УДК 004.8

Т. О. Савчук, Ю. В. Горобець

УДОСКОНАЛЕНИЙ МЕТОД ВИЯВЛЕННЯ КЛЮЧОВИХ СЛІВ У WEB-ТЕКСТІ

Вінницький національний технічний університет, м. Вінниця

Анотація. У роботі запропоновано удосконалення методу виявлення ключових слів і словосполучень у web-тексті. Розглянуто такі основні етапи формування множини ключових слів та словосполучень з метою знайти шляхи підвищення швидкодії індексації та реферування web-текстів, як токенізація вихідного тексту, виключення стоп-слів, відсікання основ і закінчень із тексту, формування ключових слів та словосполучень з вихідного тексту. Запропоноване удосконалення базується на використанні словника предметної області, складеного експертом. Словник формується з урахуванням частоти повторень ключових слів та словосполучень у web-тексті, що покращить їх релевантність. Було виконано порівняння якості виявлених ключових слів та словосполучень в україномовних та англомовних web-текстах системами «Експертні КС», «OpenCalais», «Extractor», а також системою, що базується на запропонованому методі з використанням словника, за такими метриками як повнота, точність та F-міра. Аналіз показав, що запропонований удосконалений метод виявлення ключових слів та словосполучень в україномовних і англомовних web-текстах, надасть можливість виявляти релевантні слова та словосполучення з підвищенням їх F-міри на 9.5%, а повноти та точності - на 15%.

Ключові слова: Natural Language Processing: Text Mining; Keywords Extraction; виявлення термінів; виявлення ключових слів; обробка природної мови; комп'ютерна лінгвістика.

Аннотация. В работе предложено усовершенствование метода выявления ключевых слов и словосочетаний в web-тексте. Рассмотрены следующие основные этапы формирования множества ключевых слов и словосочетаний с целью найти пути повышения быстродействия ин дек-сации и реферирования web-текстов, токенизация исходного текста, исключение стоп-слов, отсечение основ и окончаний из текста, формирование ключевых слов и словосочетаний из исходного текста. Предложенное совершенствование базируется на использовании словаря предметной области, составленного экспертом. Словарь формируется с учетом частоты повторений ключевых слов и словосочетаний в web-тексте, улучшит их релевантность. Было выполнено сравнение качества выявле-ных ключевых слов и словосочетаний в украиноязычных и англоязычных web-текстах системами «Экспертные КС», «Open-Calais», «Extractor», а также системой, базирующейся на предложенном методе с использованием словаря, по таким метрикам как полнота, точность и F-мера. Анализ показал, что предложенный усовершенствованный метод выявления ключевых слов и словосочетаний в украиноязычных и англоязычных web-текстах, позволит выявлять релевантные слова и словосочитания с повышением их F-меры на 9.5%, а полноты и точности - на 15%.

Ключевые слова. Natural Language Processing: Text Mining; Keywords Extraction; изъятие сроков; изъятие ключевых слов; обработка естественного языка; компьютерная лингвистика.

Abstract. The paper proposes an improvement of the method of extracting key words and phrases in the web-text. The following main stages of the formation of a plurality of key words and phrases are considered in order to find ways to increase the speed of indexing and refereeing web texts, to accurate source text, exclude stop words, cut off bases and endings from the text, the formation of key words and phrases from the source text. The proposed improvement is based on the use of the vocabulary of the subject area compiled by the expert. The dictionary is formed taking into account the frequency of repetitions of keywords and phrases in the web-text, will improve their relevancy. The comparison of the quality of the revealing keywords and phrases in the Ukrainian and English language web texts with the systems Expert Review, Open-Calais, Extractor, as well as the system based on the proposed method using the dictionary, recall, accuracy and F-measure. The analysis showed that the proposed advanced method for extracting keywords and phrases in Ukrainian and English web-texts will allow to reveal relevant words and word-received with an increase of their F-measures by 9.5%, and completeness and accuracy by 15%.

Key words: Natural Language Processing: Text Mining; Keywords Extraction; withdrawal of terms; keyword extraction; natural language processing; computer linguistics.

DOI: https://doi.org/10.31649/1999-9941-2018-43-3-43-47.

Вступ

В даний час обсяги і динаміка інформації, яка підлягає обробці в літературній справі, лексикографії та інформаційному пошуку, роблять особливо актуальною задачу виявлення ключових слів і фраз, які можуть використовуватися для створення і розвитку термінологічних ресурсів, а також для ефективної обробки документів: індексування, реферування, кластеризації і класифікації [1, 2].

