

В. В. Белов
А. А. Терехов
В. И. Чистякова

Повышение пертинентности поиска в современных информационных средах

Москва
Горячая линия - Телеком
2012

УДК 658.5.012:004.78:025.4.036:004.738.52

ББК 32.973.202

Б43

Рецензенты:

доктор техн. наук *М. В. Ульянов*, профессор кафедры «Персональные компьютеры и сети» Московского государственного университета приборостроения и информатики; доктор техн. наук, профессор *Е. Е. Ковшов*, заведующий кафедрой «Управление и информатика в технических системах» ГОУ ВПО МГТУ «СТАНКИН»

Белов В. В., Терехов А. А., Чистякова В. И.

Б43 Повышение pertinентности поиска в современных информационных средах. – М.: Горячая линия – Телеком, 2012. – 158 с.: ил.

ISBN 978-5-9912-0223-7.

Книга содержит исследование способа повышения показателей pertinентности информационного поиска, основанного на концепции интерфейсной поисковой системы (ИнтПС), осуществляющей объединение и переупорядочивание откликов на запросы пользователей популярных поисковых систем сети Интернет. Формализованы описания факторов ранжирования поисковых систем сети Интернет, модифицированы существующие факторы ранжирования, предложены показатели pertinентности результатов поиска и два показателя ранговой корреляции для случая разных объёмов сопоставляемых последовательностей – обобщённый и условный. Предложена концепция поисковой системы многоальтернативного поиска и адаптивного переранжирования.

Для специалистов в области информационно-поисковых систем, будет полезна студентам и аспирантам.

ББК 32.973.202

Адрес издательства в Интернет WWW.TECHBOOK.RU

Научное издание

**Белов Владимир Викторович, Терехов Алексей Андреевич,
Чистякова Валентина Ивановна**

**ПОВЫШЕНИЕ ПЕРТИНЕНТНОСТИ ПОИСКА
В СОВРЕМЕННЫХ ИНФОРМАЦИОННЫХ СРЕДАХ**

Монография

Компьютерная верстка В. И. Чистяковой

Обложка художника В. Г. Ситникова

Подписано в печать 05.10.2011. Печать офсетная. Формат 60×88/16. Уч. изд. л. 10. Тираж 500 экз.

ISBN 978-5-9912-0223-7

© В. В. Белов, А. А. Терехов,

В. И. Чистякова, 2012

© Издательство «Горячая линия – Телеком», 2012

ВВЕДЕНИЕ

Накопленные к настоящему времени колоссальные объёмы информации в совокупности с непрерывно увеличивающимися темпами её роста определяют актуальность и значимость исследований в области информационного поиска. Бурное развитие сетевых технологий, в том числе и Интернета, способствуют значительному увеличению доступных информационных ресурсов и объёмов передаваемой информации. Зачастую это разнородная, слабо структурированная и избыточная информация, обладающая высокой динамикой обновления.

При сегодняшних объёмах доступной информации решение задач информационного поиска является приоритетным для обеспечения своевременного доступа к интересующим данным в рамках *информационной среды* (ИСр).

Концепция информационной среды впервые была предложена Ю.А. Шрейдером [83], который рассматривает информационную среду не только как проводника информации, но и как активное начало, воздействующее на её участников. *Информационная среда* – совокупность технических и программных средств хранения, обработки и передачи информации, а также социально-экономических и культурных условий реализации процессов информатизации.

В настоящее время работает ряд авторитетных международных конференций, посвящённых обсуждению вопросов информационного поиска [24], например, таких как:

- TREC (Text Retrieval Conference) – цикл конференций организован под эгидой NIST (National Institute for Standards and Technology) – одного из авторитетных органов стандартизации информационных технологий в США [110,111];
- SIGIR (Special Interest Group on Information Retrieval) – цикл конференций проводимых ACM SIGIR (ACM – Association of Computing Machinery) – международной группой специалистов по информационному поиску;

- WWW (World Wide Web) Conference – специально организованная конференция для решения задач, связанных с Интернет [107, 111, 114, 115, 117].

Высокий авторитет конференций TREC, SIGIR, WWW и участие в них ведущих исследовательских коллективов и разработчиков технологий информационного поиска во многом определяет приоритетные направления исследований и задает общие принципы развития поисковых систем.

Из отечественных конференций, посвященных вопросам информационного поиска, нужно отметить ежегодную всероссийскую конференцию «Электронные библиотеки» (RCDL) и семинар по компьютерной лингвистике «Диалог».

Также необходимо отметить ряд отечественных научных школ.

- SPBU IR Group – исследовательская группа в области информационного поиска (Санкт-Петербургский Государственный Университет).
- Исследовательский центр ИИ ИПС РАН.
- Центр информационных исследований (НИВЦ МГУ).

Кроме того, существуют коммерческие организации, занимающиеся не только вопросами исследований, но и вопросами внедрения информационных технологий. Это такие известные организации как Яндекс, Рамблер, Апорт, НейрОК, Гарант-Парк-Интернет, Галактика-Зум, ABBYY-FTR, АОТ и др.

Ряд авторитетных исследователей внесли своими научными трудами значительный вклад в развитие информационно-поисковых систем: И.С. Некрестьянов, И.Е. Кураленок, В.Ю. Добрынин, А.Г. Дубинский, А.Е. Ермаков, М.Р. Когаловский, А.В. Сокирко, G. Salton, A. Singhal, M. Mitra, S. Lawrence, P. Foltz, E. Fox, J. Cho, R. Baeza-Yates, K. Tajima, C. Van Rijsbergen, L. Gravano, J. Kleinberg, J. Sparck, D. Carmel, S. Brin, L. Page, A. Singhal., T. Haveliwala.

