Санкт-Петербургский политехнический университет Петра Великого
Институт компьютерных наук и технологий

Неделя науки 2017

Метрики в информатике. Нечеткий поиск в интернете

Выполнили:
Дивенкова Дарья(гр. 23536/3)
Мигунова Любовь (гр. 23536/3)

Научный руководитель:
Доцент каф. «Высшая математика»
Филимоненкова Надежда Викторовна

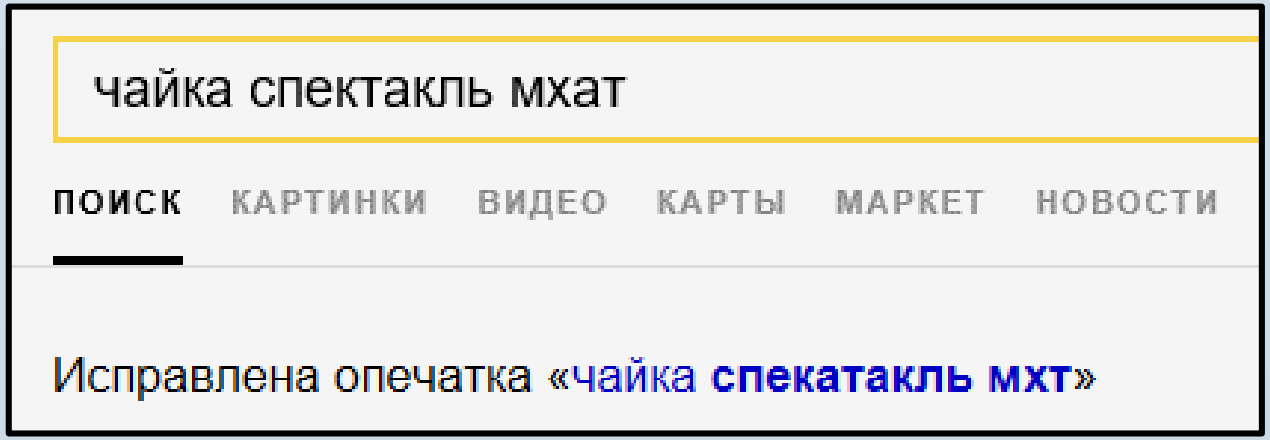
Задачи

- Изучить теорию метрик и их приложения
- Изучить работу алгоритмов нечеткого поиска
- Запрограммировать метрики Хэмминга, Левенштейна, а также триграмм для сравнения их эффективности

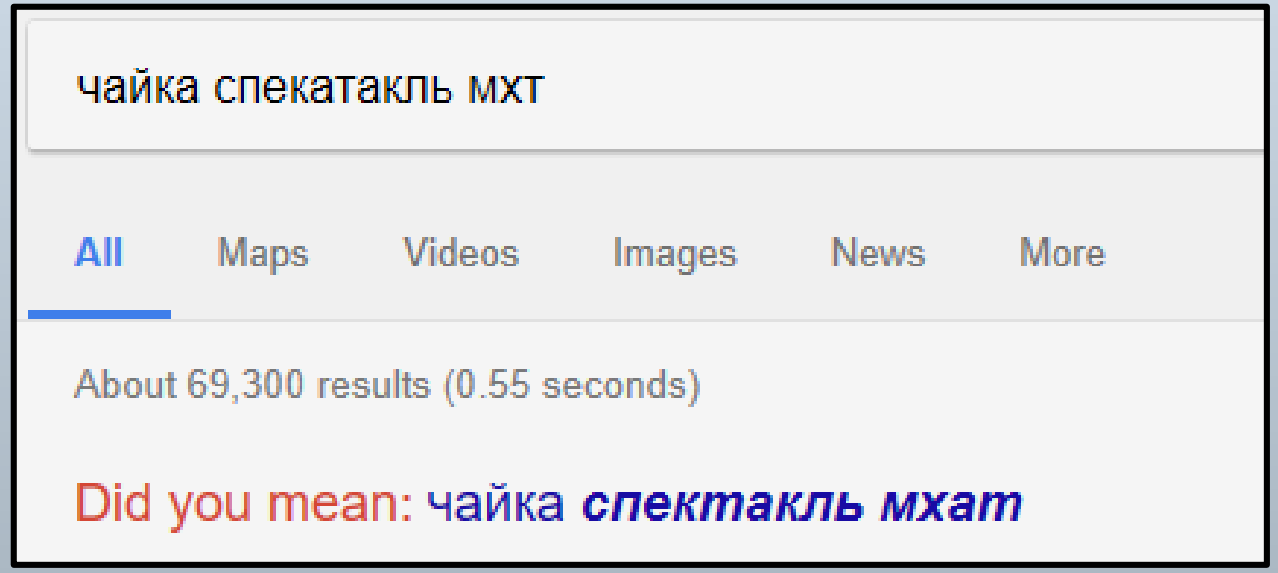
Мы рассмотрели алгоритмы нечеткого поиска как основу систем проверки орфографии и опечаток в поисковых системах. Так как их наличие негативно влияет на качество и скорость поиска, поисковые системы уделяют большое внимания распознаванию и исправлению ошибок.

