ДОДАТОК А

Перелік джерел посилання за науковими напрямами керівника та науковців кафедри програмної інженерії

1. Копаліні Диплом, URL: http://openarchive.nure.ua/bitstream/document/3272/1/KopalianiD.pdf (дата звернення: 18.04.2024).

2. Falatiuk H., Shirokopetleva M., Dudar Z. Investigation of Architecture and Technology Stack for e-Archive System. *2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S&T)*, Kyiv, Ukraine, 8–11 October 2019. 2019. URL: https://doi.org/10.1109/picst47496.2019.9061407 (дата звернення: 05.04.2024).

3. Мулєсап Диплом, URL: http://openarchive.nure.ua/bitstream/document/1168/1/MulesapPP.pdf (дата звернення: 22.06.2024).

ДОДАТОК Б

Звіт результатів перевірки на унікальність тексту в базі ХНУРЕ

UNICHECK
by Turnitin

Ім'я користувача:
Кардаш Євген Вікторович каф.ПІ

ID перевірки:
1016350549

Дата перевірки:
12.06.2024 06:54:35 EEST

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
12.06.2024 07:23:37 EEST

ID користувача:
100013622

Назва документа: **2024_М_ПІ_ІПЗм_22_6_Бєлінський_Г_А_скорочений**

Кількість сторінок: **43** Кількість слів: **5732** Кількість символів: **47153** Розмір файлу: **566.12 KB** ID файлу: **1016152766**

Виявлено модифікації тексту (можуть впливати на відсоток схожості)

4.64%
Схожість

Найбільша схожість: **1.74%** з Інтернет-джерелом (http://lingpipe-blog.com/author/mitzimorris)

3.86% Джерела з Інтернету	155	Сторінка 45
1.36% Джерела з Бібліотеки	16	Сторінка 46

0% Цитат

Вилучення цитат вимкнене

Вилучення списку бібліографічних посилань вимкнене

0%
Вилучень

Немає вилучених джерел

Модифікації

Виявлено модифікації тексту. Детальна інформація доступна в онлайн-звіті.

Замінені символи	2
Підозріле форматування	11 сторінок

ДОДАТОК В

Слайди презентації

Дослідження методів моделювання інформаційного пошуку за допомогою лінгвістичних автоматів.

Виконав:
ст. гр. ІПЗм-22-6 Бєлінський Г.А,
Керівник:
проф. каф. ПІ Шубін І.Ю.

1

Актуальність дослідження

- Лінгвістичні автомати (далі по тексту – ЛА) є ефективним інструментом для моделювання процесів інформаційного пошуку.
- ЛА можуть знаходити застосування у різних галузях, включаючи науку, техніку, медицину, соціологію та інші.
- Особливо актуальним є їх використання у веб-пошуку, де ефективність алгоритмів визначає зручність користувачів та результативність досліджень.

2

Аналіз предметної області

Мета роботи - розробка та аналіз ефективних методів моделювання інформаційного пошуку за допомогою лінгвістичних автоматів для покращення точності, релевантності та зручності пошукових систем.

Приклади Сервісів, Що Використовують Лінгвістичні Автомати в Інформаційному Пошуку:

- Google Search
- IBM Watson
- Amazon Comprehend
- Semrush
- ChatGPT (OpenAI)

3

Постановка задачі

- Огляд існуючих методів моделювання інформаційного пошуку, аналіз поточних підходів та алгоритмів, що використовуються у системах інформаційного пошуку
- Аналіз алгоритмів існуючих методів моделювання інформаційного пошуку.
- Реалізація найбільш популярних методів за допомогою стандартних технологій .NET та додатковими бібліотеками за необхідності.
- Проведення серії експериментів для оцінки ефективності методів та моделей для покращення у точності та релевантності пошукових результатів.
- Проведення екперименту зі спроби поєднання методів для досягнення переважних показників за рахунок модуляції методів реалізації ЛА.

