформаційно-довідкових системах; – визначення сфери застосування удосконаленого методу; – експериментальна перевірка удосконаленого методу.

ОБ'ЄКТ ДОСЛІДЖЕННЯ



Об’єктом дослідження є процес автоматизованої побудови баз знань для інформаційних довідкових систем.

Інформаційні довідкові системи призначені для здійснення швидкого пошуку та видачі довідкової інформації за предметною областю.

Основними завданнями інформаційної довідкової системи є:

– збір даних предметної області;

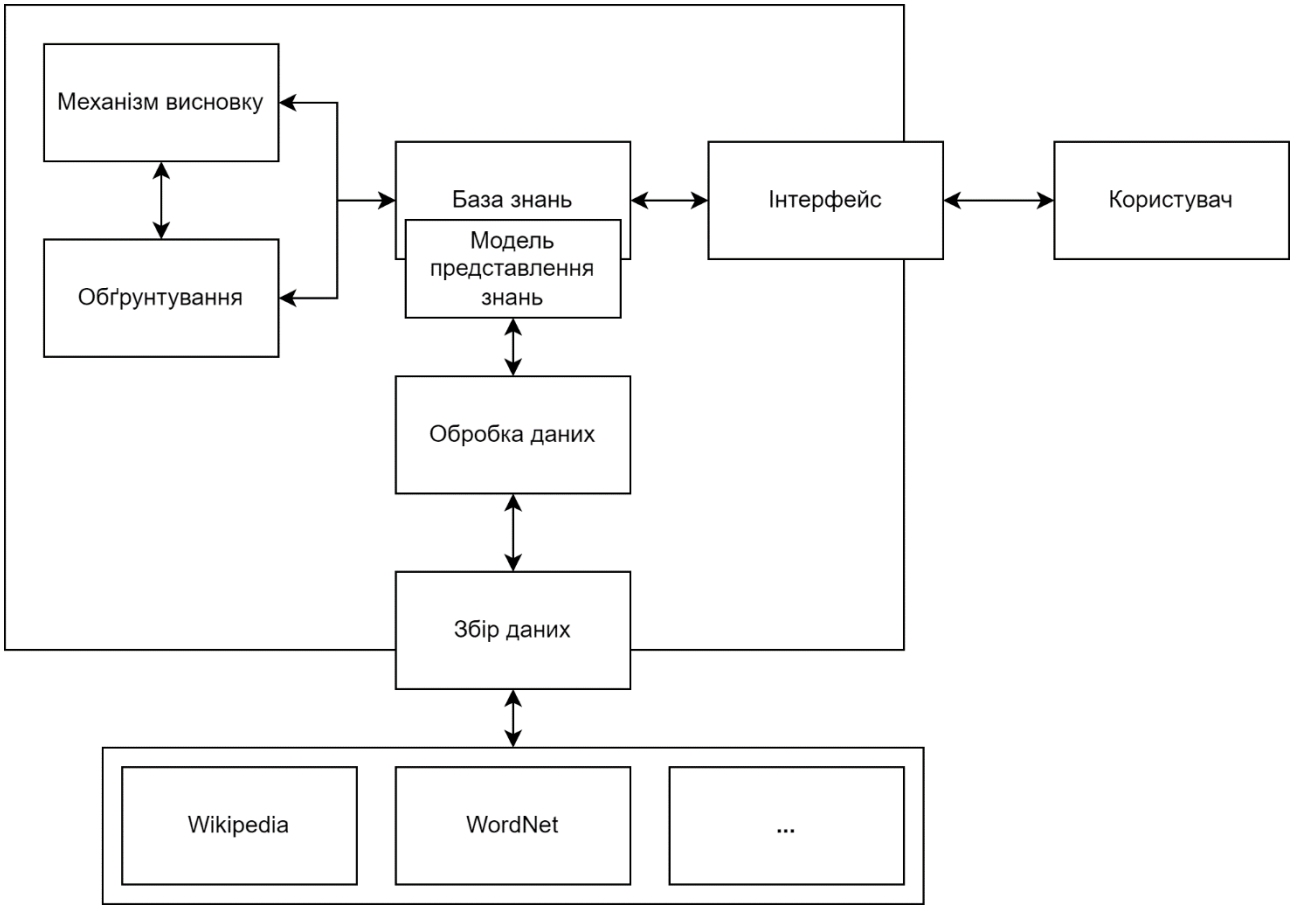
– обробка даних для представлення їх у базі знань;

– зберігання інформації;

– надання відомостей користувачам за запитами.

Інформаційні довідкові системи використовують бази знань.

