СТРУКТУРА ИНФОРМАЦИОННО-ПОИСКОВОЙ СИСТЕМЫ

к.т.н., доц. О.А. Козина, студент С.В. Лаптенко, НТУ "ХПИ", г. Харьков

Современный мир невозможно представить без использования всемирной паутины, именуемой Интернетом. Для навигации по ней используются информационно-поисковые системы (ИПС), при этом положение на рынке огромных фирм и корпораций, работающих в сфере интеллектуальных технологий (ИТ), зависит от качества и эффективности работы ИПС.

Проведен сравнительный анализ существующих ИПС, таких как: Yahoo, OpenText, Lycos, AltaVista, InfoSeek и WAI.

Показано, что в ИПС Yahoo отсутствует степень соответствия найденного документа запросу, не производится нормализация лексики, а также хороший результат поиска обеспечивается при условии наличия информации в базе данных.

ИПС OpenText является платным коммерческим продуктом, у которого ограничен размер поиска и отсутствует понятное описание по использованию.

Предложенная структура содержит в себе робот-индексировщик, блок поиска и блок формирования индекса. Архитектура блока формирования индекса устроена таким образом, чтобы поиск происходил максимально быстро, а также давал возможность оценить ценность каждого из найденных информационных ресурсов сети.

Согласно структуре ИПС разработаны: алгоритм робота-индексировщика, выполняющего индексацию интернет страниц, и алгоритм блока поиска, выполняющего сам поиск по запросу пользователя.

Аналитически разработан алгоритм ранжирования страниц и функции релевантности. Формула расчета релевантности:

$$score(D,Q) = \sum_{i=1}^{n} f_1(q_i) \times \frac{f(q_i) \times (k+1)}{f(q_i) + k \times (1 - b + b \times \frac{|D|}{avgdl})},$$

где D – количество слов в документе; Q – запрос, содержащий слова $q_1,\ldots,q_n$; $avgdl$ – средняя длина документа в коллекции; $f_1(q_i)$ – обратная документная частота слова q_i;. $f(q_i)$ – частота слова q_i; k и b – нормировочные коэффициенты.