УДК 004.853
DOI: 10.20310/1810-0198-2017-22-5-1138-1141

ФОНЕТИЧЕСКИЙ АЛГОРИТМ ПОИСКА НЕЖЕЛАТЕЛЬНОГО КОНТЕНТА

© Д.В. Лопатин, Е.С. Чиркин, А.А. Фадеева
Тамбовский государственный университет им. Г.Р. Державина
392000, Российская Федерация, г. Тамбов, ул. Интернациональная, 33
E-mail: ccs.tmb@ya.ru

Приведено описание модернизированного фонетического алгоритма поиска нежелательного контента в текстах, написанных на русском языке. Показано, что предложенный вариант фонетического алгоритма можно использовать для нечеткого текстового поиска, если размер обучающей выборки незначителен или он отсутствует.
Ключевые слова: фонетический алгоритм; поиск; контент

ВВЕДЕНИЕ

В настоящее время остро стоит проблема поиска нежелательного контента. Для детектирования вредоносного содержания можно использовать алгоритмы и методы фонетического анализа звукового состава слова.

Высококачественные фонетические алгоритмы применяются для глобальных поисковых систем, сжатия данных, криптографии, распознавания речи, преобразования генетической и молекулярной информации. Одной из актуальных задач применения фонетических алгоритмов является сопровождения БД, входящих в состав информационных систем. При фонетическом поиске для сравнения слова текста и словаря предварительно преобразовываются в форму, напоминающую их звучание или искусственный код. Слова, близкие по звучанию, будут иметь одинаковый код.

К достоинствам фонетических алгоритмов можно отнести хорошие результаты при поиске неизменяемых форм в ограниченном словаре вводимых пользователем слов, особенно с ручной доработкой словаря. Главным препятствием широкого использования фонетических алгоритмов поиска в современном русском языке является преобладание процесса словообразования с использованием аффиксов (приставка (префикс), суффикс, окончание (флексия), соединительная гласная (интерфикс), постфикс), сочетающих сразу несколько грамматических значений. Кроме того, для любого языка алгоритмы фонетического поиска особенно восприимчивы к ошибкам в начале слов, перестановкам, пропускам и добавлениям букв, ошибкам типа слияние-разделение слов, неточное написание, или используется сленг [1]. Цель работы: модернизировать известные фонетические алгоритмы для поиска нежелательного контента на русском языке.

МОДИФИКАЦИЯ АЛГОРИТМА

Современные фонетические алгоритмы Soundex, Daitch-Mokotoff Soundex, Metaphone, Caverphone, Фонетик, Русский Metaphone достаточно подробно описаны в работах [2–6]. Отметим, что классические алгоритмы разрабатывались для работы с неизменяемыми формами слов (например, фамилии).

Для фонетического поиска важно получить производящую основу слова, для этого можно применять классические процедуры предобработки контента: лемматизации и стемминга [7–11]. Лемматизация – алгоритм привидения исходного слова к нормальной форме (например, для существительных в русском языке это форма именительного падежа в единственном числе), при этом алгоритм требует предварительно составленного словаря всех возможных словоформ. Алгоритм имеет высокую скорость работы, в результате работы снижается размер словаря за счет приведения к единой форме различных словоформ. Стемминг – алгоритм усечения слов по ряду эвристических правил. В основном усечения происходят за счет ограничения количества словообразующих суффиксов и окончаний и заданных правил. К достоинствам стемминга следует отнести высокую скорость работы, снижение словаря, не требует предварительного словаря для своей работы.

С точки зрения применимости модифицированных фонетических алгоритмов поиска к задачам данной работы следует рассматривать следующие категории нежелательного текстового контента: отдельные слова, словосочетания и фразы, имеющие собственное нежелательное значение; отдельные слова (в т. ч. и несуществующие слова, т. е. просто набор букв, цифр или иных символов) и фразы (в т. ч. и скомбинированные из несуществующих слов), не имеющие собственного лексического нежелательного значения вне контекста их употребления; нежелательный контент, выраженный через эвфемизмы, аналогии, аллюзии, гиперболы, в комбинации с заумью, эрративами и другими приемами литературного языка. Все перечисленные литературные приемы умело используются в современной субкультуре для завуалирования нелитературного языка. Ситуацию осложняет тот факт, что русский язык имеет сложную орфографию, морфологическую структуру и словообразовательные средства, все это является источником ошибок при использовании лемматизации и стемминга.