4

Вибір методів для реалізації

Після порівняння різних методів моделювання інформаційного пошуку за допомогою лінгвістичних автоматів на платформі .NET було прийняте рішення реалізувати наступні методи для подальшого вивчення та тестування :

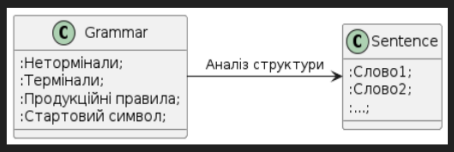
- – Метод формальних граматик
- – Метод кінцевих автоматів
- – Метод лінгвістичних правил та шаблонів

5

Формальні Граматики

Основні компоненти формальних граматик:

- Нетермінали (N) – абстрактні символи, які можуть бути розгорнуті в інші нетермінали або термінали.
- Термінали (Σ) – конкретні символи мови, які не можуть бути розгорнуті далі.
- Стартовий символ (S) – спеціальний нетермінал, з якого починається розгортання граматики.
- Продукційні правила (P) – правила виду A→β, де A – нетермінал, а β – послідовність терміналів і/або нетерміналів.
- Формальна граматика визначається як четвірка G=(N,Σ,P,S)G = (N, Σ, P, S)G=(N,Σ,P,S).



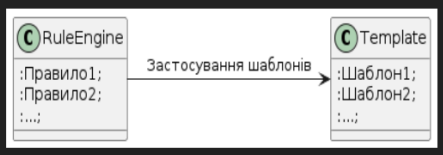
Структура контекстно-вільної граматики

6

Кінцеві Автомати

Основні компоненти кінцевих автоматів:

- Становище (Q) – множина станів, у яких може перебувати автомат.
- Алфавіт (Σ) – множина символів, які автомат може обробляти.
- Перехідна функція (δ) – функція, яка визначає наступний стан для кожної пари (поточний стан, вхідний символ).
- Початковий стан (q0) – стан, з якого автомат починає роботу.
- Множина прийнятних станів (F) – стани, у яких автомат приймає вхідний ланцюжок.
- Формально, ДКА визначається як п'ятірка M=(Q,Σ,δ,q0,F)M = (Q, Σ, δ, q0, F)M=(Q,Σ,δ,q0,F).



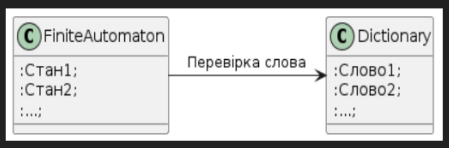
Структура кінцевого автомату

7

Лінгвістичні правила та шаблони

Приклади лінгвістичних правил та шаблонів:

- Синтаксичні шаблони: визначення фразових структур, наприклад, "іменник + дієслово".
- Семантичні шаблони: визначення семантичних зв'язків, наприклад, "іменник + прикметник", де прикметник описує іменник.
- Приклад шаблону:
- <np> ::= <det> <adj>* <noun>
- <vp> ::= <verb> <np>
- Цей шаблон описує фразову структуру простих речень.



Структура шаблонів

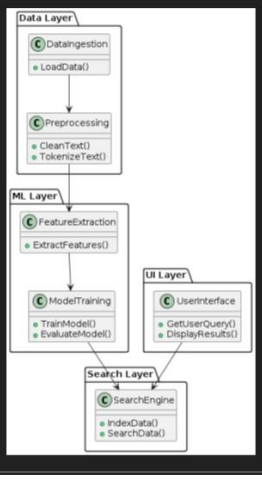
8

Засоби реалізації

Для реалізації програми було обрано мову програмування C#, вона має необхідні інструменти для реалізації визначених методів а також є популярним компонентом сучасних програмних систем.

Для випробування методів достатнім рівнем інтерфейсу користувача буде консольний інтерфейс.

Діаграма архітектури додатку на C# :



9

Данні для аналізу

Набір даний для випробування лінгвістичних автоматів був створений в форматі CSV:

text_data.csv | Program.cs

```
"Evolutionary biology studies the processes that produce the diversity of life on Earth",Science
"Renewable energy sources are essential for sustainable development",Technology
"Geology is the study of Earth's history, structure, and composition",Science
"Mobile computing is the use of computing devices while on the move",Technology
"Mathematics is the study of numbers, quantities, and shapes",Science
"Quantum computing is a revolutionary approach to computation",Technology
"Psychology is the scientific study of behavior and mental processes",Science
"Cloud computing provides on-demand access to computing resources over the internet",Technology
"Zoology is the scientific study of animals",Science
"Augmented reality overlays digital information onto the real world",Technology
"Ecology is the study of the relationships between organisms and their environments",Science
"Information technology involves the use of computers to store, retrieve, transmit, and manipulate data",Technology
"Anthropology is the study of human societies and cultures",Science
"Network security protects computer networks from unauthorized access or damage",Technology
"Oceanography is the study of the ocean and its phenomena",Science
"Computer vision enables computers to interpret and understand visual information",Technology
"Social science encompasses the study of human society and social relationships",Science
"Biotechnology applies biological systems, organisms, or derivatives to develop products or processes",Technology
"Geography examines the Earth's surface and its features",Science
"Artificial life studies life and lifelike processes through the use of computer models",Technology
"Statistics involves the collection, analysis, interpretation, and presentation of data",Science
"Information retrieval is the process of obtaining information from a collection of sources",Technology
"Political science studies the theory and practice of politics and government",Science
"Natural language processing enables computers to understand, interpret, and generate human language",Technology
"Artificial intelligence continues to advance",Technology
"Cognitive science investigates the mind and its processes",Science
"Emerging technologies shape the future",Technology
"Research in neuroscience leads to new discoveries about the brain",Science
"Advanced robotics improves efficiency and automation",Technology
"Space exploration expands our understanding of the universe",Science
"Predictive analytics uses data to forecast future trends",Technology
"Biochemistry examines the chemical processes within and related to living organisms",Science
"Internet technologies revolutionize communication",Technology
"Ecological conservation efforts protect biodiversity",Science
"Quantum mechanics revolutionizes our understanding of the microscopic world",Science
"Data mining uncovers valuable insights from large datasets",Technology
"Human genetics research advances medical treatments",Science
"Cloud-based services offer scalable and flexible computing solutions",Technology
"Evolutionary psychology explores the evolutionary basis of human behavior",Science
"Virtual assistants use artificial intelligence to perform tasks",Technology
"Marine biology studies marine organisms and ecosystems",Science
"E-commerce platforms revolutionize retail",Technology
"Anthropological research sheds light on human cultural diversity",Science
"Digital security measures protect against cyber threats",Technology
"Environmental conservation efforts promote sustainability",Science
"Blockchain technology ensures secure and transparent transactions",Technology
```

Код завантаження набору даних в програму:

```
var data = LoadData(dataPath);

// Method 1: Formal Grammars
var formalGrammarsResults = RunMethodMultipleTimes(() => RunFormalGrammars(data));
Console.WriteLine($"Formal Grammars - Time: {formalGrammarsResults.Item1}ticks, Accuracy:

// Method 2: Finite State Automata
var finiteAutomataResults = RunMethodMultipleTimes(() => RunFiniteAutomata(data));
Console.WriteLine($"Finite Automata - Time: {finiteAutomataResults.Item1}ticks, Accuracy:

// Method 3: Linguistic Rules and Patterns
var linguisticRulesResults = RunMethodMultipleTimes(() => RunLinguisticRules(data));
Console.WriteLine($"Linguistic Rules - Time: {linguisticRulesResults.Item1}ticks, Accuracy
```

10

Тестування методу формальних граматик

Фрагмент коду методу:

```
bool isScience = Regex.IsMatch(item.Text, @"\b(science|interdisciplinary|field|biology|chemistry|physics|natural science)\b", RegexOptions.IgnoreCase);
bool isTechnology = Regex.IsMatch(item.Text, @"\b(artificial intelligence|technology|subset|blockchain|cybersecurity|big data)\b", RegexOptions.IgnoreCase);

string predictedLabel = isScience ? "Science" : isTechnology ? "Technology" : "Unknown";

if (predictedLabel == item.Label)
{
    correct++;
    if (predictedLabel == "Science")
    {
        truePositives++;
    }
}
else
{
    if (predictedLabel == "Science")
    {
        falsePositives++;
    }
    else
    {
        falseNegatives++;
    }
}
```

Метрики результату тестування:

```
Formal Grammars - Time: 178352ticks, Accuracy: 0,2967032967032967, Precision:
 0,75, Recall: 0,20270270270270271, F1-score: 0,3191489361702128
```

11

Тестування методу кінцевих автоматів

Фрагмент коду методу:

```
var watch = Stopwatch.StartNew();
int correct = 0;
int truePositives = 0;
int falsePositives = 0;
int falseNegatives = 0;

foreach (var item in data)
{
    bool isScience = false;
    var words = item.Text.ToLower().Split(' ');
    foreach (var word in words)
    {
        if (word == "science" || word == "biology" || word == "chemistry" || word == "physics")
        {
            isScience = true;
            break;
        }
    }

    string predictedLabel = isScience ? "Science" : "Technology";
```

