УДК 621.7

Сравнительный анализ поисковых систем

Вандич В.Л., Матрунчик Ю.Н., Дрозд А.В.
Белорусский национальный технический университет

В настоящее время интернет переполнен информацией. Очень часто для того, чтобы найти нужную информацию приходится потратить достаточно много времени. Поиск информации в сети осуществляется с помощью поисковых систем разных видов.

Поисковая система – программно-аппаратный комплекс с веб-интерфейсом, предоставляющий возможность поиска информации в Интернете. Популярные всеязычные поисковые системы: Google (70,91%), Baidu (16,51%), Yahoo! (5,95%).

Большинство «русскоязычных» поисковых систем индексируют и ищут тексты на многих языках. Отличаются же они от «всеязычных» систем, индексирующих все документы подряд, тем, что в основном индексируют ресурсы, расположенные в доменных зонах, где доминирует русский язык или другими способами ограничивают своих роботов русскоязычными сайтами. К ним относят: Яндекс (61,3 %), Mail.ru (8,5%), Рамблер (1,9%).

Почти каждая поисковая система состоит из трех основных компонентов: веб-паук, индексатор, алгоритм поиска и оценки результатов.

Веб-паук – это специальная программа, основная задача которой – переходить по гиперссылкам «паутины» сайтов и скачивать полученные таким образом странички для второй компоненты – индексной базы.

Индексатор – это обработчик скачанных веб-пауком страниц. Он извлекает оттуда все слова и складывает их в поисковую базу (индексную базу). При этом индексатор записывает, где именно было найдено то или иное слово, и эта информация потом используется в поиске.

Алгоритм поиска – это главное умение любой поисковой системы. От алгоритма зависит эффективность полученного результата – то есть насколько быстро и точно пользователь найдет то, что его интересует. Таким образом, когда пользователь вводит свой запрос, поисковая система ищет ответ в своей индексной базе и выводит результаты в соответствии со своим алгоритмом поиска.

Для хорошей работы поисковой системы важны все три компоненты. Причем каждая из них на самом деле весьма сложна, и ее работа подчиняется огромному количеству всевозможных правил, которые к тому же постоянно корректируются. Это способствует высокой релевантности – степени соответствия документа запросу.