УДК 004.424.62/43

В. Ермакова, Т.В. Селиверстова

ИССЛЕДОВАНИЕ ФОНЕТИЧЕСКИХ АЛГОРИТМОВ С ПОМОЩЬЮ ОЦЕНКИ JARO

Аннотация. Рассмотрены основные фонетические алгоритмы и их эффективность при варьировании входных параметров. Разработана программа, реализующая алгоритмы Soundex, Daitch-Mokotoff Soundex, MetaPhoneRU и анализирующая схожесть результатов с помощью метрик похожести.

Ключевыеслова: ФОНЕТИЧЕСКИЙАЛГОРИТМ, НЕЧЕТКИЙПОИСК

Постановка проблемы. Поиск и извлечение информации из современных мощных баз данных может усложняться по причинам, связанным с неточностью строки запроса. При работе с фамилиями можно выделяют три основные причины, по которым возникает данная проблема: ошибки при печати (в базе данных или же в строке поиска), неточности транслитерации и отсутствие четкого представления о написании фамилии.

Для решения этой проблемы используют нечеткий поиск. Он охватывает множество областей, таких как: проверка орфографии, поиск в глобальных поисковиках, сжатие данных, распознавание речи, криптография. Нечеткий поиск в контексте фонетических алгоритмов дает такие возможности, как: поиск слова по звучанию, поиск подобных по написанию слов, поиск слов с опечатками, поиск в транслитерованных словах. Важная часть поиска по словам – поиск по фамилиям. Практическая ценность решения этой задачи в наше время высока, так как существует повсеместное использование баз данных. Поэтому актуальной является практическая задача поиска по фамилии, при отсутствии уверенности в правильности ее написания.

Постановка задачи. Целью работы является исследования результатов работы основных алгоритмов фонетического поиска: Soundex, Daitch-Mokotoff Soundex, MetaPhone, путем вычисления метрики Jaro.

Изложение основных материалов исследования. Существует несколько основных фонетических алгоритмов: Soundex, Daitch-Mokotoff Soundex, MetaPhone, MetaPhoneRU, Caverphone, NYSIIS. Алгоритм Soundex наиболее универсален, однако, даёт наибольшее число совпадений при сомнительной фонетической схожести. Алгоритм Caverphone был разработан с уклоном в новозеландские фамилии. Алгоритм NYSIIS – для жителей Нью-Йорка [1].

Для работы со славянскими фамилиями наиболее целесообразно использовать следующие три алгоритма: Soundex (универсален и прост в исполнении), Daitch-Mokotoff Soundex (адаптирован авторами для русских фамилий) и MetaPhoneRU – доработанный Soundex Петром Каньковски для русского языка [2].

Рассмотрим выбранные алгоритмы более детально.

Порядок действий в Soundex:

1. Преобразуем фамилию в верхний регистр
2. Запоминаем первую букву, удаляем ее из рабочей строки
3. Заменяем оставшуюся строку по таблице 1.
4. Прибавляем букву

Таблица 1

Преобразования букв по алгоритму Soundex

Исходные символы	Конечный символ
А Е Ё И О У Х Ъ Ы Ь Э Ю Я	
Б П	1
Г Ж Й	4
Д Т	6
В Ф	2
З	5
Л	7
М Н	8
Р	9
К С Ц	3
Ц Ш Щ	@

Порядок действий в MetaPhoneRU:

1. Преобразуем фамилиюв верхний регистр
2. Заменяем латинские буквы схожего написания на русские (например, Н/eɪtʃ/ на Н /эн/)

 ISSN 1562-9945

3. Преобразуем фамилиюв нижний регистр
4. Буквосочетания «тс» и «дс» заменяем на «ц».
5. Происходит замена окончаний по таблице 2.

Таблица 2

Замена окончаний

Исходные окончания	Конечный символ
–УК, –ЮК	0
–ИНА	1
–ИК, –ЕК	2
–НКО	3
–ОВ, –ЕВ, [–ИЕВ, –ЕЕВ]	4
–ЫХ, –ИХ	5
–АЯ	6
–ИЙ, –ЫЙ	7
–ИН	8
–ОВА, –ЕВА, [–ИЕВА, –ЕЕВА]	9
–ОВСКИЙ	@
–ЕВСКИЙ	#
–ОВСКАЯ	$
–ЕВСКАЯ	%

6. Удаляем повторы (например, «нн» превращается в «н»).
7. Оглушаем согласные в слабой позиции (табл. 3)

Таблица 3

Замена согласных

Исходные символы	Конечный символ
П, Б	П
С, З	С
Т, Д	Т
Ф, В	Ф
К, Г	К

8. Удаляем повторы
9. Замена гласных по таблице 4

Таблица 4

Замена гласных

Исходные символы	Конечный символ
О, Ы, А, Я	А
Ю,У	У
Е, Ё, Э, И	И

10. Удаление «ь», «ъ», «–».

