

Високопродуктивна довідково-інформаційна система на основі Wikipedia

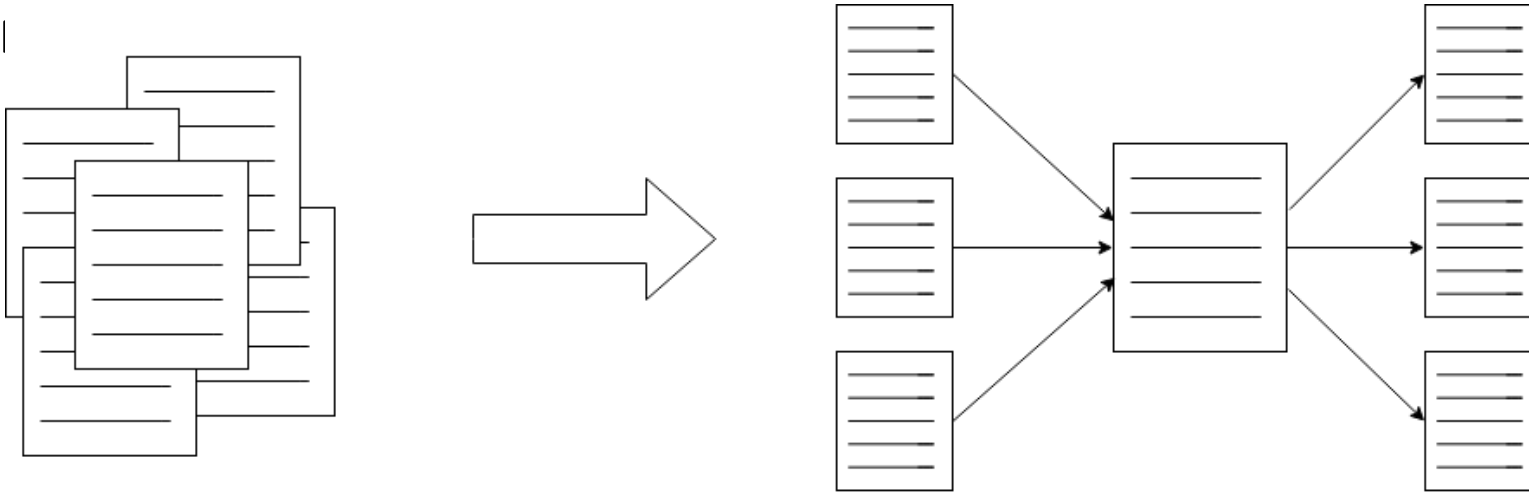
Автор: Ковальов Е.А.

Науковий керівник: к.т.н., доц. Титенко С.В.

Актуальність теми

Вікі-енциклопедії, зокрема Wikipedia, є дуже потужним засобом збереження знань людства та можуть бути використані для самоосвіти.

Проте вони не структурують інформацію за порядком вивчення. Таку проблему пропонується вирішити впорядковуючи статті енциклопедії за допомогою |



Мета роботи

Створити систему, що виконує побудову дидактичних зв'язків між статтями вікі-енциклопедій, забезпечивши високу швидкість її роботи.

У якості демонстрації результатів роботи, виконати у повному обсязі індексацію англomовної Вікіпедії як найбільшого у світі вікі-проекту.

Дидактичні зв'язки

Стаття, яку слід вивчити раніше, дидактично
передус тій, яку слід вивчити після неї:



Дидактичні зв'язки

Стаття, яку слід вивчити раніше, дидактично передує тій, яку слід вивчити після неї:



Етапи побудови зв'язків:



Множина зв'язаних з цільовою статтей

Нехай:

c_0 — цільова стаття.

$Linked(c_0)$ — множина таких статей c , що існує шлях зі статті c_0 в статтю c .

$Related(c_0)$ — множина таких статей c , що c_0 та c мають спільні категорії.

$DirectLinked(c_0)$ — множина таких статей c , що c_0 безпосередньо посилається на c .