СХЕМА ІНФОРМАЦІЙНОЇ ДОВІДКОВОЇ СИСТЕМИ



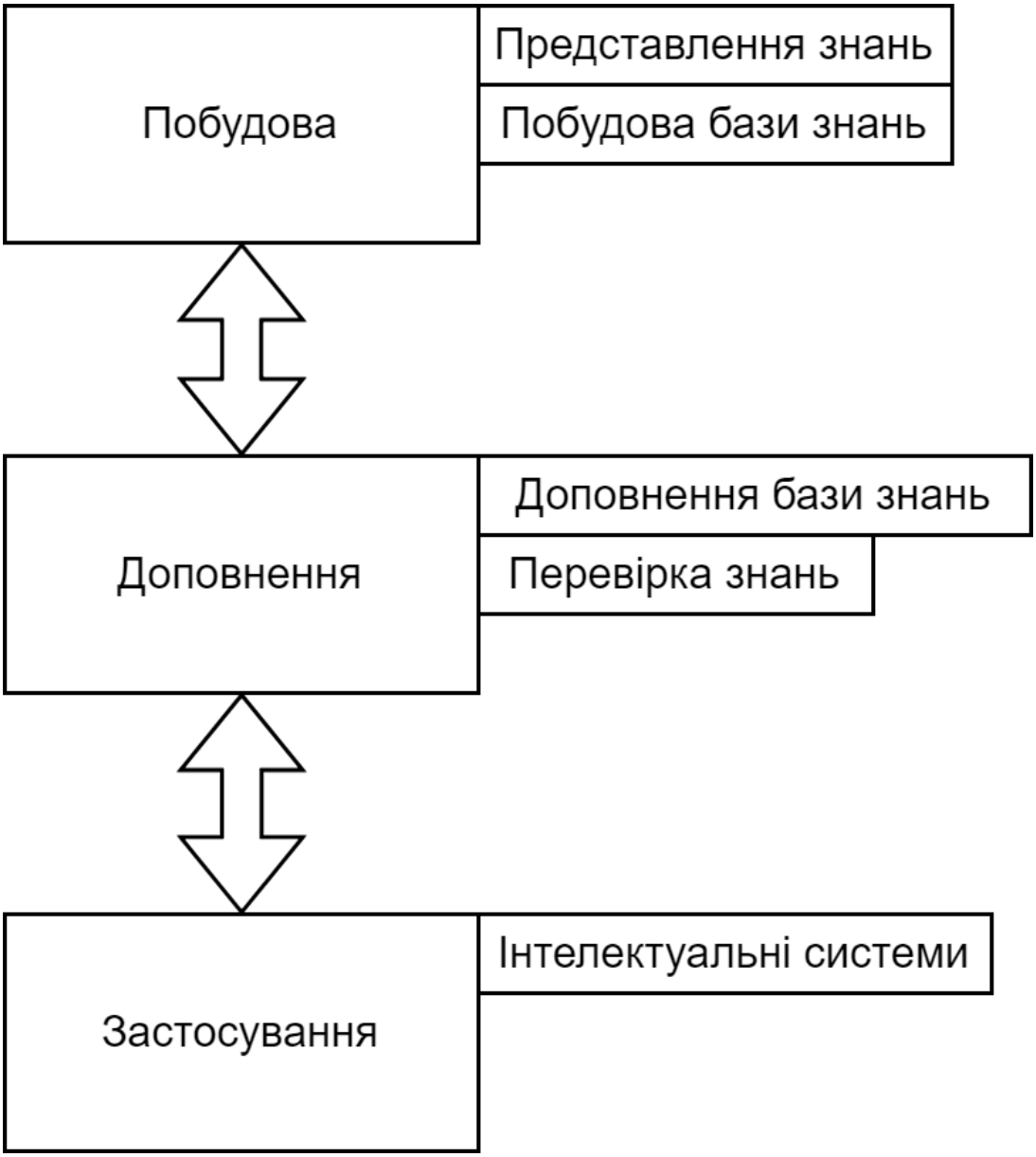
Елементи системи:

База знань містить мета-інформацію (класи сутностей, зв’язки між сутностями). Над інформацією з неї проводиться логічний висновок.

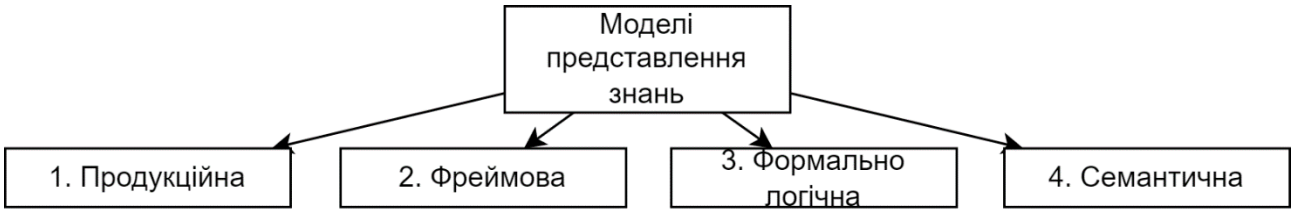
Механізм висновку автоматизовано виконує висновок зі знань. Він може синтезувати нові зв’язки, які явно не відображені в базі знань.

Для автоматизованої побудови бази знань виконується обробка даних зі зовнішніх джерел. Для цього вхідні дані розбиваються на елементи. між якими встановлюється зв’язок. Далі елементи зв’язують з існуючими знаннями.

ЗАДАЧІ АВТОМАТИЗОВАНОЇ ПОБУДОВИ БАЗ ЗАНАНЬ



МОДЕЛІ ПРЕДСТАВЛЕННЯ ЗНАНЬ



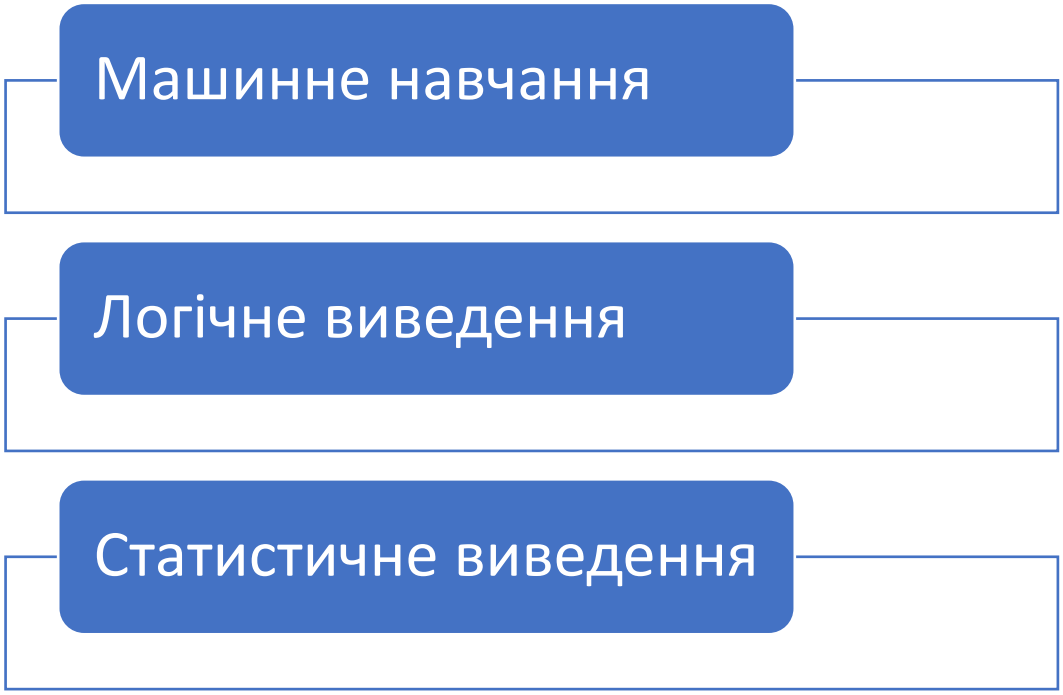
Продукційна модель впроваджує опис через правила. Він складається з умови та дії.

Фреймова модель вводить слоти, що є атрибутами об’єкта. До слотів зіставляються їх значення.

Формально-логічна модель впроваджує опис через правила, але, на відміну від продукційної моделі, вона такж описує зв’язки між сутностями.

Семантична модель описує знання через орієнтований граф, де вершини – поняття, а дуги – відношення між поняттями.

