УДК 004.021 + 004.93

В. Ф. Хісматулліна
Ю. Ю. Іванов

ФОНЕТИЧНИЙ ПОШУК НА ОСНОВІ АЛГОРИТМУ METAPHONE

Вінницький національний технічний університет

Анотація
У роботі проаналізовано особливості фонетичного пошуку на основі алгоритму MetaPhone.
Ключові слова: фонетичний пошук, алгоритм фонетичного кодування, код-ключ, метрика подібності, алгоритм MetaPhone.

Abstract
In this paper have been analyzed some features of phonetic search based on the MetaPhone algorithm.
Keywords: phonetic search, phonetic coding algorithm, code-key, similarity metric, MetaPhone algorithm.

Вступ

Задача фонетичного пошуку зводиться до формування пошукових індексів за фонетичним ключем. Розробка та дослідження подібних алгоритмів, в яких необхідно порівнювати акустичні дані з текстовими зразками (розпізнавання мовлення, коригування орфографії, пошук в базах даних, ідентифікація користувачів), є важливою задачею сучасної комп'ютерної науки. Дослідження та формалізація даних алгоритмів є актуальними задачами.

Метою роботи є розв'язання задачі фонетичного пошуку за допомогою алгоритму фонетичного кодування MetaPhone.

Результати дослідження

Алгоритм фонетичного кодування – це алгоритм індексування слів за звучанням, який на основі послідовності літер і правил вимови перетворює їх в текст для подальшого порівняння. Існує низка фонетичних алгоритмів SoundEx, Daitch-Mokotoff SoundEx, MetaPhone, CaverPhone, NYSIIS, PolyPhone та інші. Дані класичні алгоритми розроблялися для роботи з незмінними формами слів (наприклад, прізвища). Алгоритми фонетичного кодування включають в себе не тільки алгоритми для порівняння слів, але і алгоритми визначення відстані між словами при пошуку за звучанням. На практиці найбільшого поширення набули алгоритми обчислення відстані Левенштейна, Хеммінга, Джаро, Джаро-Вінклера та на основі *N*-грам. Для фонетичного пошуку важливо отримати основу слова. Тому застосовують класичні процедури передобробки контенту: лематизацію і стемінг [1, 2].

Metaphone – ще один алгоритм фонетичного кодування слів з урахуванням основних правил англійської мови, розроблений в 1990 році [3]. На виході алгоритм дає код змінної довжини, який складається з букв. Алгоритм включає в себе 16 кроків. У 2000 році, була розроблена друга версія даного алгоритму, яка отримала назву Double MetaPhone, в якому, на відміну від першої версії, що застосовується тільки для англійської мови, враховуються особливості вимови слів, запозичених з інших мов. Результатом роботи даного алгоритму є два коди – по одному для кожного варіанта вимови. Хоча Double MetaPhone має переваги перед алгоритмом MetaPhone, він має деякі обмеження. Зокрема, зустрічаються слова з різними вимовами і однаковим кодом, наприклад, "Alice", "Elsa" і "Ullos" кодуються як "ALS" [4]. У 2009 році з'явилася третя комерційна версія алгоритму під назвою MetaPhone 3. Він почав підтримувати запозичені слова з більшої кількості мов. Алгоритм включає велику кількість правил, а об'єм коду на мові Java складає більше 7 тисяч рядків. Визначено, що MetaPhone 3 збільшує точність ототожнення слів з 89% (Double MetaPhone) до 98% (MetaPhone 3) [5].

Висновки

Основні алгоритми фонетичного кодування запропоновані досить давно, але дослідження подібних алгоритмів ніколи не припинялося. Однак нових результатів в цій області отримано не було, а основним підсумком усіх досліджень є покращення ефективності базових алгоритмів. У даній роботі розглянуто особливості фонетичного пошуку з використанням алгоритму MetaPhone. Даний

алгоритм у поєднанні з алгоритмами обчислення подібності між словами може значно покращити якість пошукових результатів у базах даних.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Abdulhayoglu M.A. Use of ResearchGate and Google CSE for Author Name Disambiguation / M.A. Abdulhayoglu, B. Thijs // Scientometrics. – Budapest: Akademiai Kiado, 2017. – P. 1965-1985.
2. Выхованец В.С. Обзор алгоритмов фонетического кодирования / В.С. Выхованец, Ц. Ду, С.А. Сакулин // Управление большими системами. – Вып. 73. – М., 2018. – С. 67-94.
3. Lawrence P. Hanging on the Metaphone / P. Lawrence // Computer Language. – 1990. – Vol. 7. – № 12. – P. 39-44.
4. Каньковски П. “Как ваша фамилия”, или русский MetaPhone / П. Каньковски // Программист. – 2002. – Вып. 8. – С. 36-39.
5. Впровадження технології оптимізації індексування вузькоспеціалізованих термінів на базі фонетичного алгоритму Metaphone [Електронний ресурс] / В.Л. Бурячок, М.М. Гаджиєв, В.Ю. Соколов, П.М. Складанний, Л.В. Кузьменко. – Режим доступу: http://journals.uran.ua/eejet/article/download/181943/182455.

Хісматулліна Валентина Фанілівна — студентка групи 1ІСТ-19м, факультет комп’ютерних систем і автоматики, Вінницький національний технічний університет, м. Вінниця.

Іванов Юрій Юрійович — канд. техн. наук, доцент кафедри автоматизації та інтелектуальних інформаційних технологій, Вінницький національний технічний університет, м. Вінниця, e-mail: Yura881990@i.ua.

Hismatullina Valentina F. — student, Faculty of Computer Systems and Automation, Vinnytsia National Technical University.

Ivanov Yurii Yu. — Cand. Sc. (Eng), Senior Lecturer, Faculty of Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: Yura881990@i.ua.