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. Сниппеты 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. Именование 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

Глава 9

Обратная связь по релевантности и расширение запроса

Во многих коллекциях одно и то же понятие может выражаться разными словами. Это явление, известное как *синонимия* (synonymy), влияет на полноту поиска в большинстве информационно-поисковых систем. Например, пользователи хотели бы, чтобы запросу `aircraft` соответствовало также слово `plane` (но только в смысле самолет, а не *столярный рубанок*), а запросу `thermodynamics` (термодинамика) — слово `heat` (тепло) в соответствующем контексте. Пользователи часто стараются самостоятельно разрешить эту проблему, уточняя запросы (см. раздел 1.4). В этой главе мы рассмотрим способы, с помощью которых система может сама уточнить запрос либо автоматически, либо с участием пользователя.

Методы решения этой задачи разделяются на две основные категории: глобальные и локальные. Глобальные методы предусматривают расширение или новую формулировку запроса независимо от запроса и возвращаемых результатов, так что изменения в формулировке запроса приводят к появлению нового запроса, соответствующего другим семантически близким терминам. К глобальным относятся следующие методы.

- Расширение/новая формулировка запроса с помощью специального тезауруса или тезауруса WordNet (раздел 9.2.2)
- Расширение запроса с помощью автоматической генерации тезауруса (раздел 9.2.3)
- Методы, похожие на приемы исправления опечаток (см. главу 3)

Локальные методы изменяют запрос с учетом документов, найденных по исходному запросу. К локальным относятся следующие методы.

- Обратная связь по релевантности (раздел 9.1.1)
- Обратная связь по псевдорелевантности, известная также как *слепая обратная связь по релевантности* (раздел 9.1.6)
- (Глобальная) неявная обратная связь по релевантности (раздел 9.1.7)

В этой главе мы коснемся всех упомянутых подходов, но сосредоточим свое внимание лишь на методе обратной связи по релевантности, который оказался наиболее распространенным и успешным.

9.1. Обратная связь по релевантности и псевдорелевантности

Идея *обратной связи по релевантности* (Relevance Feedback — RF) заключается в привлечении пользователя к процессу поиска, чтобы улучшить итоговый список результатов. В частности, пользователь сообщает системе о релевантности документов в первоначальном списке результатов. Вкратце эта процедура выглядит следующим образом.

- Пользователь делает (короткий, простой) запрос.
- Система возвращает первоначальный список найденных результатов.
- Пользователь отмечает некоторые из найденных документов как релевантные или нерелевантные.
- Система определяет улучшенное представление информационной потребности, основываясь на обратной связи с пользователем.
- Система выводит на экран уточненный набор найденных результатов.

Метод RF может предусматривать одну или несколько итераций. В основе этого процесса лежит идея, согласно которой пользователь не в состоянии сформулировать точный запрос, не зная хорошо содержания коллекции, но может оценить документы. Поэтому целесообразно выполнить несколько таких итераций, чтобы уточнить запрос. В рамках этого сценария метод RF может способствовать эволюции информационной потребностей пользователя. Просмотр некоторых документов может помочь пользователю уточнить свои представления об информации, которую он ищет.

Ярким примером метода RF является поиск изображений. Результаты поиска изображений легко просмотреть, но именно в этой области пользователю трудно сформулировать свой запрос словами, но легко указать, какие изображения являются релевантными или нерелевантными. После того как пользователь введет исходный запрос `bike` на странице

`http://nayana.ece.ucsb.edu/imsearch/imsearch.html`,

он получит первоначальный список результатов (в данном случае — изображений). На рис. 9.1, *а* пользователь выбрал изображения, которые считает релевантными. Они используются для уточнения запроса, в то время как остальные изображения на новую формулировку запроса не влияют. На рис. 9.1, *б* показаны новые результаты, ранжированные после выполнения итерации по методу RF.

На рис. 9.2 приведен пример использования метода RF для текстового поиска; в данном случае пользователь хочет узнать о новых применениях космических спутников.

9.1.1. Алгоритм Роккио для обратной связи по релевантности

Алгоритм Роккио (Rocchio algorithm) — классический алгоритм для реализации метода RF. Он инкорпорирует модель обратной связи по релевантности в модель векторного пространства, описанную в разделе 6.3.