Результатом буде:

$$Result = (Linked(c_0) \cap Related(c_0)) \cup DirectLinked(c_0)$$

Факти і правила дидактичного впорядкування

Стаття c_1 дидактично передуює статті c_2 якщо:

1. У вступі статті c_2 існує посилання на статтю c_1 .

$$(c_2, c_1) \in IntroLinks \rightarrow conceptBefore_{(c_1, c_2)} < CF = 0.4 >$$

2. В основній частині тексту статті c_2 існує посилання на статтю c_1 .

$$(c_2, c_1) \in MainContentLinks \rightarrow conceptBefore_{(c_1, c_2)} < CF = 0.2 * \left(1 - \frac{Position}{Length}\right) >$$

3. Назва статті c_2 містить у собі назву статті c_1 .

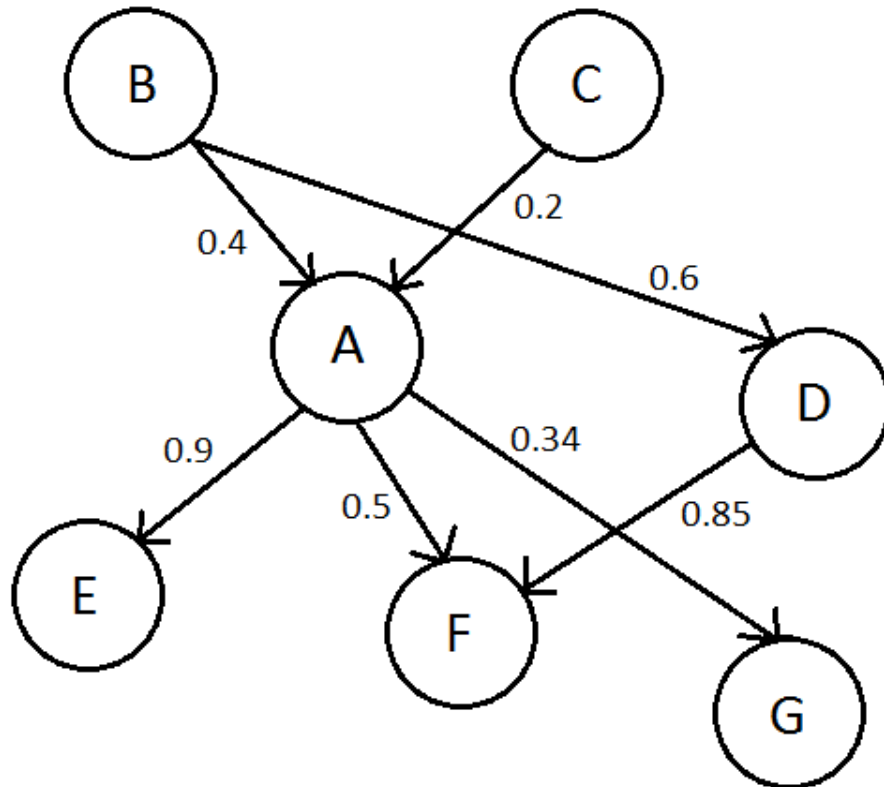
$$(c_2, c_1) \in TitleInclusions \rightarrow conceptBefore_{(c_1, c_2)} < CF = 0.9 >$$

4. Текст статті c_2 містить у собі назву статті c_1 .

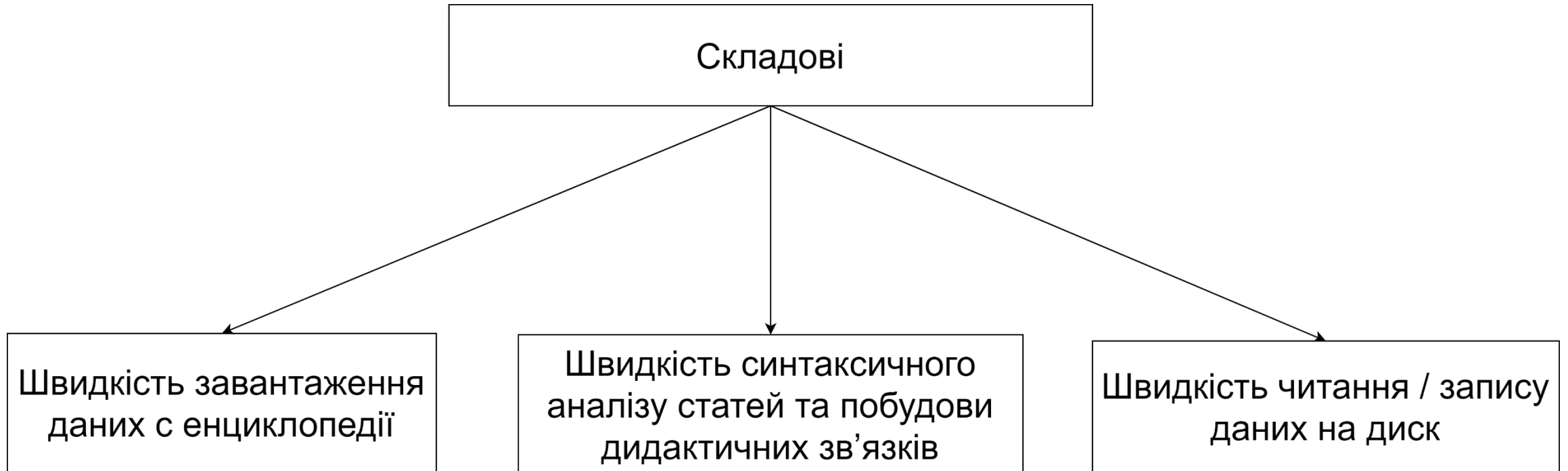
$$(c_2, c_1) \in TitleContentInclusions \rightarrow conceptBefore_{(c_1, c_2)} < CF = 0.1 * \left(1 - \frac{Position}{Length}\right) >$$

