Створення інтелектуальної системи для концентрації веб-інформації

Качур В.В.

проф., к.т.н Месюра В.І.

Вступ

Інтернет - віртуальний інформаційний простір, унікальний й універсальний за своїми властивостями і функцій. Це, насамперед, засіб відкритого зберігання та розповсюдження інформації: наукової, ділової, пізнавальної та розважальної.

В даний час в світі відбувається великий інформаційний переворот. Телебачення втрачає свої позиції, так як неможливо за ефірний час оглядати всю інформацію, яка цікавить телеглядачів. З кожним днем ми все більше використовуємо новини Інтернету як постійне джерело інформації, так як інтернет дозволяє фільтрувати непотрібні новини, залишаючи тільки той контент, який цікавий читачеві .

ІНТЕРНЕТ-РЕСУРС - сукупність інтегрованих програмно-апаратних та технічних засобів, а також інформації, яка призначена для публікації в мережі Інтернет та яка відображається у певній текстовій, графічній або звуковий формах. Інтернет-ресурс має доменне ім'я (Uniform Resource Locator) - унікальний електронну адресу, що дозволяє ідентифікувати Інтернет-ресурс, а також здійснювати доступ до Інтернет-ресурсу.

Недоліки які найчастіше присутні при роботі з інтернет ресурсом:

- незрозуміла, важка для розуміння, «карта сайту»;
- громісткий інтерфейс;
- відсутня, чи погано організована класифікація інформації, відсутність категорій;
- нечітке форматування інформації, та ін.

Об’єктом дослідження є процес отримання інформації з веб-ресурсів.

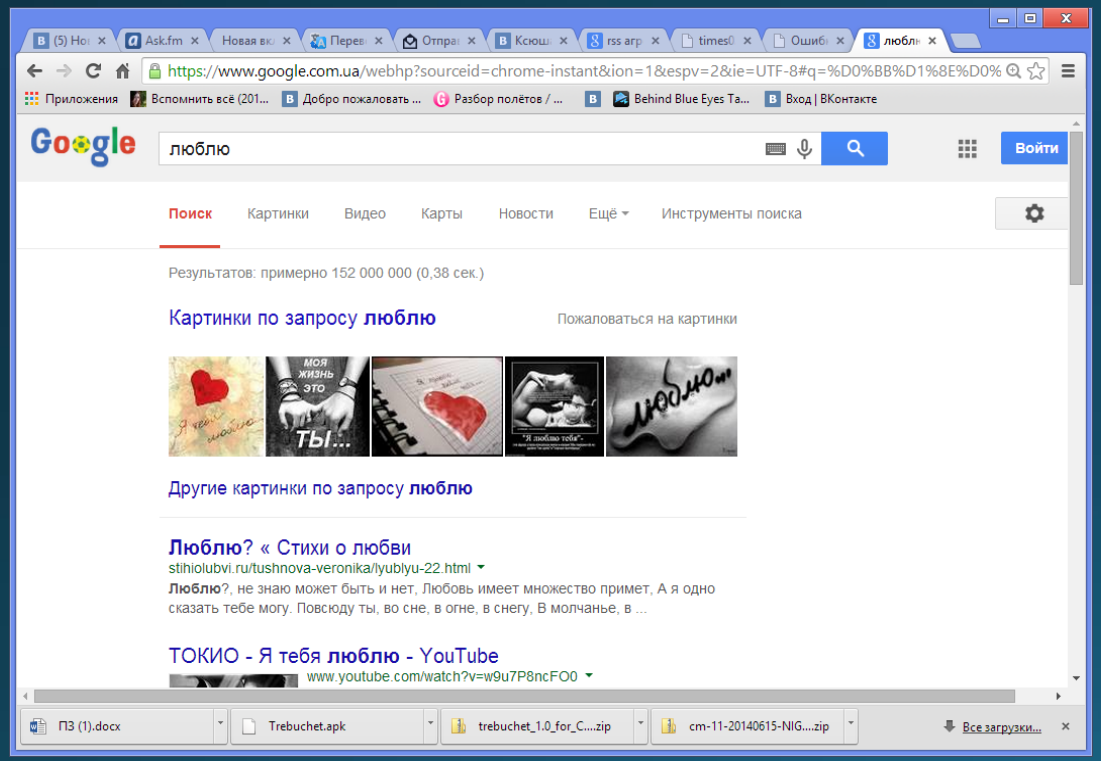
Предметом дослідження є моделі та програмні засоби отримання інформації з веб-ресурсів.