Теория. Мы хотим найти вектор запроса $\vec{q}$, максимально близкий к релевантным документам и минимально похожий на нерелевантные документы. Если C_r — это множество релевантных документов, а C_{nr} — нерелевантных, то целью наших поисков является следующий вектор.[1]

$$\vec{q}_{opt} = \arg\max_{\vec{q}} \left[sim(\vec{q}, C_r) - sim(\vec{q}, C_{nr}) \right] \tag{9.1}$$

[1] В этом равенстве выражение $\arg\max_x f(x)$ означает значение x, при котором функция $f(x)$ достигает максимума. Аналогично выражение $\arg\min_x f(x)$ означает значение x, при котором функция $f(x)$ достигает минимума.

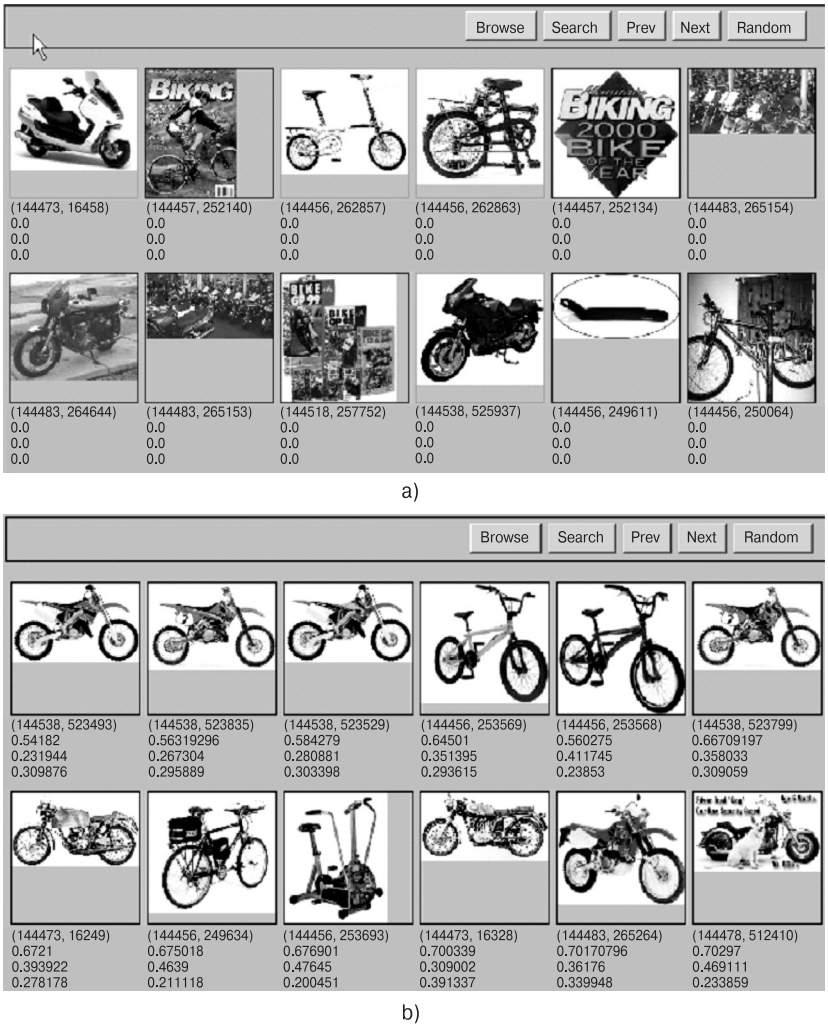

Рис. 9.1. Метод поиска изображений с помощью метода RF: а) пользователь просматривает первоначальный список результатов поиска по запросу `bike`*, выбирает в качестве релевантных первое, третье и четвертое изображения в первом ряду и четвертое — в нижнем ряду, а затем возвращает их в качестве обратной связи; б) пользователь получает уточненный набор результатов. Точность существенно повысилась. Снимок с сайта http://nayana.ece.ucsb.edu/imsearch/imsearch.html (Newsam et al., 2001)*

Здесь величина *sim* определена равенством (6.10). Если в качестве меры близости используется косинусная мера сходства, то оптимальный вектор запроса $\vec{q}_{opt}$ для разделения релевантных и нерелевантных документов определяется следующим равенством.