Кінцевий результат

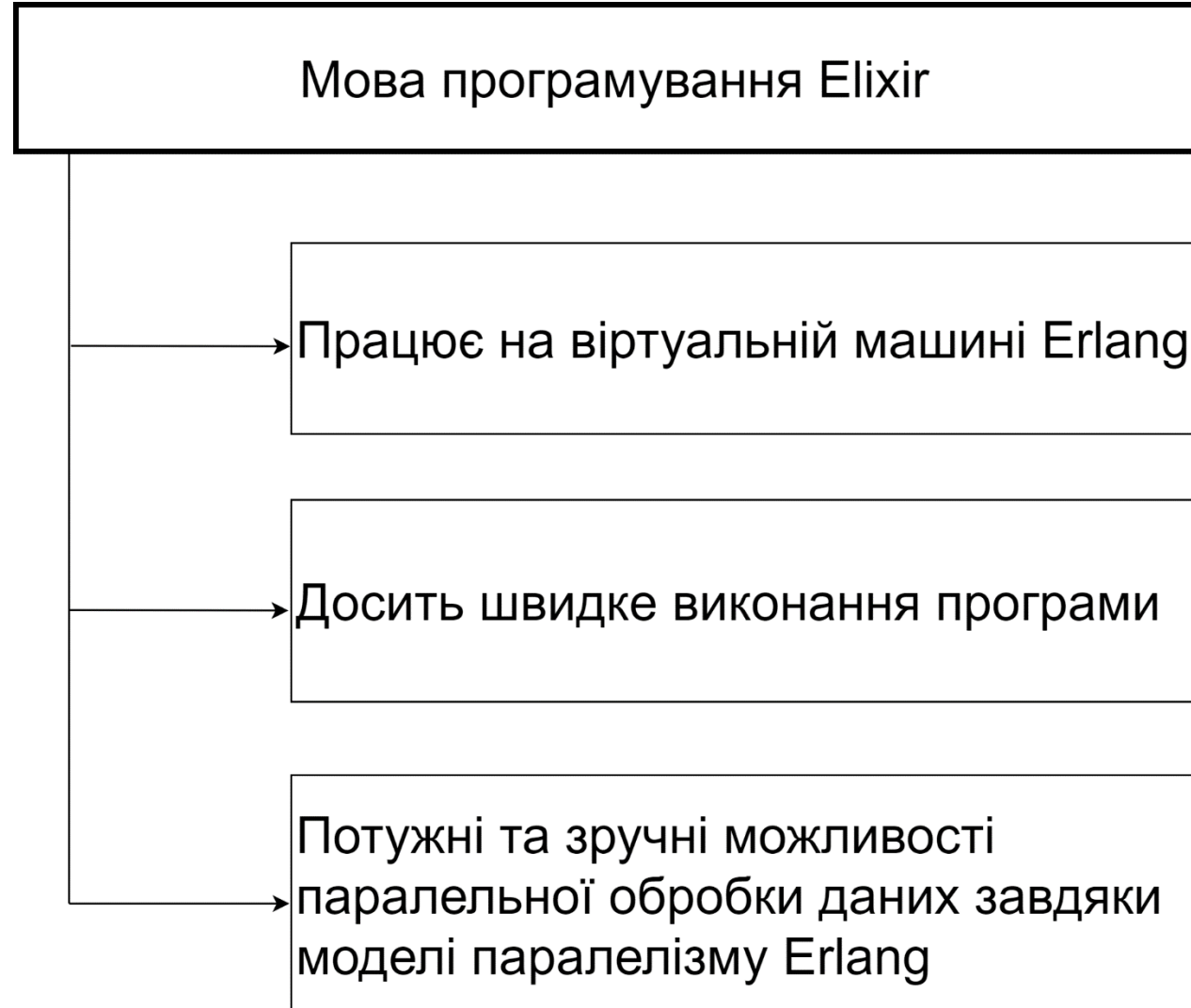
Результатом є зважений орієнтований граф, вершини якого – статті, а ребра – дидактичні зв'язки. Кожному зв'язку відповідає фактор впевненості у тому, що зв'язок існує.



