Розробка веб-ресурсу для пошуку новин

Олександр Сергійович Ковальов
Кафедра інформатики та прикладної математики,
Криворізький державний педагогічний університет,
пр. Гагаріна, 54, м. Кривий Ріг, 50086, Україна
kovalyov.aleksandr86@gmail.com

Анотація. *Метою дослідження* є проектування та реалізація новинного порталу з використанням методів аналізу тексту. *Задачами дослідження* є: 1) проаналізувати існуючі новинні порталів та методи аналізу тексту, що допомагають визначити стилістичну приналежність тексту та ступінь схожості за змістом різних текстів; 2) на підставі аналізу існуючих рішень визначити метод аналізу тексту та сформувати вимоги до новинного порталу та його функцій; 3) розробити серверну частину інформаційного портал, який буде використовувати латентно-семантичний аналіз для визначення оптимального пошуку новин. *Об'єктом дослідження* є процес функціонування новинного порталу. *Предметом дослідження* є методи аналізу тексту, визначення тематики тексту та визначення ступені схожості за змістом різних текстів в реалізації новинного порталу. В роботі проведено аналіз, узагальнення та систематизація досліджень з методів аналізу текстів у розробці новинного порталу. *Результати дослідження* планується узагальнити для формування рекомендацій щодо реалізації новинного порталу.

Ключові слова: новинний портал; аналіз; текст.

O. S. Kovalov. Development of a Web-resource for news search

Abstract. The *aim* of the study is to design and implement a news portal, using existing methods of text analysis. The *objectives of the study* is: 1) to analyze the existing news portals and methods of text analysis that help determine the stylistic affinity of the text, and the degree of similarity in the content of various texts; 2) to determine the method of analysis of the text to be used based on the analysis of existing solutions, and to form the requirements to the news portal and its functions; 3) to develop a server part of an information portal that will use latent semantic analysis to determine the optimal news search. The *object of research* is the process of functioning of the news portal. The *subject of research* is the methods for analyzing the text, defining the subject matter of the text and determining the degree of similarity in the content of various texts at the news portal implementation. The analysis, generalization and systematization of research on the methods for analyzing texts in the news portal development was done in the paper. The *results of the*

study are planned to be generalized to formulate recommendations for the news portal design.

Keywords: news portal; analysis; text.

Affiliation: Department of Computer Science and Applied Mathematics, Kryvyi Rih State Pedagogical University, 54, Gagarin Ave., Kryvyi Rih, 50086, Ukraine.

E-mail: kovalyov.aleksandr86@gmail.com

Існує велика кількість Інтернет-ресурсів, які вчасно та якісно інформують читача свіжими новинами. Але жоден з них не може інформувати у повному обсязі новинами з усіх сфер життєдіяльності. Є сайти, які спеціалізуються на економічних новинах, або швидше за інших інформують про надзвичайні події. Є сайти, які максимально детально розкажуть про спорт або політику. Тому користувачу доведеться регулярно переглядати одразу декілька ресурсів, щоб тримати руку на пульсі подій з усіх сфер. Також перегляд одразу декількох джерел дає користувачу можливість порівнювати, як різні ресурси висвітлюють одну й ту саму новину, як швидко вони реагують на події у світі, та робити для себе певні висновки. Саме тому новинний портал, який містить у собі новини з різних джерел, є дуже корисним для людини, яка прагне бути у курсі всіх подій.

Переглянувши усі існуючі рішення цієї проблеми, ми зупинилися на Інтернет-ресурсі Anews (https://www.anews.com/ua/). На головній сторінці даного новинного порталу публікуються заголовки та зображення останніх новин із різних джерел (рис. 1). Для перегляду треба обрати ту новину, що нас зацікавила. Після переходу на сторінку новини бачимо заголовок, зображення, короткий текст новини та гіперпосилання на первинне джерело. Повний текст новини є доступним лише при переході на сайт первинного джерела, що є не дуже зручним, адже кожного разу, для того, щоб прочитати новину, користувачу потрібно переходити на інший сайт, адаптуватись під інший дизайн та структуру сайту, що створює певні незручності для людини.

Ще один новинний портал, який було проаналізовано – Flipboard (https://flipboard.com/). Щоб почати читати новини на цьому порталі, спочатку необхідно увійти у свій профіль однієї з соціальних мереж, а саме: Facebook, Twitter, Google+, або зареєструватися через електронну пошту, тому дане рішення також не є оптимальним (рис. 2).

Аналіз проведених на сьогодні досліджень вказує на те, що задля коректного функціонування новинного порталу потрібно використовувати один з методів аналізу тексту. Найбільш ефективним методом аналізу текстів, що допомагають визначити тематичну

приналежність тексту, а також визначити ступінь схожості кількох різних текстів, є латентно-семантичний аналіз [1].