$$\vec{q}_{opt} = \frac{1}{|C_r|}\sum_{\vec{d}_j \in C_r} \vec{d}_j - \frac{1}{|C_{nr}|}\sum_{\vec{d}_j \in C_{nr}} \vec{d}_j \qquad (9.2)$$

а) Query: New space satellite applications

б)
\+ 1. 0.539, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer
\+ 2. 0.533, 07/09/91, NASA Scratches Environment Gear From Satellite Plan
3. 0.528, 04/04/90, Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes
4. 0.526, 09/09/91, A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget
5. 0.525, 07/24/90, Scientist Who Exposed Global Warming Proposes Satellites for Climate Research
6. 0.524, 08/22/90, Report Provides Support for the Critics Of Using Big Satellites to Study Climate
7. 0.516, 04/13/87, Arianespace Receives Satellite Launch Pact From Telesat Canada
\+ 8. 0.509, 12/02/87, Telecommunications Tale of Two Companies

в) 2.074 new 15.106 space
30.816 satellite 5.660 application
5.991 nasa 5.196 eos
4.196 launch 3.972 aster
3.516 instrument 3.446 arianespace
3.004 bundespost 2.806 ss
2.790 rocket 2.053 scientist
2.003 broadcast 1.172 earth
0.836 oil 0.646 measure

г)
* 1. 0.513, 07/09/91, NASA Scratches Environment Gear From Satellite Plan
* 2. 0.500, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer
3. 0.493, 08/07/89, When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own
4. 0.493, 07/31/89, NASA Uses 'Warm' Superconductors For Fast Circuit
* 5. 0.492, 12/02/87, Telecommunications Tale of Two Companies
6. 0.491, 07/09/91, Soviets May Adapt Parts of SS-20 Missile For Commercial Use
7. 0.490, 07/12/88, Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers
8. 0.490, 06/14/90, Rescue of Satellite By Space Agency To Cost $90 Million

*Рис. 9.2. Пример использования метода RF на текстовой коллекции: а) исходный запрос; б) пользователь отмечает релевантные документы (помечены плюсом); в) запрос расширяется до 18 терминов с указанными весами; г) пользователю возвращается уточненный список результатов. Символом * отмечены документы, которые на этапе обратной связи по релевантности пользователь отметил как релевантные*

Иначе говоря, оптимальный запрос — это вектор разницы между центроидами релевантных и нерелевантных документов (рис. 9.3). Однако это наблюдение не очень полезно, поскольку полное множество релевантных документов неизвестно — именно его мы ищем.

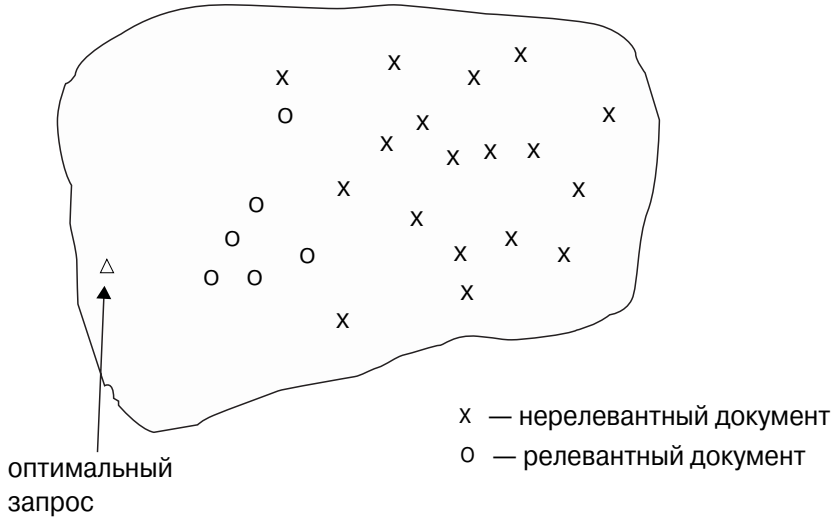

Рис. 9.3. Оптимальный по Роккио запрос для разделения релевантных и нерелевантных документов

Алгоритм Роккио (1971). Описанный механизм обратной связи по релевантности был реализован в системе Дж. Солтона SMART около 1970 года и стал известен благодаря этой системе. Пусть у нас есть запрос пользователя и — частично — знание о релевантности документов. Алгоритм предлагает использовать модифицированный запрос $\vec{q}_m$.