В качестве решения данной проблемы и снижения вычислительных затрат можно предложить следующее для предобработки контента:

1) для процедуры стемминга отказаться от обработки множества существующих суффиксов и создать правила преобразования (удаления) префикса (приставки);

2) отказаться от лемматизации и усечения окончаний, вместо этого рассматривать только определенное количество символов основы слова (это называется непроизводная основа) (4…8), отметим, что усечение слов характерно для классических фонетических алгоритмов.

В настоящей работе используются следующие правила фонетического анализа звукового состава слова (показана только часть вариантов преобразования):

– удаление мягкого и твердого знака, непроизносимых символов или их группы (часто так маскируют негативное содержание, имеет смысл сразу считать такое слово негативным);

– замена транслитерации (H – П, R – Я, SH – Ш, Y – У);

– замена сходных цифр, латинских и кириллических букв (сочетаний) (A, B, C, E, H, M, O, P, T, X);

– замена символов схожего на буквы начертания или смысла (/7 – П, @ – > А, >|< – Ж, >К – Ж, /\ – Л, тт – П);

– замена известных эвфемизмов (3.14 – ПИ);

– преобразование групп согласных (СТН – СН, ТС – Ц, ДС – Ц);

– оглушение согласных в соответствии с правилами русского языка (Б – П, Д – Т, З – С, В – Ф, Ж – Ш, Г – К);

– замена шипящих (Щ, Ч – Ш);

– замена гласных букв как в безударном слоге (О, Я, Ё – А);

– замена гласных букв (например, Е, Ы, Ё, Э – И);

– преобразование повторяющихся символов (НН – Н, АААААА – А);

– возврат к вышеописанным процедурам преобразования, чтобы исключить возможные повторы.

РЕЗУЛЬТАТЫ

Применение правил к исходному слову (ЗЕМЛЯ) приводит к формированию «вырожденного» фонетического образа (СИМЛА). В табл. 1 представлены исходные формы, фонетический образ (с основой до 8 символов), числовой код образа и процентное отношение сходства к основному слову (ЗЕМЛЯ).

Для поиска слова в словаре целесообразно использовать числовое представление (код) фонетического образа. Код можно формировать через хеш-функцию, в этом случае подобный код можно использовать как индекс в хеш-таблице. Или посредством оригинальной функции, которая в зависимости от частоты употребления букв в русском языке преобразует символы строки в коэффициенты полинома (область применения – бинарный поиск, поиск к ближайшему из словаря, интерполирование и т. д.).

