И.С. Злыгостев

ОБРАБОТКА ТЕКСТА В ПОИСКОВЫХ СИСТЕМАХ

Объемы цифровой информации значительно выросли в последние годы. Расширились возможности коммуникационных средств взаимодействия между носителями цифровых данных. Значительному социальному кругу людей открылся доступ к глобальным и локальным компьютерным сетям. Стала актуальной задача поиска необходимой пользователю информации на распределенных носителях сети.

В таких условиях возник спрос на программные реализации продуктов, организующих поиск необходимой информации в сети по запросу, сформулированному пользователем. Реализация подобных возможностей возлагается на поисковые системы (ПС). Кроме того, растет спрос на специалистов в данной области. Как следствие, возникла необходимость в разработке образовательного контента по вопросам реализации ПС и алгоритмам, используемым в них.

В компьютерной сети информация чаще всего представляется в виде файлов. Скорость доступа к ним зависит от носителя информации и коммуникационных свойств сети. Как правило, она достаточно низка в сравнении со скоростью доступа к файлам внутри одного узла сети. В силу высоких требований пользователей к быстрому получению ответов на запросы к ПС необходимо оптимизировать алгоритмы поиска путем создания поисковых образов документов на стороне ПС. Такой подход ускоряет скорость поиска в ПС.

Большинство существующих ПС сводит поиск по всем файлам сети к поиску информации, представленной на естественном языке. Информация ищется по её имени или текстовому описанию на естественном языке. В свою очередь, обработка текста на естественном языке в ПС производится в процессе создания поискового образа документа при его индексации.

Словарь и грамматические правила естественного языка не всегда бывают формализованы. Подходы к решению задач с нечеткими данными и не формализуемыми алгоритмами решения являются предметом искусственного интеллекта.

Качество и скорость поиска в ПС во многом зависит от качества ее индексного файла. Индекс является промежуточным звеном между коллекцией документов, по которым ПС осуществляет поиск, и поисковым механизмам. Индексом в ПС является база данных поисковых образов документов, полученная в результате их индексации. Для организации оптимального по скорости для поиска доступа к индексу используются метрические деревья и trie-деревья, интегрируются разнообразные хеш-функции. В частности, индексация во многих современных ПС и электронных каталогах документов основана на технологии инвертирования.

По своей структуре инвертированные файлы аналогичны предметному указателю книги, состоят из словаря и списков вхождений слов в документы коллекции. В процессе индексации производится последовательный просмотр термов документа. Перед занесением рассматриваемого терма в

инвертированный индекс производится его нормализация. Далее слово поступает в фильтр стоп-слов. Если слово не было отсеяно как малозначимое при фильтрации, то оно заносится в инвертированный индекс документа.

Некоторые ПС осуществляют нормализацию термов на основе словарей, в которых они пытаются найти каждое индексируемое слово и сопоставить с его нормальной формой. В этом случае в поисковый индекс записываются места появления в тексте не просто найденные слова, а их нормализованные формы. Если индексируемое слово не встречается в словаре, то применяется морфологические методы его нормализации. В процессе нормализации выявляется основа слова. Она и заносится в индекс документа.

Поиск, осуществляемый по индексу ПС, составленному из нормализованных форм слов, как правило, более точный и быстрый. В фильтре стоп-слов отсеиваются термы, не являющиеся словами. Также отсеиваются часто встречающиеся во многих текстах слова (союзы, предлоги, частицы). Чем меньше объем индекса, тем быстрее по нему осуществляется поиск. Чем полнее индекс отражает содержание текста, тем точнее результаты поиска.

Отметим, что в основе базовых подходов оптимального уменьшения информации, содержащейся в индексе, лежит первый и второй закон Зипфа. Законы сформулированы для текстов на естественном языке. В них установлена обратная зависимость между частотой вхождения слова в документ и рангом этой частоты. График зависимости в декартовой системе координат представляется на положительной оси координат в виде гиперболы (рис.):

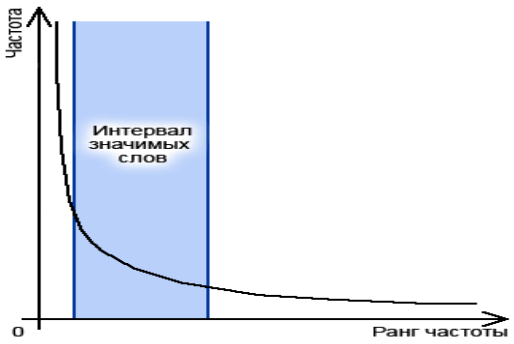

Рис. Интервал значимых слов на диаграмме, построенной по первому закону Зипфа

Наши исследования показывают, что наиболее значимые слова документа имеют ранг, лежащий в средней части диаграммы. Слова, которые попадаются слишком часто, в основном оказываются предлогами, местоимениями и т.п. Редкие слова, чаще всего, не имеют решающего семантического значения в контексте документа. Именно интервал значимых слов необходимо вносить в поисковый индекс коллекции документов.