$$\vec{q}_m = \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j \qquad (9.3)$$

Здесь q_0 — оригинальный вектор запроса, D_r и D_{nr} — множества известных релевантных и нерелевантных документов соответственно, α, β и γ — веса каждого слагаемого. Эти величины управляют вкладом множества оцененных документов и первоначального запроса. Если у нас много оцененных документов, то мы можем увеличить веса β и γ. Начиная с запроса q_0, новый запрос перемещается на определенное расстояние в направлении центроида релевантных документов и на некоторое расстояние от центроида нерелевантных документов. Этот новый запрос можно использовать для поиска документов в стандартной модели векторного пространства (см. раздел 6.3). Мы можем уйти из положительного квадранта векторного пространства, вычитая вектор нерелевантного документа. В алгоритме Роккио отрицательные веса терминов игнорируются. Иначе говоря, вес такого термина полагается равным нулю. Эффект применения обратной связи по релевантности показан на рис. 9.4.

Обратная связь по релевантности может улучшить как полноту, так и точность. Однако на практике было показано, что она является наиболее полезной для повышения полноты в тех ситуациях, когда полнота критична. Частично это объясняется тем, что метод Роккио расширяет запрос (т.е. добавляются новые термины), а частично — контекстом использования. Когда пользователи хотят достичь высокой полноты, они готовы потратить время на просмотр результатов и повторно выполнить поиск. Положительная обратная связь также оказывается более ценной, чем отрицательная, поэтому в большинстве информационно-поисковых систем принято устанавливать параметры так, чтобы $\gamma < \beta$. Разумно выбрать следующие параметры: $\alpha = 1$, $\beta = 0{,}75$ и $\gamma = 0{,}15$. Многие системы, на-

пример система поиска изображений, продемонстрированная на рис. 9.1, допускают только положительную обратную связь, что эквивалентно установке $\gamma = 0$. В качестве альтернативы можно использовать как отрицательную обратную связь только нерелевантные документы с высоким рангом (т.е. в равенстве (9.3) $|D_{nr}| = 1$). Несмотря на то что многочисленные экспериментальные результаты сравнения разных вариантов обратной связи по релевантности не позволяют сделать окончательные выводы, некоторые исследователи считают, что именно этот вариант, получивший название *Ide dec-hi*, является наиболее эффективным или по крайней мере наиболее непротиворечивым.

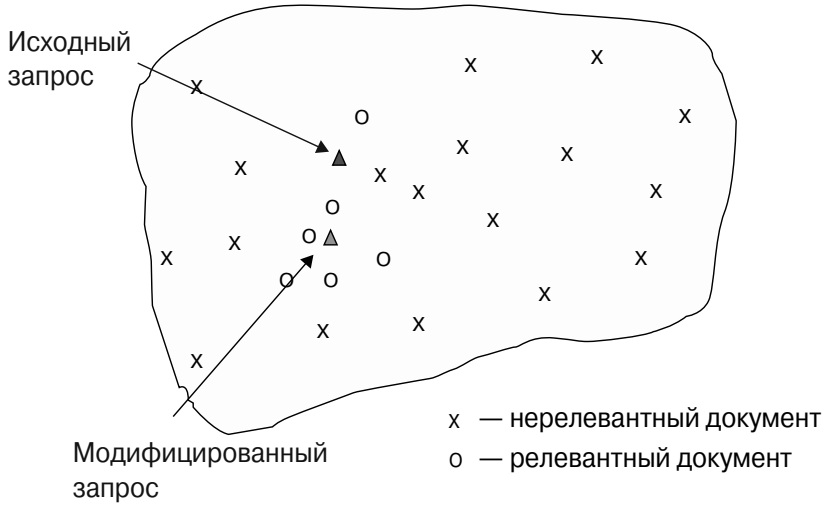

Рис. 9.4. Применение алгоритма Роккио. Некоторые документы помечены как релевантные или нерелевантные, а вектор исходного запроса перемещается в ответ на сигнал обратной связи

? ***Упражнение 9.1.*** При каких условиях модифицированный запрос q_m в равенстве (9.3) совпадает с исходным запросом q_0? Можно ли утверждать, что во всех других вариантах запрос q_m ближе к центроиду релевантных документов, чем запрос q_0?

