МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

РЕАЛИЗАЦИЯ И АНАЛИЗ АЛГОРИТМОВ ФОНЕТИЧЕСКОГО КОДИРОВАНИЯ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 441 группы
направления 02.03.03 «Математическое обеспечение и администрирование информационных систем»
факультета компьютерных и информационных технологий
Маниной Дарьи Романовны

Научный руководитель
зав. кафедрой к.ф.-м.н, доцент _____ Огнева М.В.

Зав. кафедрой
к.ф.м.н., доцент _____ Огнева М.В.

Саратов 2020

ВВЕДЕНИЕ

Существуют предметные области, в которых задача хранения безошибочных документов в электронной форме имеет особую значимость. Порой ошибка или опечатка в документе может признать его недействительным, что повлечет за собой неприятные ситуации или даже порождение конфликтов.

С другой стороны, отсутствие опечаток влияет на отклик поисковых систем в интернете. Для того, чтобы найти какую-либо информацию используются поисковые системы. Такие системы позволяют из множества текстов отбирать релевантные, соответствующие определённому запросу. Запрос представляет собой одно или несколько ключевых слов, которые, как предполагается, содержатся в искомом документе. На этом этапе начинают возникать сложности, так как в запросе могут содержаться разного рода ошибки, которые пользователь допустил по случайности, либо вследствие своей неграмотности. В результате, поисковая система выдаст огромное количество ссылок, большинство из которых не отвечают, запросу и являются информационным мусором.

Для предотвращения вышеперечисленных ситуаций используются алгоритмы фонетического кодирования. Задачей таких алгоритмов является нахождение элементов, которые в наибольшей степени близки по написанию к запросу. Такой вид поиска учитывает возможные ошибки и опечатки пользователей, допущенные ими при вводе запросов.

Алгоритмы фонетического кодирования предназначены для определения схожести слов по звучанию. Они устраняют для пользователя необходимость знать правильное написание каждого термина, с которым он работает. Алгоритмы фонетического кодирования разделены на алгоритмы для сравнения слов, фонетические алгоритмы, и алгоритмы определения расстояния между словами — фонетические расстояния.

Фонетические алгоритмы представляют собой группировку слов со схожим произношением с помощью закодированной строки, в которую они

преобразуются на основе последовательности букв слова и правил произношения. По закодированным строкам двух различных слов можно делать выводы о близости этих слов по звучанию, то есть смотреть насколько совпадают или близки получившиеся закодированные последовательности. Большинство фонетических алгоритмов предназначены для английского языка, но некоторые из них адаптированы и для других языков, в том числе и для русского. Для этого нужно, чтобы алгоритм учитывал правила фонетического кодирования языка и фонетические особенности языка.

Фонетическое расстояние применяется в алгоритмах нечеткого поиска. Оно определяет близость строк по написанию с помощью метрики — функции расстояния, которая сопоставляет двум строкам некоторое число, по которому можно судить об их различии. Происходит определение сходства слов по произношению путем подсчета расстояния между словами по написанию. Такие алгоритмы не берут во внимание языки запроса и множества элементов, по которым ведётся поиск — для них важна только операция сравнения символов.

Алгоритмы фонетического кодирования являются основой для построения современных систем проверки орфографии, которые используются в текстовых редакторах, системах оптического распознавания символов и поисковых системах, вроде Google или Yandex. Например, такие алгоритмы используются для функций, которые выдают пользователю сообщение «возможно вы имели в виду…» в поисковых системах.

В работе рассматриваются аспекты отдельного и совместного использования фонетических алгоритмов и алгоритмов нечеткого поиска.

Целью данной работы является программная реализация фонетических расстояний, фонетических алгоритмов и их совместного использования, а также анализ эффективности реализованных вариантов.

Данная цель определила **следующие задачи**:

1. Рассмотреть расстояния Левенштейна и Дамерау–Левенштейна.
2. Оптимизировать расстояние Левенштейна.

3. Реализовать расчет и сравнить время выполнения метрик Левенштейна и Дамерау–Левенштейна.
4. Рассмотреть и реализовать метод N-грамм.
5. Рассмотреть коэффициенты схожести по Левенштейну и Серенсена–Дайса.
6. Адаптировать и реализовать алгоритм Metaphone для русского языка.
7. Реализовать алгоритм Polyphone.
8. Разработать вариант совместного использования алгоритма Metaphone для русского языка, расстояния Дамерау–Левенштейна и меры схожести по Левенштейну.
9. Разработать вариант совместного использования алгоритма Polyphone, метода N-грамм и коэффициента Серенсена–Дайса.
10. Проанализировать эффективность использования отдельного и совместного применения алгоритмов, при помощи сравнения по времени и количеству исправленных ошибок.
11. Сделать выводы.