Существует широкий спектр предлагаемых решений и перспективных направлений исследований в области информационного поиска, начиная от построения глобальных распределенных информационных структур и поисковых систем, заканчивая элементарными на первый взгляд вопросами анализа документов при помощи латентно семантического анализа [94, 96, 97]. Все они, без-

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ.....	3
----------------------	----------

Глава 1. ПРОБЛЕМЫ ПОИСКА ИНФОРМАЦИИ В СОВРЕМЕННОЙ ИНФОРМАЦИОННОЙ СРЕДЕ.....	9
--	----------

1.1. Предварительные замечания	9
1.2. Поиск информации в документальных системах	11
1.2.1. Понятие документальных систем.....	11
1.2.2. Общая функциональная структура документальных информационно-поисковых систем	14
1.3. Семантический поиск и технология Semantic Web	16
1.3.1. Технология Semantic Web	16
1.3.2. Формализация и обработка знаний на основе онтологического подхода.....	18
1.4. Интеллектуальные поисковые системы	20
1.4.1. Принципиальный алгоритм работы системы	21
1.4.2. Концептуальная архитектура интеллектуальных поисковых систем	22
1.5. Поиск в сети Интернет.....	25
1.5.1. Компоненты поисковых систем	25
1.5.2. Повышение затрат и потенциальные опасности при использовании поисковых роботов.....	29
1.6. Основные результаты.....	32

Глава 2. МЕТОДИКА ЭКСПЕРИМЕНТАЛЬНОЙ ОЦЕНКИ ПЕРТИНЕНТНОСТИ РЕЗУЛЬТАТОВ ПОИСКА	33
---	-----------

2.1. Предварительные замечания	33
2.2. Классификация поисковых запросов.....	34
2.2.1. Классификация поисковых запросов по многословности.....	34
2.2.2. Классификация по чёткости формулировки	34
2.2.3. Классификация по конкурентности запроса	35
2.2.4. Классификация на основе частотности запроса.....	36
2.2.5. Классификация по коммерческой привлекательности запроса.....	36
2.2.6. Классификация по целям пользователей.....	37

2.3. Характеристики поисковых систем Интернет, механизмы обеспечения релевантности и пертинентности.....	38
2.3.1. Статические факторы ранжирования.....	39
2.3.2. Ссылочное ранжирование	40
2.3.3. Внутренние факторы ранжирования.....	42
2.3.4. Влияние собственных ресурсов поисковых машин.....	44
2.3.5. Персонализация поиска.....	46
2.4. Методика определения пертинентности поиска при помощи экспертных оценок	46
2.4.1. Количественные оценки пертинентности.....	46
2.4.2. Описание эксперимента	51
2.4.3. Список определений и обозначений при проведении эксперимента	53
2.5. Основные результаты.....	70

Глава 3. МЕТОДЫ ВЫЧИСЛЕНИЯ ПОКАЗАТЕЛЕЙ ССЫЛОЧНОЙ АВТОРИТЕТНОСТИ

СТРАНИЦ И САЙТОВ В СЕТИ ИНТЕРНЕТ	72
3.1. Предварительные замечания	72
3.2. Определение PageRank	73
3.3. Методы вычисления PageRank.....	79
3.3.1. Итерационный метод расчёта PageRank.....	79
3.3.2. Матричный метод расчёта PageRank	80
3.3.3. Недостаток итерационных методов расчёта PageRank.....	81
3.3.4. Функциональный метод расчёта PageRank	82
3.3.5. Специфика функционального метода	83
3.3.6. Предлагаемый метод расчёта PageRank	85
3.4. Недостатки вычисления авторитетности страницы с помощью алгоритма расчёта классического показателя PR	89
3.5. Понятие SolidPageRank.....	91
3.6. Преимущества Solid PageRank	99
3.7. Инструментарий для реализации предложенного метода... ..	99
3.8. Основные результаты.....	102

Глава 4. ИНТЕРФЕЙСНАЯ ПОИСКОВАЯ СИСТЕМА

СЕТИ ИНТЕРНЕТ	104
4.1. Предварительные замечания	104

4.2. Концепция интерфейсной поисковой системы	104
4.2.1. Персонализированный поиск в Google	105
4.2.2. Сервисы социальных закладок в сети Интернет как источник определения пертинентности поиска ..	105
4.2.3. Структура интерфейсной поисковой системы	106
4.2.4. Методика формирования выдачи ИнтПС	108
4.3. Реализация многоальтернативного поиска и последующего адаптивного переранжирования	111
4.3.1. Текущая и специальная оценка показателей качества ИнтПС	111
4.3.2. Контроль и прогнозирование оценок пертинентности	116
4.3.3. Хранение оценок качества ИнтПС в виде временных рядов. Определение алгебраических операций над временными рядами	125
4.3.4. Ситуации, возникающие в процессе решения задачи идентификации статистического материала	129
4.3.5. Формирование консолидированного временного ряда	134
4.4. Идентификация структуры фрагмента сети Интернет	135
4.4.1. Предварительные замечания	135
4.4.2. Алгоритм построения матрицы смежности для произвольного фрагмента сети Интернет	138
4.5. Методика определения пертинентности поиска на основе программ AltoSearch и SearchAnalyzer	140
4.5.1. Общий алгоритм расчёта оценок пертинентности	140
4.5.2. Программа AltoSearch	140
4.5.3. Программа SearchAnalyzer: аннотация	141
4.5.4. Результаты опытной эксплуатации первой версии интерфейсной поисковой машины	142
4.6. Основные результаты	142
ЗАКЛЮЧЕНИЕ	144
СПИСОК ЛИТЕРАТУРЫ	147