МЕТОДЫ И АЛГОРИТМЫ НЕЧЕТКОГО ПОИСКА ТЕКСТОВОЙ ИНФОРМАЦИИ

Е.А. Рубашко, С.С. Куликов

Белорусский государственный университет информатики и радиоэлектроники, Минск, Беларусь, e.rubashko@gmail.com, ollaniel@gmail.com

Abstract. Approximate string matching, or fuzzy search, provides availability of finding documents which have strings that match a pattern approximately. The closeness of a match is measured in terms of the number of primitive operations to convert the string into an exact match; it's the edit distance between the string and the pattern. Approximate string matching algorithms are classified into two categories: on-line and off-line. On-line techniques do searching without an index unlike off-line technique.

В процессе подготовки или проведения дистанционных курсов часто возникает необходимость в поиске схожих текстов, определении факта плагиата при выполнении работ и иных операций, в которых невозможно строгое сравнение текстовых величин по их значениям. В таком случае эффективным является применение нечёткого поиска.

Под нечётким поиском или поиском по сходству подразумевается поиск текстового документа, содержащего поисковый шаблон с учётом нескольких возможных различий. Среди возможных различий могут быть вставка, замена, перестановка или удаление символов.

Для определения меры схожести слов используется специальная метрика. В случае нечёткого поиска в качестве метрики удобно выбрать функцию расстояния между двумя строками, которая показывает минимальное число операций редактирования для преобразования одной строки в другую. Примерами таких метрик могут служить расстояние Хемминга, Левенштейна и Дамерау-Левенштейна.

Расстояние Левенштейна в качестве единичной операции редактирования определяет вставку, удаление или замену одного символа в строке. Перестановка символов рассматривается как две операции. В отличие от расстояния Левенштейна, расстояние Дамерау-Левенштейна определяет перестановку как единичную операцию. Расстояние Хемминга рассматривает только операции замены, а также используется только для строк одинаковой длины.

Для сопоставления поискового шаблона исходным данным существуют два принципиально разных подхода: поиск без индексации (on-line поиск) и поиск с индексацией (off-line поиск). В процессе on-line поиска каждый поисковый шаблон ищется в заранее неизвестном исходном тексте непосредственно, при этом исходный текст не требует предварительной обработки.

Примером on-line поиска служит обычный линейный поиск. Он предполагает последовательное применение заданной метрики к словам из исходного текста. Метод можно оптимизировать, если использовать метрику с ограничением.

Главным недостатком методов on-line поиска является малая скорость. Для увеличения скорости используется предварительная обработка (индексация). Индекс строится по словарю, составленному из слов исходного текста. Структура индекса может представлять собой совокупность значений "поисковой шаблон – документ – частота встречаемости поискового шаблона в документе". Индекс вместе со словарём хранятся в оперативной памяти. Наиболее известными методами нечёткого поиска с индексацией являются метод расширенной выборки, метод N-грамм, а также хеширование по сигнатуре.

Метод расширенной выборки, или метод spell checker'a, сводит нечёткий поиск к множественному поиску на совпадение. Для поискового шаблона строятся всевозможные “ошибочные” варианты, отличающиеся от исходной строки не более чем на k символов. Данный метод применяет в случае, когда размер словаря невелик и скорость поиска не является основным фактором.

Для оптимизации данного метода можно генерировать не все возможные “ошибочные” варианты, а только те из них, которые являются более естественными для человека: грамматические ошибки с учётом расположения клавиш клавиатуры.

Метод N-грамм основан на следующем предположении: если строки совпадают с учётом нескольких отличий, то с большой вероятностью у них будет хотя бы одна общая подстрока длины N. Данная подстрока называется N-граммой.

При построении индекса все слова из исходного текста разбиваются на N-граммы, после чего каждое слово попадает в списки для N-грамм, входящих в данное слово. При поиске искомая строка также разбивается на N-граммы, для каждой из которых происходит просмотр списка слов, содержащих данную N-грамму.

Метод хеширования по сигнатуре для построения индекса использует хеш-функцию (сигнатуру) строки. Для каждого значения хеша существуют списки слов, хеш которых равен данному значению. В процессе поиска строки просматриваются те списки слов, для которых значение хеша имеет не более k различий с хешем поискового шаблона.

Вычисление значения хеш-функции происходит следующим образом: весь алфавит разбивается на группы символов, и каждой группе соответствует один бит значения хеш-функции. Бит равен единице в том случае, если слово содержит символы из соответствующей группы алфавита. Порядок символов в слове значения не имеет. Пример вычисления хеш-функции для слова “анапестоидный” приведен на рисунке 1:

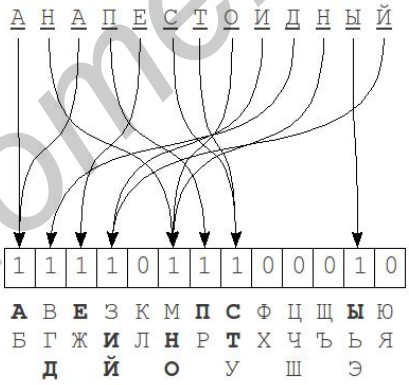

Рисунок 1 – Вычисление хеш-функции для слова “анапестоидный”

Применение рассмотренных алгоритмов нечёткого поиска в системах дистанционного обучения позволяет не только строить эффективные системы проверки орфографии, но и выполнять более сложные операции – такие как поиск текстов схожей тематики или обнаружение плагиата в выполненных слушателями работах.

Литература

1. Информационный поиск и поиск по сходству [Электронный ресурс]. – Электронные данные. – Режим доступа: http://www.itman.narod.ru

2. Нечеткий поиск в тексте и словаре [Электронный ресурс]. – Электронные данные. – Режим доступа: http://habrahabr.ru