ПОБУДОВА ЗАНАНЬ

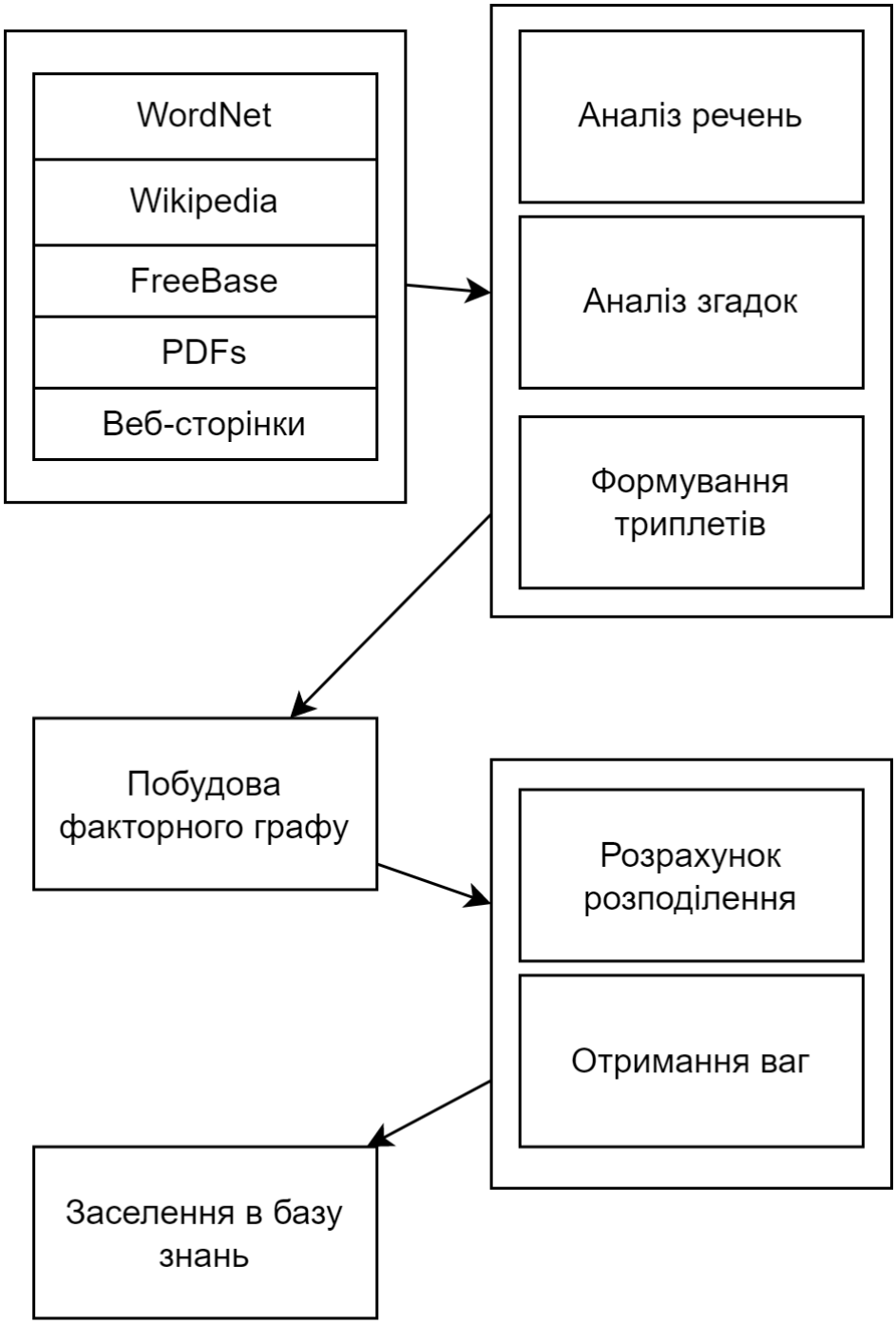


Методи машинного навчання використовують закономірності на великих масивах даних. Такі методи добре працюють з великим масивом даних, але складно виявити помилки, якщо вони трапляються.

Методи логічного виведення використовують дану множину правил та фактів для логічного висновку на основі зіставлення сутностей фактів та правил. У таких методах легше знаходити та виправляти помилки, але вони обмежені заданим набором правил.

Методи статистичного виведення використовують частотні показники фактів та їх зв’язків у множині. Такі методи працюють з великим масивом даних та знаходять найбільш імовірні факти як найбільш часті у тексті.