Результатом алгоритмов нечеткого поиска является автоматическое исправление опечаток, а также сообщение «Возможно, вы имели в виду...».

Пример автозамены



Пример подсказки



Типы ошибок в поисковых запросах

- Ошибка в отдельном слове – ***буквaед***
- Ошибка слитно-раздельного написания – ***театр натаганке***
- Неверная раскладка клавиатуры – ***ghbdtn*** **(привет)**
- Транслитирация – ***тревелер*** **(traveler),** ***баттерфляй*** **(butterfly)**

ТИПЫ ОШИБОК В ПОИСКОВЫХ ЗАПРОСАХ

Зависимость исправлений от контекста

Анализируя ошибки в запросах можно заметить, что большинство из них являются очевидным опечатками и исправляются не зависимо от словарного окружения.
По статистике, контекстно-независимых исправлений оказалось 74 %, что подтверждает предположение об однозначности исправлений для большинства словарных ошибок.

Среди контекстно-зависимых исправлений большую часть составляют ошибки в коротких словах, допускающие разные, зависящие от контекста, исправления.

- Ошибки в коротких словах: *«ко**нс**туция» – «ко**нс**титуция», «сбо**р**ик» – «сбо**рн**ик».*

- Ошибки с учетом контекста: *«**сво** законов» – «**свод** законов»* и «**сво** машина» – «**своя** машина»

Метрика

Многие методы нечеткого поиска основаны на вычислении расстояния между словами(метрики).

Множество X называется ***метрическим пространством***, если для всех его элементов определена такая числовая функция двух аргументов, что для любых x,y,z$\in X$ выполняются три аксиомы:
I. $\rho(x,y) \geq 0$, причем $\rho(x,y) = 0 \Leftrightarrow x=y$;
II. $\rho(x,y) = \rho(y,x)$ (симметричность);
III. $\rho(x,y) \leq \rho(x,z) + \rho(z,y)$ (неравенство треугольника).

Функцию $\rho(x,y)$ называют ***метрикой*** или ***расстоянием*** между x и y

Описание подхода. Этапы нечеткого поиска

Введенное слово

Котрина

Список слов, отобранных фильтрацией
Картина
Катрина
Катерина
Витрина
…

- **фильтрация** - отбрасываются слова, которые заведомо далеки от введенного слова в заданной метрике: оставляются слова, удаленные от введенного слова на расстояние не больше некоторого порогового значения ε.

- **сравнение** – среди оставленных слов находится ближайшее к введенному слову в данной метрике, оно и предлагается как исправление введенного слова.

Основные метрики, используемые для нечеткого поиска

	Описание	Использование
Расстояние Хэмминга	число позиций, в которых соответствующие символы двух слов одинаковой длины различны.	слова одинаковой длины; Локальные орфографические ошибки
Расстояние Левенштейна	минимальное число вставок, замен и удалений символов необходимых для того, чтобы преобразовать первую строку во вторую.	слова разной длины; Локальные орфографические ошибки
Триграммы, n-граммы	числом совпадающих символьных n - грамм в обоих строках	орфографические ошибки, редактированный текст

Расстояние Хэмминга

Ричард Хэмминг (1915-1998) — американский математик, работы которого в сфере теории информации оказали существенное влияние на компьютерные науки и телекоммуникации.