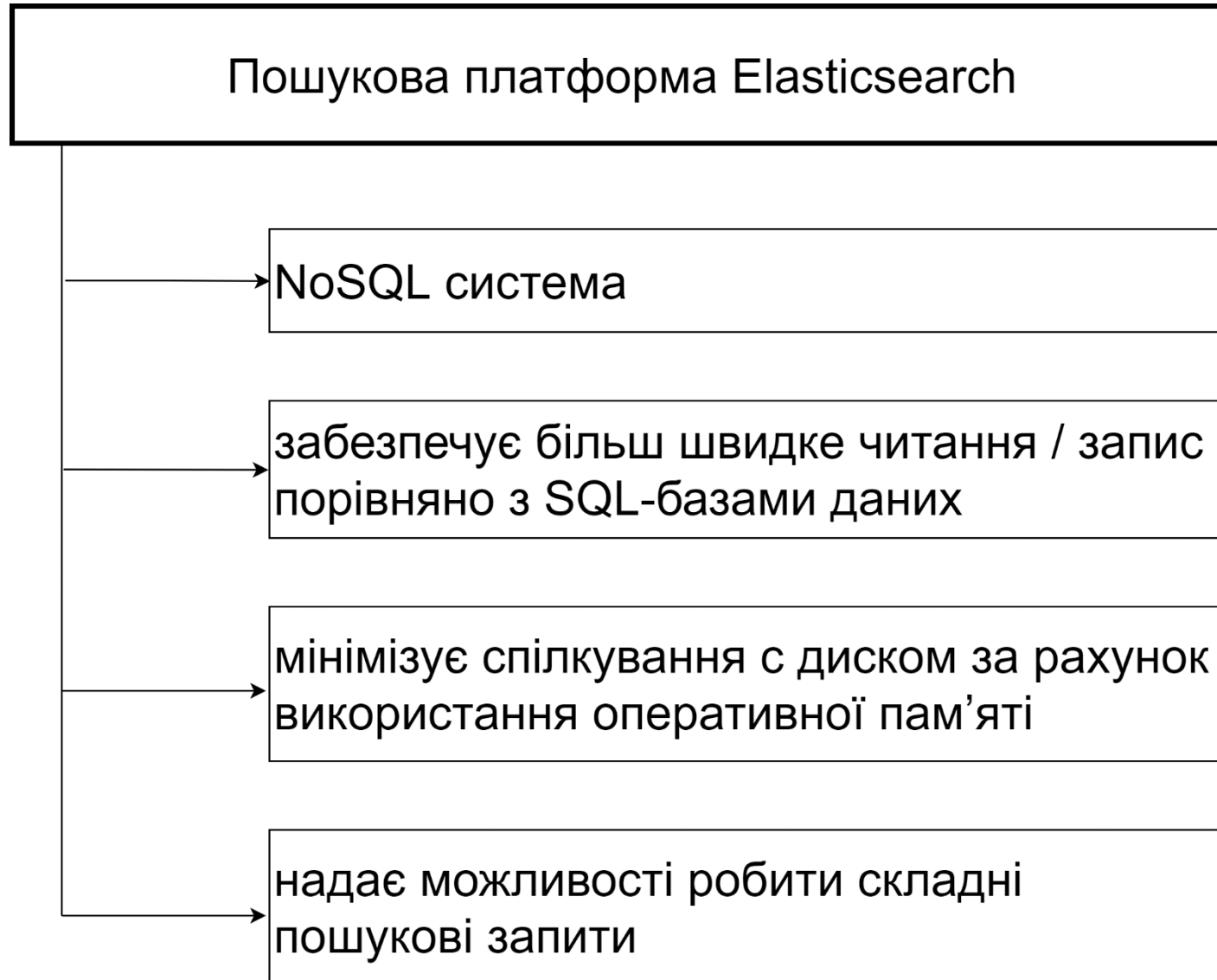
Швидкість роботи системи



Засоби розробки



Сховище даних



Недоліки:

- є дуже чутливим до апаратного забезпечення

Завантаження даних

- уповільнюється передачею даних по мережі
- можливо робити лише один запит до API енциклопедії одночасно
- є найбільш слабким місцем з точки зору швидкодії системи

Завантаження даних

- уповільнюється передачею даних по мережі
- можливо робити лише один запит до API енциклопедії одночасно
- є найбільш слабким місцем з точки зору швидкодії системи

Методи оптимізації:

- відокремлення даного етапу від інших

Завантаження даних

- уповільнюється передачею даних по мережі
- можливо робити лише один запит до API енциклопедії одночасно
- є найбільш слабким місцем з точки зору швидкодії системи

Методи оптимізації:

- відокремлення даного етапу від інших
- найбільш оптимальне використання API

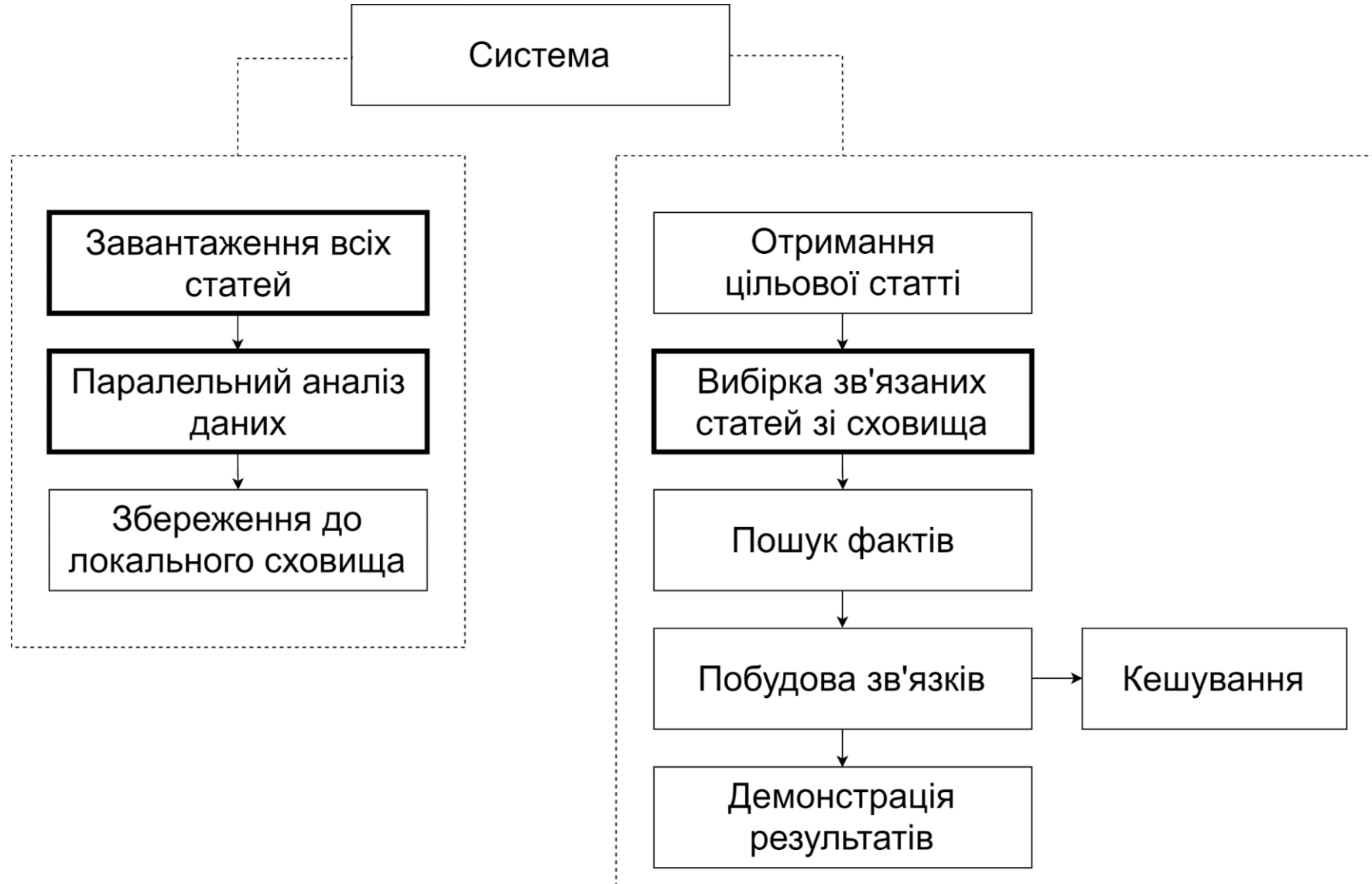
Завантаження даних

- уповільнюється передачею даних по мережі
