Р. М. Сергеенко

АНАЛИЗ МЕТОДОВ ПОИСКА ПО СХОДСТВУ В ТЕКСТЕ И СЛОВАРЕ

Рассматриваются методы и алгоритмы поиска по сходству в тексте и словаре: расстояние Левенштейна, расстояние Дамерау-Левенштейна, алгоритм расширения выборки и метод N-грамм. Производится анализ метрик и алгоритмов поиска по сходству в тексте и словаре.

Введение

Приближенное сопоставление строк является постоянной проблемой во многих областях информатики: приложениях для поиска текста, вычислительной биологии, паттернам распознавания, обработке сигналов и т. д. В данной публикации проанализированы метрики и алгоритмы нечеткого поиска.

I. Анализ метрик для поиска текста по сходству

Наиболее известными метриками являются – расстояние Левенштейна и Демерау-Левенштейна. Асимптотическая сложность расстояния Левенштейна и расстояния Дамерау-Левенштейна – $O(mn)$, где m и n – длины сравниваемых строк. Отличие метрик в том, что расстояния Дамерау-Левенштейна имеет транспозицию(перестановку двух соседник букв). В таблице 1 приведены результаты поиска слова в естественном английском тексте 1.2млн слов. Количество совпадений – количество найденных слов с учетом ошибок.

Таблица 1 – Результаты скорости поиска

	Значение расстояния	Кол-во совпадений	Время поиска
Точный поиск	-	70	134 ms
Расст. Левенштейна	2	135313	2078 ms
Расст. Дамерау-Левенштейна	2	135313	2223 ms

По полученным результатам можно сделать вывод о том, что использование расстояния Левенштейна при поиске увеличает время поиска, однако количество найденных совпадений покрывает данный недостаток.

II. Алгоритмы нечеткого поиска

Проанализируем алгоритмы нечеткого поиска с использованием индексации - алгоритм расширения выборки и метод N-грамм.

Алгоритм расширения выборки — алгоритм, основанный на сведении задачи о нечетком поиске к задаче о точном поиске[2]. Асимптотическая сложность алгоритма – $O((m|A|)^k * m * log(n))$.

Метод N-грамм - это алгоритм, использующий вероятностные методы для прогнозирования слова, содержащего подстроку длины N. Сравнение этих алгоритмов приведено в таблице 2. Эффективность – отношение числа вычислений расстояния редактирования по отношению к общему числу записей в индексе.

Таблица 2 – Сравнение алгоритмов

Название алгоритма	Эффективность	Скорость	Итог
Алгоритм расширения выборки	Высокая (на малых размерах индекса)	Высокая (При размерах словоря до 100тыс)	Не очень эффктивен, так как при 2 ошибках и более размер словаря огромен
Метод N-грамм	Высокая	Средняя (зависит от длины строк)	Эффективен (Можно разбивать хеш-таблицы)

III. Выводы

Алгоритм расширения выборки не очень эффективен на больших размерах индекса, однако при расстоянии радактирования равном 1, время поиска в памяти не изменяется при увеличении размера индекса и горазго эффективнее классического перебора. Метод n-грамм – один из наиболее эффективных алгоритмов при использовании хеш-таблиц, а в случае индекса, загруженного в оперативную память, время поиска отличается на 1-2 порядка.

Список литературы

1. Двоичные коды с исправлением выпадений, вставок и замещений символов. / В. И. Левенштейн // –1965
2. Интеллектуальные системы / И. А. Бессмертный, А. Б. Нугуманова, А. В. Платонов // Учебник и практикум для СПО. – 2018. – С. 178.

Сергеенко Роман Михайлович, магистрант кафедры интеллектуальных информационных технологий БГУИР, romansergeenko@gmail.com.

Научный руководитель: Степанова Маргарита Дмитриевна, кандидат технических наук, старший научный сотрудник, доцент кафедры ИИТ, stepanova@bsuir.by.