ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ІНДИВІДУАЛЬНОГО РАНЖУВАННЯ ТЕКСТІВ НА ОСНОВІ ІНТЕЛЕКТУАЛЬНОГО МОДУЛЯ КАТЕГОРИЗАЦІЇ

Хрущак В.В. 1КН-14мі
Науковий керівник:
к.т.н., професор Месюра В.І.

АКТУАЛЬНІСТЬ

АКТУАЛЬНІСТЬ

Щоб уникнути "потопу" від інформації потрібно автоматично обробляти та фільтрувати інформаційний потік за допомогою комп’ютерних можливостей. Одним з напрямків, який може допомогти у пошуку та відборі корисної інформації – це рекомендаційні системи.

АКТУАЛЬНІСТЬ

Інша проблема полягає у тому, що більшість текстової інформації в мережі не має чітко сформованої теми, «хмар тегів», тощо. Тому постає ще одна задача – створення рекомендаційної системи, яка, за допомогою алгоритмів діставання інформації, змоглаб ранжувати такі тексти написані природною мовою.

МЕТА

Підвищити показники ефективності розв’язання задачі індивідуального ранжування за рахунок розробки модифікованих методів рекомендацій, що використовують алгоритми добування даних для аналізу текстів написаних мовою.

ЗАВДАННЯ

- провести аналіз проблеми розв’язання задачі індивідуального ранжування текстів;
- розглянути існуючі способи вирішення задачі індивідуального ранжування текстів;
- розробити модифіковані методи індивідуального ранжування текстів за допомогою використання інтелектуального модулю категоризації;
- розробити програмний засіб, що реалізує інформаційну технологію індивідуального ранжування текстів на основі інтелектуального модуля категоризації.

ОБ’ЄКТ, ПРЕДМЕТ ТА МЕТОДИ ДОСЛІДЖЕННЯ

Об’єкт: процес індивідуального ранжування текстів.

Предмет: метод індивідуального ранжування на основі модулю категоризації.

Методи:

- методи аналізу;
- методи моделювання;
- методи об’єктно-орієнтованого проектування.

НАУКОВА НОВИЗНА

- ✓ дістав подальшого розвитку модифікований метод фільтрації вмісту на основі алгоритму латентно-семантичного аналізу;
- ✓ удосконалено модель системи інтелектуального модуля категоризації, що дозволяє використовувати його, як універсальний інструмент аналізу схожості текстів, написаних природною мовою;
- ✓ вперше розроблено модель модифікованого методу фільтрації вмісту для розв’язання задачі індивідуального ранжування текстів за рахунок використання інтелектуального модулю категоризації, що використовує алгоритми добування даних, а саме метод латентно-семантичного аналізу для кластеризації та порівняння текстів написаних природною мовою.

ПРАКТИЧНЕ ЗНАЧЕННЯ

Розроблено новий, модифікований алгоритм роботи програмної реалізації системи індивідуального ранжування за рахунок чого досягається підвищення швидкості ранжування і відповідності результатів очікуванням. Розроблений алгоритм, на відміну від інших не залежить від об’ємів бази даних, або кількості відомої інформації про текст, що є об’єктом ранжування.

АНАЛІЗ ІСНУЮЧИХ МЕТОДІВ

- Фільтрація вмісту
- Колаборативні методи

Недоліки:

- Проблема "Холодного старту"
- Проблема "Напівавтоматичної фільтрації«
- Ігнорування семантики

МЕТОД LSA

Переваги:

- найкращий для виявлення латентних зв’язків;
- частково вирішує проблеми полісемії та омонімії;
- можна застосовувати, як з навчанням, так і без нього;
- використовує матрицю використань, що заснована на частотних характеристиках лексичних одиниць.

Недоліки:

- повільна швидкість обчислень;
- проблема полісемії і синонімії.

