Министерство образования и науки Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

# РЕАЛИЗАЦИЯ И СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ НЕЧЕТКОГО ПОИСКА

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Елисеевой Елизаветы Дмитриевной

Научный руководитель:
Зав. кафедрой информатики
и программирования, к.ф.-м.н. _____ М.В. Огнева
(подпись, дата)

Зав. кафедрой:
к.ф.-м.н. _____ М.В. Огнева
(подпись, дата)

Саратов 2018

# ВВЕДЕНИЕ

**Актуальность темы.** Задача анализа алгоритмов нечёткого поиска на сегодняшний момент актуальна, так как область применения данных алгоритмов невероятно велика и разнообразна. Они применяются для форм заполнения информации на сайтах и в полноценных поисковых системах. Например, такие алгоритмы используются для функций наподобие «Возможно вы имели в виду …» и для обнаружения опечаток в полях ввода программ. Так же алгоритмы нечеткого поиска используются для распознавания рукописных символов, которые с массовым распространением устройств с сенсорным экраном активно используется для обеспечения удобства ввода. Введённый символ преобразуется в комбинацию цифр в зависимости от последовательности произведённых жестов, и полученная комбинация сравнивается со значениями, заранее известными для всех символов используемого алфавита, записанными в таблицу. Символ, для которого совпадение будет самым полным, и считается распознанным. Для определения полноты совпадения используются алгоритмы нечёткого поиска. Также данные алгоритмы активно используются для сравнения генов, белков, хромосом в биоинформатике; в полнотекстовом поиске; для фильтрации спама; для исправления ошибок; для распознавания лиц, речи, радужки глаз; для дедупликации адреса; для проверки на плагиат текста и музыки; для распознавания объектов в реальном времени и во многих других областях. Именно поэтому так важно знать особенности основных применяемых на сегодняшний момент алгоритмов, чтобы для конкретной ситуации была возможность выбрать максимально эффективный из них. **Цель бакалаврской работы**: изучение основных метрик нечеткого поиска, и их применение для поиска в тексте, исправления орфографических ошибок и выравнивания генетических последовательностей.

**Задачи:**

1. Изучить расстояния Левенштейна и Дамерау-Левенштейна.
2. Изучить алгоритм Вагнера-Фишера, привести теоретическую оценку сложности.
3. Реализовать расчет метрик Левенштейна и Дамерау-Левенштейна
4. Изучить алгоритмы сопоставления строк Нидлмана-Вунша и Смита-Ватермана, привести теоретическую оценку сложности.
5. Реализовать алгоритмы Нидлмана-Вунша и Смита-Ватермана.
6. Сравнить время выполнения алгоритмов Нидлмана-Вунша и Смита-Ватермана.
7. Показать применение алгоритмов нечеткого поиска для сопоставления генетических последовательностей.

8. Изучить алгоритм Байетса-Йетс-Гоннета с модификациями Ву-Манбера для нечеткого поиска и привести теоретическую оценку сложности.
9. Реализовать алгоритм Байетса-Йетс-Гоннета с модификациями Ву-Манбера
10. Сравнить время выполнения поиска всех вхождений подстрок, допускающих k ошибок, выполненного алгоритмами Байетса-Йетс-Гоннета с модификациями Ву-Манбера и наивным поиском.
11. Изучить алгоритмы нечеткого поиска в словаре SymSpell и n-грамм.
12. Реализовать нечеткий поиск в словаре методом n-грамм.
13. Сравнить время выполнения алгоритмов SymSpell и n-грамм.

**Практическая значимость бакалаврской работы.** В ходе выполнения практической части бакалаврской работы были реализованы алгоритмы и метрики нечеткого поиска и проведен сравнительный анализ алгоритмов, после чего были сделаны выводы о преимуществах и недостатках рассматриваемых алгоритмов.

**Структура и объём работы**. Бакалаврская работа состоит из введения, пяти разделов, заключения, списка использованных источников и пяти приложений. Общий объем работы – 64 страницы, из них 38 страниц – основное содержание, включая 9 рисунков и 20 таблиц, список использованных источников информации – 29 наименований.

## КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

**Первый раздел «Редакционное расстояние»** посвящен обзору метрик нечеткого поиска: расстояние Левенштейна и расстояние Дамерау-Левенштейна, а также методов для их вычисления.

В подразделе «Расстояние Левенштейна» дается определение понятия редакционное расстояние, приводится историческая справка о его первом упоминании, выводится формула для подсчета расстояния Левенштейна.