- можливо робити лише один запит до API енциклопедії одночасно
- є найбільш слабким місцем з точки зору швидкодії системи

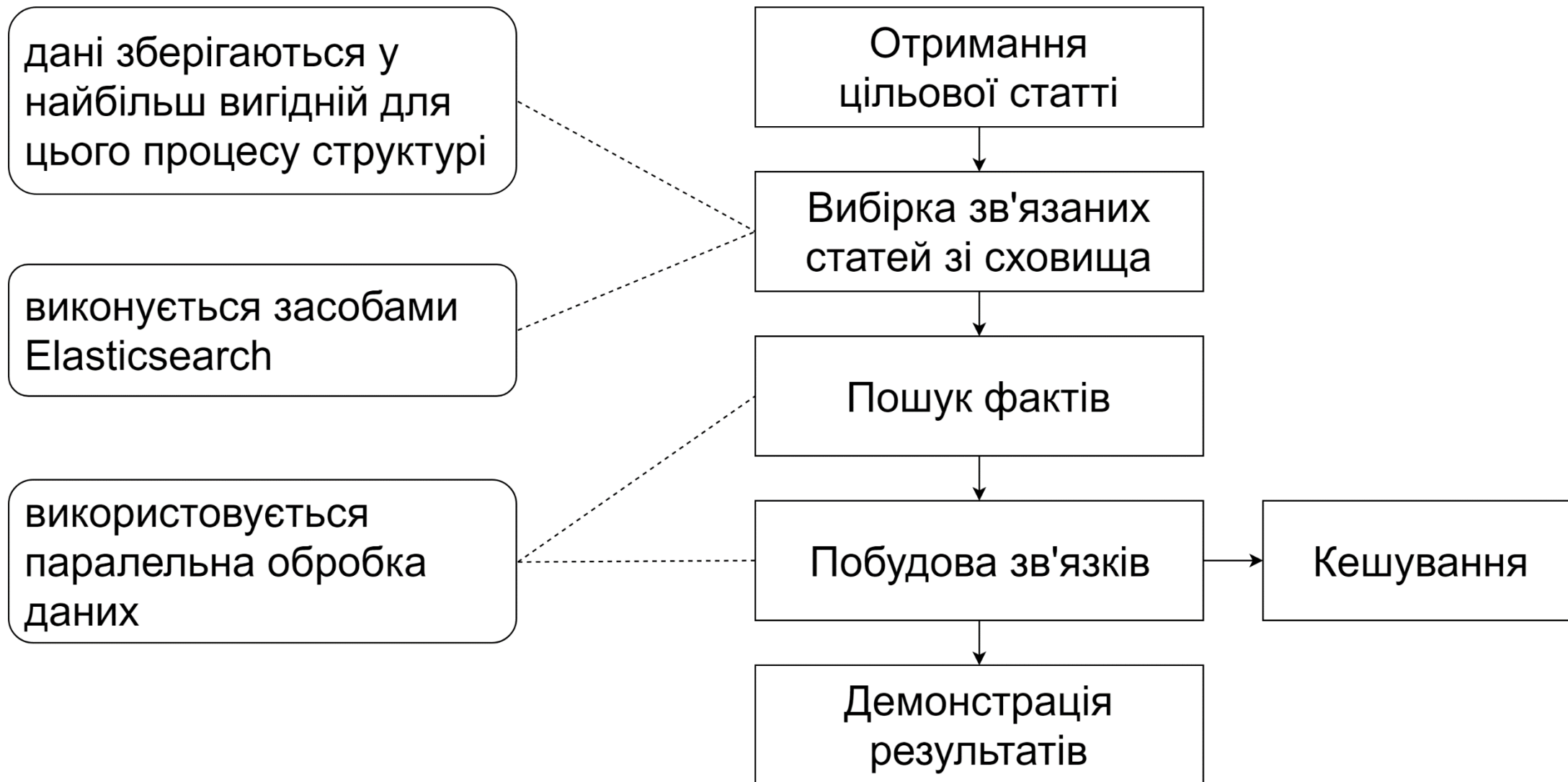
Методи оптимізації:

- відокремлення даного етапу від інших
- найбільш оптимальне використання API
- паралельний аналіз даних під час очікування на відповідь від API

Принцип роботи системи



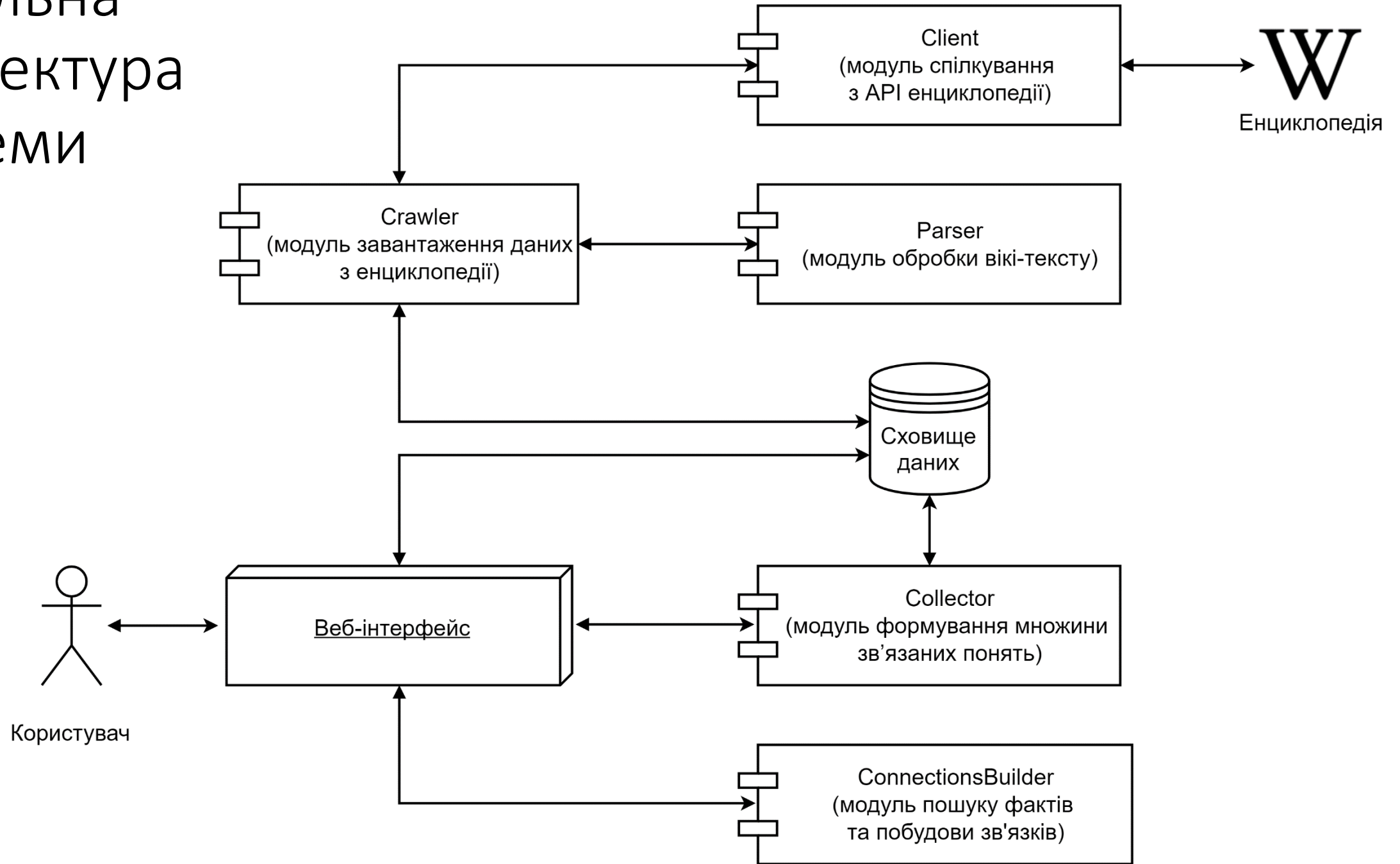
Побудова дидактичних зв'язків (жирним виделить)



