

Содержание

Таблица обозначений	10
Предисловие	17
Благодарности	21
Глава 1. Булев поиск	23
1.1. Пример информационного поиска	24
1.2. Первая попытка создать инвертированный индекс	28
1.3. Обработка булевых запросов	31
1.4. Сравнение расширенной булевой модели и ранжированного поиска	35
1.5. Библиография и рекомендации для дальнейшего чтения	38
Глава 2. Лексикон и списки словопозиций	41
2.1. Схематизация документа и декодирование последовательности символов	41
2.2. Определение лексикона терминов	44
2.3. Быстрое пересечение инвертированных списков с помощью указателей пропусков	57
2.4. Словопозиции с координатами и фразовые запросы	60
2.5. Библиография и рекомендации для дальнейшего чтения	66
Глава 3. Словари и нечеткий поиск	69
3.1. Поисковые структуры для словарей	69
3.2. Запросы с джокером	72
3.3. Исправление опечаток	76
3.4. Фонетические исправления	82
3.5. Библиография и рекомендации для дальнейшего чтения	84
Глава 4. Построение индекса	85
4.1. Основы аппаратного обеспечения	85
4.2. Блочное индексирование, основанное на сортировке	87
4.3. Однопроходное индексирование в оперативной памяти	91
4.4. Распределенное индексирование	93
4.5. Динамическое индексирование	96
4.6. Другие типы индексов	99
4.7. Библиография и рекомендации для дальнейшего чтения	101

Глава 5. Сжатие индекса	103
5.1. Статистические характеристики терминов в информационном поиске	104
5.2. Сжатие словаря	108
5.3. Сжатие инвертированного файла	113
5.4. Библиография и рекомендации для дальнейшего чтения	123
Глава 6. Ранжирование, взвешивание терминов и модель векторного пространства	127
6.1. Параметрические и зонные индексы	128
6.2. Частота термина и взвешивание	134
6.3. Модель векторного пространства для ранжирования	137
6.4. Варианты функций tf-idf	143
6.5. Библиография и рекомендации для дальнейшего чтения	149
Глава 7. Ранжирование в полнофункциональной поисковой системе	151
7.1. Эффективное ранжирование	151
7.2. Компоненты информационно-поисковой системы	159
7.3. Влияние операторов языка запросов на ранжирование в векторном пространстве	162
7.4. Библиография и рекомендации для дальнейшего чтения	164
Глава 8. Оценка информационного поиска	165
8.1. Оценка информационно-поисковой системы	165
8.2. Стандартные тестовые коллекции	167
8.3. Оценка неранжированных результатов поиска	168
8.4. Оценка ранжированных результатов поиска	171
8.5. Оценка релевантности	177
8.6. Более широкая точка зрения: качество системы и ее полезность для пользователя	181
8.7. Снимпеты	183
8.8. Библиография и рекомендации для дальнейшего чтения	185
Глава 9. Обратная связь по релевантности и расширение запроса	189
9.1. Обратная связь по релевантности и псевдорелевантности	189
9.2. Глобальные методы для переформулирования запроса	200
9.3. Библиография и рекомендации для дальнейшего чтения	204
Глава 10. XML-поиск	207
10.1. Основные концепции языка XML	209
10.2. Проблемы, связанные с XML-поиском	213
10.3. Модель векторного пространства для XML-поиска	217
10.4. Оценка XML-поиска	221
10.5. Методы XML-поиска, ориентированные на текст и на данные	225
10.6. Библиография и рекомендации для дальнейшего чтения	227

Глава 11. Вероятностная модель информационного поиска	231
11.1. Основы теории вероятностей	232
11.2. Принцип вероятностного ранжирования	233
11.3. Бинарная модель независимости	234
11.4. Вероятностные модели и некоторые модификации	241
11.5. Библиография и рекомендации для дальнейшего чтения	245
Глава 12. Языковые модели для информационного поиска	247
12.1. Языковые модели	247
12.2. Модель правдоподобия запроса	252
12.3. Сравнение языкового моделирования с другими подходами к информационному поиску	258
12.4. Расширения языковых моделей	259
12.5. Библиография и рекомендации для дальнейшего чтения	260
Глава 13. Классификация текстов и наивный байесовский подход	263
13.1. Классификация текстов	266
13.2. Наивная байесовская классификация текстов	267
13.3. Модель Бернулли	272
13.4. Свойства наивной байесовской модели	274
13.5. Выбор признаков	279
13.6. Оценка классификации текстов	287
13.7. Библиография и рекомендации для дальнейшего чтения	293
Глава 14. Классификация в векторном пространстве	295
14.1. Представление документов и меры близости в векторном пространстве	297
14.2. Метод Роккио	298
14.3. Метод k ближайших соседей	302
14.4. Линейные и нелинейные классификаторы	307
14.5. Классификация с несколькими классами	311
14.6. Компромисс между смещением и дисперсией	314
14.7. Библиография и рекомендации для дальнейшего чтения	321
Глава 15. Метод опорных векторов и машинное обучение на документах	323
15.1. Метод опорных векторов: случай линейно разделимых классов	323
15.2. Расширения модели опорных векторов	330
15.3. Проблемы, связанные с классификацией текстовых документов	338
15.4. Методы машинного обучения для поиска по запросу	344
15.5. Библиография и рекомендации для дальнейшего чтения	349
Глава 16. Плоская кластеризация	353
16.1. Кластеризация в информационном поиске	354
16.2. Формулировка задачи	358
16.3. Оценивание кластеризации	359
16.4. Метод K -средних	363

16.5. Кластеризация, основанная на моделях	370
16.6. Библиография и рекомендации для дальнейшего чтения	376
Глава 17. Иерархическая кластеризация	379
17.1. Агломеративная иерархическая кластеризация	380
17.2. Кластеризация методами одиночной и полной связи	383
17.3. Агломеративная кластеризация на основе усреднения по группе	390
17.4. Кластеризация методом центроидов	392
17.5. Оптимальность агломеративной иерархической кластеризации	393
17.6. Нисходящая кластеризация	396
17.7. Именованье кластеров	397
17.8. Вопросы реализации	399
17.9. Библиография и рекомендации для дальнейшего чтения	401
Глава 18. Разложение матриц и латентно-семантическое индексирование	403
18.1. Обзор сведений из линейной алгебры	403
18.2. Матрицы “термин–документ” и сингулярные разложения	407
18.3. Малоранговые аппроксимации	409
18.4. Латентно-семантическое индексирование	411
18.5. Библиография и рекомендации для дальнейшего чтения	417
Глава 19. Основы поиска в вебе	419
19.1. Основы и история	419
19.2. Характеристики веба	421
19.3. Реклама как экономическая модель	426
19.4. Опыт пользователей поисковых систем	428
19.5. Размер индекса и оценка его размера	430
19.6. Нечеткие дубликаты и алгоритм шинглов	434
19.7. Библиография и рекомендации для дальнейшего чтения	438
Глава 20. Обход и индексирование веба	439
20.1. Обзор	439
20.2. Обход веба	440
20.3. Распределение индексов	449
20.4. Серверы проверки ссылочной связности	450
20.5. Библиография и рекомендации для дальнейшего чтения	453
Глава 21. Анализ ссылок	455
21.1. Веб как граф	455
21.2. Метод PageRank	457
21.3. Порталы и авторитетные источники	466
21.4. Библиография и рекомендации для дальнейшего чтения	472
Библиография	473
Предметный указатель	506