СХЕМА ТИПОВОГО ПРОЦЕСУ ПОБУДОВИ БАЗИ ЗНАНЬ ІНФОРМАЦІЙНО-ДОВІДКОВИХ СИСТЕМ

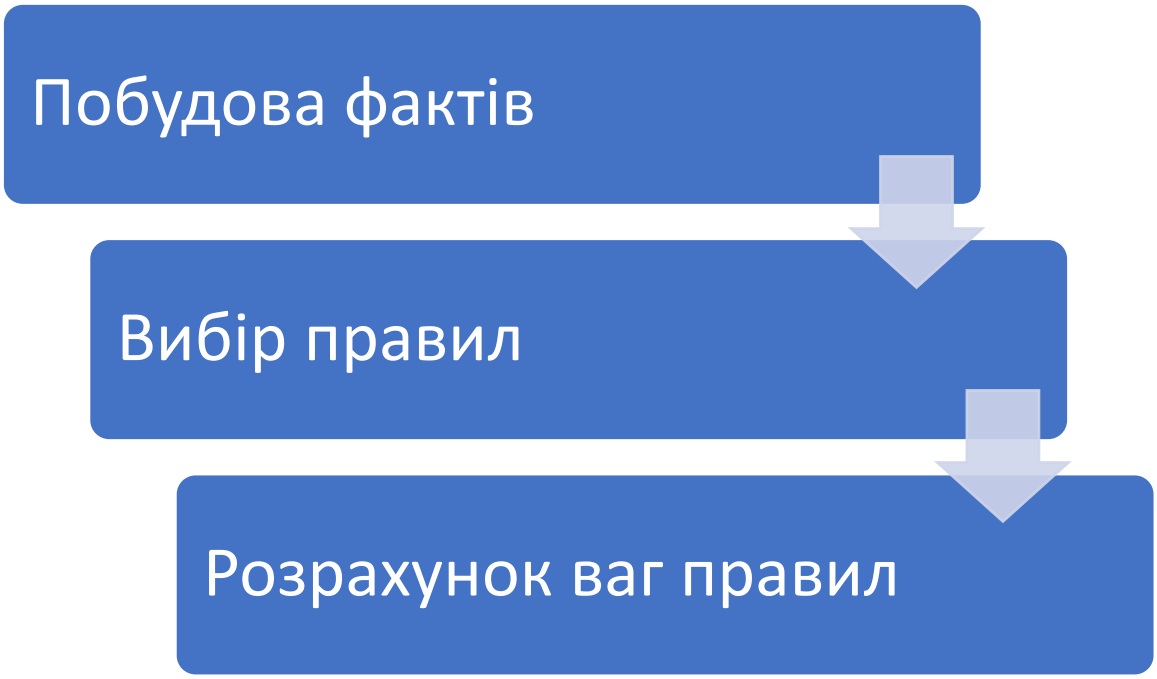


ПРОБЛЕМА ДОСЛІДЖЕННЯ

Проведене дослідження предметної області показало, що автоматизований збір та обробка інформації з великої кількості джерел є необхідною умовою для ефективного розвитку інтелекутальних систем.

Актуальність даної задачі є наслідком невідповідності можливостей існуючих методів автоматизованої побудови баз знань для інформаційно-довідкових систем, які використовують додаткові зовнішні бази типових правил, та практичними потребами до процесу автоматизованої побудови баз знань безпосереднього аналізу документів.

КЛЮЧОВІ ЕТАПИ МЕТОДУ АВТОМАТИЗОВАНОЇ ПОБУДОВИ БАЗ ЗНАНЬ З ВИКОРИСТАННЯМ ФАКТОРНОГО ГРАФУ



На першому етапі метод використовує системи вилучення інформації на основі аналізу текстів для отримання вхідних фактів.

На другому етапі метод відбирає правила із бази Sherlok.

На 3 етапі метод використовує SQL запити для формування факторного графу. Запити можна розбити на 2 категорії: ті, що формують множину фактів на основі виведення з правил, та ті, що формують матрицю, що описує факторний граф. Після побудови факторного графу розраховується розподілення за Гіббсом.

УДОСКОНАЛЕНИЙ МЕТОД ПОБУДОВИ БАЗ ЗНАНЬ

Етап 1 Побудова фактів за допомогою систем вилучення інформації на основі аналізу текстів.

Крок 1.1 Вибір документу

Крок 1.2 Побудова стеммів (стеммінг)

Результатом кроку є множина основ слів

$$\boldsymbol{Q} = \{\boldsymbol{q_i}, \boldsymbol{p_l}\}. \quad (1)$$

де p_l – це розділ тексту, q_i – це слово

Крок 1.3 Виділення триплетів (сутність, властивість, відношення)

$$\boldsymbol{T} = \{(\boldsymbol{o_i}, \boldsymbol{r_k}, \boldsymbol{s_h}) : \boldsymbol{o_i}, \boldsymbol{r_k}, \boldsymbol{s_h} \in \boldsymbol{Q}\}. \quad (2)$$

де $\boldsymbol{o_i}$ – це об’єкт, $\boldsymbol{r_k}$ – це відношення, $\boldsymbol{s_h}$ - це суб’єкт

Крок 1.4 Побудова фактів на основі виділених триплетів

$$F = \{f_j : f_j = (o_{j,i}, r_{j,k}, s_{j,h}), \forall (m \neq j), o_j, i \neq o_{m,i} \; or \; r_{j,k} \neq r_{m,k} \; or \; s_{j,h} \neq s_{m,h}\}. \quad (3)$$

Етап 2 Побудова правил, що визначають зв’язки між фактами.

Етап 3 Розрахунок ваг правил

Крок 3.1 Формування факторного графу

Крок 3.2 Розрахунок розподілення за методом Гіббсу

Крок 3.3 Формування зважених правил

Крок 3.4 Упорядкування правил за вагою

ЕТАП 2 УДОСКОНАЛЕНОГО МЕТОДУ ПОБУДОВИ БАЗ ЗНАНЬ

Етап 2 Побудова правил, що визначають зв'язки між фактами.

Крок 2.1 Вибір розділу документу (p_l)

Крок 2.2 Вибір фактів в рамках розділу документу.

$$F_l = \{ f_j : (o_{j,i}, r_{j,k}, s_{j,h}) \in p_l \}. \quad (4)$$

Узагальнене представлення правила можна визначити як:

$$G = \{ g_z(f_j, f_m) \}, \quad (5)$$

$$g_z(f_j, f_m) = (s_{j,i}, r_{j,k}, s_{m,h}). \quad (6)$$

Загальна кількість правил тоді буде визначатися як:

$$G = G^{(1)} \cup G^{(2)}. \quad (7)$$

Крок 2.3 Формування множини розміщень фактів. Множина розміщень по два та три елемента у кортежі буде мати вигляд:

$$A_1 = \{(f_j, f_m) \in p_l \times p_l\}, \quad (8)$$

$$A_2 = \{(f_j, f_m) \in p_l \times p_l\}, \quad (9)$$

$$A_3 = \{(f_j, f_m, f_n) \in p_l \times p_l\}. \quad (10)$$

де f_j, f_m, f_n є фактами, q_i є множиною усіх фактів з розділу документу.

Крок 2.4 Формування правил на основі розміщень фактів. Правила першого, другого та третього типів можна визначити так:

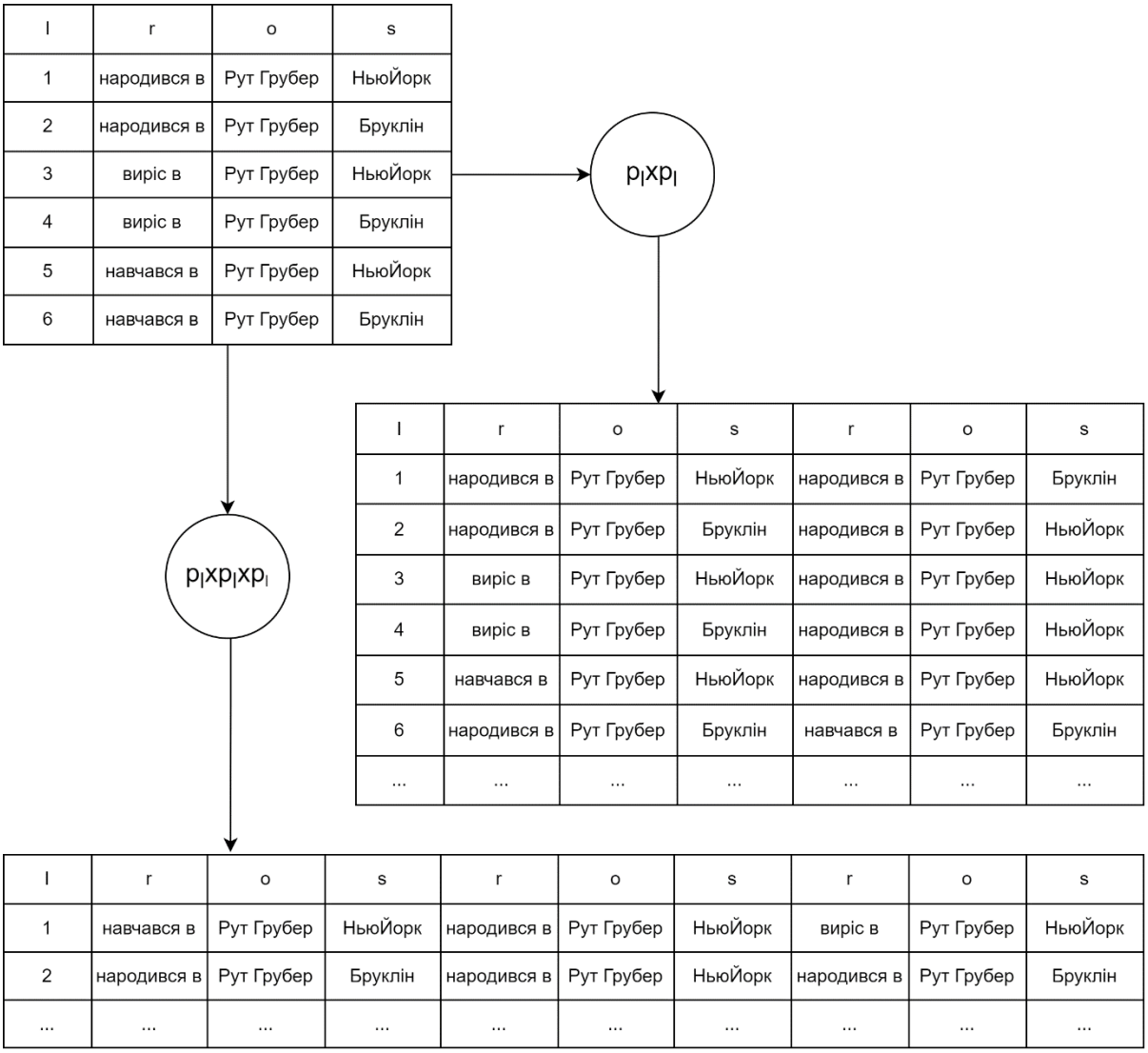
$$G^{(1)} = \{g_z^{(1)}(f_j, f_m) : o_{j,i} = o_{m,i}, r_{j,k} = r_{m,k}\}, \quad (11)$$

$$G^{(2)} = \{g_z^{(2)}(f_j, f_m) : o_{j,i} = o_{m,i}, r_{j,k} \neq r_{m,k}\}, \quad (12)$$

$$G^{(3)} = \{g_z^{(3)}(f_j, f_m, f_n) : o_{j,i} = o_{m,i}, o_{j,i} = o_{n,i}, s_{j,h} \neq s_{n,h}, s_{j,h} \neq s_{m,h}, r_{j,k} \neq r_{m,k} \neq r_{n,k}\}. \quad (13)$$

Ми говоримо, що якщо o_j співпадає з s_j, то таке правило можна вважати обмеженням і його вага буде ∞.