Актуальність

Існує потужна множина програмного забезпечення для виявлення ключових слів у web-тексті, але лише одиниці з них мають підтримку української мови [3]. Багато з їх числа закриті для загального використання і поширюються на комерційній основі. Отже, існують проблеми в виборі відповідного інструменту для виявлення ключових слів з україномовних web-текстів.

На сьогодення поширений підхід до виявлення ключових слів у web-тексті базується на використанні статистичного аналізу [4]. При цьому, при формуванні множини ключових слів, їх значний відсоток є нерелевантним. Тому, актуальною є задача усунення означенного недоліку, що може бути досягнутою за рахунок виявлення ключових слів із сукупності токенів з використанням словників предметних областей [5]. Це дозволить виключити нерелевантні слова з множини ключових слів

Виявлення ключових слів у web-текстів з використанням словника

Процес виявлення ключових слів у web-тексті складається з таких чотирьох етапів (рис. 1):

1. Після отримання тексту web-документа (Текст), він підлягає процедурі токенізації, яка дозволяє отримати масив токенів - слів з тексту в тому порядку, в якому вони зустрічаються в тексті (Т-текст).

2. Для зменшення надлишковості отриманого масиву слів, проводиться виключення стоп-слів і знаків пунктуації (Т'-текст). Під стоп-словами розуміються слова, які зустрічаються практично у всіх текстах і не несуть змістового навантаження (прийменники, сполучники, вигуки і ін.) [4].

3. Відсікання основ, закінчень, суфіксів з вихідного тексту.

4. Розрахунок частоти появи слів у вихідному web-тексті, на основі яких виявляється множина ключових слів.

Одним з важливих етапів виявлення ключових слів в web-тексті є аналіз частоти їх появи. При цьому, аналізу підлягає саме частота слів у web-тексті, а не їх належність до предметної області. Використання словників предметної області, складених експертом, дозволить, при формуванні множини ключових слів враховувати їх відповідність предметній області, термінологічні особливості ключових слів, їх корисність, що сприятиме підвищенню швидкодії індексації, реферування web-текстів.

Щоб усунути цей недолік, запропоновано використання словників предметної області заздалегідь складених експертом, з метою точнішого виявлення ключових слів предметних областей з урахуванням специфічної для них термінів і термінологічних сполучень. це покращить релевантність виявлених ключових слів, і як наслідок цього пришвидшить індексацію і реферуванню цих текстів.

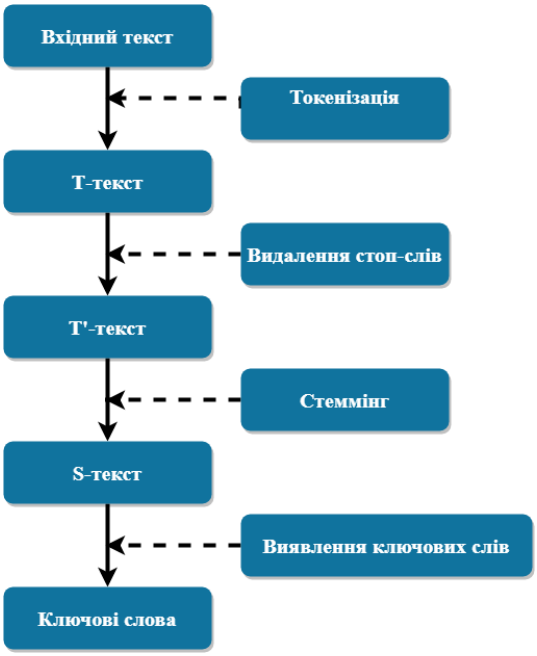

Рисунок 1 – Основні етапи процесу виявлення ключових слів

Алгоритм роботи запропонованого методу для виявлення ключових слів з використанням словника предметної області включає такі кроки:

1. Текст, в якому виявляються ключові слова, і словник предметної області після розбиття на токени, піддається стеммінгу [6].

2. Формування початкової множини ключових слів предметної області (S-текст).

3. Послідовний пошук слів і термінологічних сполучень, притаманних предметній області, з використанням словника, а також підрахунок їх повторень.

4. Формування множини ключових слів web-тексту.

Розглянемо приклад аналізу речення предметної області «операційні системи»:

«Процеси - це одна з найстаріших і найбільш важливих абстракцій, властивих операційній системі».

Після виконання процедур токенізації, стеммінгу та послідовного пошуку слів у словнику предметної області, було отримано такий S-текст:

«Процес, одним із варіантів, сам, старий, має важливе значення, абстракції, притаманний, операційні методи».

Представлений S-текст є початковою множиною ключових слів.