Упражнение 9.2. Почему положительная обратная связь в информационно-поисковых системах считается более полезной, чем отрицательная? Почему использование только одного нерелевантного документа может быть более эффективным, чем использование нескольких?

Упражнение 9.3. Допустим, что исходный запрос пользователя имеет вид *cheap CDs cheap DVDs extremely cheap CDs*. Пользователь сравнивает два документа: d_1 и d_2. Он считает документ d_1, содержащий фразу *CDs cheap software cheap CDs*, релевантным, а документ d_2, содержащий фразу *cheap thrills DVDs*, нерелевантным. Допустим, что мы используем обычную частоту термина (без масштабирования и без документной частоты). Нормировать вектор не обязательно. Как изменится вектор запроса после поступления сигнала обратной связи Роккио, как указано в равенстве (9.3)? Будем считать, что $\alpha = 1$, $\beta = 0{,}75$ и $\gamma = 0{,}25$.

Упражнение 9.4 **[*].** Омар (Omar) реализовал систему веб-поиска, основанную на методе RF, в которой обратная связь по релевантности использует только слова в

заголовке страницы (для эффективности). Пользователь собирается ранжировать три результата. Первый пользователь, Джинксинг, послал запрос

banana slug

и получил в ответ три заглавия:

banana slug Ariolimax columbianus

Santa Cruz mountains banana slug

Santa Cruz Campus Mascot

Джинксинг считает два первых документа релевантными, а третий — нерелевантным. Допустим, что поисковая система Омара использует частоту термина и не применяет нормализацию длины векторов и обратную документную частоту. Предположим также, что параметры алгоритма Роккио выбраны так: $\alpha = \beta = \gamma = 1$. Покажите, как выглядит окончательный запрос. (Элементы вектора перечислите в алфавитном порядке.)

9.1.2. Вероятностная обратная связь по релевантности

Кроме изменения веса термина запроса в векторном пространстве, существуют другие способы организации обратной связи по релевантности. Например, если пользователь указал несколько релевантных и нерелевантных документов, то можно построить классификатор. В частности, основой классификатора может стать наивная байесовская вероятностная модель. Пусть R — булева индикаторная переменная, отражающая релевантность документа. Тогда мы можем оценить величину $P(x_t = 1)$, т.е. вероятность того, что термин t встретится в документе, в зависимости от того, релевантный это документ или нет.

$$\begin{aligned} \hat{P}\left(x_t = 1 \middle| R = 1\right) &= |VR_t| / |VR|, \\ \hat{P}\left(x_t = 0 \middle| R = 0\right) &= \left(df_t - |VR_t|\right) / \left(N - |VR|\right). \end{aligned} \tag{9.4}$$

Здесь N — общее количество документов, df_t — количество документов, содержащих термин t, VR — множество известных релевантных документов и VR_t — подмножество документов из VR, содержащих термин t. Хотя множество известных релевантных документов, скорее всего, является небольшим подмножеством множества всех релевантных документов, если предположить, что само множество релевантных документов представляет собой небольшое подмножество всех документов, то приведенные выше оценки вполне разумны. На основе этих оценок можно предложить новый способ изменения весов терминов запроса. Более подробно вероятностные подходы будут исследованы в главах 11 и 13. В частности, применение этого подхода к обратной связи по релевантности будет изложено в разделе 11.3.4. Пока заметим, что равенств (9.4) еще недостаточно для изменения весов терминов. Эти равенства используют статистику по коллекции и информацию о распределении термина в документах, считающихся релевантными, но не информацию о конкретном запросе.

9.1.3. Когда обратная связь по релевантности эффективна?

Успех метода RF зависит от определенных предположений. Во-первых, пользователь должен иметь достаточно знаний, чтобы сформулировать исходный запрос, хоть в какой-то мере близкий к искомым документам. Это условие является необходимым в любой

информационно-поисковой системе, но следует отметить несколько проблем, которые метод RF сам по себе устранить не может.