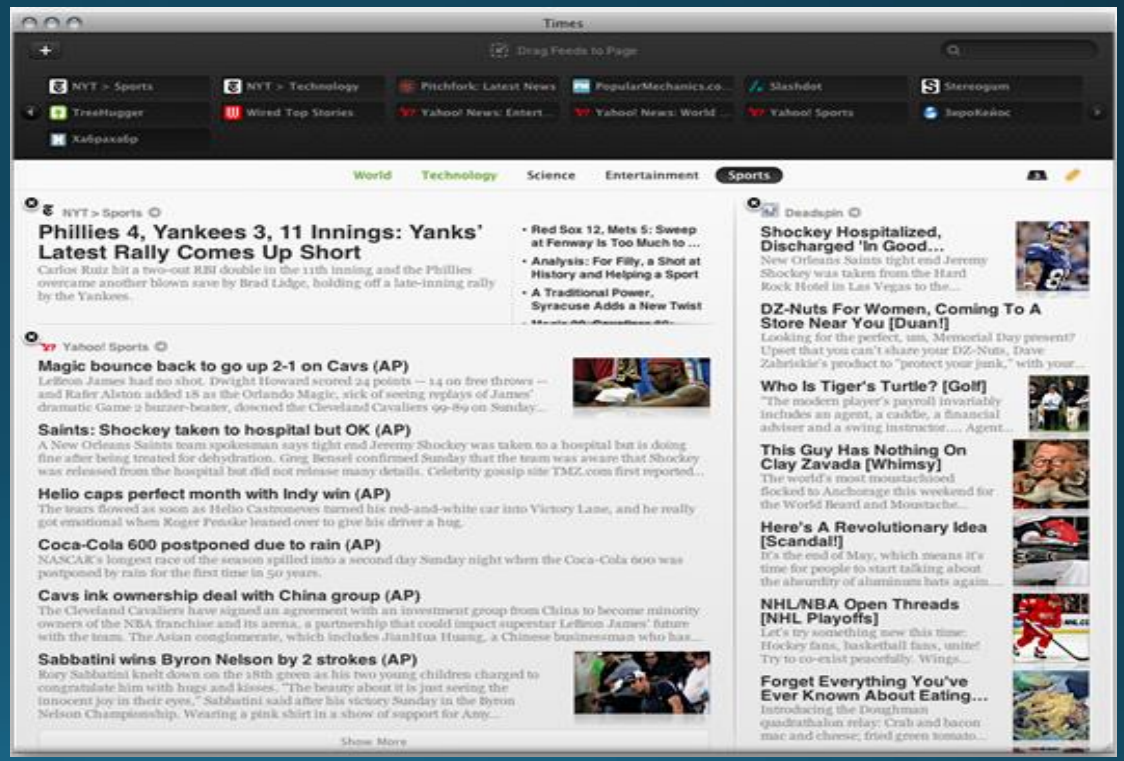
Метою дослідження є зменшення часу доступу користувачем до інформації, яка знаходиться на різних Інтернет-ресурсах, а також досягнення хороших показників правильності класифікації отриманої інформації шляхом розроблення інтелектуальної системи для концентрації веб-інформації.

Основні задачі:

- дослідження методів отримання веб-інформації;
- побудова схем роботи модулів інтелектуальної системи для концентрації веб-інформації;
- об’єктно-орієнтоване проектування інтелектуального модуля для концентрації веб-інформації;
- програмна реалізація інтелектуального концентратора веб-інформації.
- тестування та аналіз результатів.