УДОСКОНАЛЕНИЙ МЕТОД ПОБУДОВИ БАЗ ЗНАНЬ

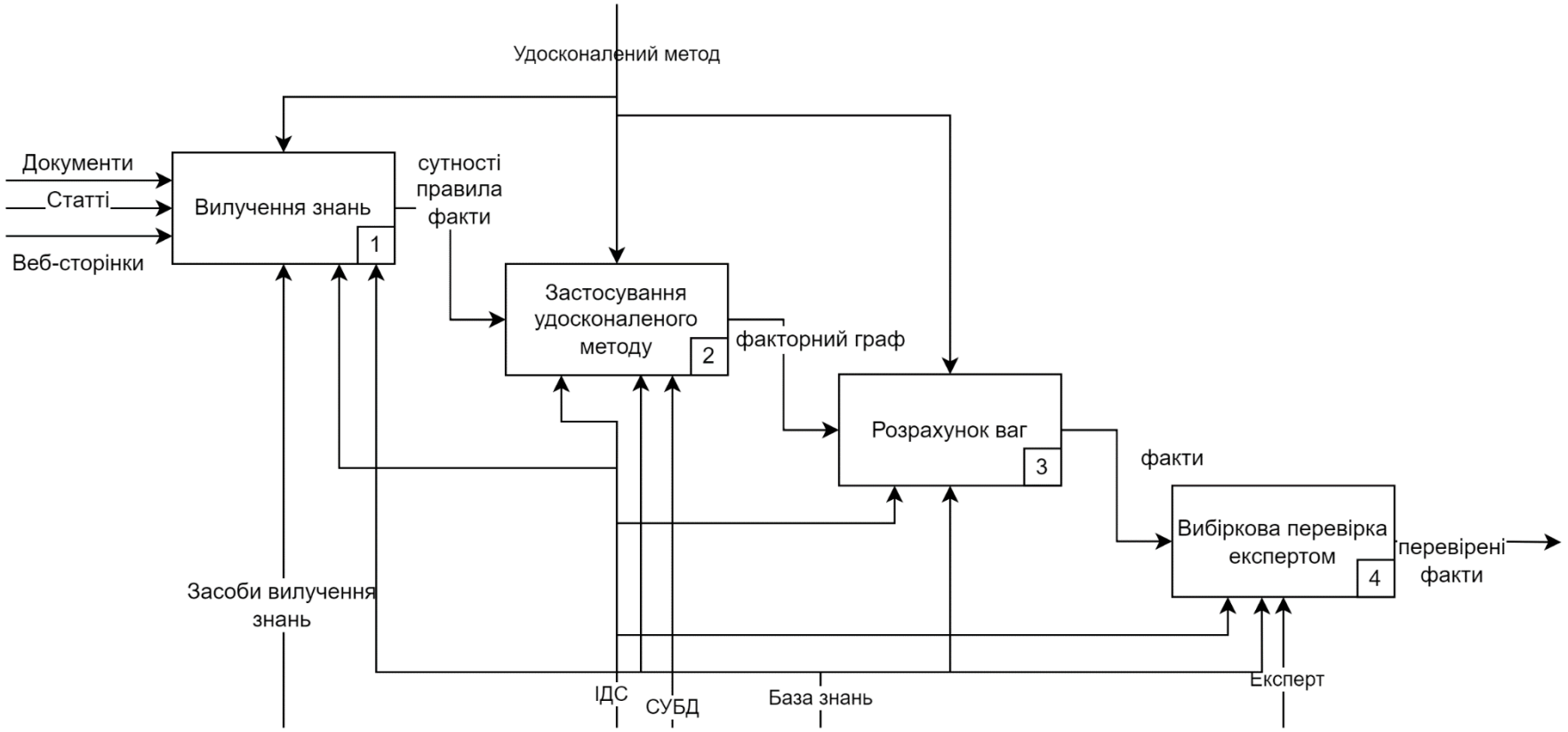


I	r	o	s
1	народився в	Рут Грубер	НьюЙорк
2	народився в	Рут Грубер	Бруклін
3	виріс в	Рут Грубер	НьюЙорк
4	виріс в	Рут Грубер	Бруклін
5	навчався в	Рут Грубер	НьюЙорк
6	навчався в	Рут Грубер	Бруклін

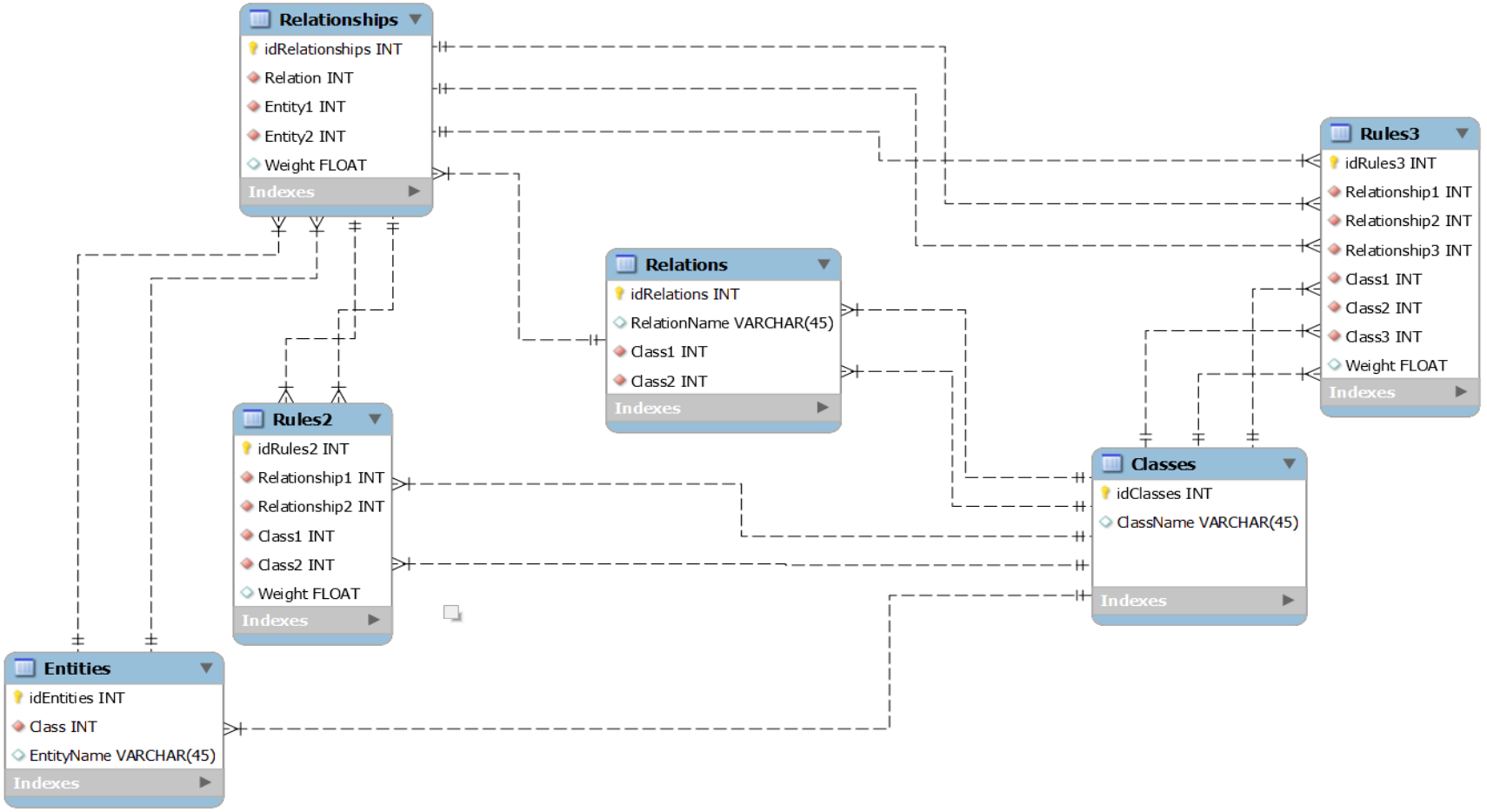
I	r	o	s	r	o	s
1	народився в	Рут Грубер	НьюЙорк	народився в	Рут Грубер	Бруклін
2	народився в	Рут Грубер	Бруклін	народився в	Рут Грубер	НьюЙорк
3	виріс в	Рут Грубер	НьюЙорк	народився в	Рут Грубер	НьюЙорк
4	виріс в	Рут Грубер	Бруклін	народився в	Рут Грубер	НьюЙорк
5	навчався в	Рут Грубер	НьюЙорк	народився в	Рут Грубер	НьюЙорк
6	народився в	Рут Грубер	Бруклін	навчався в	Рут Грубер	Бруклін
...	...	...	...	...	...	...

I	r	o	s	r	o	s	r	o	s
1	навчався в	Рут Грубер	НьюЙорк	народився в	Рут Грубер	НьюЙорк	виріс в	Рут Грубер	НьюЙорк
2	народився в	Рут Грубер	Бруклін	народився в	Рут Грубер	НьюЙорк	народився в	Рут Грубер	Бруклін
...	...	...	...	...	...	...	...	...	...