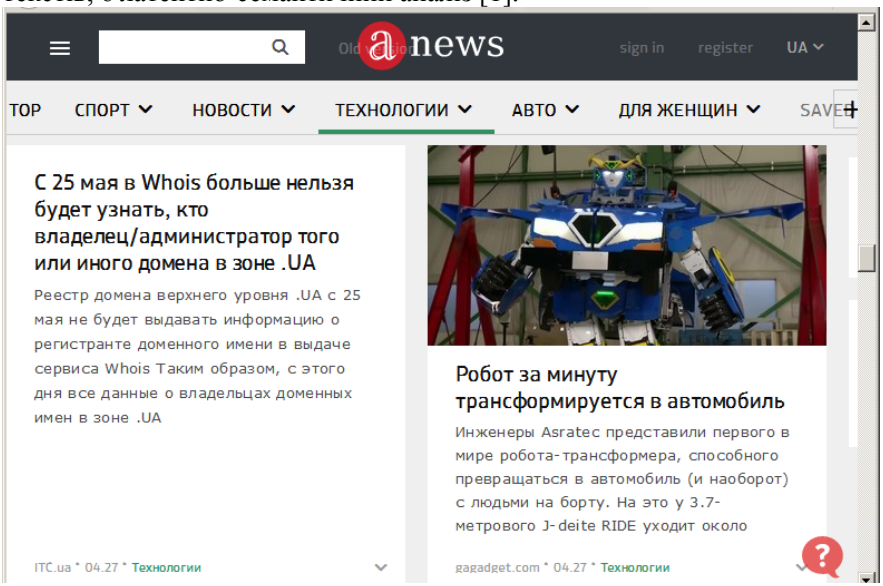

Рис. 1. Портал Anews

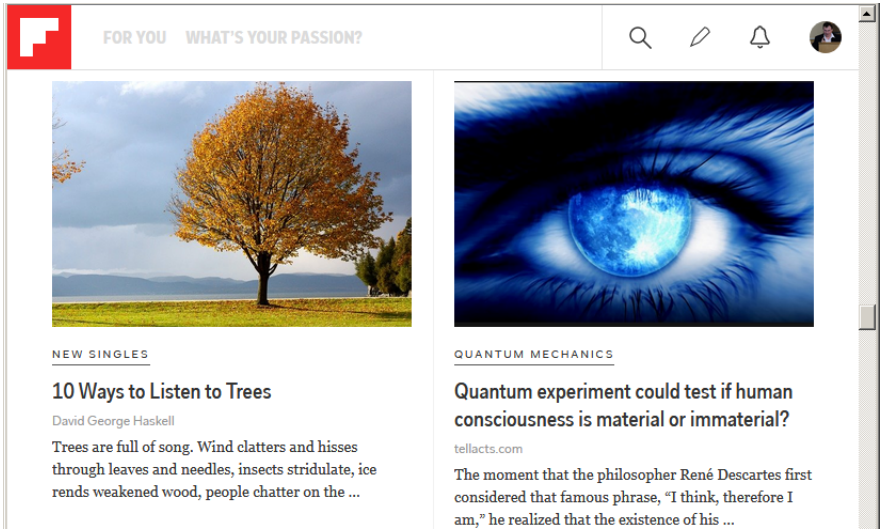

Рис. 2. Портал Flipboard

Основними цілями використання латентно-семантичного аналізу при

розробці новинного порталу є:

– забезпечення читача свіжими новинами з усіх сфер життєдіяльності;

– отримання можливості пропонувати читачу новини тієї ж тематики, що і новина, яку він переглядає;

– отримання можливості пропонувати читачу переглянути ту саму новину з інших джерел задля порівняння та можливого отримання нових подробиць.

Латентно-семантичний аналіз, що використано для реалізації новинного порталу, має наступні переваги:

– метод є найкращим для виявлення латентних залежностей усередині множини документів;

– метод може бути застосований як з навчанням, так і без навчання (наприклад, для кластеризації);

– висока точність результатів [2, с. 4].

Для реалізації розробки серверної частини новинного порталу було обрано Node.js – платформу з відкритим кодом, що характеризується такими властивостями:

– асинхронна однопотокова модель виконання запитів;

– неблокуючий ввід/вивід;

– система модулів CommonJS;

– рушій JavaScript Google V8 [3].