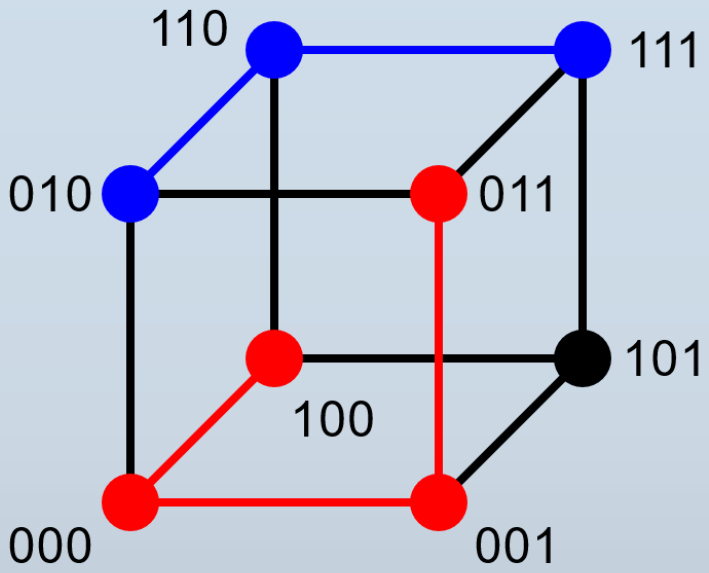
Расстояние Хэмминга - число позиций, в которых соответствующие символы двух слов одинаковой длины различны.

Расстояние Хэмминга – метрика в строгом математическом смысле

ρ(ма**ш**ина, ма**л**ина)=1

ρ(**1**01**1**, **0**01**0**)=2

Расстояние Хэмминга. Булев Куб



расстояние 3: 100 → 011
расстояние 2: 010 → 111

Для геометрического представления расстояния Хэмминга для двоичных векторов используется **булев куб** - регулярный двудольный граф, у которого вершинами являются наборы куба и ребрами соединяются соседние в смысле расстояния Хемминга вершины, и только они.

Расстояние Левенштейна

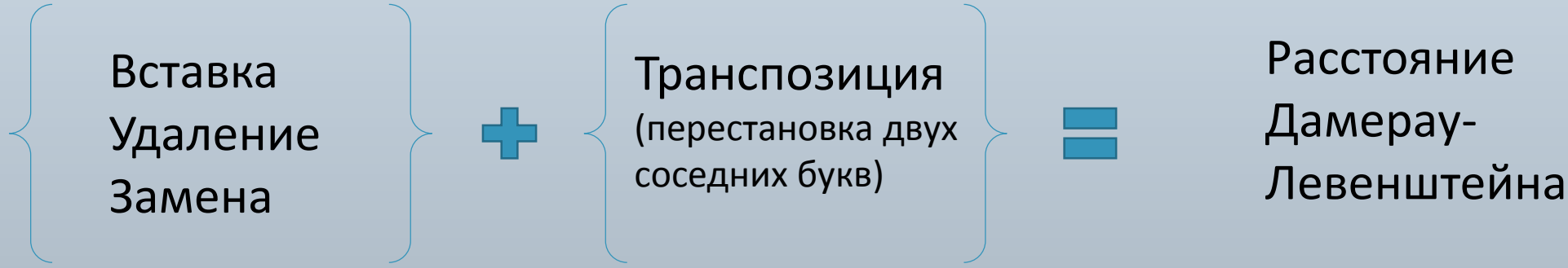
Левенштейн, Владимир Иосифович(1935-2017) — советский и российский математик, доктор физико-математических наук.

Расстояние Левенштейна позволяет сравнить строки различной длины с такими искажениями как вставки, замены и удаления символа. Расстояние Левенштейна равно числу минимальному числу операций, необходимых для преобразования одной строки в другую.