Словник предметної області піддається процедурам токенізації і стеммінгу, в результаті яких словникові поняття зберігаються хеш-таблицею. Як ключ таблиці, використовується отримана основа слова, або основи слів, розділені пробілами, якщо поняття складається з декількох слів. Значення, на яке вказує ключ - це безпосередньо саме словникове поняття. Після завантаження та обробки тексту (отриманої початкової множини S-тексту), виявляються ключові слова. Для цього проводиться послідовний пошук елементів множини S-тексту і формування структури FS_x, де *x* - номер проаналізованого елемента [7].

В основі структури FS_x лежить багатовимірний масив, де нульовий рівень (нульова позиція) структури представлений одновимірним масивом з довжиною в одне слово, і на кожному наступному рівні потужність даного масиву збільшується на одиницю [5]. Тим самим досягається можливість виявлення ключових слів та словосполучень з S-тексту.

Кожен новий прочитаний елемент додається в нульовий рівень початкового одновимірного масиву. При цьому, раніше записані у структуру FS_x ключові слова та словосполучення, просуваються на нижчі рівні.

В процесі послідовного пошуку ключових слів у web-тексті виконується аналіз поточних значень структури FS_x в словнику предметної області. Якщо значення знайдено, проводиться запис його позиції в структурі FS_x. Якщо знайдене значення на наступному кроці не збільшиться, то, винайдений елемент слід додати у множину ключових слів. Аналогічно аналізується весь web-текст з метою формування множини ключових слів та словосполучень.

З метою перевірки ефективності запропонованого методу виявлення ключових слів у web-тексті, було проаналізовано тексти українською та англійською мовою потужністю від 19000 до 20000 слів. В результаті були отримані множини в середньому потужністю відповідно 85 і 99 ключових слів, що склало близько 3% від загального обсяг web-тексту. У таблиці 1 представлені перші 30 елементів наборів ключових слів і словосполучень з текстів українською та англійською мовою з урахуванням кількості їх повторень у web-тексті. Для врахування частоти повторення ключового слова *W (kw)* доцільно використати формулу [5]:

$$W(kw) = \frac{N_W(kw) \times Freq(kw)}{N_i} \times 100\ \%,$$

де kw – ключове слово, N_w(kw) – кількість слів в kw, Freq (kw) – кількість повторень kw у web-тексті, N_i – кількість елементів S-тексту

Ефективність запропонованого методу було оцінено відносно «ідеального» алгоритму виявлення ключових слів у тексті, що базується на оцінці експертів [5]. Результати аналізу ключових слів текстів українською та англійською мовою представленні у таблиці 2. Експертом було сформовано множину з 20 ключових слів та словосполучень без урахування їх повторень у web-документі (табл. 2).

Для оцінки якості сформованих множин ключових слів та словосполучень з тексту, використовуються такі метрики: повнота, точність, F-міра [8]. Було проаналізовано якість виявлення ключових слів та словосполучень з web-тексту системами «Експертні КС» [9], «OpenCalais» [10], »Extractor» [11], а також системою, що базується на запропонованому методі з використанням словника предметної області (таблиця 3).

Таблица 1 – Результати аналізу тексту українською та англійською мовами з використанням удосконаленого методу

Ключові слова (укр. мов)	Вага	Ключові слова (англ. мов)	Вага
Процес	6,03	Process	6,56
Центральний процесор	2,11	Operatingsystem	1,34
Операційна система	1,21	Programcounter	0,91
Дочірній процес	0,78	Memory	0,79
Лічильник команд	0,72	Interrupt	0,67
Переривання	0,60	Systemcall	0,61
Системний виклик	0,48	UNIX	0,46
Створення процесу	0,48	Windows	0,39
Файл	0,42	Disk	0,39
Батьківський процес	0,42	Scheduler	0,36
UNIX процеси	0,36	Multiprogramming	0,30
Адресний простір	0,36	Addressspace	0,30
Послідовний процес	0,30	Operatingsystemdesign	0,27
Таблиця процесу	0,30	Signal	0,24
Стан процесу	0,30	Stackpointer	0,24
Ієрархія процесу	0,30	Childprocess	0,24

Продовження таблиці 1

Диск	0,27	Processtable	0,24
Windows	0,24	Processtermination	0,24
UNIX	0,24	Systemprocess	0,24
Подія	0,24	Processor	0,18

Таблиця 2 – Результати аналізу тексту українською та англійською мовою експертом

№	Ключові слова	№	Ключові слова
1.	Процес	11.	Переривання
2.	створення процессу	12.	вектор переривання
3.	блокування процессу	13.	операційна система
4.	завершення процессу	14.	Багатозадачність
5.	батьківський процесс	15.	Псевдо паралельність
6.	дочірній процесс	16.	Центральний процесор
7.	дескриптор процессу	17.	UNIX
8.	таблиця процесів	18.	Windows
9.	Адресний простір	19.	Системний виклик
10.	Планування	20.	Демон

