УДК 004.031.42

В. Р. Констянтинів
Тернопільський національний технічний університет імені Івана Пулюя, Україна

ОГЛЯД МЕТОДІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ТЕКСТОВИХ ДАНИХ

V.R. Konstiantyniv

SURVEY OF TEXT MINING TECHNIQUES

1. Процес інтелектуального аналізу тексту

Переважна більшість інформації у WEB (до 80%) – звичайний неструктурований текст [1]. Отримання значущої інформації – це і є задача інтелектуального аналізу тексту, який є процесом отримання якісної інформації з напів- та неструктурованих даних [2]. Інформація може бути отримана спеціальним програмним забезпеченням шляхом виявлення певних шаблонів чи трендів на основі статистичної обробки тексту (рис. 1).

Процес аналізу відбувається відповідно до наступної послідовності кроків.

1. Збір (gathering) текстових документів з різних джерел з наступним виконанням попередньої обробки (preprocessing) (усунення реклами; токенизація (tokenization), тобто розділення його на складові частини; усунення незначущих слів (артиклів, сполучників тощо); вирівнювання (stemming), тобто зведення всіх однокореневих слів до основної форми і, зменшення загальної кількості слів; перетворення тексту до вигляду, придатного для подальшого аналізу, наприклад, у вигляді гістограми частоти появи слів).

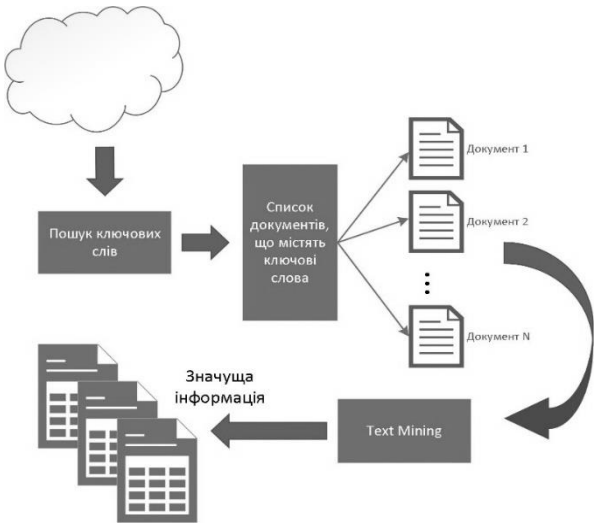

Рисунок 1. Різниця між пошуком ключових слів та інтелектуальним аналізом тексту

2. Вибірка потрібних та усунення непотрібних властивостей тексту.

3. Застосування одного чи декількох методів інтелектуального аналізу тексту з метою отримання патернів (шаблонів).

4. Оцінювання патернів відповідно до критеріїв пошуку [3].

На основі аналізу літературних джерел можна виділити наступні технології інтелектуального аналізу тексту.

2 Виділення інформації з тексту

Виділення інформації [4, 1] означає отримання структурованих даних з неструктурованого тексту для отримання набору сутностей (імена, локації, взаємозв'язки тощо).

Використовується для відслідковування новин; дописів користувачів соціальних мереж з метою отримання їх думки стосовно певної події; отримання інформації з повідомлень електронної пошти; отримання особистих даних з профілів користувачів; отримання інформації в цифрових онлайн бібліотеках; виділення таблиць з тексту.

2 Узагальнення

Узагальнення текстових даних [6] є процесом створення стислого представлення великої кількості даних. Під час узагальнення здійснюється пошук найбільш релевантних до теми речень і, як наслідок, зменшення об'єму текстових даних без

втрати інформації. Такий метод може використовуватись для: обробки інформації з агрегаторів новин; генерування звітів; сумісної роботи з поштовими клієнтами; виділення інформації про події.

3 Класифікація

Класифікацією називається процес знаходження певних ознак класу (поміток) серед частково структурованих даних. Нехай, до прикладу, менеджер хоче встановити, чи купуватиме замовник комп'ютер. Тоді класами для покупок будуть значення "Так" і "Ні". Під час класифікації створюється класифікатор на відповідних тестових даних з наступним означенням міток класів. Класифікація тексту [7] є процесом присвоєння категорії новому текстовому документові. Таким чином можна розподілити текстові документи відповідно до належності певним класам.

Області використання такої обробки тексту: бізнес; медицина; юриспруденція; суспільні науки.

4 Кластеризація

Кластеризація – це процес групування елементів даних споріднених типів у окремі кластери [5, 1]. Таким чином, кластеризація зменшує часові затрати при пошуку інформації. Інформація, що не належить до жодного кластеру, вважається нерелевантною.

Можливі області застосування такої методики: розпізнавання образів; аналіз зображень; таксонометрія; виділення тематики з тексту.

Таким чином, у даному огляді описано і класифіковано наявні на сьогодні методики інтелектуального аналізу текстових даних. Серед них – отримання (виділення) інформації, узагальнення, кластеризація. Всі вони можуть бути застосовані відповідно до задач, які ставляться при інтелектуальному аналізу текстових даних. Інструментальна реалізація цих методик лежить за межами даного дослідження, але попередній аналіз показав, що для кожної з методик можна підібрати відповідне програмне забезпечення. Проте, розробка власної системи буде хорошим досвідом в практичній реалізації однієї з цих методик.

Література

1. Unstructured data [електронний ресурс] / Wikipedia / Режим доступу: https://en.wikipedia.org/wiki/Unstructured_data (жовтень 2017)
2. Areas Dr. Shilpa Dang, Peerzada Hamid Ahmad. A Review of Text Mining Techniques Associated with Various Application Areas // International Journal of Science and Research (IJSR), ISSN 2319-7064. Volume 4, Issue 2, February 2015, pp. 2461 – 2466.
3. Atika Mustafa, Ali Akbar, and Ahmer Sultan, "Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization" // International Journal of Multimedia and Ubiquitous Engineering. Vol. 4, No. 2, April, 2009, pp. 837 – 848.
4. Baeza-Yates, B. Ribeiro-Neto, "Modern Information Retrieval", ACM, Press, Page 64, Year1999.
5. Y. Zhao, "Analysing twitter data with text mining and social network analysis," in Proceedings of the 11th Australasian Data Mining and Analytics Conference (AusDM 2013), 2013, p. 23.
6. Dr shilpa Dang, Peerzada Hamid Ahmad, "A Review of Text Mining Techniques Associated with Various Application Areas", International Journal of Science and Research, ISSN (Online): 2319-7064, Volume 4, Issue 2, 2015
7. Chauhan Shrihari R, Amish Desai, "A Review on Knowledge discovery using Text classification techniques in Text Mining", International Journal of Computer Applications (0975-8887) Volume-111-No 6,2015
8. Varsha C. Pande and A.S. Khandelwal "A Survey of Different Text Mining Techniques", IBMRD's Journal of Management & Research, ISSN: 2348-5922, Volume 3, No. 1, pp. 125-133, 2014.