УДК 621.319

М.А.Никитин (асп., каф. ИУС), В.П.Котляров, к.т.н., проф.

ПОИСКОВАЯ СИСТЕМА НА ОСНОВЕ КЛАССОВ СЛОВ

Подавляющая часть наиболее интересной и важной информации представлена в виде различных текстов и документов на естественном языке. Поэтому задача поиска слов и документов является чрезвычайно важной на сегодняшний день. Т.к. объёмы информации крайне велики и постоянно растут, необходимы эффективные средства поиска, которые бы быстро и с максимальной точностью удовлетворяли поисковые запросы пользователя. Очевидно, что для наиболее эффективной обработки текстов компьютером их необходимо переводить в формальный вид. Основная идея заключается в том, чтобы, используя словарь классов слов Тузова В.А., разработать метрику (набор алгоритмов), позволяющую эффективно сравнивать релевантность (значимость) различных документов в распределенной базе данных в зависимости от текстового запроса пользователя. Таким образом, результатом работы такой системы будет выдача документов в порядке значимости их смыслового содержания исходному запросу.

Наиболее популярные поисковые системы (Google, Yandex, Yahoo и др.) используют алгоритм индексирования и поиска по ключевым словам и имеют эффективность не более 15%. Эффективность разрабатываемой системы будет составлять 30-40%. Поиск основан на построении профайла встречаемости всех слов в документе, замене их на классы, как более общие сущности, к которым могут относиться сразу несколько слов, анализе связей между словами с использованием синтаксических и семантических словарей. Далее на основании разработанной метрики каждому предложению, содержащему искомое словосочетание, присваивается некий ранг, который говорит о том, насколько сильно совпадение. Затем подсчитывается суммарный ранг по документу и сравнивается с рангами других документов. В результате происходит сортировка документов в порядке их значимости.

Систему планируется внедрить для использования на мобильных телефонах в связи с их постоянно растущей популярностью и развитием мобильных технологий. Поэтому данная тема тем более актуальна, т.к. реализация эффективных алгоритмов поиска в распределенных базах данных решает проблемы быстрого поиска и ресурсных затрат, связанных с хранением повторяющихся данных, а также позволяет создать инструментальную систему поддержки поиска в беспроводной коммуникационной среде. На мобильном телефоне будет работать приложение, которое по беспроводной связи Bluetooth будет обращаться к PC с текстовым запросом. На компьютере запрос будет обрабатываться поисковой системой, и генерироваться ответ.

По своей значимости работа относится к новым технологическим средствам поддержки эффективного поиска слов и документов и не имеет аналогов среди существующих коммерческих систем.