ТЕХНОЛОГІЯ АВТОМАТИЗОВАНОЇ ПОБУДОВИ БАЗ ЗНАНЬ ДЛЯ ІНФОРМАЦІЙНИХ ДОВІДКОВИХ СИСТЕМ



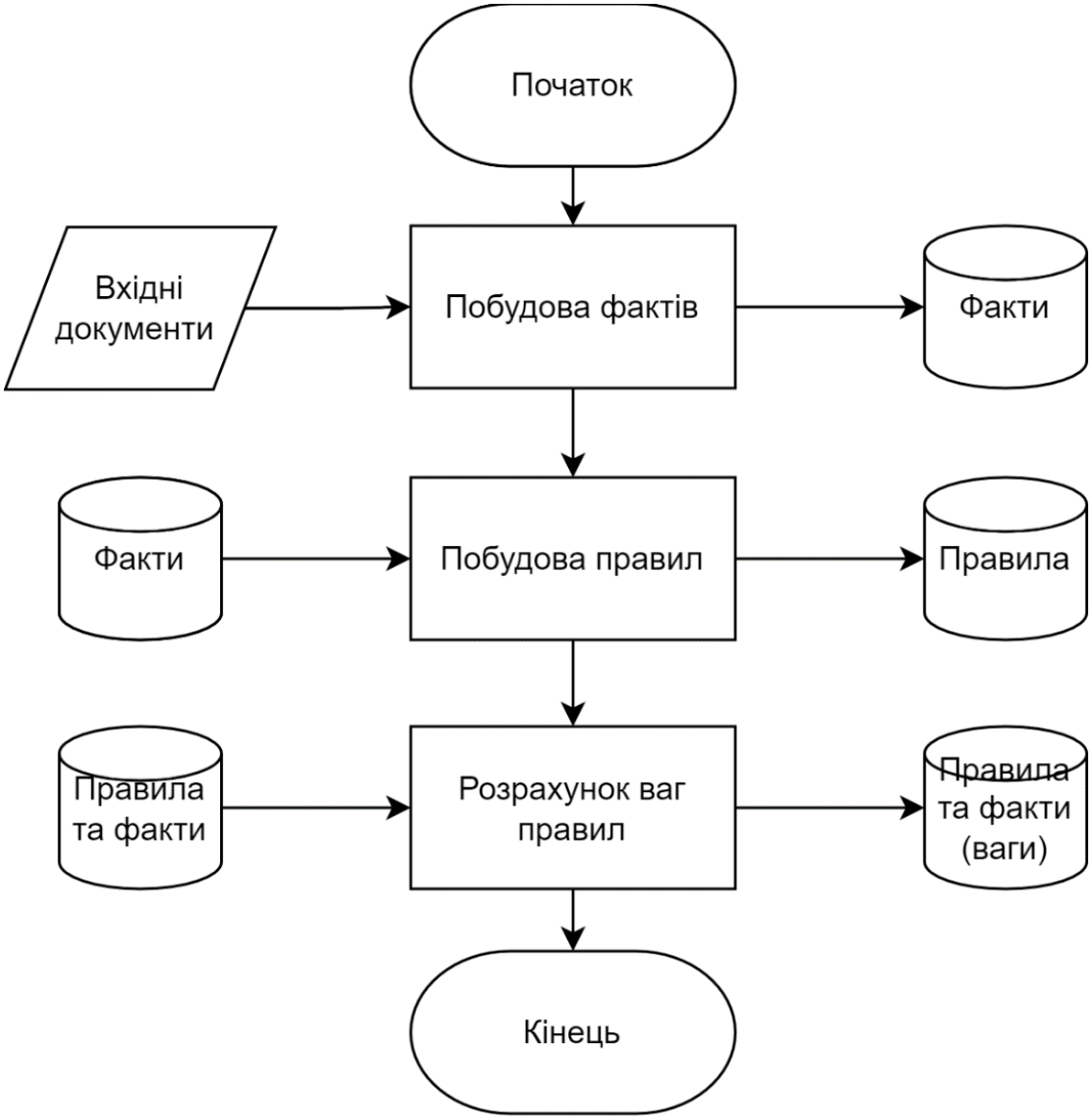
РЕЛЯЦІЙНА МОДЕЛЬ ДЛЯ ЗБЕРІГАННЯ БАЗИ ЗНАНЬ



