УДК 681.3

АНАЛИЗ МЕТОДОВ НЕЧЕТКОГО ПОИСКА

© 2018 Ю. П. Преображенский, Д. Н. Мирошник

Воронежский институт высоких технологий (г. Воронеж, Россия)

В данной работе проводится анализ основных характеристик, связанных с методами нечеткого поиска. Показана роль расстояния Левенштейна, позволяющего сделать выводы о сложности поиска.

Ключевые слова: метод, нечеткий поиск, структура данных, текстовый фрагмент, функция похожести.

В существующих условиях алгоритмы, связанные с нечетким поиском строк имеют достаточно большое распространение.

Они используются внутри в систем автоматизации перевода, для орфографических корректоров, в программах, позволяющих распознавать печатный текст и еще при построении поисковых систем [1].

Нечеткий текстовый поиск для общего случая связан с тем, что ищутся произвольные участки текстов. При этом бывают случаи, когда задачи могут быть сведены к словарному поиску.

Для действующих поисковых систем и электронных каталогов документов [2] индексацию часто связывают с технологией инвертирования**.**

Есть много общего в процессах инвертирования и формирования алфавитных указателей.

Они базируются на том, что выделяются значимые ключевые слова, а также идет составление списков того, как входят ключевые слова в тексты.

Структура данных, которая получается как результат подобного преобразования будет считаться инвертированным индексом, иногда говорят об инвертированным файле. При этом список по ключевым словам является словарем.

Для случая инвертированного файла в два этапа совершается поиск: вначале делают выборку слов запроса из словаря.

После этого происходит считывание и обработка соответствующих списков по вхождениям.

Многие из действующих электронных библиотек являются коллекциями документов, в которых есть инвертированный индекс, чтобы был быстрый доступ.

В качестве основного компонента в поисковых модулях можно рассматривать словарный поиск.

Появление ошибок и искажений может быть, и по повторно добавляемым в систему документах, и в случае пользовательских запросов.

Исходя из этого, проблема, связанная с тем, чтобы проходил эффективный нечеткий словарный поиск, является актуальной, на разных этапах: когда создается документ [3], когда идет поиск [4] по проиндексированным коллекциям.

Различные авторы рассматривали возможности осуществления нечеткого строкового поиска для случаев, когда не рассматривается предварительное индексирование. Такой случай иногда называют – on-line поиск.

Менее изученным считается словарный поиск в котором есть предварительная индексация (off-line поиск).

Есть подходы: n-граммная индексация, которая базируется на индексации фиксированной длины, разные модификации метрических деревьев, алгоритмы, связанные с поиском по абстрактным метрическим пространствам, trie-деревья (лучи). Но довольно мало исследователей, связанных со сравнительном анализом алгоритмов нечеткого словарного поиска.

В нечетком словарном поиске одной из главных является функция похожести строк.

Процесс выбора требуемой функции похожести оказывает влияние не только на характеристики качества выборки и значения скоростей поиска.

Преображенский Юрий Петрович – Воронежский институт высоких технологий, к. т. н., профессор, petrovich@vivt.ru.

Мирошник Денис Николаевич – Воронежский институт высоких технологий, аспирант.

Он связан и со сложностью реализации индексов. Разные виды искажений учитываются в хорошей функции близости по словам. Это касается удалений, замен, вставок и транспозиций символов. В самом лучшем случае учитывается и похожесть в звучании слов.

Функция Левенштейна является одной из первых предложенных мер, показывающей близость.

Расстояние Левенштейна связано с минимальным числом элементарных операций редактирования, которые требуются при преобразований одной из строк в другую.

Подобный набор содержит операции, связанные с заменой, вставкой и удалением одного символа.

Для модификации Дамерау, в множество, включающее элементарные операции, введены транспозиции символов.

При этом необходимо, чтобы для транспонированных символов не использовались другие операции, связанные с редактированием.

Подобное расстояние редактирования можно вычислить на базе методов динамического программирования.

Алгоритм характеризуется сложностью O(MN), здесь M и N – являются длинами в сравниваемых строках, а для того, чтобы найти значения расстояний необходимо провести вычисление MN элементов в так называемой матрице динамического программирования. Сложность вычисления расстояния Левенштейна–Дамерау характеризуется квадратичным порядком по размеру строк.

Многие исследователи работают над тем, чтобы создавать более эффективные алгоритмы. Возможные процедуры вычисления можно условным образом поделить по двум категориям.

В первую категорию входят алгоритмы. Они базируются на алгоритме динамического программирования, при этом, чтобы определять расстояния редактирования нет необходимости в вычислении всех MN элементов в матрице.

Во вторую категорию входят алгоритмы, базирующиеся на том, что эффективным образом применяются битовые операции.

