МЕХАНИЗМ ПОИСКА ДОКУМЕНТОВ В ИНФОРМАЦИОННО-ПОИСКОВОЙ СИСТЕМЕ

Глацкевич О.В.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Одинец Д.Н. – канд. техн. наук, доцент

По мере накопления документов в информационно-поисковых системах неминуемо возникает необходимость в организации эффективного механизма поиска информации в системе. Современная система должна отвечать таким критерием поиска, как быстрая скорость поиска и высокая степень соответствия найденных документов поисковому запросу. Для достижения этих целей эффективно использовать полнотекстовый поиск информации, выполняющий поиск документов по их содержимому.

Разрабатываемый поисковый механизм осуществляет полнотекстовый поиск документов в базе данных на основе информационно-поискового языка и соответствующих правил поиска. На основе текстовых данных выполняется построение инвертированного индекса, который представляет из себя структуру данных, состоящую из морфологического словаря, где каждому слову словаря соотносится список номеров документов, в котором это слово встречается.

Индексатор получает текст на вход, выполняет морфологические преобразования текста (приведение к лексемам или корневой основе слова, символьные фильтры, исключение стоп-слов) и сохраняет в поисковый индекс, сопоставляя каждому слову набор документов.

Также в индексе сохраняется метаинформация (позиция слова в документе, частота встречаемости слова в документе и во всей коллекции документов), используемая в алгоритме механизма поиска.

Система содержит лингвистическую базу данных, позволяющую выполнять морфологические преобразования входной строки посредством использования информации из лингвистической базы данных. Лингвистическая база данных содержит следующую информацию:

- словарь словоформ русского языка;
- словарь неинформативных слов;
- словарь ключевых слов;
- словарь синонимов;
- словарь сокращений.

После прохождения поисковой строки морфологического анализатора формируется результат пересечения списков инвертированного индекса в виде набора документов, содержащих каждое слово запроса. Сортировка документов выполняется на основе формулы ранжирования, выставляя каждому документу весовой коэффициент.

Основу формулы ранжирования в механизме поиска составляют функции BM25 (встречаемость слов запроса в каждом документе и во всей коллекции) и LCS (порядок слов), для которого в индексе для каждого слова записывается его позиция в документе. Формула ранжирования содержит ряд факторов ранжирования, которые являются характеристиками алгоритма поисковой системы, выполняющие анализ документов информационного массива на соответствие заданным требованиям с целью определения степени релевантности документов и представляют собой произвольные арифметические выражения, которые могут использовать константы, атрибуты документов, встроенные функции и логические операторы.

По результатам информационного поиска списки документов формируются по релевантности, где каждому документу присваивается весовой коэффициент, который представляет из себя числовое значение, полученное путем суммирования всех факторов ранжирования, встретившихся в документе. Документ с наибольшим весовым коэффициентом отобразится на первой позиции в списке результатов поиска.

В общем виде применяется следующая стратегия информационного поиска:

- формулировка запроса;
- морфологические преобразования информативных слов запроса;
- поиск релевантных документов;
- определение степени релевантности документов;
- ранжирование документов;
- предоставление результата поиска.

Список использованных источников:

1. Информационный портал habr. Как работают поисковые системы [Электронный ресурс]. Режим доступа: https://habr.com/ru/company/yandex/blog/464375/. – Дата доступа: 28.03.2021