- *Неправильное правописание*. Если пользователь пишет термин запроса в том виде, в котором он не встречается ни в одном документе коллекции, то метод RF вряд ли окажется эффективным. Эту проблему можно разрешить с помощью исправления ошибок, описанного в разделе 3.
- *Многоязычные информационно-поисковые системы*. Документы на разных языках в векторном пространстве находятся далеко друг от друга, а документы на одном и том же языке группируются ближе друг к другу.
- *Несогласованность словаря пользователя и словаря коллекции*. Если пользователь ищет термин `laptop`, а все документы содержат слово `notebook`, то запрос не принесет результата, и обратная связь по релевантности, скорее всего, станет неэффективной.

Во-вторых, метод обратной связи по релевантности требует, чтобы релевантные документы были похожи друг на друга, т.е. образовывали кластеры. В идеальном случае распределение термина по всем релевантным документам должно быть похожим на распределение термина во всех документах, отмеченных пользователем, а распределение термина по всем нерелевантным документам должно отличаться от распределения термина в релевантных документах. Метод работает хорошо, если все релевантные документы образуют кластер вокруг отдельного прототипа или существуют разные прототипы, словари релевантных документов сильно перекрываются и схожесть между релевантными и нерелевантными документами мала. Модель Роккио неявно интерпретирует релевантные документы как отдельный *кластер*, который моделируется с помощью центроида. Этот подход не работает, если релевантные документы образуют мультимодальный класс, т.е. состоят из нескольких кластеров в векторном пространстве. Это может произойти в следующих ситуациях.

- Подмножества документов используют разные словари, например `Burma`[2] или `Myanmar`
- Запрос, множество ответов на который разнородно (дизъюнктивно) по своей природе, например `Pop stars who once worked at Burger King`
- Общие понятия, которые часто распадаются на дизъюнкцию нескольких более конкретных понятий, например `felines`

Качественно подготовленное содержание документов коллекции часто может помочь в решении этой проблемы. Например, статья об отношении различных групп к ситуации в Бирме может содержать терминологию, которую используют разные стороны: в результате возникают связи между кластерами документов.

Обратная связь по релевантности не всегда нравится пользователям. Они часто отказываются от явной обратной связи или вообще не желают продолжать. Более того, часто по результатам поиска на основе обратной связи трудно понять, почему был найден конкретный документ.

Кроме того, обратная связь по релевантности порождает несколько практических проблем. Длинные запросы, генерируемые в результате применения метода RF, в обыч-

[2] Бирма (Burma) — прежнее название государства Мьянма (Myanmar). — *Примеч. ред.*

ных информационно-поисковых системах оказываются неэффективными. Это приводит к большим вычислительным затратам и увеличивает время отклика на запрос пользователя. Частичное решение этой проблемы можно получить, изменив веса самых важных терминов в релевантном документе, например первых двадцати наиболее часто встречающихся терминов. Согласно некоторым экспериментальным результатам использование ограниченного количества терминов может дать более хорошие результаты (Harman, 1992), хотя в другой публикации авторы утверждают, что использование большего количества терминов повышает качество найденных документов (Buckey et al., 1994*b*).

9.1.4. Обратная связь по релевантности в вебе

Некоторые системы веб-поиска предоставляют функциональность поиска похожих/связанных документов: пользователь указывает в списке результатов документ, который, по его мнению, лучше всего соответствует его информационной потребности, и запрашивает другие документы, похожие на него. Такую функциональность можно рассматривать как упрощенный вариант обратной связи по релевантности. Однако в целом метод RF редко используется в вебе. Исключением была система веб-поиска Excite, которая первоначально предоставляла полноценную обратную связь по релевантности. Однако от этой опции со временем отказались, так как она не была востребована пользователями[3]. В вебе лишь немногие люди используют расширенные возможности поиска, а большинство предпочитает ограничиться единственным запросом. Такое неприятие может объясняться двумя причинами: во-первых, суть метода RF трудно объяснить типичному пользователю и, во-вторых, основной целью этого метода является повышение полноты поиска, которая, как правило, редко интересует пользователей систем веб-поиска.

В 2000 году Спинк и др. (Spink et al., 2000) опубликовали результаты использования метода RF на поисковой системе Excite. Этот метод применялся только примерно в 4% поисковых сессий, причем большинство из них ограничивались использованием опции “More like this” (“Похожие документы”), которая сопровождала каждый результат. Около 70% пользователей просмотрели первую страницу результатов и не искали документы на следующих. При использовании метода RF результаты были улучшены примерно в двух третях случаев.