Методологические основы фонетического кодирования представлены в работах В.И. Левенштейна, Фредерика Дамерау, Петра Каньковски, В.С. Выхованца.

В теоретической части изложена общая информация о фонетических расстояниях и фонетических алгоритмах. Подробно описан их принцип работы и выделены их особенности, недостатки и достоинства.

В практической части работы рассматриваются реализация и аспекты отдельного и совместного использования фонетических алгоритмов и фонетических расстояний.

Структура и объём работы. Бакалаврская работа состоит из введения, шести разделов, заключения, списка использованных источников и семи приложений. Общий объем работы – 75 страниц, из них 33 страницы – основное содержание, включая 5 рисунков и 22 таблицы, список использованных источников информации – 23 наименования.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Фонетическое расстояние» посвящен обзору наиболее известных фонетических расстояний: расстояние Левенштейна, расстояние Дамерау−Левенштейна, расстояние на основе N-грамм и коэффициент Серенсена−Дайса.

В алгоритмах нечеткого поиска используется метрика — функция расстояния между двумя словами, позволяющая оценить степень их сходства в данном контексте. Под нечетким поиском понимается возможность найти достаточно близкое приближение к запрошенному термину или фразе.

Предполагается, что речевой и письменный строй тесно связаны друг с другом. Опираясь на этот факт, метод нечеткого поиска, используемый в алгоритмах фонетического кодирования, заменяет слова с помощью вычисления расстояния между ними посредством попарного сравнения символов. Части слов, которые близки по звучанию, также близки и в письменном виде. Благодаря метрикам, можно поставить задачу определения сходства слов по произношению путем подсчета расстояния между словами по написанию.

В подразделе «Редакционное предписание» дается определение понятия и пример для двух строк.

В подразделе «Расстояние Левенштейна» дается определение понятия и где оно применяется. Также выводится формула для подсчета расстояния Левенштейна, приводится пример расчета матрицы и нахождения расстояния Левенштейна для двух конкретных строк.

В подразделе «Расстояние Дамерау−Левенштейна» дается определение понятия, выводится формула для подсчета расстояния Дамерау−Левенштейна, приводится пример расчета матрицы и нахождения расстояния Дамерау−Левенштейна для двух конкретных строк.

В подразделе «Цена операций» дается определение понятия и объясняется, в каких случаях какую цену стоит указывать.

В подразделе «Метод N-грамм» описан принцип, на котором основывается данный метод, приводится пример разбиения слов и символов на N-граммы.

Второй раздел «Коэффициенты сходства» посвящен обзору меры схожести по Левенштейну и коэффициенту сходства Серенсена–Дайса.

Коэффициент сходства — показатель сходства сравниваемых строк. Область его применения обширна: в биологии, для количественного определения степени сходства биологических объектов, в географии, социологии, распознавании образов, поисковых системах, сравнительной лингвистике, биоинформатике, при сравнении строк.

Большинство коэффициентов нормированы и находятся в диапазоне от 0 (сходство отсутствует) до 1 (полное сходство).

В подразделе «Мера схожести по Левенштейну» выводится формула для расчета меры для двух строк.

В подразделе «Коэффициент сходства Серенсена–Дайса» дается определение понятия и выводится формула для расчета коэффициента для двух строк.

Третий раздел «Фонетические алгоритмы» посвящен фонетическим алгоритмам, адаптированным для русского языка: Metaphone и Polyphone.

Фонетические алгоритмы изначально предназначались для английского языка. Они применимы для русского языка только после транслитерации символов. Орфографические ошибки, допущенные в русских словах, обычно отличаются от орфографических ошибок в текстах на английском языке. Это все происходит из-за различий в правилах произношения и письма на разных языках. В транслитерации невозможно учесть фонетические особенности буквенных последовательностей для каждого языка. Таким образом, наиболее известные фонетические алгоритмы не так эффективны для текстов на русском языке.

В подразделе «Алгоритм Metaphone» описывается адаптация английской версии алгоритма к русскому языку и рассматриваются примеры

преобразования исходного слова в закодированные строки в соответствии с правилами и нормами языка при помощи данного алгоритма. Если закодированные строки двух слов совпадают, то эти слова считаются фонетически схожими.

В подразделе «Алгоритм Polyphone» рассматриваются примеры преобразования исходного слова в закодированные строки в соответствии с правилами и нормами языка при помощи данного алгоритма и дальнейшее преобразование в фонетический код с помощью простых чисел. Каждой букве соответствует простой числовой код. Полученный код представляет собой сумму простых чисел.