В подразделе «Расстояние Дамерау-Левенштейна» дается определение понятия расстояние Дамерау-Левенштейна, выводится формула подсчета расстояния Дамерау-Левенштейна.

В подразделе «Алгоритм Вагнера-Фишера для вычисления расстояние Левенштейна» приводится псевдокод алгоритма Вагнера-Фишера и его асимптотическая сложность.

В подразделе «Алгоритм для вычисления расстояния Дамерау-Левенштейна» приводится псевдокод алгоритма для вычисления расстояния Дамерау-Левенштейна и его асимптотическая сложность.

**Второй раздел «Выравнивание строк»** посвящен проблеме выравнивания строк, дается определение выравниванию, рассматриваются

два вида выравнивания: полное и локальное, а также рассматриваются методы выравнивания строк.

В подразделе «Алгоритм Нидлмана-Вунша» рассматривается алгоритм Нидлмана-Вунша для выравнивания последовательностей по всей длине. Приводятся основные понятия, необходимые для понимания работы алгоритма, описание алгоритма, примеры работы, псевдокод алгоритма для построения матрицы выравнивания и для построения выравнивания, а также приводится асимптотическая сложность.

В подразделе «Алгоритм Смита-Вотермана» рассматривается алгоритм Смита-Вотермана для выравнивания локальных последовательностей. Приводятся описание алгоритма, псевдокод алгоритма для построения матрицы выравнивания и для построения выравнивания, а также асимптотическая сложность.

В подразделе «Сравнение алгоритмов» приводятся основные отличия алгоритма Нидлмана-Вунша от алгоритма Смита-Вотермана.

**Третий раздел «Нечеткий поиск в тексте»** посвящен проблеме нечеткого поиска в тексте, приводятся его актуальность и методы нечеткого поиска в тексте, использующие расстояние Левенштейна как метрику для нечеткого поиска.

В подразделе «Алгоритм Байетса-Йетс-Гоннета с модификациями Ву-Манбера» приводятся описание алгортима Байетса-Йетс-Гоннета для поиска точного совпадения в тексте и алгоритма Байетса-Йетс-Гоннета с модификациями Ву-Манбера для нечеткого поиска в тексте, примеры их работы. Приводится псевдокод алгоритма Байетса-Йетс-Гоннета с модификациями Ву-Манбера и его асимптотическая сложность.

В подразделе «Наивный алгоритм» рассматривается наивный алгоритм для нечеткого поиска в тексте, приводится его псевдокод и асимптотическая сложность.

**В четвертом разделе «Нечеткий поиск в словаре»** рассматривается проблема исправления орфографических ошибок, приводятся различные подходы к решению этой проблемы, использующие расстояние Дамерау-Левенштейна для нечеткого сравнения.

В подразделе «Алгоритм SymSpell» приводится обзор библиотеки, которая предназначена для решения проблемы исправления орфографических ошибок. Основной идей этого алгоритма является генерация слов с расстоянием редактирования (только удаление) из каждого словарного слова и добавление их вместе с исходным термином в словарь.

В подразделе «Метод n-грамм» приводится определение понятия n-грамм, методы получения n-грамм, и способ применения n-грамм для нечеткого поиска в словаре.

**В пятом разделе «Сравнительный анализ алгоритмов»** приводятся результаты тестирования алгоритмов и выводы, сделанные после проведения тестирования.

Описываются:

- результаты тестирования подсчета расстояний Левенштейна и Дамерау-Левенштейна на случайно сгенерированных строках разной длины
- результаты тестирования построения матриц выравнивания и выравниваний алгоритмами Нидлмана-Вунша и Смита-Вотермана на случайно сгенерированных строках разной длины
- результаты тестирования алгоритмов Нидлмана-Вунша и Смита-Вотермана на реальных генетических последовательностях, проведение
- результаты тестирования нечеткого поиска строк разной длины, с разным количеством допустимых ошибок в художественном тексте алгоритмами Алгоритм Байетса-Йетс-Гоннета с модификациями Ву-Манбера и наивным поиском
- результаты тестирования нечеткого поиска строки с разным количеством допустимых ошибок в тексте разной длины алгоритмами Алгоритм Байетса-Йетс-Гоннета с модификациями Ву-Манбера и наивным поиском
- результаты тестирования построения индексов для словаря алгоритмами SymSpell и n-грамм
- результаты тестирования нечеткого поиска слов разной длины с разным количеством допустимых ошибок в словаре алгоритмами SymSpell и n-грамм

В подразделе «Выводы» приводятся выводы, основанные на результатах тестирования.

