МЕТОД ПОИСКА НЕЧЕТКО ЗАДАННЫХ СЛОВ

Маталыга А.А.
Кафедра инженерной психологии и педагогики
Научный руководитель: Герман О.Г., к.т.н., доцент
e-mail: kadosha1@rambler.ru

***Аннотация* — В докладе представлен оригинальный метод поиска информации, использующий методы ассоциативного и бинарного поиска.**

Ключевые слова: поиск; ассоциативная память; бинарное дерево; ошибки; опечатки.

Проблема поиска информации существует с момента появления глобальной сети Интернет. Проблема обусловлена тем, что пользователи, обычно, не имеют исчерпывающих знаний о информационном поиске. Набирая поисковый запрос, пользователи нередко ошибаются, например: пропускают или добавляют лишние буквы, пишут слова с орфографическими ошибками, пишут слова разговорным языком (включая слэнг), пишут слова не переключив клавиатуру на нужный язык.

Очевидно, что поиск информации в Интернете является больше процессом решения поисковой задачи, стоящей перед пользователем, а не просто нахождением релевантной запросу информации.

В предлагаемом методе используется аналог метода динамики средних (идея нивелирования влияния случайных отклонений при ошибках в записи ключей), полученный список поиска дает вероятностные результаты (определяемый документ не обязательно тот, поиск которого задумал клиент сайта).

Рассмотрим более детально алгоритм поиска. Когда пользователь вводит в строку запроса слово или группу слов производятся следующие действия: 1) строка запроса преобразуется в массив слов; 2) удвоенные согласные буквы заменяются только одной буквой (например: слово «удвоенный» будет трансформировано в «удвоеный», т.е. «нн» заменяется на «н»). Следующее действие: подсчитывается среднее значение ASCII-кода каждого слова и сравнивается среднее значение слова с элементом массива, т.е. сравнивается значение массива с диапазоном среднего значения ключа. Необходимо отметить, что для улучшения точности поиска нами была разработана своя таблица ASCII-кодов символов.

Первая запись входной последовательности сопоставляется с диапазоном значений корня дерева. Для каждой следующей записи ключ сначала сравнивается с диапазоном значений ключа корня, т.е. входит ли ключ записи в диапазон значений ключа корня дерева. Если он меньше чем диапазон значений ключа корня, то далее он сравнивается с диапазоном значений ключа правого потомка и т.д. до тех пор, пока потомок не будет отсутствовать. Место отсутствующего потомка занимает новая вершина, с которой сопоставляется очередная запись.

Данные действия повторяются до тех пор, пока не будет просмотрена вся входная последовательность записей.

Рассмотрим только что описанный алгоритм поиска на примере (Рис.1). В строку запроса введено слово «грепп». В начале осуществляется поиск введенного слова (ключевого слова) путем параллельного сравнения со всеми хранимыми в памяти словами. Поиск по ключу оказался безрезультатным, далее поиск автоматически продолжается по дереву. Слово «грепп» имеет ключ 1174, среднее значение ASCII-кода запроса - 234,8. Сравнивается среднее значение запроса с элементом массива, т.е. идет сравнение значения массива с диапазоном среднего значения ключа. Соответственно производится поиск диапазона значений ключа по дереву. Первая запись входной последовательности сопоставляется с диапазоном значений корня дерева.

В рассматриваемом примере ответом будет узел p4 с диапазонам значений (230; 240) в который попадает ключ искомого слова.

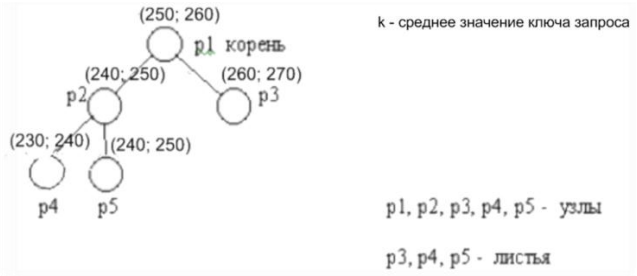

Рис. 1. Пример первого алгоритма поиска

В алгоритме «сравнений» происходит трансформация слова: исключаем все гласные буквы, поиск производится среди массива слов, которые точно так же видоизменены по степени близости. Поиск считается успешно завершенным, если видоизмененное слово найдено.

Таким образом, применение выше представленных методов позволяет уменьшить появление неудачных запросов, т.е. запросов по которым не было найдено ни одного совпадения с искомым словом.

[1] Поиск в интернете: самые популярные запросы, что чаще всего ищут в интернете, интересные запросы и опечатки [Электронный ресурс]. – Электронные данные.– Режим доступа: http://www.phorumka.ru/forum/75 - 8452-1/

[2] Кохонен, Т. Ассоциативная память / Т. Кохонен – М.: Мир, 1980. – 240 с.

[3] Власова, А.Е. Алгоритм формирования ассоциативных связей и его применение в поисковых системах / А.Е. Власова, В.И. Шабанов // Тезисы докладов международной конференции «Диалог 2003» − М., Московский государственный лингвистический университет, 2003. – 6 с.