Более важным направлением в последнее время стал анализ кликов пользователя по результатам поиска, что обеспечивает неявную обратную связь по релевантности. Использование таких данных подробно изучено в работах Йоахимса (Joachims, 2002*b*; Joachims et al., 2005). Очень успешным стало использование структуры гиперссылок в вебе (см. главу 21), которую также можно рассматривать как вид неявной обратной связи, хотя эта связь устанавливается с авторами страницы, а не с ее читателями (хотя на практике большинство авторов также являются читателями).

9.1.5. Оценка стратегий обратной связи по релевантности

Интерактивная обратная связь по релевантности позволяет получить существенный выигрыш в качестве поиска. С эмпирической точки зрения часто очень полезным оказывается даже один цикл обратной связи по релевантности. Два цикла иногда оказываются

---

[3] Функция, модифицирующая первоначальный запрос с помощью включения в него выбранного пользователем релевантного документа, присутствует в форме *[текст_старого_запроса related:URL_документа]* во многих поисковых системах, например в Google и Bing. — *Примеч. ред*.

лишь ненамного полезнее. Для успешного использования RF требуется достаточно большое количество оцененных документов, иначе процесс становится неустойчивым и может уйти в сторону от информационной потребности пользователя. Соответственно, рекомендуется иметь не менее пяти оцененных документов.

Трудно оценить эффективность метода RF в полной мере и наглядно. Очевидно, что в качестве первой стратегии можно начать с исходного запроса q_0 и построить график "точность–полнота". После одного цикла обратной связи от пользователя мы вычисляем модифицированный запрос q_m и снова строим график "точность–полнота". В ходе обоих циклов оценивается качество работы системы по всем документам в коллекции, которая допускает простое сравнение. В этом случае можно получить впечатляющий выигрыш: увеличение показателя MAP примерно на 50%. Но, к сожалению, это обман. Этот выигрыш частично объясняется тем фактом, что известные релевантные документы (оцененные пользователем) теперь получают более высокий ранг. Для корректной оценки качества сравнение следует проводить только по документам, неизвестным пользователю.

Вторая идея заключается в том, чтобы использовать в ходе второго цикла документы из *остаточной коллекции* (residual collection), т.е. множество документов за исключением оцененных как релевантные. Такая оценка кажется более реалистичной. К сожалению, измеренная эффективность часто оказывается ниже, чем для исходного запроса. Этот эффект особенно ярко проявляется, когда существует небольшое количество релевантных документов, и значительная их доля была оценена пользователем в ходе первого цикла. Относительную эффективность разных методов обратной связи по релевантности можно корректно оценить, но очень трудно обоснованно сравнить качество систем с обратной связью по релевантности и без нее, поскольку размер коллекции и количество релевантных документов до и после цикла обратной связи различаются.

Таким образом, ни один из этих методов не является вполне удовлетворительным. Третий метод предполагает работу с двумя коллекциями, одна из которых используется для исходного запроса и оценок релевантности, а вторая — для сравнительной оценки. Качество обработки запросов q_0 и q_m можно корректно сравнить с помощью второй коллекции.

Возможно, наилучшим способом оценки полезности метода RF являются эксперименты с участием пользователей, в частности, путем сравнения временны*х* показателей: насколько быстро пользователь находит релевантные документы с помощью RF по сравнению с другой стратегией (например, с новой формулировкой запроса) или как много релевантных документов пользователь находит за определенный отрезок времени. Такие оценки полезности являются наиболее корректными и близкими к реальному использованию системы.

9.1.6. Обратная связь по псевдорелевантности

Обратная связь по псевдорелевантности (pseudo relevance feedback), или *слепая обратная связь по релевантности* (blind relevance feedback), — это метод автоматического локального анализа. Он позволяет автоматизировать ту часть RF, которая выполняется вручную, так что пользователь повышает качество поиска, не вступая в дополнительное взаимодействие с системой. В рамках этого метода сначала выполняется поиск и находится исходная совокупность наиболее релевантных документов, в которой первые k документов, имеющие наибольшие ранги, *предполагаются* релевантными, а затем к ним применяется метод RF с учетом этого предположения.

