УДК 004.93:159.95

О. В. Бісікало
О. В. Яхимович

КОЛІЗІЯ ПРИ ЗНАХОДЖЕННІ КЛЮЧОВИХ СЛІВ

Вінницький національний технічний університет

Анотація
Розглядається колізія при визначенні ключових слів і можливі шляхи її вирішення, а також можливість покращити результати роботи методу визначення ключових слів для невеликих текстів.
Ключові слова: колізія, ключові слова, аналіз текстів, лінгвістичний пакет, DKPro Core, частота.

Abstract
Consider the collision in determining the keywords and possible ways to address it, and the ability to improve the results of the method for determining keywords for small texts.
Keywords: collision, keywords, text analysis, linguistic package, DKPro Core, frequency.

Вступ

Основний зміст документа (тексту) може бути представлений за допомогою певних слів, узятих безпосередньо з цього тексту. Як правило, до кожного розгорнутого тексту можна скласти цілий набір ключових слів різного обсягу (від 5 до 15 слів). Але взагалі кількість ключових слів може варіюватися в широких межах [1].

Ключовим називають таке слово в тексті, яке здатне в сукупності з іншими ключовими словами представляти зміст тексту.

Метою роботи є розробка підходу для зменшення колізії при визначенні ключових слів і його застосування для покращення результатів роботи методу визначення ключових слів для невеликих за розміром текстів.

Результати дослідження

В інформатиці та криптографії колізія хеш-функції – це рівність значень хеш-функції на двох різних блоках даних.

Колізія при знаходженні ключових слів – це рівність значень частоти для двох чи більше кандидатів в ключові слова, причому вибрати в якості ключових з них потрібно тільки частину. Здебільшого така задача актуальна для текстів невеликого розміру, таких як анотація або записи мікроблогів.

Розглянемо приклад – при визначенні ключових слів англомовного тексту на основі інструментальних засобів пакету DKPro Core [2] для тексту [3] отримано та наведено в таблиці 1 список ключових слів, що відсортовані за кількістю зв’язків (частотою) по спаданню.

Таблиця 1. – Ключові слова і кількість зв’язків для них

Слово	Кількість зв'язків (частота)	Слово	Кількість зв'язків (частота)	Слово	Кількість зв'язків (частота)	Слово	Кількість зв'язків (частота)
model	9	plan	4	challenge	2	savings	2
line	7	vehicle	4	creation	2	support	2
product	7	define	3	derive	2	variant	2
reuse	7	design	3	development	2	activity	1
variability	7	diagram	3	domain	2	adaptive	1
approach	6	extract	3	feature	2	architecture	1
base	6	identify	3	implement	2	brazilian	1

Продовження таблиці 1

Слово	Кількість зв'язків (частота)	Слово	Кількість зв'язків (частота)	Слово	Кількість зв'язків (частота)	Слово	Кількість зв'язків (частота)
software	6	mechanism	3	key	2	finally	1
increase	5	offer	3	launcher	2	hypothetic	1
management	5	reflection	3	productivity	2	large-scale	1
process	5	step	3	propose	2	object	1
aim	4	study	3	quality	2	space	1
issue	4	benefit	2	satellite	2	specific	1

Колізія виникає тоді, коли потрібно вибрати зі списку потенційних ключових слів тільки перших N слів з найбільшою частотою, які і будуть вважатися ключовими словами. Для наведеного прикладу тексту, якщо потрібно 10 ключових слів, то перших вісім вибрати легко. Це будуть слова: model, line, product, reuse, variability, approach, base, software. Тоді необхідні останні два слова потрібно вибрати між трьома словами з однаковою частотою п'ять: increase, management, process. Тому для невеликих текстів розв'язання такої колізії є актуальною задачею.

Для зменшення колізії можна використати такі підходи:

- Відсортувати слова з однаковою частотою за частотою їх появи в певному корпусі. Відносна значимість термінів в аналізованому контексті визначається за допомогою даних про частоту їх використання в якості ключових в інтернет-енциклопедії Вікіпедія. Робота алгоритму заснована на розрахунку "інформативності" кожного терміна, тобто оцінки ймовірності того, що він може бути обраний ключовим в тексті [4]. Такий підхід є досить точним, але потребує попереднього аналізу корпусу.

- Відсортувати слова з однаковою частотою за частотою їх появи в частотному словнику словоформ для даного тексту. Такий підхід, що узагальнює слова до словоформ, не потребує попередньої обробки корпусу текстів і легкий в реалізації, але точність його невелика.