## ЗАКЛЮЧЕНИЕ

В данной работе были изучены расстояния Левенштейна и Дамерау-Левенштейна, был реализован поиск расстояния Левенштейна и Дамерау-Левенштейна, оценена сложность поиска. Также были изучены и реализованы алгоритмы Нидлмана-Вунша и Смита-Вотермана для выравнивания последовательностей, оценена сложность этих алгоритмов. Был изучен и реализован алгоритм Байетса-Йетс-Гоннета с модификацией Ву-Манбера для нечеткого поиска в тексте, оценена его сложность, описан и реализован наивный алгоритм нечеткого поиска. Был изучен алгоритм SymSpell для нечеткого поиска в словаре и изучен и реализован метод n-грамм. Все алгоритмы реализованы на языке Java.

Были проведены различные тесты всех алгоритмов. Алгоритмы поиска расстояний Левенштейна и Дамерау-Левенштейна тестировались на случайно сгенерированных строках разной длины. Алгоритмы Нидлмана-Вунша и Смита-Вотермана тестировались на сгенерированных строках разной длины, выполнялся замер времени построения матрицы выравнивания и построения выравнивания, а также было показано применение этих алгоритмов для выравнивания генетических последовательностей. Алгоритм Байетса-Йетс-Гоннета с модификацией Ву-Манбера и алгоритм наивного нечеткого поиска подстроки в тексте тестировались на романе Д. Стейнбека «Гозди гнева». В романе искались строки разной длины с разным количеством допустимых ошибок, так же проводился тест с заданной искомой строкой, но с меняющимися длиной текста и количеством допустимых ошибок. Было проведено тестирование алгоритмов нечеткого поиска в словаре. Выполнялся замер времени построения индексов к словарю алгоритмом SymSpell и методом n-грамм, а также выполнялся замер времени поиска слов разной длины в словаре с разным допустимым количеством ошибок.

По результатам тестирования были сделаны выводы, что алгоритмы поиска расстояния Левенштейна и Дамерау-Левенштейна работают быстро, особенно на маленьких последовательностях. Алгоритмы Нидлмана-Вунша и Смита-Вотермана эффективно справляются каждый со своей задачей, а именно полным и локальным выравниванием строк. Алгоритм наивного нечеткого поиска в тексте работает очень медленно, алгоритм Байетса-Йетс-Гоннета с модификацией Ву-Манбера работает быстро, но у него есть ограничения по длине искомой строки. Метод n-грамм работает быстрее алгоритма SymSpell, но SymSpell выполняет сортировку выдаваемых результатов, что важно при исправлении орфографических ошибок.

Алгоритмы нечеткого поиска эффективны и используются во многих областях, где требуется нечеткое сравнение. Расстояния Левенштейна и Дамерау-Левенштейна являются одними из наиболее популярных метрик нечеткого сопоставления. Алгоритмы выравнивания строк используются в биоинформатике для сопоставления последовательностей генов, белков и хромосом. Это помогает ученым легко находить схожие участки последовательностей и выявлять функциональные, структурные или эволюционные взаимосвязи. Алгоритмы нечеткого поиска в тексте находят все вхождения искомой строки в текст с заданным количеством ошибок. Алгоритмы нечеткого поиска в словаре используются в орфографических корректорах, программах распознавания печатного текста, в поисковых системах и других областях.

## СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ


1. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов, Докл. АН СССР, 1965, том 163, номер 4
2. Гасфилд Д. Строки, деревья и последовательности в алгоритмах: Информатика и вычислительная биология / Пер. с англ. И.В.Романовского – Спб.: Невский Диалект, БХВ-Петербург, 2003, – 654 с: ил.
3. Damerau F.J. A Technique for Computer Detection and Correction of Spelling Errors, Communications of the ACM, ACM, March 1964.
4. Boytsov L. Indexing Methods for Approximate Dictionary Searching, Journal of Experimental Algorithmics, Association for Computing Machinery (ACM), May 2011.
5. Lowrance R., Wagner R.A. An Extension of the String-to-String Correction Problem, J ACM, April 1975.
6. Wagner R.A., Fischer M.J. The String-to-string Correction Problem // Journal of ACM. 1974. Vol. 21. No. 1.
7. Кормен Т.Х. Алгоритмы: построение и анализ — 3-е изд. — М.: «Вильямс», 2013. — с. 440. — ISBN 978-5-8459-1794-2
8. Navarro G. A Guided Tour to Approximate String Matching, Dept. of Computer Science, University of Chile Blanco Encalada 2120 – Santiago – Chile
9. Sellers P.H. (1980) Theory and Computation of Evolutionary. Distances: Pattern Recognition» Journal of Algorithms 1, 1980.
10. Needleman S.B., Wunsch C. D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins, Journal of Molecular Biology, 1970.

