ИСПОЛЬЗОВАНИЕ АЛГОРИТМОВ НЕЧЕТКОГО ПОИСКА ПРИ ПАРОЛЬНОЙ АУТЕНТИФИКАЦИИ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Дубинецкий В. В.

Ярмолик В. Н. – д.т.н., профессор

С развитием науки и увеличением объема информации растет необходимость в удобном и быстром поиске. В первых информационных системах поиск был основан либо на 100%-ом совпадении введенного пользователем слова или строки с оригиналом (как слово «Пассажирский» в словосочетании «станция Минск-Пассажирский»), либо на вхождении поисковой строки в исходный текст (как «Пассажир» или «саж» в словосочетании «станция Минск-Пассажирский»). Во втором случае получили широкое распространение алгоритмы, такие как алгоритм Укконена, в которых на вхождение проверяется только начало слова (по запросу «Мин» слово «Минск-Пассажирский» будет найдено, а по запросу «жир» - нет), которые применяются в системах с автодополнением текста.

В настоящее время широко распространяются программные системы и средства, поиск информации в которых основан не на простом совпадении введенного пользователем слова или строки с оригиналом или его частью, а на их схожести. В них вычисляется редакционное расстояние (расстояние Левенштейна[1] или его модификация[2]) между поисковой строкой и строкой оригинала или просто установление факта их схожести (без подсчета расстояния). Широкое применение получили метод динамического программирования Вагнера и Фишера, метод N-грамм, алгоритм Bitap и др.

Рассмотрим глубже понятие редакционного расстояния.

Расстояние Левенштейна (редакционное расстояние или дистанция редактирования) — это минимальное количество операций вставки одного символа, удаления одного символа и замены одного символа на другой, необходимых для превращения одной строки в другую.[1]

В расстоянии Дамерау–Левенштейна (модификации расстояния Левенштейна) к операциям вставки, удаления и замены добавляется операция транспозиции (перестановки двух соседних символов).[2]

Из определений следует, что все операции равнозначны и их вес (цена) при подсчете расстояния одинаков и равен единице. Однако в общих алгоритмах нахождения редакционного расстояния они могут быть произвольными.

МОДИФИКАЦИЯ РЕДАКЦИОННОГО РАССТОЯНИЯ

Проанализировав основные ошибки при вводе текста, было решено расширить список операций при подсчете редакционного расстояния:

– операция вставки одного символа, ее цена w_B;

– операция удаления одного символа, ее цена $w_У$;

– операция замены одного символа на другой, ее цена $w_З$;

– пустая операция, или операция совпадения (символы обеих строк совпадают), ее цена w_C (обычно равна нулю и добавлена для универсальности алгоритма);

– операция локального сдвига одного символа на соседний по устройству ввода (клавиатуре), ее цена $w_{ЛС}$;

– операция группового сдвига (замены последовательности из более чем N символов на соседние по устройству ввода (клавиатуре) в одном направлении), ее цена $w_{ГС}$, цена сдвига каждого символа группы $w_{ГСС}$, (обычно равна нулю и добавлена для универсальности алгоритма);

– операция смены языка ввода всего текста, ее цена $w_{СЯ}$;

– операция смены языка ввода одного символа, ее цена $w_{СЯС}$, начисляется один раз для группы подряд идущих символов;

– операция смены регистра ввода всего текста, ее цена $w_{СР}$;

– операция смены регистра ввода одного символа, ее цена $w_{СРС}$, начисляется один раз для группы подряд идущих символов;

– операция удаления продублированного символа , ее цена $w_Д$;

– операция перестановки двух соседних символов местами, ее цена $w_П$.

ПАРОЛЬНАЯ АУТЕНТИФИКАЦИЯ И НЕТОЧНЫЙ ПОИСК

Для проверки особенностей алгоритма была сгенерирована случайная выборка из 30000 слов, и для каждого из восьми слов длиной от одного до восьми символов было рассчитано расстояние до каждого слова из выборки. Параметры алгоритма: $w_B = 0.8$, $w_У = 0.9$, $w_З = 1$, $w_C = 0$, $w_{ЛС} = 0.5$, N = 2, $w_{ГС} = 0.2$, $w_{ГСС} = 0$, $w_{СЯ} = 0.1$, $w_{СЯС} = 0.7$, $w_{СР} = 0.1$, $w_{СРС} = 0.7$, $w_Д = 0.1$, $w_П = 0.1$. Результаты приведены на рис.1.