Існуючі методи отримання веб-інформації

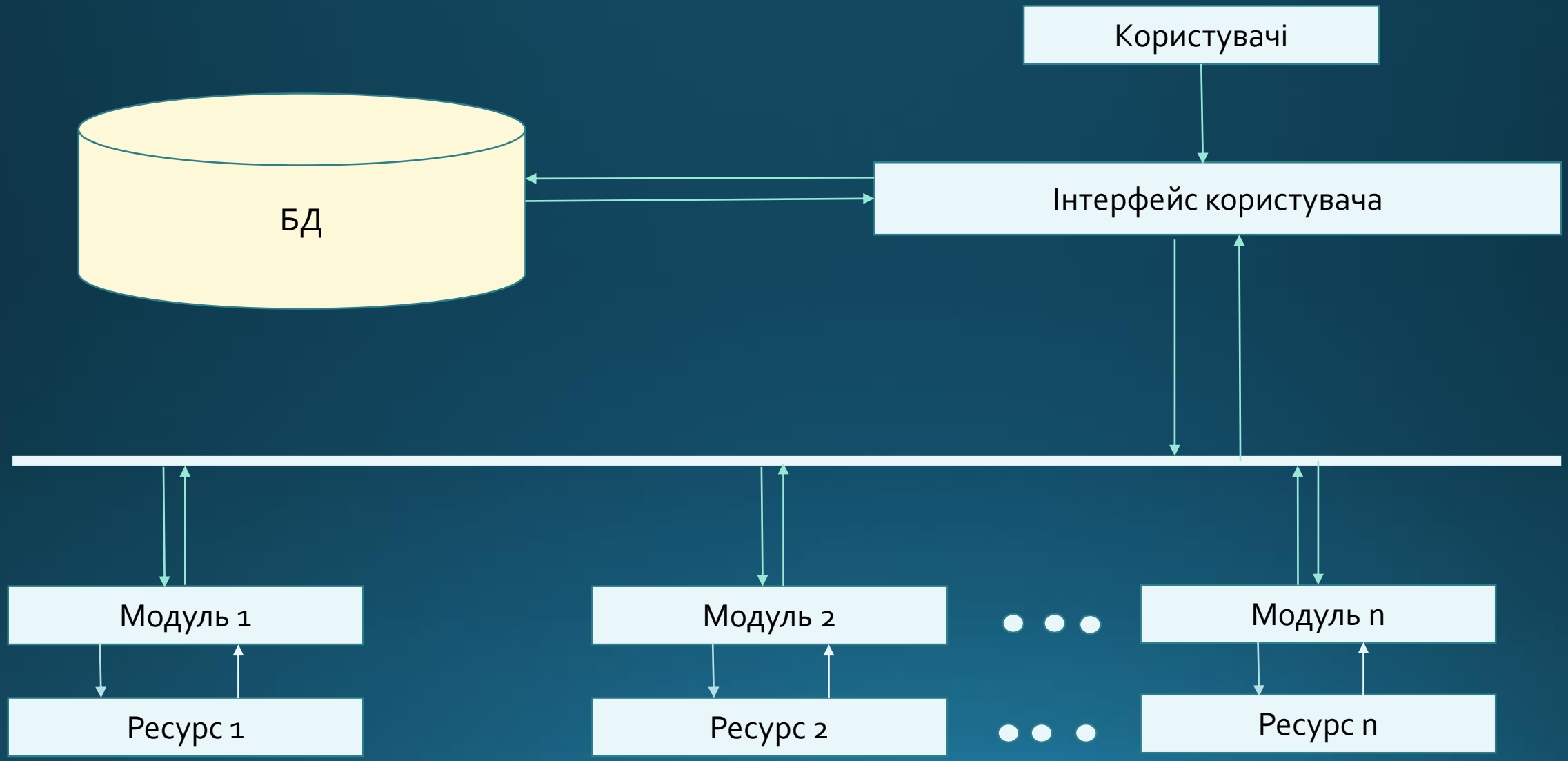




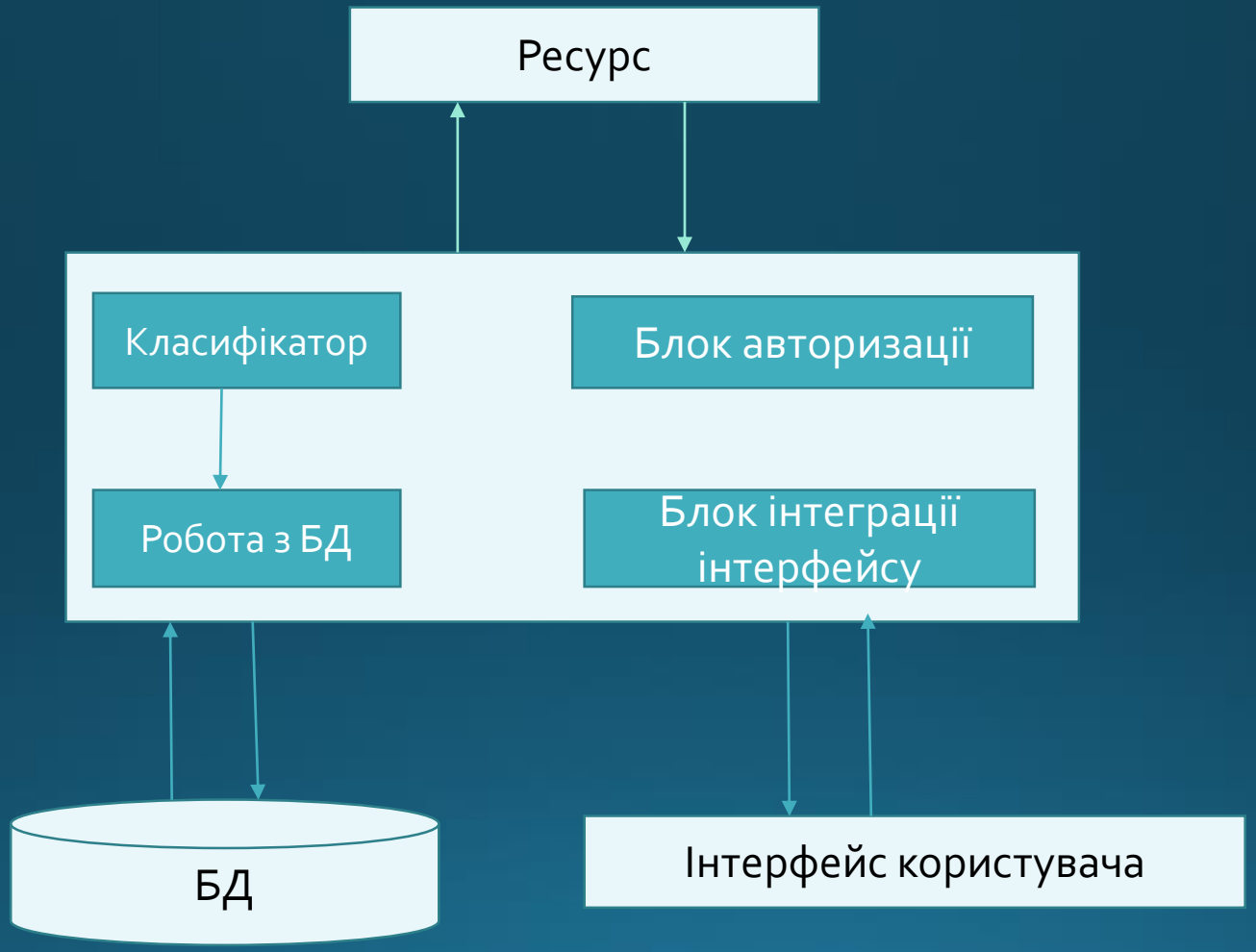
Комбінований метод дозволить:

- автоматизовано отримувати інформацію в різних веб-ресурсів;
- швидко аналізувати тисячі веб сторінок, відділяти технічну інформацію від «людської», відділяти необхідне і фільтрувати зайве;
- забезпечить можливість повного контролю інформації;
- дозволить відображати дані в необхідному вигляді.

Структурна схема інтелектуальної системи



Модуль для роботи з ресурсом



Класифікатор Інформації

Нехай у нас є текст d. Крім того, є класи C, до одного з яких ми повинні віднести текст. Нам необхідно знайти такий клас c, при якому його ймовірність для даного тексту була б максимальна. Математично це записується так:

$$C_{MAP} = \arg\max_{c \in C} P(c \mid d)$$

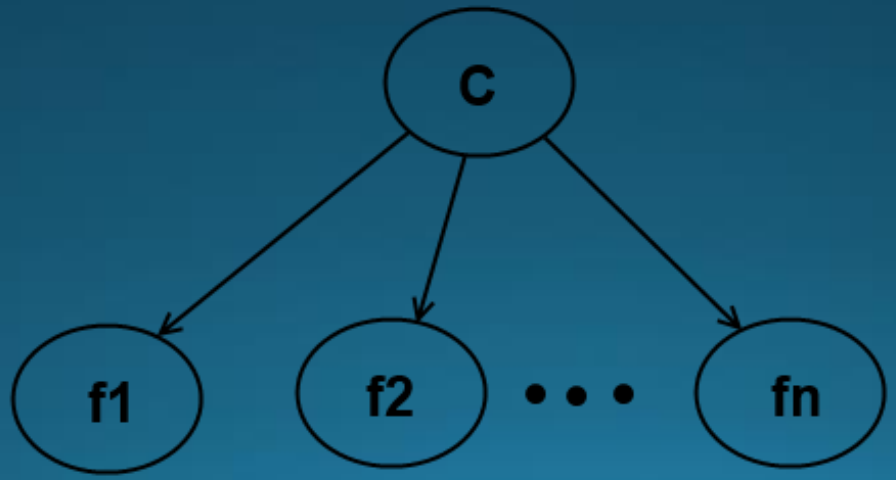
Обчислити P (C | D) складно. Але можна скористатися теоремою Байєса і перейти до непрямих ймовірностей:

$$C_{MAP} = \arg\max_{c \in C} \frac{P(d \mid c)P(c)}{P(d)}$$

Модель наївного байєсівського класифікатора приймає два допущення, від того вона така і наївна:

1. порядок проходження ознак об'єкта не має значення;
2. ймовірності ознак не залежать одне від одного при даному класі:

$$P(f_i \cap f_i \mid c) = P(f_i \mid c)P(f_j \mid c)$$



Лексикографічний аналіз вхідної послідовності символів:

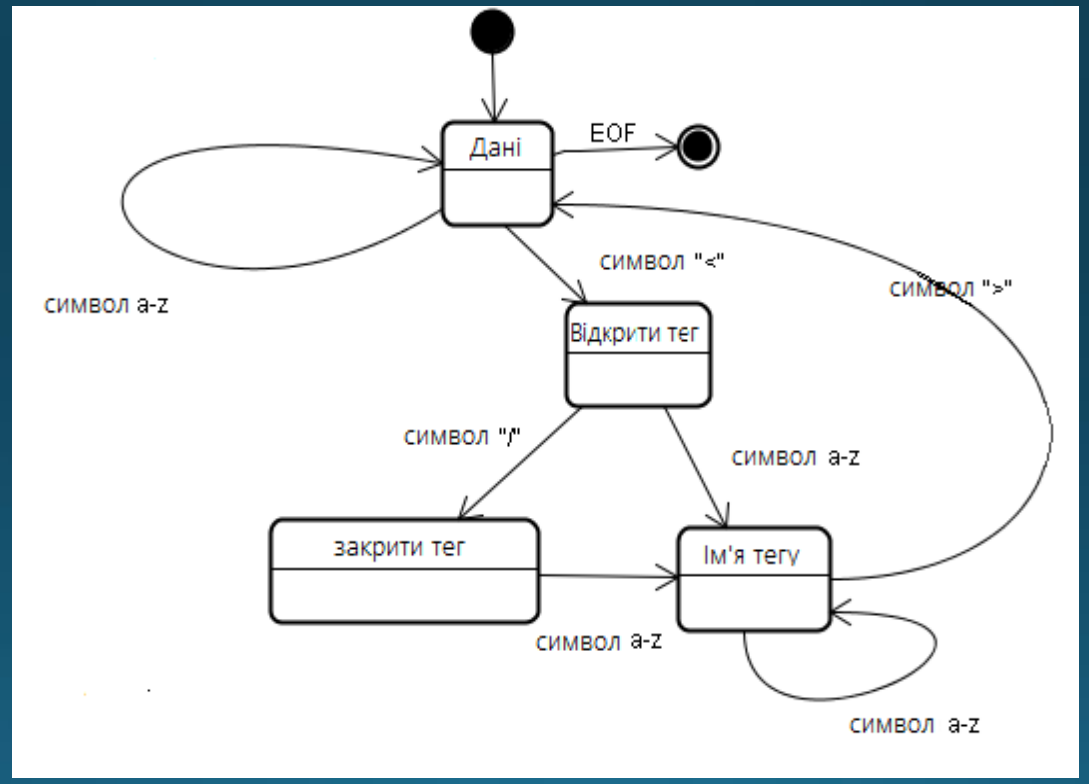
<html>

<body>

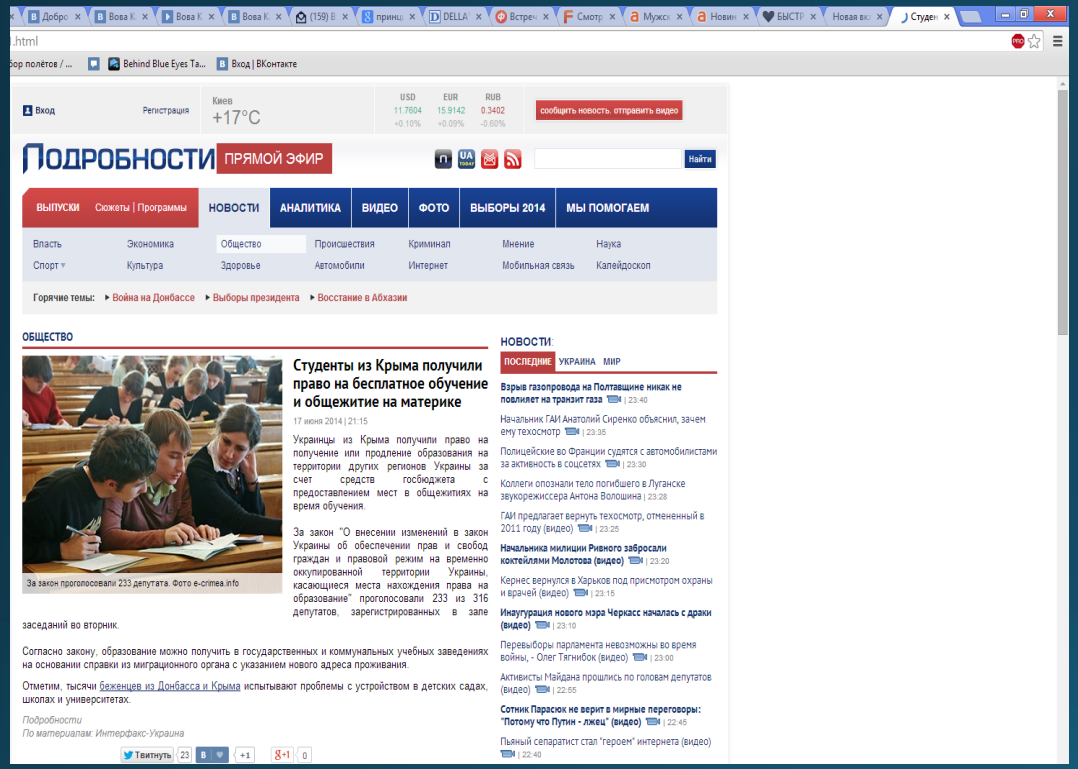
Привіт світ

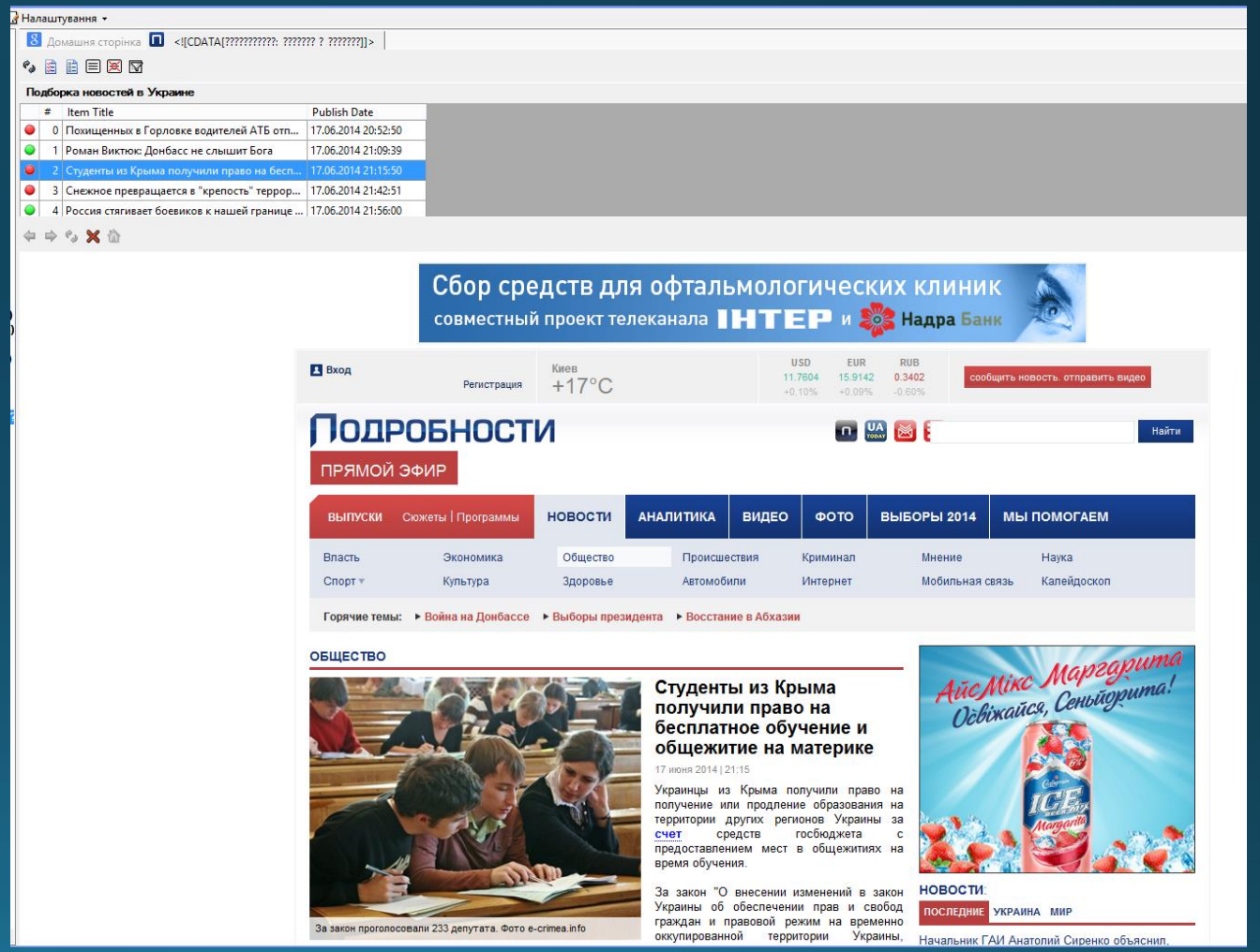
</ body >

</ HTML>



Результат роботи інтелектуальної системи





Результати тестування інтелектуальної системи:

Спосіб доступу	Час доступу (сек.)			Середнє значення (сек.)
	Новина 1	Новина 2	Новина 3	
RSS- агрегатор	10	12	9	10,33
Веб- серфінг	20	32	15	22,33
Веб-концентратор	10	11	9	10

Засіб концентрації	Кількість новин	Правильність класифікації
BazQux Reader	<100	100%
	100-500	97%
	>500	92%
Система для концентрації веб-інформації	<100	100%
	100-500	99%
	>500	97%

Економічне обгрунтування доцільності розробки інтелектуальної системи:

В дипломному проекті були проведені відповідні економічні розрахунки, що підтверджують економічну доцільність розробки інтелектуальної системи для концентрації веб-інформації, оскільки ціна реалізації складає 1182,69 і являється дешевшою, ніж аналог на 867,31 грн.
Термін окупності інтелектуальної системи складає 0,15 року.
Загальні витрати на розробку нового програмного продукту складають 22607 грн.

Основні результати роботи:

1. Здійснено аналіз предметної області.
2. Проведено огляд існуючих методів отримання веб-інформації.
3. Визначено необхідність створення нових підходів для вирішення проблеми отримання інформації з веб-ресурсів.
4. Розроблено інтелектуальну частину для вирішення проблем класифікації новин.
5. Розроблені детальні схеми роботи системи.
6. Здійснено тестування інтелектуальної системи.
7. Розглянуті питання розробки системи з економічної точки зору.

Дякую за увагу!