Результуюча продуктивність системи

- повне завантаження та обробка англomовної Вікіпедії було виконане за близько 70 годин
- при наявності потужного апаратного забезпечення очікуваний середній час виконання етапу побудови зв'язків – 100 - 1000 мілісекунд

Детальна Архітектура системи



Результат

Високопродуктивна довідково-інформаційна система на основі Wikipedia

World War II

Вивчати

World War II

[Читати статтю](#)

Для попереднього ознайомлення:

World war (0.99)Вивчати

World War I (0.97)Вивчати

Soviet Union (0.96)Вивчати

Europe (0.96)Вивчати

United Nations (0.95)Вивчати

China (0.85)Вивчати

Asia (0.82)Вивчати

Germans (0.75)Вивчати

United Kingdom (0.73)Вивчати

Також читайте:

World War II by country (0.99)Вивчати

Strategic bombing during World War II (0.99)Вивчати

Outline of World War II (0.99)Вивчати

Home front during World War II (0.98)Вивчати

Diplomatic history of World War II (0.98)Вивчати

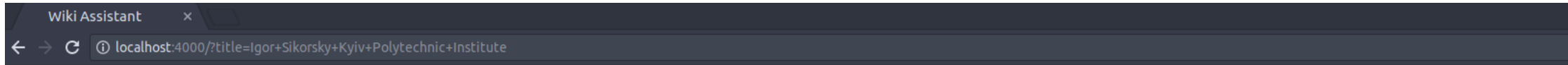
Air warfare of World War II (0.98)Вивчати

World War III (0.98)Вивчати

Allies of World War II (0.98)Вивчати

Romanian Navy during World War II (0.97)Вивчати

Результат



Високопродуктивна довідково-інформаційна система на основі Wikipedia

Igor Sikorsky Kyiv Polytechnic Institute

Вивчати

Igor Sikorsky Kyiv Polytechnic Institute

[Читати статтю](#)

Для попереднього ознайомлення:

[Kiev](#) (0.96)

Вивчати

[Ukraine](#) (0.96)

Вивчати

[Igor Sikorsky](#) (0.92)

Вивчати

[Engineer](#) (0.74)

Вивчати

[Webometrics Ranking of World Universities](#) (0.4)

Вивчати

Також читайте:

[Military Institute of Telecommunications and Information Technologies](#) (0.4)

Вивчати

[Kyiv National University of Construction and Architecture](#) (0.19)

Вивчати

[Sergei Korolev](#) (0.19)

Вивчати

[Vladimir Chelomey](#) (0.19)

Вивчати

[Valeriya Gontaryeva](#) (0.18)

Вивчати

[EPAM Systems](#) (0.32)

Вивчати

[Dmitri Mendeleev](#) (0.25)

Вивчати

[Nikolay Yegorovich Zhukovsky](#) (0.2)

Вивчати

[Stephen Timoshenko](#) (0.18)

Вивчати

[Evgeny Paton](#) (0.17)

Вивчати

[List of universities, colleges, and research institutions in Kiev](#) (0.12)

Вивчати

Висновки

- розроблено високопродуктивну систему, що виконує побудову дидактичних зв'язків між статтями вікі-енциклопедій для полегшення вивчення інформації
- оптимізовано процеси використання API енциклопедії, обробки та збереження даних
- виконано повну обробку англomовної Вікіпедії за прийнятний час
- забезпечено високу швидкість побудови дидактичних зв'язків за наявності потужного апаратного забезпечення

Дякую за увагу!