Таблиця 3 –Результат аналізу метрик множини ключових слів та словосполучень

Система	Мова	Кількість КС	Повнота (R)	Точність (P)	F-міра (F = 2RP / (R + P))
Експертні КС	Укр	20	0.6	0.5	0.55
OpenCalais	Англ.	12	0,2 (4/20)	0,3 (4/12)	0,24
Extractor	Англ.	7	0,2 (4/20)	0,6 (4/7)	0,30
Система що базується на запропонованого метод	Укр.	15	1	0.9	0.95
OpenCalais	Укр.	21	0.5 (4/20)	0,5 (4/12)	0,5
Extractor	Укр.	5	0,3 (4/20)	0,4 (4/7)	0,34
Система що базується на запропонованого метод	Англ.	15	0.4	0.6	0.48

Як видно з таблиці 3, значення показника F-міри при застосуванні підходу, що базується на використанні словника предметної області є максимальним для україномовних web-текстів і складає 0.95. Аналогічне значення показника F-міри для англомовних web-текстів відповідає середньому значенню означеної метрики, для інших систем (0.26). Таким чином повнота, точність і як наслідок F-міра, множини ключових слів і словосполучень, враховані у F-мірі покращуються.

Висновки

Отже, запропоновано удосконалений метод виявлення ключових слів та словосполучень в україномовних і англомовних web-текстах, що базується на використанні словника предметної області, що надасть можливість виявляти лише релевантні слова та словосполучення, а також підвищити швидкодію індексації і реферування web-текстів. При цьому F-міра підвищиться на 9.5%, а повнота і точність на 15%.

Список літератури

1. Bracewell, D. B., Ren F. Multilingual Single Document Keyword Extraction for Information Retrieval. Proceedings of NLP-KE, 2005, pp. 517-522.

2. Большакова Є. І., Клишінскій Е. С., Ланде Д. В., Носков А. А., Пєскова О. В., Ягунова Є. В. Автоматична обробка текстів на природній мові і комп'ютерна лінгвістика: навч. посібник. М .: MIEM, 2011. 272 с.

3. Hasan K. Automatic Keyphrase Extraction: A Survey of the State of the Art / K. Hasan, V. Ng // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. – 2011. – Vol 1. – pp. 1262-1273.

4. Dictionary Based Annotation at Scale with Spark, SolrTextTagger and OpenNLP [Електронний ресурс] / Sujit Pal // Spark Summit 2015. Europe. - URL: https://spark-summit.org/eu-2015/events/dictionary-based-annotation-at-scale-with-spark-solrtexttagger-and-opennlp.

5. Плющ М.Я Граматика української мови: У 2-ч. Ч. 1. Морфеміка. Словотвір. Морфологія: Підручник. — К.: Вища шк., 2005. — 286 с

6. Dostal M. Automatic Keyphrase Extraction Based on NLP and Statistical Methods. Proceedings of the Dateso 2011: Annual International Workshop on Databases, Texts, Specifications and Objects. Pisek, Czech Republic, 2011, pp. 140-145.

7. The Porter Stemming Algorithm – Porter's homepage. [Електронний ресурс]. – Режим доступу: http://tartarus.org/~martin/ PorterStemmer/. – Назва з титул. екрану

8 Агєєв, М. Додаток А. Офіційні метрики РОМІП 2010 / М. Агєєв, І. Кураленок, І. некрестьянам // Праці РОМІП'2010. СПб .: Изд-во НУ ЦСМ. -2010. - с. 172-187.

9. Extractor [Електронний ресурс] – Режим доступу до ресурсу:https://extractor.com/.

10. OpenCalasis [Електронний ресурс] – Режим доступу до ресурсу: https://opencalasis.com/

11. Експертні КС [Електронний ресурс] – Режим доступу до ресурсу: https://expertcs;ua/.

Стаття надійшла: 13.11.18.

Відомості про авторів

Савчук Тамара Олександрівна – Ph D,професор кафедри комп'ютерних наук Вінницького національного технічного університету.

Горобець ЮрійВолодимирович – магістрант кафедри комп'ютерних наук Вінницького національного технічного університету.

Т. О. Савчук, Ю.В. Горобец

УСОВЕРШЕНСТВОВАНЫЙ МЕТОД ВЫЯВЛЕНИЯ КЛЮЧЕВЫХ СЛОВ В WEB-ТЕКСТЕ

Винницкий национальный технический университет, г. Винница

T. O. Savchuk, Y. V. Horobets

IMPROVED METHOD OF EXTRACTION OF KEYWORDS IN THE WEB-TEXT

Vinnytsia National Technical University, Vinnytsia