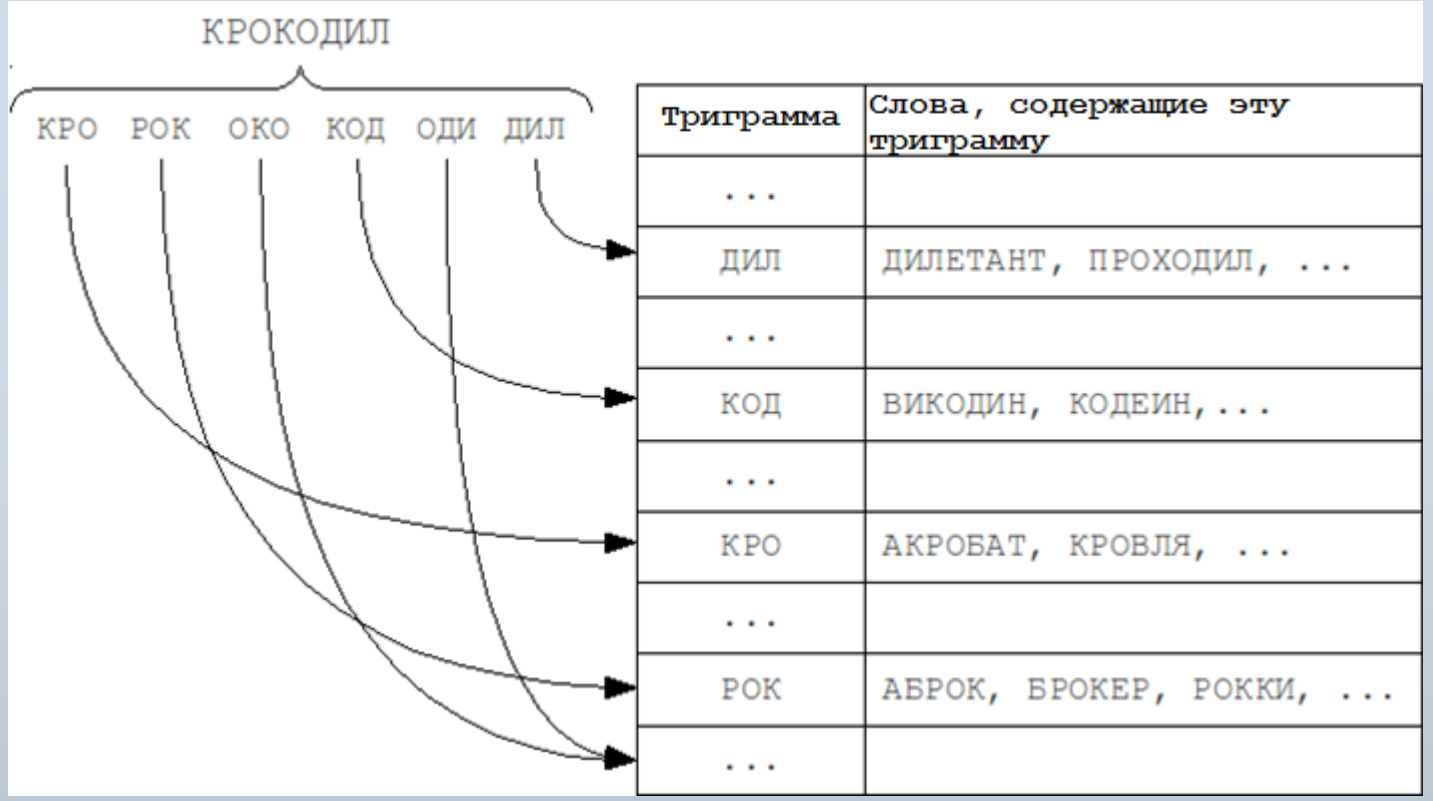
Расстояние Левенштейна – метрика в строгом математическом смысле.

Расстояние Левенштейна. Преимущества и модификации

- Относительная несложность в реализации
- Возможность использования для любых алфавитов
- Возможность модификации



Триграммы, n-граммы



Триграмма	Слова, содержащие эту триграмму
...	
ДИЛ	ДИЛЕТАНТ, ПРОХОДИЛ, ...
...	
КОД	ВИКОДИН, КОДЕИН,...
...	
КРО	АКРОБАТ, КРОВЛЯ, ...
...	
РОК	АБРОК, БРОКЕР, РОККИ, ...
...	

N-грамма — последовательность из n элементов. С семантической точки зрения, это может быть последовательность звуков, слогов, слов или букв. Является метрикой в широком смысле.

Триграмма - последовательность из 3 элементов.