11. Wing-Kin, Sung. Algorithms in Bioinformatics: a Practical Introduction. Boca Raton: Chapman & Hall/CRC Press, 2010, ISBN 9781420070330.
12. Smith T.F., Waterman M.S. Identification of Common Molecular Subsequences, Journal of Molecular Biology, 1981, 147.
13. Baeza-Yates R.A., Gonnet G.H. A New Approach to Text Searching, Communications of the ACM, October 1992, 35(10).
14. Fast Text Searching With Errors, Department of Computer Science, University of Arizona, Tucson, June 1991, AZ 85721.
15. Manber U., Wu S. Fast Text Search Allowing Errors, Communications of the ACM, October 1992, 35(10).
16. Андерсон Д.А. Дискретная математика и комбинаторика. : Пер. с англ. – М. : Издательский дом «Вильямс», 2004 – 960 с. : ил. Парал. тит. англ. ISBN 5-8459-0498-6 (рус.)
17. 1000x Faster Spelling Correction algorithm (2012) [Электронный ресурс] URL: https://medium.com/@wolfgarbe/1000x-faster-spelling-correction-algorithm-2012-8701fcd87a5f (дата обращения: 15.05.2018)
18. Fast approximate string matching with large edit distances in Big Data (2015) [Электронный ресурс] URL: https://medium.com/@wolfgarbe/fast-approximate-string-matching-with-large-edit-distances-in-big-data-2015-9174a0968c0b (дата обращения: 15.05.2018)
19. SymSpell vs. BK-tree: 100x faster fuzzy string search & spell checking [Электронный ресурс] URL: https://towardsdatascience.com/symspell-vs-bk-tree-100x-faster-fuzzy-string-search-spell-checking-c4f10d80a078 (дата обращения: 15.05.2018)
20. SymSpell [Электронный ресурс] URL: https://github.com/wolfgarbe/SymSpell#compound-aware-multi-word-spelling-correction (дата обращения: 15.05.2018)
21. Manning C.D., Schütze H. Foundations of Statistical Natural Language Processing, MIT Press: 1999. ISBN 0-262-13360-1.
22. Welch T. A Technique for High-Performance Data Compression, Computer, 1984, 17 (6).
23. Lempel A., Ziv J. Compression of individual sequences via variable-rate coding // IEEE Transactions on Information Theory[en]. — 1978. — Т. 24, № 5. — С. 530–536.
24. Java SymSpell Realization [Электронный ресурс] URL: https://github.com/Lundez/JavaSymSpell (дата обращения: 17.05.2018)
25. Octopus insularis voucher STM_35 16S ribosomal RNA gene, partial sequence; mitochondrial [Электронный ресурс]____ URL:

https://www.ncbi.nlm.nih.gov/nuccore/MF040878.1 (дата обращения: 28.05.2018)
26. Octopus vulgaris mitochondrial COX1 gene for cytochrome c oxidase subunit 1, partial cds [Электронный ресурс] URL: https://www.ncbi.nlm.nih.gov/nuccore/LC043307.1 (дата обращения: 28.05.2018)
27. Homo sapiens ADP-ribosylation factor-like 2 binding protein, mRNA (cDNA clone MGC:104924 IMAGE:6728613), complete cds [Электронный ресурс] URL: https://www.ncbi.nlm.nih.gov/nuccore/BC094878.1 (дата обращения: 28.05.2018)
28. Elephant endotheliotropic herpesvirus 4 isolate North American NAP22 polymerase processivity factor (U27) gene, partial cds; protein ORF-J (E35A), envelope glycoprotein N (U46), envelope glycoprotein O (U47), and envelope glycoprotein H (U48) genes [Электронный ресурс] URL: https://www.ncbi.nlm.nih.gov/nuccore/KT832488.1 (дата обращения: 28.05.2018)
29. A text file containing 479k English words for all your dictionary/word-based projects e.g: auto-completion / autosuggestion [Электронный ресурс] URL: https://github.com/dwyl/english-words