В алгоритме Polyphone сравнение идет не только по закодированной строке, но и по коду. Если полученные закодированные строки и коды двух слов идентичны, значит, слова фонетически похожи.

Четвертый раздел «Анализ алгоритмов» посвящен реализации рассмотренных алгоритмов и проведению анализа эффективности использования при помощи сравнения по времени и количеству исправленных ошибок.

Все рассмотренные на практике алгоритмы были реализованы на языке C#.

Для всех тестов был взят словарь с русскими словами размером 100000 слов. Была проведена подготовка данных для тестирования: из словаря выбирались рандомные отрезки слов различной длины, являющиеся запросами, и слова в них заменялись на ошибочные. Для первых трех тестов, в которых количество слов с ошибками равнялось 10, 50 и 100, слова были намеренно изменены на ошибочные вручную путем применения операций удаления, вставки, замены или транспозиции символов. Это было сделано для того, чтобы приблизить эксперимент к реальным условиям, так как автоматическая замена букв в словах, может не так эффективно показать результаты экспериментов.

В разделе приводятся результаты тестирования алгоритмов по времени и количеству исправленных ошибок и выводы, сделанные после проведения тестирования для:

1. расстояния Левенштейна;
2. расстояния Дамерау–Левенштейна;
3. метода N-грамм;
4. алгоритма Metaphone для русского языка;
5. алгоритма Polyphone.

Пятый раздел «Аспекты совместного использования алгоритмов» посвящен объединению фонетических расстояний и фонетических алгоритмов. Совмещения выбирались неслучайно и заключались в том, что это позволяет взять достоинства каждого и уменьшает зависимость от недостатков.

Для первого объединения были выбраны: расстояние Дамерау–Левенштейна, фонетический алгоритм Metaphone и метрика схожести по Левенштейну.

Для следующего объединения выбраны фонетический алгоритм Polyphone, метод N-грамм и коэффициент Серенсена–Дайса.

Шестой раздел «Выводы» обобщает полученные результаты тестирования в таблице, в которой для каждого алгоритма приведены точность исправления ошибок и среднее время выполнения.

Благодаря таблице видно, что совместное использование фонетических алгоритмов и фонетических расстояний позволяет достичь более релевантных и точных результатов. Их совместная интеграция дает возможность проводить более глубокий и затрагиваемый различные аспекты анализ информации, что позволяет повысить качество научных исследований.

ЗАКЛЮЧЕНИЕ

В данной работе были рассмотрены метрики Левенштейна и Дамерау–Левенштейна, N-граммная метрика, коэффициент схожести по Левенштейну, коэффициент Серенсена–Дайса, фонетические алгоритмы Metaphone и Polyphone, совместное использование алгоритмов Metaphone с расстоянием Дамерау–Левенштейн и метрикой схожести по Левенштейну, алгоритмов Polyphone с методом N-грамм и коэффициентом Серенсена–Дайса. Для каждого из вышеперечисленных вариантов отдельного и совместного использования алгоритмов была выполнена программная реализация и произведен анализ эффективности на основе сравнения по времени и по количеству исправленных ошибок в словах.

По результатам полученных экспериментов можно сделать вывод, что совместное использования фонетических алгоритмов и метрик, обеспечивает более эффективный поиск за сравнительно короткое время, в отличии от использования алгоритмов по отдельности.

По тематике бакалаврской работы были представлены доклады:

1. «Алгоритмы нечеткого поиска в обучении» на X Всероссийской научно-практической конференции «Информационные технологии в образовании» «ИТО-Саратов-2018», Саратов, 1-2 ноября 2018 года.

2. «Исправление опечаток и ошибок в словах с помощью нечеткого поиска» на Студенческой международной научно-практической конференции «Научное сообщество студентов XXI столетия. Технические науки», Новосибирск, 2019 год. По результатам конференции выдан диплом лауреата за лучшую научную работу по результатам интернет-голосования.

Доклады опубликованы в материалах конференций.

Основные источники информации:

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

2. Левенштейн В. И. Двоичные коды с исправлением выпадений, вставок и замещений символов//Доклады Академий Наук СССР, 1965. С. 845-848.
3. Damerau F.J. A Technique for Computer Detection and Correction of Spelling Errors//Communications of the ACM. С. 171-176.
4. КАНЬКОВСКИ П. «Как ваша фамилия?» или русский MetaPhone //Программист, 2002. С. 36–39.
5. PARAMONOV V.V., SHIGAROV A.O., RUZHNIKOV G.M. et al. Polyphon: An Algorithm for Phonetic String Matching in Russian Language // Int. Conference on Information and Software Technologies, Druskininkai, Lithuania, October 13–15, 2016. – Springer International Publishing, 2016. С. 568–579.