Таблица 1

Результат работы оригинального фонетического алгоритма

Исходная форма	Фонетический образ	Оригинальный код	%
ЗЕМЛЯ	СИМЛА	100CD0220A000000	0
З@3emlRть	СИМЛИТ	100CD0220A400000	2
ЗЕМЛЕВЕДЕНИЕ	СИМЛИФИТ	100CD0220C700C40	20
земЛЕВЛадЕлЕц	СИМЛИФЛА	100CD0220C70220A	20
ЗЕМЛЕВЛАДЕНИЕ	СИМЛИФЛА	100CD0220C70220A	20
Zemledelec	СИМЛИТИЛ	100CD0220C400C22	19
ЗЕМЛЕКОП	СИМЛИКАП	100CD0220C0C0A00	17
ЗЕМЛЕКОПНЫЙ	СИМЛИКАП	100CD0220C800A60	20
ЗЕМЛЕМЕР	СИМЛИМИР	100CD0220CD00C30	22
ЗЕМЛЕМЕРНЫЙ	СИМЛИМИР	100CD0220CD00C30	22
зиМЛИП@ШИЦ	СИМЛИПАШ	100CD0220C600A90	19
ЗЕМЛЕПОЛЬЗОВАНИЕ	СИМЛИПАЛ	100CD0220C600A22	19
Zemlepr0x0dec	СИМЛИПРА	100CD0220C300A0A	18
ЗЕМЛЕРОЙКА	СИМЛИРАИ	100CD0220C300A0C	18
ЗемлеС0С	СИМЛИСАС	100CD0220C100A10	17
ЗЕМЛЕТРЯСЕНИЕ	СИМЛИТРА	100CD0220C40300A	18
ЗЕМЛЕУСТРОЙСТВО	СИМЛИУС	100CD0220C0B1000	17
ЗЕМЛИСТЫЙ	СИМЛИСТИ	100CD0220C104070	17
ЗЕМЛИЦА	СИМЛИЦА	100CD0220C110A00	18
ЗЕМЛИШКА	СИМЛИШКА	100CD0220C90800A	20
ЗЕМЛЯНЕ	СИМЛАНИ	100CD0220A200C00	1
ЗЕМЛЯЧЕСТВО	СИМЛАЧИС	100CD0220AD00C10	2
ЗЕМЛЯЧОК	СИМЛАЧАК	100CD0220CD00A80	22
ПРИЗЕМЛЕНИЕ	СИМЛИНИИ	100CD0220C200C0C	18
ПРИЗЕМЛЁННОСТЬ	СИМЛИНА	100CD0220C200A00	18
ПРИЗЕМЛИТЬСЯ	СИМЛИЦ	100CD0220C110000	17
ПРИЗЕМЛЯТЬ	СИМЛАТ	100CD0220A407000	2

Таблица 2

Величина ошибки в тестовой выборке

Размер обучающей выборки, примеров	Величина ошибки, %	
	Наивный поиск	Фонетический поиск
50	20,79	10,11
100	10,16	3,72
500	2,12	2,64
1000	1,01	2,55
2000	0,52	2,61
3000	0,33	2,72
4000	0,21	2,61
5000	0,20	2,54
6000	0,16	2,03
7000	0,15	1,45

Для проверки эффективности работы предложенного алгоритма были получены словари фонетических образов на основе размеченных корпусов нежелательного контента, полученных в работе [12], содержащих порядка 2500 упоминаний четырех сущностей «популярного» нелитературного языка. Размеченные документы разделялись однородно на обучающую (использовалась для формирования корпусов нежелательного контента) и тестовую выборки в пропорции 70:30. Источником документов были комментарии пользователей с ряда развлекательных ресурсов. Видно, что на начальном этапе количество ошибок в тестовой выборке в зависимости от размера обучающей выборки меньше, чем при наивном поиске (табл. 2). Таким образом, предложенный модифицированный вариант фонетического алгоритма можно использовать для решения задачи нечеткого текстового поиска (фильтрации нежелательного контента), если размер обучающей выборки незначителен или он отсутствует.