Метрики результату тестування:

```
Finite Automata - Time: 28398ticks, Accuracy: 0,5714285714285714, Precision:
0,8, Recall: 0,25, F1-score: 0,38095238095238093
```

12

Тестування методу лінгвістичних правил

Фрагмент коду методу:

```
foreach (var item in data)
{
    // Simple linguistic rules for demonstration purposes
    bool isScience = Regex.IsMatch(item.Text, @"\b(biology|chemistry|physics|science|field|natural)\b", RegexOptions.IgnoreCase);
    bool isTechnology = Regex.IsMatch(item.Text, @"\b(technology|ai|artificial intelligence|blockchain|cybersecurity|data)\b", RegexOptions.IgnoreCase);

    string predictedLabel = isScience ? "Science" : isTechnology ? "Technology" : "Unknown";

    if (predictedLabel == item.Label)
    {
        correct++;
        if (predictedLabel == "Science")
        {
            truePositives++;
        }
    }
    else
    {
        if (predictedLabel == "Science")
        {
            falsePositives++;
        }
        else
        {
            falseNegatives++;
        }
    }
}
```

Метрики результату тестування:

```
Linguistic Rules - Time: 7688ticks, Accuracy: 0,31868131868131866, Precision:
 0,7142857142857143, Recall: 0,2112676056338028, F1-score: 0,3260869565217391
```

13

Аналіз тестування

Таблиця, яка зображує результативні данні тестування методів:

Метод	Час виконання (ticks)	Точність (Accuracy)	Точність (Precision)	Повнота (Recall)	F1-міра (F1-score)
Формальні граматики	23847	0.297	0.750	0.203	0.319
Кінцеві автомати	28973	0.571	0.800	0.250	0.381
Лінгвістичні правила	6020	0.319	0.714	0.211	0.326

Скріншот виводу роботи програми:

```
Formal Grammars - Time: 178352ticks, Accuracy: 0,2967032967032967, Precision:
 0,75, Recall: 0,20270270270270271, F1-score: 0,3191489361702128
Finite Automata - Time: 28398ticks, Accuracy: 0,5714285714285714, Precision:
0,8, Recall: 0,25, F1-score: 0,3809523809523093
Linguistic Rules - Time: 7688ticks, Accuracy: 0,31868131868131866, Precision:
 0,7142857142857143, Recall: 0,2112676056338028, F1-score: 0,3260869565217391
```

14

тоду лінгвістични...

Аналіз тестування

З огляду на таблицю виводів можна зробити наступні висновки:

Метод кінцевих автоматів є найкращим методом з огляду на точність, точність (precision), повноту (recall) та F1-міру, хоча він дещо повільніший за інші методи.

Якщо швидкість виконання є критичною, метод лінгвістичних правил може бути прийнятним вибором, оскільки він найшвидший і має прийнятні показники точності та F1-score.

Для завдань, де точність і збалансованість важливіші за час виконання, метод кінцевих автоматів буде найбільш ефективним вибором.

15

Похідний гібридний метод

Після аналізу роботи та логіки методів було винайдено можливість для комбінації модулів методів : основна ідея полягає в тому, що спочатку використовуються формальні граматики для попередньої ідентифікації ключових слів та фраз, а потім результати уточнюються за допомогою кінцевих автоматів:

$$\begin{cases} \text{Science} & \text{якщо } (\text{FG}(T) = \text{Science}) \text{ або } (\text{FA}(T) = \text{Science}) \\ \text{Technology} & \text{якщо } (\text{FG}(T) = \text{Technology}) \text{і } (\text{FA}(T) \neq \text{Science}) \\ \text{Unknown} & \text{в іншому випадку} \end{cases}$$

де:

- T — текст, який потрібно класифікувати,
- $\text{FG}(T)$ — результат класифікації за допомогою формальних граматик,
- $\text{FA}(T)$ — результат класифікації за допомогою кінцевих автоматів.