Порядок действий в Daitch-MokotoffSoundex:

1. Слово транслитерируется
2. Буквы заменяются согласно таблице

Таблица 5

Преобразования букв по алгоритму Daitch-MokotoffSoundex

Исходные буквосочетания	В начале	За гласной	Остальное
AI, AJ, AY, EI, EY, EJ, OI, OJ, OY, UI, UJ, UY	0	1	
AU	0	7	
1	2	3	4
1	2	3	4
IA, IE, IO, IU	1		
EU	1	1	
A, UE, E, I, O, U, Y	0		
J	1	1	1
SCHTSCH, SCHTSH, SCHTCH, SHTCH, SHCH, SHTSH, STCH, STSCH, STRZ, STRS, STSH, SZCZ, SZCS	2	4	4
SHT, SCHT, SCHD, ST, SZT, SHD, SZD, SD	2	$	$
CSZ, CZS, CS, CZ, DRZ, DRS, DSH, DS, DZH, DZS, DZ, TRZ, TRS, TRCH, TSH, TTSZ, TTZ, TZS, TSZ, SZ, TTCH, TCH, TTSCH, ZSCH, ZHSH, SCH, SH, TTS, TC, TS, TZ,	4	4	4

 ISSN 1562-9945

ZH, ZS			
SC	2	4	4
DT, D, TH, T	3	3	3
CHS, KS, X	5	%	%
S, Z	4	4	4
CH, CK, C, G, KH, K, Q	5	5	5
MN, NM		^	^
M, N	6	6	6
FB, B, PH, PF, F, P, V, W	7	7	7
H	5	5	
L	8	8	8
R	9	9	9

В процессе исследования был реализован программный продукт «PhSA – PhoneticSearchAlgorithms», автоматизирующий действия по сравнению строк и систематизированию результатов. Для создания программного продукта был использован язык C# и база данных AccessDB. В его состав входит главная («PhSA») и вспомогательная программа («PhSAHelper»). Задача главной программы – производить поиск в базе данных, ранее сформированной вспомогательной программой. Вспомогательная программа снабжена функциями: добавления ключей для быстрого поиска по каждому алгоритму, удаления случайных символов и пробелов перед записью, работы с двойными фамилиями. Таким образом, при использовании вспомогательной программы «PhSAHelper» создаётся новая база данных, содержащая исходный столбец с фамилиями и еще 3 столбца с ключами, соответствующими каждому алгоритму.

Тестирование работы фонетических алгоритмов обработки славянских фамилий производилось на данных телефонной книги Днепропетровской области.

Для исследования фонетических алгоритмов с помощью программного продукта «PhSA» необходимо ввести фамилию в строку поиска, после чего выбрать режим поиска по базе ключей (в противном случае, генерация ключа будет происходить «на лету»), и, непосредственно, сам алгоритм. После нажатия на кнопку «Искать!» отображаются результаты поиска (рис. 1). Разработанный программный

продукт «PhSA» содержит справочную информацию об используемых алгоритмах (рис. 2).

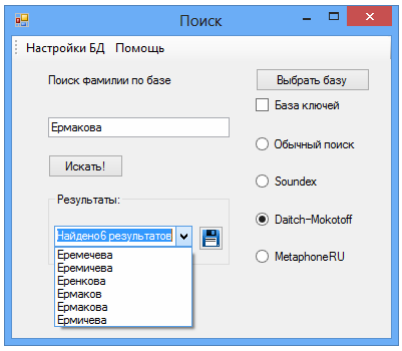

Рисунок 1 – Интерфейс «PhSA», с результатами поиска

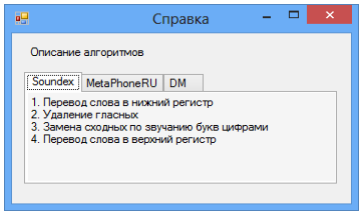

Рисунок 2 – Интерфейс «PhSA», справочная информация

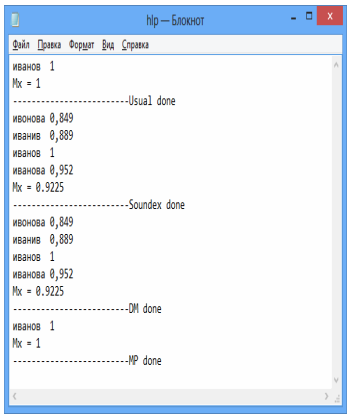

Рисунок 3 – Текстовый документ, созданный в результате поиска

Программный продукт «PhSA» предоставляет возможность сохранения результатов поиска, оценки Jaro[4] и математического ожидания оценки Jaro в файл, формат записи в который приведен на рис. 3.