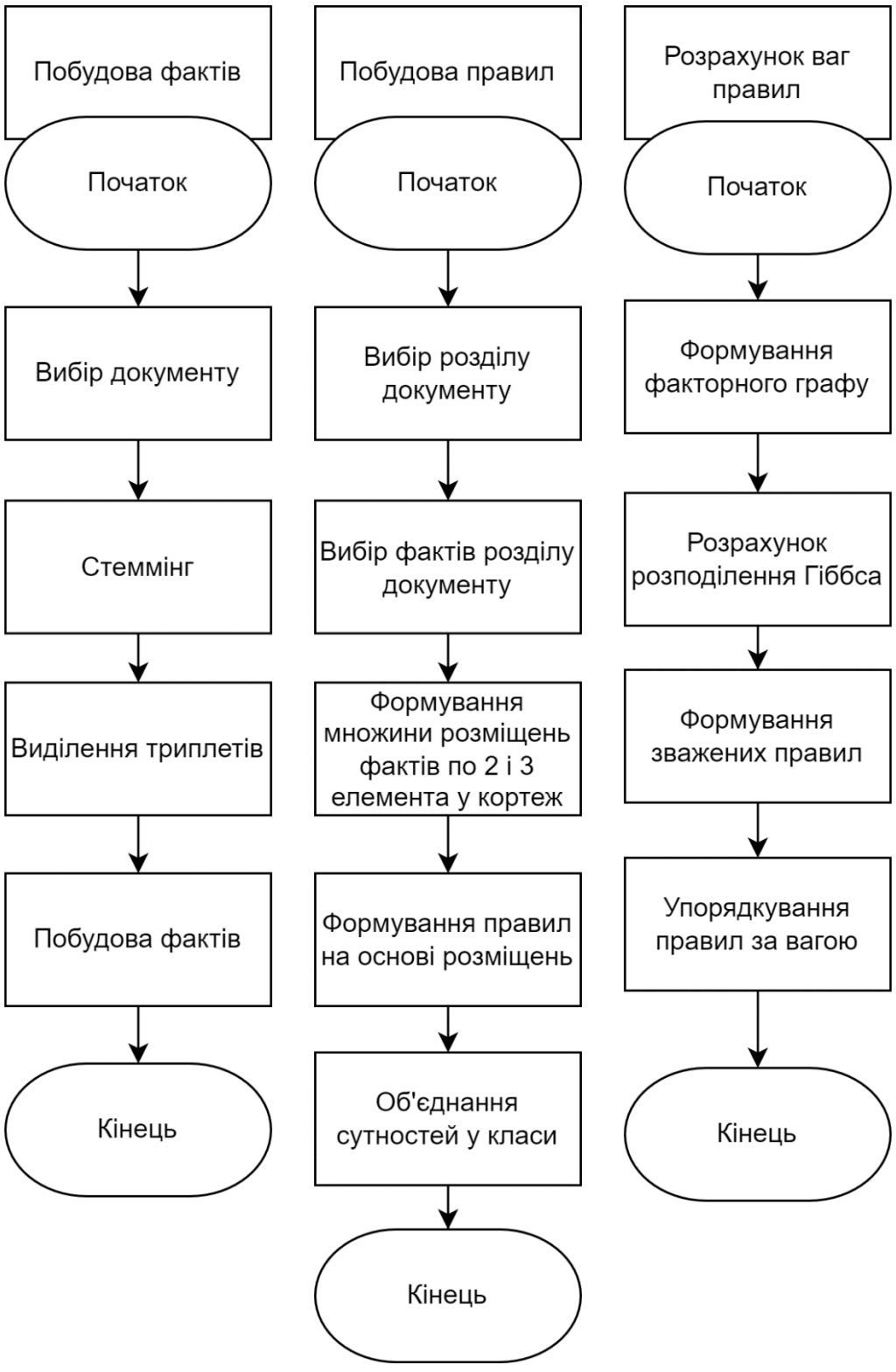
МОДЕЛЬ ПРЕДСТАВЛЕННЯ ЗНАНЬ

Сутності	Класи	Зв’язки	Взаємозв’язки
Рут Грубер	W (Письменник) = {Рут Грубер}	народився в(W, P), народився в(W, C)	народився в(Рут Грубер, Нью-Йорк)
Нью-Йорк	C (Місто) = {Нью-Йорк}	жив в(W, P), жив в(W, C)	народився в(Рут Грубер, Бруклін)
Бруклін	P (Місце) = {Бруклін}	знаходиться в(P, C)	
Правила			
∀x ∈ W ∀y ∈ P (жив в(x, y) ← народився в(x, y))			
∀x ∈ W ∀y ∈ C (жив в(x, y) ← народився в(x, y))			
∀x ∈ P ∀y ∈ C ∀z ∈ W (знаходиться в(x, y) ← жив в(z, x) ∧ жив в(z, y))			
∀x ∈ P ∀y ∈ C ∀z ∈ W (знаходиться в(x, y) ← народився в(z, x) ∧ народився в(z, y))			
∞ ∀x ∈ C ∀y ∈ C ∀z ∈ W (народився в(z, x) ∧ народився в(z, y) → x = y)			

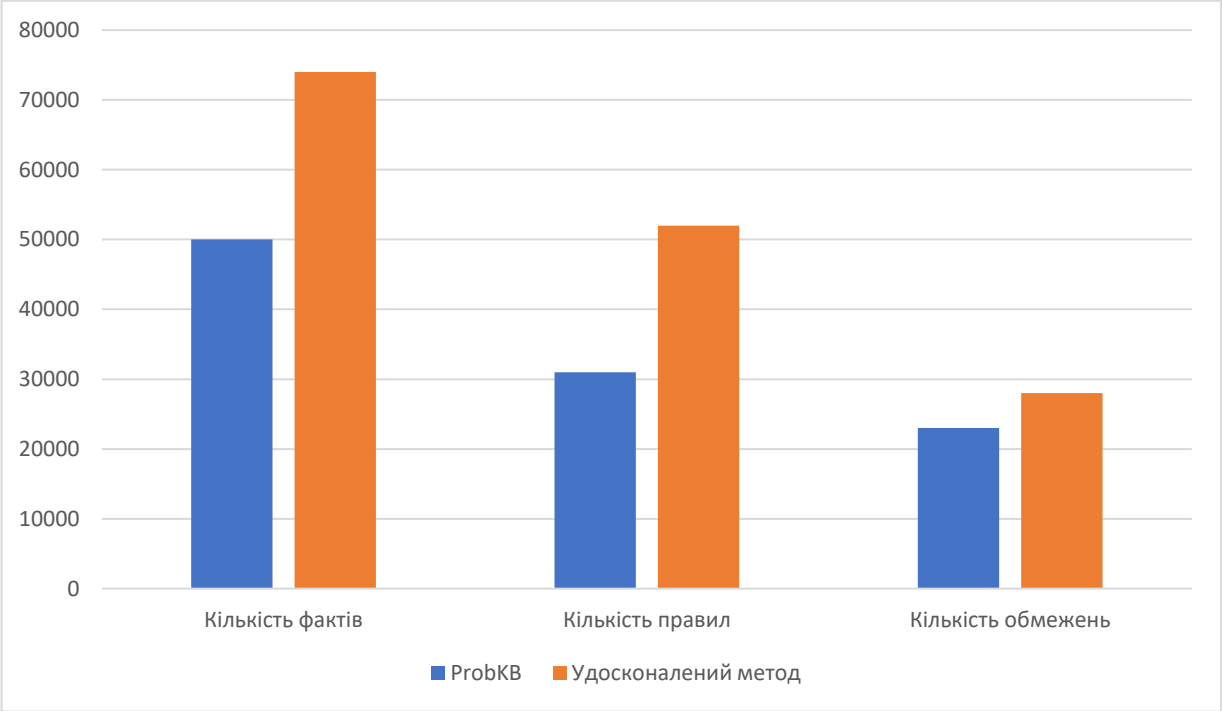
АЛГОРИТМ ПОБУДОВИ ПРАВИЛ



АЛГОРИТМ ПОБУДОВИ ПРАВИЛ



ЕКСПЕРИМЕНТАЛЬНА ПЕРЕВІРКА



ВИСНОВКИ

В результаті виконання кваліфікаційної роботи було проведено аналіз методів автоматизованої побудови бази знань для інформаційно-довідкових систем.

В результаті роботи було удосконалено метод автоматизованої побудови баз знань на основі зважених правил для інформаційно-довідкової системи.

Була проведена експериментальна перевірка та виконаний порівняльний аналіз системи з використанням удосконаленого методу та методу автоматизованої побудови баз знань на основі факторного графу. За результатами експерименту було виявлено, що удосконалений метод може виявляти більше правил, ніж використовується в базовому методі, а також може виводити більше фактів.

Результати магістерської роботи представлені у матеріалах 25-го міжнародного молодіжного форуму «Радіоелектроніка та молодь у XXI столітті» 2021р.