Для керування модулями використовується пакетний менеджер npm. Для збору необхідної інформації з веб-сторінок було використано npm-пакет X-ray [4]. Для реалізації латентно-семантичного аналізу було використано алгоритм стемінгу – скорочення слова до основи шляхом відкидання допоміжних частин, таких як закінчення чи суфікс. Цей алгоритм було опубліковано Мартіном Портером у 1980 році [5]. На етапі сингулярного розкладу частотної матриці було використано js-бібліотеку numeric.js. Для зберігання необхідної інформації було обрано базу даних PostgreSQL – об'єктно-реляційну СУБД, яка підтримує багато типів даних, зокрема JSON, який є необхідним для зручного зберігання даних.

Продемонструємо деякі результати роботи серверної частини новинного порталу. Із масиву результатів було обрано найбільш показові приклади: спочатку виведено дві новини, а далі – косинусний коефіцієнт їх схожості (рис. 3). З результатів видно, що аналіз працює досить добре і дійсно визначає за заголовками новин, наскільки ці новини схожі.

У результаті було проаналізовано існуючі рішення даного питання, обрано оптимальний метод аналізу тексту, розроблено та реалізовано серверну частину веб-сервісу для оптимального пошуку новин. Наступним етапом буде розробка та реалізація клієнтської частини

новинного порталу.

```
Порошенко официально принял обращение к Вселенскому патриарху об автокефальной церкви в Украине
Порошенко: Автокефалия украинской церкви - это окончательная независимость от России
0.9313

Порошенко: Автокефалия украинской церкви - это окончательная независимость от России
Российский террор в Украине: Порошенко рассказал о сорванных планах Кремля
0.9072

Российский террор в Украине: Порошенко рассказал о сорванных планах Кремля
Из-за пошлины на импорт серной кислоты в Украине может вырасти курс доллара - промышленники
0.1832
```

Рис. 3. Демонстрація роботи серверної частини новинного порталу

Список використаних джерел

1. Landauer T. Introduction to Latent Semantic Analysis / Thomas Landauer, Peter W. Foltz, Darrell Laham // Discourse Processes. – 1998. – Vol. 25. – Issue 2-3: Quantitative Approaches to Semantic Knowledge. – P. 259-284. – DOI: 10.1080/01638539809545028.

2. Минаев В. А. Методы выявления латентной и негативной информации в текстовых документах [Электронный ресурс] / В. А. Минаев, И. Д. Королев, И. А. Кисленко // Технологии техносферной безопасности. – 2016. – № 5 (69). – 8 с. – Режим доступа : http://agps-2006.narod.ru/ttb/2016-5/35-05-16.ttb.pdf.

3. Node.js [Electronic resource] / Bunyk. – 25 April 2018. – Access mode : https://uk.wikipedia.org/w/index.php?title=Node.js&oldid=22499025.

4. x-ray - npm [Electronic resource]. – 2017. – Access mode : https://www.npmjs.com/package/x-ray.

5. Porter M. F. An algorithm for suffix stripping [Electronic resource] / M. F. Porter // Program. – 1980. – Vol. 14. – No 3. – P. 130-137. – DOI: https://doi.org/10.1108/eb046814. – Access mode : http://www.cs.odu.edu/~jbollen/IR04/readings/readings5.pdf.

References (translated and transliterated)

1. Landauer T. Introduction to Latent Semantic Analysis / Thomas Landauer, Peter W. Foltz, Darrell Laham // Discourse Processes. – 1998. – Vol. 25. – Issue 2-3: Quantitative Approaches to Semantic Knowledge. – P. 259-284. – DOI: 10.1080/01638539809545028.

2. Minaev V. A. Metody vyiavleniia latentnoi i negativnoi informatcii v tekstovykh dokumentakh [Methods for latent and negative information detection in text documents] [Electronic resource] / V. A. Minaev, I. D. Korolev, I. A. Kislenko // Tekhnologii tekhnosfernoi bezopasnosti. – 2016. – № 5 (69). – 8 s. – Access mode : http://agps-2006.narod.ru/ttb/2016-5/35-05-16.ttb.pdf.

3. Node.js [Electronic resource] / Bunyk. – 25 April 2018. – Access mode : https://uk.wikipedia.org/w/index.php?title=Node.js&oldid=22499025.

4. x-ray - npm [Electronic resource]. – 2017. – Access mode : https://www.npmjs.com/package/x-ray.

5. Porter M. F. An algorithm for suffix stripping [Electronic resource] / M. F. Porter // Program. – 1980. – Vol. 14. – No 3. – P. 130-137. – DOI: https://doi.org/10.1108/eb046814. – Access mode : http://www.cs.odu.edu/~jbollen/IR04/readings/readings5.pdf.

Received: 14 April 2018; in revised form: 25 April 2018 / Accepted: 28 April 2018