СПИСОК ЛИТЕРАТУРЫ

1. *Binstock A., Rex J.* Practical Algorithms for Programmers. Boston: Addison-Wesley, 1995. 577 p.
2. *Moyal A., Aharonson V., Tetariy E., Gishri M.* Phonetic Search Methods for Large Speech Databases. Heidelberg: Springer-Verlag, 2013. 53 p.
3. Soundex Coding. NARA and Daitch-Mokotoff Soundex. URL: http://www.jewishgen.org/infofiles/soundex.html (accessed: 15.07.2017)
4. *Philips L.* Hanging on the Metaphone // Computer Language. 1990. V. 7. № 12. P. 38-45.
5. *Hood D.* Caversham Project Occasional Technical Paper. 2004. URL: http:// caversham.otago.ac.nz/files/working/ctp150804.pdf (accessed: 15.07.2017)
6. *Каньковски П.* «Как ваша фамилия», или Русский MetaPhone // Программист. 2002. № 8. С. 36-39.
7. *Tomlinson S.* Lexical and Algorithmic Stemming Compared for 9 European Languages with Hummingbird SearchServerTM at CLEF 2003 // Peters C., Gonzalo J., Braschler M., Kluck M. (eds.) Comparative Evaluation of Multilingual Information Access Systems. CLEF 2003. Lecture Notes in Computer Science. Berlin: Springer, 2004. V. 3237. P. 286-300.
8. *Губин М.В., Морозов А.Б.* Влияние морфологического анализа на качество информационного поиска // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: труды 8 Всерос. науч. конф. Суздаль, 2006. С. 95-100.
9. Russian stemming algorithm. URL: http://snowball.tartarus.org/algorithms/russian/stemmer.html (accessed: 15.07.2017)
10. Вероятностный морфологический анализатор русского и украинского языков. URL: http://www.keva.ru/stemka/stemka.html (дата обращения: 15.07.2017)
11. *Segalovich I.* A fast morphological algorithm with unknown. URL: http://download.yandex.ru/company/iseg-las-vegas.pdf (accessed: 15.07.2017)
12. *Чиркин Е.С., Лопатин Д.В.* Подходы к нечеткому поиску нежелательного контента на веб-странице // Вестник Тамбовского университета. Серия Естественные и технические науки. Тамбов, 2016. Т. 21. № 6. С. 2358-2365. DOI: 10.20310/1810-0198-2016-21-6-2358-2365.

БЛАГОДАРНОСТИ: Работа выполнена при финансовой поддержке РФФИ (проект № 15 17 08378).

Поступила в редакцию 30 августа 2017 г.

Лопатин Дмитрий Валерьевич, Тамбовский государственный университет им. Г.Р. Державина, г. Тамбов, Российская Федерация, кандидат физико-математических наук, доцент, директор Центра компьютерной безопасности, e-mail: +79107540080@ya.ru

Чиркин Евгений Сергеевич, Тамбовский государственный университет им. Г.Р. Державина, г. Тамбов, Российская Федерация, инженер Центра компьютерной безопасности, e-mail: ccs.tmb@ya.ru

Фадеева Ангелина Александровна, Тамбовский государственный университет им. Г.Р. Державина, г. Тамбов, Российская Федерация, кандидат филологических наук, доцент кафедры зарубежной филологии и прикладной лингвистики, e-mail: ccs.tmb@ya.ru

Для цитирования: *Лопатин Д.В., Чиркин Е.С., Фадеева А.А.* Фонетический алгоритм поиска нежелательного контента // Вестник Тамбовского университета. Серия Естественные и технические науки. Тамбов, 2017. Т. 22. Вып. 5. С. 1138-1141. DOI: 10.20310/1810-0198-2017-22-5-1138-1141

For citation: Lopatin D.V., Chirkin E.S., Fadeeva A.A. Foneticheskiy algoritm poiska nezhelatel'nogo kontenta [Phonetic search algorithm of inappropriate content]. *Vestnik Tambovskogo universiteta. Seriya Estestvennye i tekhnicheskie nauki – Tambov University Reports. Series: Natural and Technical Sciences*, 2017, vol. 22, no. 5, pp. 1138-1141. DOI: 10.20310/1810-0198-2017-22-5-1138-1141 (In Russian, Abstr. in Engl.).

UDC 004.853
DOI: 10.20310/1810-0198-2017-22-5-1138-1141

PHONETIC SEARCH ALGORITHM OF INAPPROPRIATE CONTENT

© D.V. Lopatin, E.S. Chirkin, A.A. Fadeeva
Tambov State University named after G.R. Derzhavin
33 Internatsionalnaya St., Tambov, Russian Federation, 392000
E-mail: ccs.tmb@ya.ru

The upgrade phonetic algorithm for searching for inappropriate content in texts written in Russian is described. It is shown that the proposed variant of the phonetic algorithm can be used for search in fuzzy text if the size of the training sample is insignificant or it is absent.
Keywords: phonetic algorithm; search; content

