УДК 004.93

Использование обратных индексов для полнотекстового поиска по большим объемам речевых данных

О. Е. Петров, В.И.Кабаров
(Университет ИТМО, г. Санкт-Петербург)
Научный руководитель – д. т. н., Ю. Н. Матвеев
(Университет ИТМО, г. Санкт-Петербург)

Полнотекстовый поиск является одной из ключевых технологий, значительно повлиявшей на развитие сети Интернет. Популярные поисковые системы дают возможность быстро получить нужную информацию, что стимулирует рост объемов информации в сети, прежде всего текстовой. Однако, прогресс в технологиях дал толчок для роста объемов не только текстовой информации, но и разного рода медиа — видео и аудио данных. Поиск по медиа данным традиционно решается с помощью сопровождающей метаинформации: ключевым словам, названию, источнику записи, списка участников и прочему. С развитием технологии автоматического распознавания речи стало возможным осуществлять поиск в том числе и по фразам, произнесенным на записи. Целый ряд алгоритмов решает задачу поиска ключевых слов в фонограммах, содержащих речь, но одна из ключевых для поиска по большим объемам данных особенностей опускается — полученные результаты необходимо ранжировать по релевантности перед выдачей пользователю. В данной статье описывается механизм полнотекстового поиска как по речевым данным, так и по текстовым, используя единый индекс данных.

Описанный в статье подход включает в себя два этапа. На первом шаге выполняется преобразование речи в текстовый формат с помощью системы автоматического распознавания речи. После этого на основе полученного текста строится обратный индекс, дающий возможность осуществлять эффективный поиск.

Вместо обычных словных сеток, полученных в результате декодирования, предполагается использовать более компактное представление словной сети — сети спутывания. Такие сети строятся дополнительным шагом с использованием MBR-декодирования. В статье представлен подход, дающий возможность строить обратные индексы поиска на основе сетей спутывания, что в свою очередь, дает возможность использовать классические подходы организации поисковых индексов.

Благодаря своим свойствам сети спутывания можно считать обобщением обычного текста. В данной статье рассмотрен алгоритм построения обратного индекса на основе сети спутывания, а также алгоритм поиска слов и словосочетаний в таком индексе. Предложенные методы позволяют переиспользовать готовые решения для полнотекстового поиска, такие как Lucene, в задачах поиска по аудиоданным.