Триграммы

Формула, для расчета метрики

$$\rho = \frac{(a+b)}{2*m}$$

m – число совпадающих триграмм,

a – общее число триграмм в первой строке,

b – общее число триграмм в второй строке

Пример расчета расстояния с помощью триграмм

Строка 1: **ALEXANDRE**
Строка 2: **ALEKSANDER**

ρ =(a+b)/2*m = (9+10)/2*3 = 3,16

(Совпадают **_AL**, **ALE**, **AND**)

Сравнение эффективности метрик по времени

Данные, представленные в таблице, были получены в результате реализации метрик Хэмминга, Левенштейна и Триграмм на языке программирования C# в среде разработки MS Visual Studio. Разделитель в словах стоит для наглядности работы метрики Хэмминга (программа обрезает наиболее длинную строку).

String 1	String 2	Hamming Distance	Time, ticks	Levenshtein Distance	Time, ticks	Trigrams Distance	Time, ticks
психоанализ	психзоонааа|лиз	7	432	4	1290	2,625	2352
щелкунчи|к	шекунчик	7	420	2	1354	1,625	2128
Навуходоносо|р	новходоносср	10	459	5	1220	2,3	1808
контрреволюционный	кнотьревоалеционый	9	430	6	1574	3,2	1872
бомбардир_бомбардировал_барышень_бранденбур|га	бонбанди_рбомбандирава_лбарышеньранденбурга	19	577	10	1864	1,6	2458
pharmaceutical	farmacetical	12	431	3	1302	1,57	2306

Сравнение эффективности метрик по времени. Анализ результатов

Исходя из данных таблицы можно сделать вывод, что метрика Левенштейна наиболее эффективна. Она хоть и дольше, чем метрика Хэмминга, но зато работает с разной длиной строк, что чаще всего применимо в реальности. В отдельных случаях, когда длина введенной с опечатками строки равна длине "правильной" строки, то эффективнее будет метрика Хэмминга.

Метрика Хэмминга – маломощный инструмент для нечеткого поиска в силу условия равенства строк.

Эталонная строка	Входные строки	Расстояние Левенштейна	Триграммы
Возможно, что если вычислить расстояние триграмм_	Возможно,тчо если вычислитть расстаояние триграм_	5	1,2
	возжнол,тчо еслии вычисмВлиттьраст аояние треграм	15	2,24
Возможно, когда мы вычислим расстояние до Марса_	Возможно,тчо если вычислитть расстаояние триграм_	19	2,1
	возжнол,тчо еслии вычисмВлиттьраст аояние треграм	27	4,18

Сравнения эффективности распознавания текста. Анализ результатов

Для демонстрации корректности работы метрик Левенштейна и триграмм мы рассматриваем две похожие эталонные строки: «Возможно, что если вычислить расстояние триграмм_ » и «Возможно, когда мы вычислим расстояние до Марса_». Затем вводим искажение первого эталона: сначала искажаем мало, затем сильнее. Искаженные строки мы сравниваем с обоими эталонами и видим, что даже сильно искаженную строку обе метрики ставят ближе к эталону №1, чем к эталону №2.
Таким образом, на этом конкретном примере видно, что обе метрики распознают текст правильно.

Алгоритмы нечеткого поиска за пределами интернета

- Текстовые редакторы
- Базы данных
- Обработка массивов данных
- Сопоставление документов
- Выявление плагиата
- Сравнение генов, хромосом, белков, аминокислот

Список литературы

- Филимоненкова Н. В. Конспект лекций по функциональному анализу: учеб. пособие, – СПб., 2015. – 10 с.
- Панина М. Ф., Байтин А. В., Галинская И. Е. «Автоматическое исправление опечаток в поисковых запросах без учета контекста» Яндекс, Москва, Россия
- Желудков А. В., Макаров Д. В., Фадеев П. В. «Особенности алгоритмов нечёткого поиска», - МГТУ им. Н.Э. Баумана, Россия, 2014.
- Реализация нечеткого поиска // Хабрахабр. Дата обновления: 02.11.2017. – URL: https://habrahabr.ru/post/123320/
- Нечёткий поиск в тексте и словаре // Хабрахабр. Дата обновления: 02.11.2017. – URL: https://m.habrahabr.ru/post/114997/