МОДЕЛЬ МЕТОДУ LSA

1. Матриця використань

$$t_i^T \rightarrow \begin{matrix} & d_j \\ & \downarrow \end{matrix} \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{bmatrix}$$

2. Розкладання матриці використань

$$A[u_1 \dots u_n] = [v_1, \dots, v_m] \begin{bmatrix} w_1 & & \\ & \ddots & \\ & & w_r \end{bmatrix} \Longrightarrow AU = VW,$$

або

$$A = VWU^*$$

3. Апроксимація вихідної матриці (пониження рангу)

$$\|A - B\|_2 \geq \|(A - B)z\|_2 = \|Az\|_2 = \sqrt{\sum_{l=1}^{k+1} |\alpha_l|^2 \, w_l} \geq w_{k+1}.$$

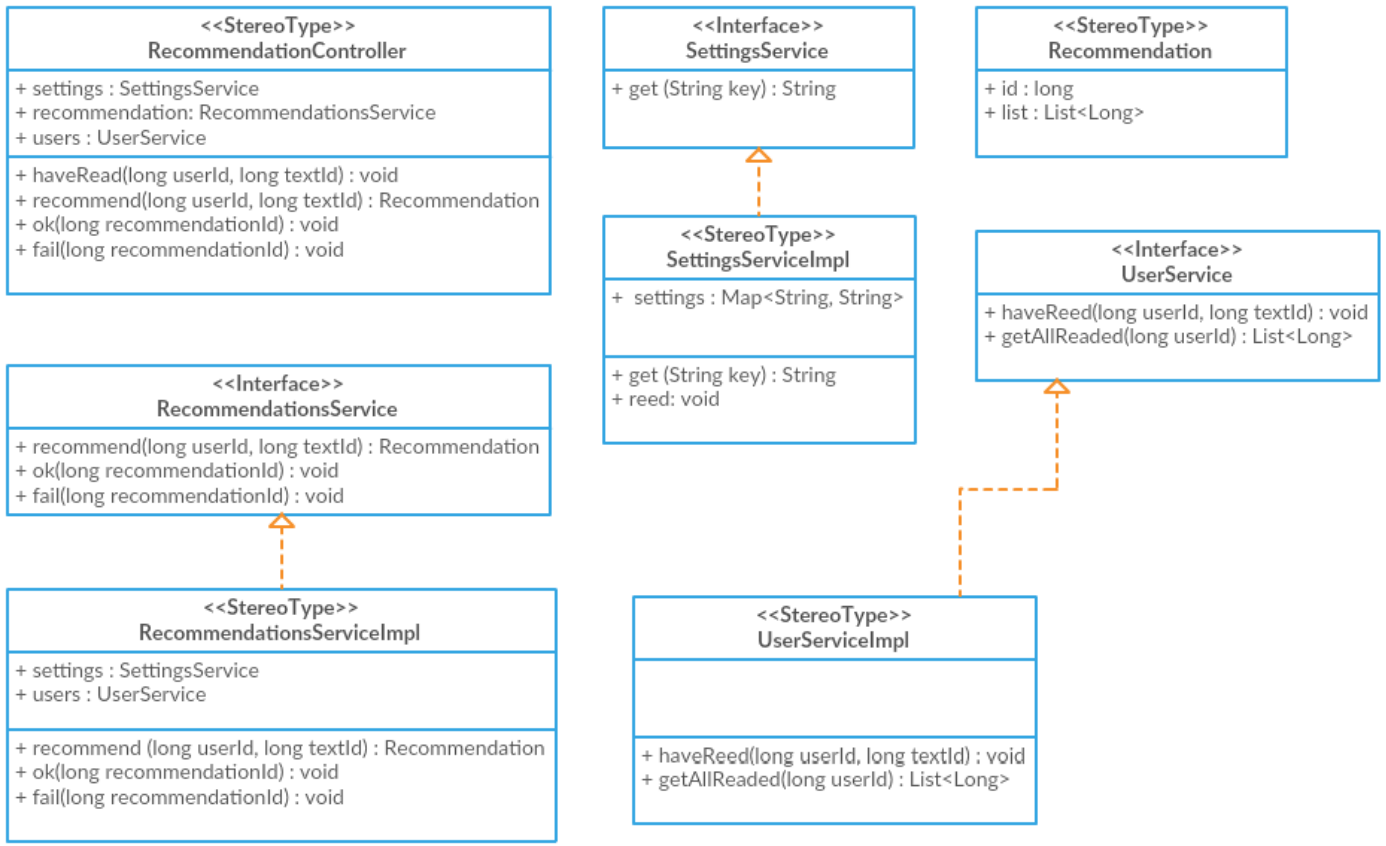
АЛГОРИТМ МОДУЛЮ

МОДЕЛЬ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ

- Розбиваємо тексти на частини (абзаци, або розділи).
- Будуємо матрицю використань слів в цих частинах.
- Виконуємо алгоритм LSA до нашої матриці.
- Групуємо групи у кластери.
- Визначаємо схожість текстів на основі кількості кластерів, що мають в собі частини з різних текстів та наближеності цих елементів до центру кластеру.

