УДК 004.021 + 004.93

О. І. Черешнюк
Ю. Ю. Іванов
О. М. Бевз
В. В. Кабачій

ОГЛЯД АЛГОРИТМІВ ФОНЕТИЧНОГО КОДУВАННЯ

Вінницький національний технічний університет

Анотація

У роботі проаналізовано алгоритми фонетичного кодування і метрики, які використовують для оцінювання фонетичної подібності слів.

Ключові слова: фонетика, алгоритм фонетичного кодування, метод еквівалентних перетворень, метрика подібності слів.

Abstract

In this paper have been analyzed a few phonetic coding algorithms and metrics, that are used to evaluate the phonetic similarity of words.

Keywords: phonetics, phonetic coding algorithm, equivalent transform method, word similarity metric.

Вступ

У сучасному світі взаємодія між людиною і машиною стала буденністю. Завдяки технічному прогресу вона не обмежується простим натисканням клавіш і виведенням інформації на екран. Користувач взаємодіє з пристроями різними способами: за допомогою сенсорів, жестів і голосових команд. Зараз всі ці можливості активно використовуються в повсякденному житті, наприклад існують мультимедійні додатки для вивчення іноземних мов. Для перевірки орфографії у таких програмах використовують швидкі фонетичні алгоритми, які формують однакові коди для слів із схожим звучанням (вимовою), що дозволяє здійснювати порівняння й індексацію множини таких слів на основі їх фонетичної подібності [1].

Метою даної роботи є огляд та аналіз алгоритмів фонетичного кодування і метрик, які використовують для оцінювання фонетичної подібності.

Результати дослідження

Алгоритм SoundEx запатентований у 1918 році для роботи з англійською мовою, тому існує досить сильна залежність від заданої мови. У наш час розроблено багато модифікацій даного алгоритму для різних мов: китайської, іспанської, перської, арабської, малайської тощо [1].

Алгоритм NYSIIS розроблений у 1970 році для використання в однойменній інформаційній системі «*New York State Identification and Intelligence System*». Цей алгоритм дає трохи кращі результати порівняно з алгоритмом *SoundEx*, використовуючи більш складні правила перетворення вхідного слова в код. Також враховуються особливості написання слів та функціонування голосних звуків під час їх вимови. У ході дослідження алгоритмів фонетичного кодування встановлено, що даний алгоритм більш всього підходить для кодування прізвищ [1].

Для врахування специфіки вимови слів у різних мовах розроблений *алгоритм Daitch-Mokotoff SoundEx*, названий за прізвищами авторів, який використовує більшу довжину коду. Даний алгоритм має складніші правила перетворення вхідного слова в код. Як і в алгоритмі *NYSIIS*, у формуванні результуючого коду беруть участь не лише поодинокі букви, але і їх послідовності [2].

Metaphone – ще один алгоритм фонетичного кодування слів з урахуванням основних правил англійської мови, розроблений у 1990 році. Він відрізняється від попередніх алгоритмів тим, що реалізує детальніші правила кодування. Ще одна відмінність полягає в тому, що літери не розбиваються на групи і не кодуються цифрами. На виході алгоритм дає код змінної довжини, який складається з букв [3].

Спеціально для російської мови розроблений алгоритм фонетичного кодування *Polyphone*. У ньому враховані морфологічні, фонологічні, фонетичні та історичні аспекти вимови слів. Даний алгоритм можна досить легко модифікувати для роботи з українською мовою [1].

Основним методом, реалізованим у розглянутих алгоритмах фонетичного кодування, є метод еквівалентних перетворень слова за звучанням, при якому частина слова, яка належить певній множині, замінюється її кодом або типовим представником. При цьому помітно, що частини слів з однієї множини, близькі за звучанням, також близькі за написанням.

При введенні відповідної метрики на словах можна поставити задачу визначення їх схожості за звучанням шляхом обчислення відстані між ними за написанням. Цей підхід використовується в іншому класі алгоритмів фонетичного кодування, в якому обчислення фонетичного коду слова замінюється попарним порівнянням слів та обчисленням відстані між ними в певному метричному просторі. Зазвичай можна застосувати ряд метрик:

1. *Відстань Левенштейна* – це міра відмінності двох слів, яка враховує мінімальну кількість операцій вставки, видалення і заміни, які необхідні для перетворення одного слова в інше [4].

2. *Відстань на основі N-грам*. *N*-грамою називається послідовність з *N* елементів (букв). Для визначення фонетичної схожості двох слів обчислюється кількість загальних *N*-грам (зазвичай $N = 3$). Очевидно, що такий розрахунок відстані між словами дає кращий результат для більш довгих слів, ніж для коротких. Даний алгоритм має досить низьку обчислювальну складність [1, 5].

3. *Відстань Джаро* – це мінімальна кількість однобуквених змін, яку необхідно виконати для перетворення одного слова в інше. Чим менша відстань Джаро, тим більше схожі слова. Слід зазначити, що обчислювальна складність визначення даної метрики найвища з розглянутих [1, 6].

Висновки

У даній роботі проведено огляд алгоритмів фонетичного кодування і метрик, які використовують для оцінювання фонетичної подібності слів. Представлено основні ідеї, які можна реалізувати у програмному забезпеченні.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Выхованец В.С. Обзор алгоритмов фонетического кодирования / В.С. Выхованец, Ц. Ду, С.А. Сакулин // Управление большими системами. – М., 2018. – С. 67-94.

2. Soundexing and Genealogy by Gary Mokotoff [Web Resource]. – Access mode: http://www.avotaynu.com/soundex.htm.

3. Lawrence P. Hanging on the Metaphone / P. Lawrence // Computer Language. – 1990. – V. 7 (12). – P. 39-44.

4. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов / В.И. Левенштейн // Доклад АН СССР. – 1965. – Выпуск 163 (4). – С. 845-848.

5. Jurafsky D. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition / D. Jurafsky, J.H. Martin. – Pearson Prentice Hall, 2009. – 988 p.

6. Jaro M.A. UNIMATCH – a Computer System for Generalized Record Linkage Under Conditions of Uncertainty / M.A. Jaro // Spring Joint Computer Conference. – Anaheim (USA), 1972. – P. 523-530.

Черешнюк Олексій Ігорович — студент групи 2СІ-16б, 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.

Бевз Олександр Миколайович — канд. техн. наук, доцент кафедри автоматизації та інтелектуальних інформаційних технологій, Вінницький національний технічний університет, м. Вінниця.

Кабачій Владислав Володимирович — канд. техн. наук, доцент кафедри автоматизації та інтелектуальних інформаційних технологій, Вінницький національний технічний університет, м. Вінниця.

Chereshnyuk Olexiy I. — student, Faculty of Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia.

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

Bevz Alexander M. — Cand. Sc. (Eng), Senior Lecturer, Faculty of Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia.

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