REFERENCES

1. Binstock A., Rex J. *Practical Algorithms for Programmers*. Boston, Addison-Wesley, 1995, 577 p.
2. Moyal A., Aharonson V., Tetariy E., Gishri M. Phonetic Search Methods for Large Speech Databases. Heidelberg, Springer-Verlag, 2013, 53 p.
3. *Soundex Coding. NARA and Daitch-Mokotoff Soundex*. Available at: http://www.jewishgen.org/infofiles/soundex.html (accessed 15.07.2017).
4. Philips L. Hanging on the Metaphone. *Computer Language*, 1990, vol. 7, no. 12, pp. 38-45.
5. Hood D. *Caversham Project Occasional Technical Paper. 2004*. Available at: http://caversham.otago.ac.nz/files/working/ctp150804.pdf (accessed 15.07.2017).
6. Kankovski P. «Kak vasha familiya», ili Russkiy MetaPhone [“What is your surname?”, or Russian MetaPhone]. *Programmist*, 2002, no. 8, pp. 36-39. (In Russian).
7. Tomlinson S. Lexical and Algorithmic Stemming Compared for 9 European Languages with Hummingbird SearchServerTM at CLEF 2003. In: Peters C., Gonzalo J., Braschler M., Kluck M. (eds.). *Comparative Evaluation of Multilingual Information Access Systems. CLEF 2003. Lecture Notes in Computer Science*. Berlin, Springer, 2004, vol. 3237, pp. 286-300.
8. Gubin M.V., Morozov A.B. Vliyanie morfologicheskogo analiza na kachestvo informatsionnogo poiska [The influence of morphology analysis on the quality of information search]. *Trudy 8 Vserossiyskoy nauchnoy konferentsii «Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii»* [Proceedings of 8th All-Russian Scientific Conference “Electronic Libraries: Prospective Methods and Technologies, Electronic Collections”]. Suzdal, 2006, pp. 95-100. (In Russian).
9. *Russian stemming algorithm*. Available at: http://snowball.tartarus.org/algorithms/russian/stemmer.html (accessed 15.07.2017).
10. *Veroyatnostnyy morfologicheskiy analizator russkogo i ukrainskogo yazykov* [Serendipitous morphological analyzer of Russian and Ukrainian languages]. (In Russian). Available at: http://www.keva.ru/stemka/stemka.html (accessed 15.07.2017).
11. Segalovich I. *A fast morphological algorithm with unknown*. Available at: http://download.yandex.ru/company/iseg-las-vegas.pdf (accessed 15.07.2017).
12. Chirkin E.S., Lopatin D.V. Podkhody k nechetkomu poisku nezhelatel'nogo kontenta na veb-stranitse [Approaches to fuzzy search of inappropriate content on the webpage]. *Vestnik Tambovskogo universiteta. Seriya Estestvennye i tekhnicheskie nauki – Tambov University Reports. Series: Natural and Technical Sciences*, 2016, vol. 21, no. 6, pp. 2358-2365. (In Russian). DOI: 10.20310/1810-0198-2016-21-6-2358-2365.

ACKNOWLEDGEMENTS: The work is fulfilled under financial support of Russian Foundation for Basic Research (project no. 15 17 08378).

Received 30 August 2017

Lopatin Dmitrii Valerevich, Tambov State University named after G.R. Derzhavin, Tambov, Russian Federation, Candidate of Physics and Mathematics, Associate Professor, Director of Computer Security Center, e-mail: 79107540080@ya.ru

Chirkin Evgeniy Sergeevich, Tambov State University named after G.R. Derzhavin, Tambov, Russian Federation, Engineer of Computer Security Center, e-mail: ccs.tmb@ya.ru

Fadeeva Angelina Aleksandrovna, Tambov State University named after G.R. Derzhavin, Tambov, Russian Federation, Candidate of Philology, Associate Professor of Foreign Philology and Applied Linguistics Department, e-mail: ccs.tmb@ya.ru