В другом подходе к задачам ускорения вычислений расстояний редактирования осуществляется выбор более легким образом вычислимой функции похожести.

Исследователи предложили и описали разные модификации по n-граммным расстояниям, связанные с подсчетом числа общих подстрок с фиксированной длиной.

В существующих условиях известно большое число альтернативных функций близости, при этом расстояние Левенштейна–Дамерау наиболее точным образом будет соответствовать интуитивному понятию похожести.

Функция Левенштейна может быть обобщена с тем, чтобы при ее использовании точнее оценивались фонетические похожести слов.

Есть также алгоритмы, в которых поиск идет только относительного элементарного расстояния редактирования.

Но это реализуется лишь для случая, если функция Левенштейна может быть применена как фильтра.

Если объем выборки, которая получается на базе простых алгоритмов, не очень большой, тогда по каждой найденной строке можно сделать уточнение расстояния до поискового образца, при помощи более качественной и ресурсоемкой функции.

Ускорение поиска на основе сходства является целью индексации списка слов. Алгоритмы детерминированные, если , в них можно найти все строки словаря для заданной окрестности поискового терминаи. Поскольку понятие меры близости само определено неточно, не всегда имеет смысл выборка всех слов в заданной окрестности поискового шаблона [5].

Существуют рандомизированные алгоритмы, на основе которых идет поиск большей части строк, но нет гарантии, что всех.

В качестве примера можно привести поиск слов, которые имеют такое же значение функции soundex, что и в искомом слове.

Довольно часто в рандомизированных алгоритмах осуществляется индексация по значениям нескольких хеш-функций. В каждой из хеш-функций идет преобразование слов в числовые значения.

Рандомизированные алгоритмы применяют индексацию по точному равенству, в этой связи они довольно эффективны, а также они сильным образом отличаются с точки зрения полноты и точностей выборки [6, 7].

Это значит, что проводить сравнение подобных алгоритмов необходимо большей частью по качеству выборки [8].

Если есть тривиальное индексирование, когда обработка запроса идет путем последовательного перебора, нет необходимости в дополнительном преобразовании.

В качестве недостатка подобного подхода можно назвать низкую эффективность.

ЛИТЕРАТУРА

1. Львович, И. Я. Основы информатики / И. Я. Львович, Ю. П. Преображенский, В. В. Ермолова. – Воронеж,: Воронежский институт высоких технологий (Воронеж). – 2014. – 339 с.

2. Кострова, В. Н. Оптимизация распределения ресурсов в рамках комплекса общеобразовательных учреждений / В. Н. Кострова, Я. Е. Львович, О. Н. Мосолов // Вестник Воронежского государственного технического университета. – 2007. – Т. 3. – № 8. – С. 174-176.

3. Львович, Я. Е. Автоматизированное проектирование технологических процессов и систем производства РЭС / Я. Е.Львович, В. Н. Фролов, Н. П. Меткин. – М.: Издательство «Высшая Школа», 1991. – 463 с.

4. Львович, Я. Е. Оптимизационное моделирование ресурсоэффективности системы высшего образования по результатам мониторинго-рейтингового оценивания / Я. Е. Львович, А. А. Михель // Экономика и менеджмент систем управления. – 2014. – Т. – 11. – № 1-1. – С. 144-149.

5. Львович, Я. Е. Оптимизация перераспределения инвестиций на развитие ИКТ в регионе с использованием экспертных знаний / Я. Е. Львович, Д. А. Недосекин // Вестник Воронежского государственного технического университета. – 2013. – Т. 9. – № 4. – С. 82-83.

6. Куташов, В. А. Оптимизация диагностики и терапии аффективных расстройств при хронических заболеваниях / В. А. Куташов, Я. Е. Львович, И. В. Постникова. – Воронеж, Издательство «Научная книга», 2009, 228 с.

7. Завьялов, Д. В. О применении информационных технологий / Д. В. Завьялов // Современные наукоемкие технологии. – 2013. – № 8-1. – С. 71-72.

8. Кондрашов, А. В. Модели и алгоритмы решения на основе вербального анализа данных в продуктах на платформе 1с: предприятие / Кондрашов А. В., Попова Н. А. // Моделирование, оптимизация и информационные технологии. – 2018. – Т. 6. – № 2 (21). – С. 220-229.

ANALYSIS OF METHODS OF FUZZY SEARCH

© 2018 Yu. P. Preobrazhenskiy, D. N. Miroshnik

Voronezh Institute of High Technologies (Voronezh, Russia)

This paper analyzes the main characteristics associated with fuzzy search methods. The role of the distance of Levenshtein, which allows to draw conclusions about the complexity of the search, is shown

Key words: method, fuzzy search, data structure, text fragment, similarity function.