АЛГОРИТМ СИСТЕМИ

АРХІТЕКТУРА



ТЕСТОВА ВИБІРКА

1. Стаття про Java Core.
2. Стаття про історію літакобудування.
3. Стаття про мову програмування D.
4. Стаття про Java Servlets.
5. Новина про випуск нової моделі літака Boeing.
6. Стаття про особливості розробки авіаційного програмного забезпечення.
7. Стаття про аеропорт в Абу-Дабі.
8. Стаття про історію розвитку автомобільних коробок передач.
9. Стаття про програмне забезпечення в нових автомобілях компанії Tesla Motors.

ПРИКЛАД РОБОТИ

Id	Опис	
1	Java Core	
2	Історія літакобудування	
3	Мова програмування D	
4	Java Servlets	
5	Літак Boeing	
6	Авіаційне ПЗ	
7	Аеропорт в Абу-Дабі	
8	Коробки передач	
9	ПЗ Tesla	

Система сама створить тестового користувача, та задасть для нього необхідні параметри.

Від тексту:

8

Прочитані тексти, через кому:

2,6

Ранжувати

Відповідь:


```
{
  "id": 1447667173643,
  "list": [5,9,7,4,1,3]
}
```

ТЕСТОВІ ВИПАДКИ

1. Не прочитано жодного тексту, на основі тексту 1 "Java Core" будемо шукати подібні.
2. Ранжування від тексту 1 "Java Core", в прочитаних текстах є текст 7 "Аеропорт Абу-Дабі".
3. Ранжування від тексту 8 “Коробки передач", в прочитаних текстах є тексти: 2 "Літакобудування" та 6 "Авіаційне ПЗ".
4. Такі ж, як і в тесті 3, але вкажемо, що попередній результат ранжування був не успішним
5. Початковий текст, від якого відбувається ранжування – 3 "Мова програмування D" і немає жодного вже прочитаного тексту.

АНАЛІЗ РЕЗУЛЬТАТІВ

Перший тест:

№	Експерт	Система індивідуального ранжування
1	4 Java Servlets	4 Java Servlets
2	3 Мова програмування D	3 Мова програмування D
3	9 ПЗ Tesla	6 Авіаційне ПЗ
4	6 Авіаційне ПЗ	9 ПЗ Tesla

Другий тест:

№	Експерт	Система індивідуального ранжування
1	4 Java Servlets	4 Java Servlets
2	6 Авіаційне ПЗ	3 Мова програмування D
3	3 Мова програмування D	6 Авіаційне ПЗ
4	9 ПЗ Tesla	5 Літак Boeing

П’ятий тест:

№	Експерт	Система індивідуального ранжування
1	1 Java Core	1 Java Core
2	2 Java Servlets	9 ПЗ Tesla
3	6 Авіаційне ПЗ	6 Авіаційне ПЗ
4	9 ПЗ Tesla	2 Java Servlets

ЕКОНОМІЧНА ЧАСТИНА

- Ціна реалізації – 12685 грн
- Абсолютна ефективність інвестицій – 41633 грн.
- Термін окупності – 4,3 року.

РЕЗУЛЬТАТИ

- Дістав подальшого розвитку модифікований метод фільтрації вмісту на основі латентно-семантичного аналізу, за рахунок використання стемінгу та словників синонімів та багатозначних слів.
- Удосконалено модель системи інтелектуального модуля категоризації, що дозволяє використовувати його, як універсальний інструмент аналізу схожості текстів, написаних природною мовою, за рахунок створення окремого програмного інтерфейсу доступу до алгоритму латентно-семантичного аналізу.
- Розроблено модель модифікованого методу фільтрації вмісту для розв’язання задачі індивідуального ранжування текстів за рахунок використання інтелектуального модулю категоризації, що використовує алгоритми добування даних, а саме метод латентно-семантичного аналізу для кластеризації та порівняння текстів написаних природною мовою.
- Розроблено програмне забезпечення, що реалізує інформаційну технологію індивідуального ранжування текстів на основі інтелектуального модуля категоризації.
- Розроблене програмне забезпечення може бути використане як для розв’язання конкретних задач індивідуального ранжування текстів, так і для перевірки результатів роботи інших методів при розв’язанні даної задачі.

Дякую за увагу!