- Перевіряти список ключових слів на зв'язність, тобто враховувати парні залежності для різних типів речень [5]. Вибирати ключовими нові слова, які мають більшу сумарну кількість зв’язків з тими словами, що раніше потрапили до списку ключових. Цей підхід не потребує корпусу і його можна реалізувати засобами того ж самого лінгвістичного пакету DKPro Core.

- Вибирати спочатку іменники, потім дієслова, а потім інші частини мови. Оскільки головні члени речення зазвичай бувають іменниками та дієсловами, то вибиратися будуть саме ті слова, які потенційно можуть належати до множини ключових. Також, для іменників можна спочатку вибирати власні назви, тому що одним з запитань, на які повинні відповідати ключові слова: з якими назвами організацій, персон, географічних областей тощо асоціюється стаття [6].

У розробленому авторами методі визначенні ключових слів англомовного тексту на основі інструментальних засобів пакету DKPro Core [2] враховуються парні залежності для різних типів речень, причому визначати частини мови можна засобами DKPro Core [7]. Тому, комбінуючи два останні підходи для зменшення колізії можна покращити результати знаходження ключових слів для невеликих за розміром текстів.

Пропонується для комбінованого підходу спочатку перевіряти ключові слова з однаковою частотою на зв'язність. На другому етапі, якщо у блоці потенційних ще залишилися ключові слова з однаковою частотою, вибираються спочатку іменники, потім дієслова, а потім інші частини мови. Наступні експериментальні дослідження мають підтвердити доцільність використання такого підходу.

Висновки

Колізія виникає тоді, коли декілька кандидатів в ключові слова мають однакову частоту, а серед них потрібно визнати ключовими меншу кількість слів, що особливо актуально для невеликих за розміром текстів.

Комбінований підхід для зменшення колізії можна використовувати як додатковий модуль, що покращить результати знаходження ключових слів для відомого методу визначення ключових слів англомовного тексту на основі інструментальних засобів пакету DKPro Core, а також для інших алгоритмів знаходження ключових слів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Ершов Ю. С. Выделение ключевых слов в русскоязычных текстах / Ю. С. Ершов // Молодежный научно-технический вестник. – М.: ФГБОУ ВПО "МГТУ им. Н.Э. Баумана", 2014. – № ФС77-51038. – С. 70-79.

2. Bisikalo O.V. Method of determining of keywords in English texts based on the DKPro Core / Bisikalo, O.V., Wójcik, W., Yahimovich, O.V., Smailova, S. // Proceedings of SPIE 10031, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2016. - Wilga, Poland 28 September 2016. - DOI:10.1117/12.2249225

3. Burgareli, L. A. (2009, Jul.-Dec.). Variability management in software product lines using adaptive object and reflection. Journal of Aerospace Technology and Management, V. 1, № 2. Available: http://www.jatm.com.br/papers/vol1_n2/JA-TMv1n2_thesis_abstracts.pdf. Last accessed 12.03.2017.

4. Коршунов А. В. Извлечение ключевых терминов из сообщений микроблогов с помощью Википедии / Коршунов А. В. // Труды Института системного программирования РАН. – 2011. – №20. – С. 102-115.

5. Бісікало О.В. Формальні методи образного аналізу та синтезу природно-мовних конструкцій: монографія / О. В. Бісікало. – Вінниця: ВНТУ, 2013. – 316 с.

6. Абрамов Е. Г. Подбор ключевых слов для научной статьи / Е. Г. Абрамов // Научная периодика: проблемы и решения. – 2011. – № 1(2). – С. 35-40.

7. Natural Language Processing: Integration of Automatic and Manual Analysis [Електронний ресурс]. – Режим доступу: http://tuprints.ulb.tu-darmstadt.de/4151/1/rec-thesis-final.pdf – Назва з екрану.

Олег Володимирович Бісікало — доктор технічних наук, професор, декан факультету комп'ютерних систем і автоматики, Вінницький національний технічний університет, Вінниця.

Олександр Вікторович Яхимович — аспірант кафедри автоматики та інформаційно-вимірювальної техніки, факультет комп'ютерних систем і автоматики, Вінницький національний технічний університет, Вінниця, e-mail: yahimovich.olexandr@gmail.com.

Oleg V. Bisikalo — Doctor of Engineering, Professor, Dean of Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia

Alexander V. Yahimovich — Department Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia, email: yahimovich.olexandr@gmail.com.