Лепшокова М.Р. Информационно-поисковые системы, модели поиска текстовой информации. // Проблемы информатики в образовании, управлении, экономике и технике: Сб. статей VIII Всерос. научно-техн. конф. – Пенза: ПДЗ, 2008. – С. 84-85.

ИНФОРМАЦИОННО-ПОИСКОВЫЕ СИСТЕМЫ, МОДЕЛИ ПОИСКА ТЕКСТОВОЙ ИНФОРМАЦИИ

М.Р. Лепшокова

Карачаево-Черкесский университет им.У.Д. Алиева

Информационный поиск – самостоятельное направление исследований, изучающее вопросы поиска документов, обработки результатов поиска, а также целый ряд смежных вопросов: моделирования, классификации, кластеризации и фильтрации документов, проектирования архитектур поисковых систем и пользовательских интерфейсов, языки запросов и т.д. Документ – это порция информации, которой оперируют информационно-поисковые системы.

Информационно-поисковые системы появились на свет достаточно давно. Теории и практике построения таких систем посвящено множество статей, основная масса которых приходится на конец 70-х – начало 80-х годов. Нельзя сказать, что с появлением Интернет и бурным вхождением его в практику информационного обеспечения появилось нечто принципиально новое, чего не было раньше. На сегодняшний день нет другого способа быстрого поиска данных, кроме поиска по ключевым словам.

Ранние информационно-поисковые системы и методы поиска разрабатывались и тестировались на относительно небольших однородных коллекциях. Современные условия поиска и соответственно требования к информационно-поисковым системам претерпели значительные изменения. Главным образом, эти условия и требования связаны с развитием Интернета, который имеет свои специфические черты и особенности: динамика, взаимосвязи, свободная публикация, избыточность, неконтролируемое качество, пользователи, доступ, многоязычность.

Способы поиска можно разделить на две большие группы:

1. Библиографический поиск или поиск «по каталогу».
2. Тематический поиск или поиск «по тексту».

В состав типичной ДИПС входят, как правило, 4 основные подсистемы:

1) подсистема ввода и регистрации;
2) подсистема обработки;
3) подсистема хранения;
4) подсистема поиска.

Так как текстовые документы, поступающие на вход системы, могут быть представлены как в бумажном, так и в электронном виде, подсистема ввода и регистрации решает следующие основные задачи:

- создание электронных копий бумажных документов;
- обеспечение подключения к каналам доставки электронных документов;
- распознавание, а при необходимости и преобразование формата электронных документов;
- присвоение электронным документам уникальных идентификаторов (регистрация), а также ведение таблицы синхронизации имен.

Все поступающие документы без внесения в них каких-либо изменений направляются в подсистему хранения для сохранения в базе документов, которая представляет собой простую совокупность файлов, распределенную по каталогам жесткого диска. Однако такой тип представления базы документов характеризуется двумя недостатками: неэффективностью использования дискового проектирования; низкой скоростью доступа при большом количестве файлов.

Одним из ключевых понятий, характеризующим выбор того или иного метода анализа текстовой информации, а также реализацию конкретного варианта поиска, является модель поиска.

Модель поиска текстовой информации характеризуется четырьмя параметрами:

- представлением документов и запросов;
- критерием смыслового соответствия;
- методами ранжирования результатов запроса;
- механизмом обратной связи, обеспечивающим оценку релевантности пользователем.

К простейшим моделям поиска относится модель дескрипторного поиска и модель, основанная на Дублинском ядре. Наиболее распространенные модели – это булевая модель, модель нечетких множеств, пространственно-векторная и вероятностная модели.

Важным фактором является вид представления информации в программе-интерфейсе. Различают два типа интерфейсных страниц: страницы запросов и страницы результатов поиска. При обзоре интерфейсов и средств поиска нельзя пройти мимо процедуры коррекции запросов по релевантности. Релевантность – это мера соответствия найденного системой документа потребности пользователя. Различают формальную релевантность и реальную.

Одной из частных задач информационного поиска является задача поиска по документу-образцу.

Сегодня ИПС являются наиболее мощным механизмом поиска сетевых информационных ресурсов Интернет.