Р. Н. Селин, **С. Ю. Беседина** (Ростов-на-Дону, ФГНУ НИИ «Спецвузавтоматика»). **Алгоритмы индексации больших массивов документов.**

В настоящее время часто встает задача обеспечения быстрого поиска в больших массивах текстовых документов по их содержанию. Эта задача может быть решена при помощи создания полнотекстового индекса документов, учитывающих разные словоформы. Поиск с учетом словоформ обеспечивается путем сохранения в индексной структуре не слов целиком, а основ (стемм), которые могут быть получены алгоритмом морфологического усечения (стемминга) [1]. Сохранение только основ значимых слов имеет дополнительное преимущество в виде сокращения размера индекса. Элемент полнотекстового индекса определим как тройку (w_i, d_j, p_k^{ij}), где w_i — слово, d_j — номер документа, p_k^{ij} — k-я позиция i-го слова в j-м документе. По словам w_i строится классический B-tree индекс [2], обеспечивающий быструю выборку тройки по основе слова (стемме), что хорошо реализуется в SQL СУБД.

Построение индекса производится по схеме: парсинг исходного документа → удаление семантически незначимых слов (стоп-слов) → разбиение текста на слова → морфологическое усечение → построение индексной структуры. Поиск по запросу пользователя: Разбор поискового запроса → разбиение на слова → морфологическое усечение → последовательность стемм → выборка документов и оценка релевантности.

Последовательность стемм для поиска обозначим w_{i_k}, $k = 1, \ldots, n$. Выборку состоит из n шагов. На первом шаге выбираются документы, со словом w_{i_1}, на втором среди них выбираются слова w_{i_2} и так далее. Значение релевантности уточняется на каждом шаге выборки. Для оценки релевантности будем использовать функцию $\varphi(d_j, w_{i_k})$, определяемую так: $\varphi(d_j, w_{i_1}) = 1$; $\varphi(d_j, w_{i_k}) = \varphi(d_j, w_{i_{k-1}}) + 10\varphi_0(d_j, w_{i_k}) + 1$. Функция $\varphi_0(d_j, w_{i_k})$ определяет количество пар позиций в документе d_j удовлетворяющих условию $p_{i_k} = p_{i_{k-1}} + 1$, другими словами, что в документе d_j слова со стеммами $w_{i_{k-1}}$ w_{i_k} следуют подряд.

Таким образом наибольшую оценку релевантности будут получать документы в которых встречаются наиболее длинные последовательности слов из запроса пользователя.

СПИСОК ЛИТЕРАТУРЫ

1. Snowball–Small string-handling language [электронный ресурс] 2002–2010 — режим доступа: http://snowball.tartarus.org.
2. *Том Кайт.* Oracle для профессионалов. Пер. с англ. — ТомКайт–СПб.: ООО «ДиаСофтЮП», 2003–672 с.