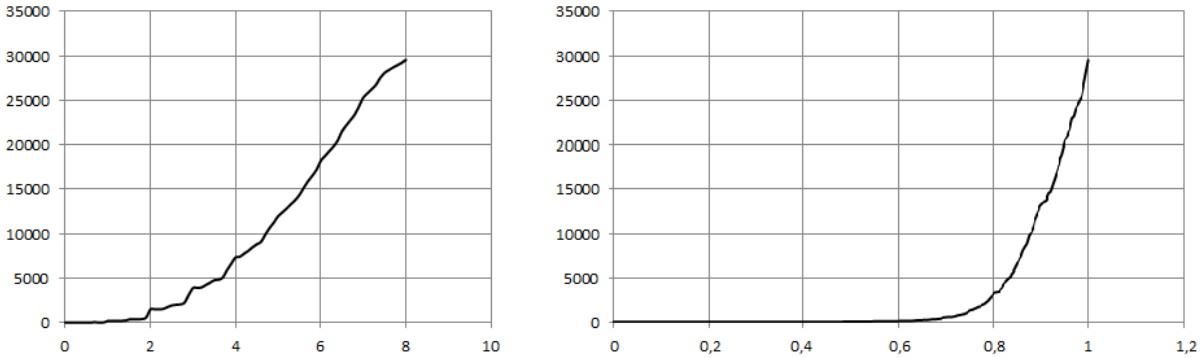

Рисунок 1– Зависимость количества успешных авторизаций от абсолютного и относительного редакционного расстояния

Относительное редакционное расстояние – это отношение редакционного (абсолютного) расстояния к длине пароля. Для определения сходства двух слов удобнее использовать именно его, т.к. одно и то же абсолютное расстояние для паролей различной длины имеет различный смысл (редакционное расстояние, равное пяти, для слов длиной шесть символов говорит больше об их различии, а для слов длиной 20 – об их сходстве).

Графики показывают, сколько различных слов будут распознаны алгоритмом как правильный пароль при заданном пороговом абсолютном или относительном расстоянии. Например, если одинаковыми считаются все слова с относительным расстоянием менее 0.6, то успешных авторизаций будет 97 (точка {0.6; 97} на втором графике). Это означает, что при оптимальном алгоритме подбора пароля взломщиком время взлома сократится в 97 раз по сравнению со строгой парольной аутентификацией. Чтобы этого не произошло, пароль при нечеткой аутентификации должен иметь сложность в 97 раз выше, чем при строгой. Например, при использовании в пароле только английских строчных букв, длина пароля должна быть на $\sqrt[26]{97} \approx 1.19 < 2$ символа больше, чем при строгой аутентификации.

Отсюда следует, что при выборе порогового значения редакционного расстояния необходимо учитывать требования к паролю (минимальная длина, обязательное наличие цифр, прописных букв и спецсимволов и др.), и наоборот, при выборе требований к паролю учитывать пороговое расстояние.

Использовать неточный поиск можно во многих случаях, однако наиболее удобно именно при вводе пароля, когда вводимый текст нельзя проверить на отсутствие ошибок и опечаток.

Однако применение редакционного расстояния для проверки пароля накладывает ряд ограничений:

– Для сравнения введенной строки с паролем, метод хранения последнего должен позволять восстанавливать его значение. Из-за этого нельзя хранить в системе только результат применения к паролю односторонних функций (хеш-значение пароля) –самый распространенный метод, обеспечивающий высокий уровень безопасности.

– Невозможно использовать данный метод в системах с высокими требованиями к защите (где нарушение безопасности может повлечь финансовый ущерб или людские потери).

– Требования к паролю (минимальная длина, наличие цифр, прописных букв и спецсимволов и др.) должны быть более жесткими, чем в обычной системе. Доступ в систему предоставляется, даже если пароль введен не точно, т.е. вместо ввода пароля можно ввести любое слово из группы слов, которые близки (в смысле редакционного расстояния) к паролю. Простая атака полным перебором для одинаковых паролей займет в разы меньше времени, чем при строгой парольной аутентификации.

Список использованных источников:

1. Расстояние Левенштейна [Электронный ресурс]. – 2013. – Режим доступа: http://ru.wikipedia.org/wiki/Расстояние_Левенштейна Дата доступа: 05.01.13
2. Damerau–Levenshtein distance [Электронный ресурс]. – 2013. – Режим доступа: http://en.wikipedia.org/wiki/Damerau–Levenshtein_distance Дата доступа: 01.04.13