# УСОВЕРШЕНСТВОВАННЫЙ МЕТОД НЕЧЕТКОГО ПОИСКА СЛОВ

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

*Маталыга А.А.*

*Герман О.Г., к.т.н., доцент*

В данной статье 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.

Таким образом, целесообразно при разработке алгоритма поиска учитывать типичные ошибки в поисковых запросах.

В предлагаемом нами 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.

Поиск считается успешно завершенным, если ключ искомого элемента входит в диапазон значений узла. Если поиск завершается неудачей, т.е. ключ не вошел в диапазоны, приписанные узлам дерева, то переходим к алгоритму «сравнений».

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.

Если и этот алгоритм завершился неудачей, т.е. искомое слово не было найдено, то применяем разработанный нами алгоритм «Вилка». Данный алгоритм заключается в сравнении отдельных букв в слове со словом эталоном (ключом), т.е. производится поиск (сравнение) по 1, 3 и 5 (или 2, 4, 6 в зависимости от успеха поиска) буквам в слове. Рассмотрим этот алгоритм на примере: пусть введено слово «гироскоп». Производится сначала поиск по «первой вилке» (1, 3, 5 позиция букв в слове), т.е. сравниваются слова начинающиеся на первую, потом на третью и пятую буквы. В случае успешного поиска выводится ответ (найденное слово), в случае неудачи применяется «вторая вилка», т.е. производиться поиск слов совпадающих со второй, четвертой и шестой позициями букв в слове.

Как мы видим, соединение ассоциативного поиска с поиском по дереву позволяет улучшить алгоритм ассоциативного поиска, что, в свою очередь, ведет к уменьшению затрат времени на получение необходимого пользователю объема информации.

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.

Список использованных источников:

1. Бойцов, Л. М. Классификация и экспериментальное исследование современных алгоритмов нечеткого словарного поиска / Л. М. Бойцов //Труды 6ой Всероссийской научной конференции “Электронные библиотеки: перспективные методы и технологии, электронные коллекции” – Пущино, Россия, 2004.

2. 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.

3. Кохонен, Т. Ассоциативная память / Т. Кохонен – М.: Мир, 1980. – 240 с.

4. Прохождение и поиск по бинарным деревьям [Электронный ресурс]. – Электронные данные.– Режим доступа: http://rk6.bmstu.ru/electronic_book/posapr/zadanpo/bintree.htm, свободный.