УДК 004.021

МОДЕЛИ ИНФОРМАЦИОННОГО ПОИСКА, ИССЛЕДОВАНИЕ БУЛЕВСКОЙ МОДЕЛИ

Безуглый Е.Н. Аноприенко А.Я.
Донецкий национальный технический университет

1. Введение

В связи с большим ростом объемов информационных носителей становится возможным запоминание все больших объемов данных. Постоянно увеличивающиеся объемы информации облегчают работу человека. На основе большего количества данных вычислительным машинам становится легче прогнозировать, анализировать и изучать новые факты, связанные с различными сферами деятельности человека. Однако помимо получения новых данных, как и прежде, остается актуальным вопрос поиска адекватной информации.

Процесс поиска информации состоит их нескольких этапов:

- Индексирование данных
- Анализ запроса
- Работа модели поиска на основе результатов запроса
- Ранжирование результатов

Основной идей построение поисковых систем было создание индекса, представляющего собой некую базу данных, в которой ставиться соответствие между существующими частями документов (словами, термами) и самими документами, т.е. там, где эти слова встречаются.

На основе этих индексов действуют несколько моделей поиска.

Рассматриваются две модели поиска: булевская модель и векторная модель.

2. Булевская модель

Булевская модель является самой простой моделью [1], основанной на булевской алгебре. Согласно этой модели, выполняется поиск в индексе, представленный как матрица документов и термов (слов), и над множествами результатов поиска выполняются основные логические операции: И, ИЛИ, НЕ.

Полученный результат подвергается ранжированию.

3.Векторная модель

Основной идей этой модели является представление документов и запросов в виде векторов. Каждому терму в документе и запросу ставится в соответствии неотрицательный вес. Таким образом, каждый запрос и документ может быть

представлен в виде к-мерного вектора.

Dj = (w1j,w2j,wkj), где Dj – j-й документ,

wij – вес, i-го терма в j-м документе

Близость запроса к документу может быть вычислена, например, как скалярное произведение соответствующих векторов описаний [2]. Существует множество способов представления весов. Одним из таких может служить нормализирована частота данного терма в документе, - отношение количества повторений терма к общему количеству слов в документе.

4. Исследование булевской модели

Булевская модель в чисто теоретическом виде не используется. Существует большое множество модификаций и улучшений. В качестве матрицы индекса используется бинарное дерево. Для универсальности поиска используется стеммер английского языка — модуль для определения начальной формы слова. Может быть использован также и тагер — модуль для анализа части слова на основе синтаксиса и морфологии и смыслового употребления.

В качестве пространства для поиска было выбрано более 18 000 текстовых файлов, содержащие тексты на различные темы.

Индексация этих данных занимает 53 секунды, из них 21секунда — операции ввода-вывода.

В таблице 4.1 представлены результаты исследования работы булевской модели.

Табл. 4.1

И		**ИЛИ**		**НЕ**	
запрос,слов	время,сек	запрос,слов	время,сек	запрос,слов	время,сек
3	0,25	3	0,51	3	1,67
6	0,43	6	1,86	6	9,67
8	0,45	9	2,10	9	13,43

Табл. 4.1 Результаты исследования работы булевской модели

Литература

1. Baeza-Yates, Berthier Ribeiro-Neto Modern Information Retrieval, Pearson Education Canada, - 2007, - 2 изд. - 608с.
2. Gobinda G. Chowdhury Introduction to Modern Information Retrieval, Facet, - 2004. - 474 с.
3. Дубинский А.Г. Некоторый вопросы применения векторной модели представления документов в информационном поиске // Управляющие системы и машины. - 2001. - №4. - С. 77-78.

ЗАЯВКА НА ДОПОВІДЬ

на міжнародну студентську науково-технічну конференцію
"Інформатика та комп'ютерні технології"

1. **ВНЗ** Донецький національний технічний університет
2. **Секція** Бази даних і знанньь
3. **Назва доповіді** МОДЕЛІ ІНФОРМАЦІЙНОГО ПОШУКУ, ДОСЛІДЖЕННЯ БУЛІВСЬКОЇ МОДЕЛІ
4. **Автор доповіді** Безуглий Євген Миколайович (курс 5, СП-08м)
5. **Факультет** обчислювальної техніки та інформатики
6. **Науковий керівник** Анопрієнко Олександр Якович
   вчене звання Доцент
   науковий ступінь кандидат технічних наук
   посада Декан факультету обчислювальної техніки та інформатики
7. **Адреса автора** 83054, м. Донецьк, вул. Челюскінців 184а / 602
8. **E-mail автора** dit-2006@ya.ru
9. **E-mail керівника** **anoprien@cs.dgtu.donetsk.ua**
10. **Телефон автора** 80501614379

УДК 004.021
Безуглий Є.М.
Донецький національний технічний університет
МОДЕЛІ ІНФОРМАЦІЙНОГО ПОШУКУ, ДОСЛІДЖЕННЯ БУЛЕВСКОЇ МОДЕЛІ
Науковий керівник: декан факультету ВТІ Анопрієнко О.Я.