С целью исследования работы фонетических алгоритмов было проведено их тестирование на двух группах фамилий (таблица 1). Группа 1 – содержит фамилии из списка «25 общерусских фамилий» [3], группа 2 – редкие, либо принадлежащие другим этническим группам.

Результаты приведенные в таблице 6 подтверждают эффективность работы алгоритмов Soundex и Daitch-Mokotoff Soundex для распространенных славянских фамилий. Так же из данных приведенных в таблице 1 следует, что алгоритм MetaPhoneRU «склонен» отыскивать точные совпадения и предоставляет меньшее количество подобных, с точки зрения алгоритма, фамилий. Так же проводилось исследование работы фонетических алгоритмов на фамилиях из славянской группы разной длины (таблица 7). Данные приведенные в таблице 2 демонстрируют различия работы алгоритмов при различной длине запроса и позволяют сделать вывод о том, что ключ для длинных фамилий оказывается «уникальным», т.е. ему находится меньшее количество соответствий в базе по сравнению с более короткими строками запроса.

 ISSN 1562-9945

Таблица 6

Результат работы фонетических алгоритмов для фамилий различных этнических групп (N– количество фамилий соответствующих ключу, сформированному каждым из алгоритмов; J– математическое ожидание оценки Jaro)

Группа фамилий	Запрос	Алгоритм					
		Soundex		Daitch-Mokotoff Soundex		MetaPhoneRU	
		N	J	N	J	N	J
1	Иванов	4	0,922	4	0,922	1	1
	Васильев	16	0,632	12	0,679	2	0,956
	Петров	10	0,735	10	0,735	1	1
	Смирнов	3	0,917	3	0,917	1	1
2	Ояпер	4	0,561	5	0,544	1	1
	Мороз	5	0,698	9	0,590	1	1
	Мухомор	2	0,603	2	0,603	1	1
	Сорока	14	0,575	18	0,554	1	1

Таблица 7

Результат работы фонетических алгоритмов для фамилий разной длины

Длина фамилии	Запрос	Алгоритм					
		Soundex		Daitch-Mokotoff Soundex		MetaPhoneRU	
		N	J	N	J	N	J
до 3-х букв	Дуб	5	0,817	19	0,570	1	1
	Сом	23	0,366	47	0,318	1	1
	Чиж	5	0,565	12	0,429	1	1
до 9-ти букв	Ермакова	4	0,924	7	0,776	1	1
	Васильев	16	0,632	12	0,679	2	0,956
	Смирнов	3	0,917	3	0,917	1	1
больше 9-ти букв	Воскобойникова	2	0,988	2	0,988	1	1
	Константинопольский	1	1	1	1	1	1
	Космодемьянская	1	1	1	1	1	1

Выводы

1. Для исследования работы фонетических алгоритмов Soundex, Daitch-Mokotoff Soundex, MetaPhone был разработан программный продукт «PhSA – PhoneticSearchAlgorithms», автоматизирующий действия по сравнению строк и систематизированию результатов.

2. Исследование работы фонетических алгоритмов на фамилиях различных этнических групп и славянских фамилиях различной длины показали, что алгоритмы Soundex иDaitch-Mokotoff Soundex дают схожие результаты, при высокой вычислительнойзатратности алгоритма Daitch-Mokotoff Soundex по сравнению с Soundex, а алгоритм MetaPhoneRU генерирует такой ключ, на который, чаще всего, приходится один вариант результата поиска, и, как следствие, имеет высокую оценку Jaro.

ЛИТЕРАТУРА

1. Фонетические алгоритмы // habrhabr.ru ежедн. интернет-изд. 2011. 4 марта. URL: http://habrahabr.ru/post/114947/
2. Каньковски П. «Как ваша фамилия», или Русский MetaPhone // Программист. 2002. №8. С. 36-39.
3. Балановская Е. В., Соловьева Д. С., Балановский О. П. и др. «Фамильные портреты» пяти русских регионов // Медицинская генетика. 2005.№ 1. С. 2-10
4. EcolePolytechnique de Louvain, Universitґecatholique de Louvain, Belgium «Mahalanobis distance, Jaro-Winkler distance and nDollar in UsiGesture» URL:http://www.vinc.be/www-pdf/10-11_SINF2356-Mahalanobis-JaroWinckler-nDollar.pdf

 ISSN 1562-9945