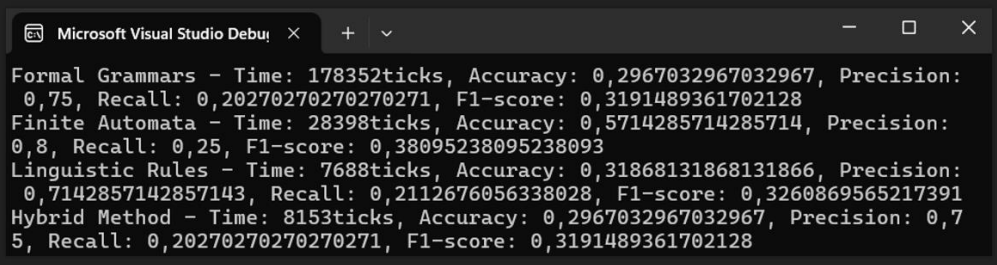
Фрагмент коду гібридного методу:

```
var watch = Stopwatch.StartNew();
int correct = 0;
int truePositives = 0;
int falsePositives = 0;
int falseNegatives = 0;

foreach (var item in data)
{
    // Step 1: Formal Grammars
    bool isScienceGrammar = Regex.IsMatch(item.Text, @"\b(science|interdisciplinary|field|biology|chemistry|physics|natural science)
    bool isTechnologyGrammar = Regex.IsMatch(item.Text, @"\b(artificial intelligence|technology|subset|blockchain|cybersecurity|big

    // Step 2: Finite State Automata for refinement
    bool isScienceAutomata = false;
    var words = item.Text.ToLower().Split(' ');
    foreach (var word in words)
    {
        if (word == "science" || word == "biology" || word == "chemistry" || word == "physics")
        {
            isScienceAutomata = true;
            break;
        }
    }

    // Combine results
    string predictedLabel = (isScienceGrammar || isScienceAutomata) ? "Science" : (isTechnologyGrammar ? "Technology" : "Unknown");
```

16

Оцінка результатів гібридого методу

Повний вивід метрик результатів усіх наявних методів:



З огляду на дані в порівнянні з іншими методами – можна зробити висновок про переваги та недоліки гібридного методу:

17

Висновок про продуктивність методу

Гібридний метод, поєднуючи формальні граматики та кінцеві автомати, пропонує збалансований підхід до класифікації тексту, забезпечуючи високу точність при помірних витратах часу на виконання.

Це робить його ефективним вибором для завдань, де потрібна висока якість класифікації, але залишає швидкодію.

Метод	Час виконання (ticks)	Точність (Accuracy)	Точність (Precision)	Повнота (Recall)	F1-міра (F1-score)
Формальні граматики	23847	0.297	0.750	0.203	0.319
Кінцеві автомати	28973	0.571	0.800	0.250	0.381
Лінгвістичні правила	6020	0.319	0.714	0.211	0.326
Гібридний метод	8565	0.297	0.750	0.203	0.319

18

Висновки

В кваліфікаційній роботі представлено результати, що відповідають меті дослідження, а саме – дослідження методів моделювання інформаційного пошуку за допомогою лінгвістичних автоматів. Було проаналізовано предметну галузь та існуючі підходи до вирішення задачі інформаційного пошуку за допомогою лінгвістичних автоматів. Згідно до поставленої мети кваліфікаційної роботи було виконано основні етапи розробки моделі лінгвістичного автомату для вирішення задач інформаційного пошуку.

Проведене дослідження та розробка гібридного методу підтвердили його ефективність та доцільність застосування у сучасних системах інформаційного пошуку. Він дозволяє поєднати кращі риси різних підходів, забезпечуючи високу точність, ефективність та швидкість виконання, що є ключовими факторами для успішної реалізації систем інформаційного пошуку в різних галузях.

19

ДОДАТОК Г

Експертний висновок результатів перевірки кваліфікаційної роботи на відповідність оформлення вимогам ДСТУ 3008: 2015

Експертний висновок результатів перевірки кваліфікаційної роботи

студент (посада) | програмної інженерії (кафедра) | ІПЗм-22-6 (група)

Бєлінський Георгій Андрійович

(прізвище, ім'я, по батькові)

Зауваження

Пункт ДСТУ 3008-2015	Зміст пункту	Сторінка кваліфікаційної роботи
1	2	3
	7.1 Загальні положення	
	7.3 Нумерація сторінок звіту	
	7.4 Нумерація розділів, підрозділів, пунктів, підпунктів	
	7.5 Рисунки	
	7.6 Таблиці	
	7.7 Переліки	
	7.8 Примітки	
	7.9 Виноски	
	7.10 Формули та рівняння	
	7.11 Посилання	
	7.13 Список авторів	
	7.14 Скорочення та умовні познаки	
	7.15 Додатки	

зауважень немає

Експерт _____ (підпис) Олена ОЛІЙНИК (прізвище, ініціали)

14.06.2024