В большинстве случаев этот автоматический метод работает хорошо[4]. Опыт показывает, что он работает лучше, чем глобальный анализ (раздел 9.2). Было показано, что метод повышает качество выполнения заданий дорожки TREC ad hoc (рис. 9.5). Однако автоматический процесс не лишен недостатков. Например, если запрос имеет вид `copper mines` (медные рудники) и в нескольких документах, занимающих первые места, речь идет о рудниках в Чили, то возможен дрейф запросов в направлении документов о Чили.

	Точность на уровне k = 50	
Взвешивание термина	Без RF, %	ПсевдоRF, %
lnc.ltc	64,2	72,7
Lnu.ltu	74,2	87,0

Рис. 9.5. Результаты, свидетельствующие о том, что метод обратной связи по псевдорелевантности существенно улучшает качество поиска. Эти результаты получены на системе Cornell SMART в рамках эксперимента TREC 4 (Buckley et al., 1995). В этом исследовании сравнивались две разные схемы нормализации длины (l и L, см. рис. 6.15). Метод обратной связи по псевдорелевантности сводился к добавлению двадцати терминов к каждому запросу

9.1.7. Неявная обратная связь по релевантности

В качестве базы для обратной связи можно использовать косвенные свидетельства вместо явных оценок релевантности. Этот метод часто называется *неявной обратной связью по релевантности* (implicit relevance feedback). Неявная обратная связь менее надежна, чем явная, но более полезна, чем обратная связь по псевдорелевантности, не учитывающая мнение пользователей. К тому же пользователи часто отказываются участвовать в явной обратной связи, а собрать данные большого объема о неявной обратной связи в случае большой системы, например системы веб-поиска, несложно.

В контексте веба был предложен метод DirectHit, идея которого состоит в том, чтобы присваивать более высокий ранг тем документам, которые пользователи выбирают для просмотра чаще. Иначе говоря, предполагается, что клики по ссылкам на страницы являются показателями релевантности этих страниц запросу. В этом подходе делается несколько предположений, например что сниппеты документов в результатах поиска (с помощью которых пользователи выбирают, на какие страницы перейти) являются индикаторами релевантности этих документов. В оригинальной системе поиска DirectHit данные о кликах собирались глобально и не были связаны с конкретными пользователями или запросами. Этот метод является одной из разновидностей общего подхода *анализа кликов* (clickstream mining). В настоящее время очень похожий метод используется для ранжирования рекламных объявлений, соответствующих поисковым запросам в вебе (глава 19).

[4] Важно не забывать об упомянутой выше дополнительной нагрузке на систему, которую создает обратная связь по псевдорелевантности: предполагается не просто собрать большое количество новых терминов, но и затем автоматически задать еще один запрос системе. — *Примеч. ред.*

9.1.8. Резюме

Как показали исследования, метод RF позволяет очень эффективно повысить релевантность результатов. Для его успешного использования необходимы запросы, для которых существует достаточно много релевантных документов. Полная обратная связь по релевантности является обременительной для пользователя, а ее реализация в большинстве информационно-поисковых систем не очень эффективна. Во многих случаях достичь аналогичного улучшения можно с помощью других методов интерактивного поиска, затратив меньше усилий.

Помимо основного сценария поиска по произвольному запросу, обратная связь по релевантности используется в следующих ситуациях.

- Отслеживание изменяющихся информационных потребностей (например, марки автомобилей, которыми интересуется пользователь, со временем изменяются).
- Поддержка информационных фильтров (например, для получения новостей). Такие фильтры рассматриваются в главе 13.
- Активное обучение (active learning) (нахождение примеров, которые полезно классифицировать вручную, чтобы сократить затраты на формирование обучающей выборки).

? ***Упражнение 9.5.*** Какие значения весов α, β и γ в алгоритме Роккио соответствуют команде “найти похожую страницу”?

Упражнение 9.6 **[*].** Назовите три причины, по которым метод обратной связи по релевантности редко используется в поисковых веб-системах.

9.2. Глобальные методы для переформулирования запроса

В данном разделе мы кратко обсудим три глобальных метода расширения запроса: путем подсказок и помощи пользователю с использованием тезауруса, создаваемого вручную, и с использованием тезауруса, создаваемого автоматически.

9.2.1. Словарные инструменты для переформулирования запроса

Для того чтобы пользователь мог определить, насколько был успешен его поисковый запрос, существуют разные методы поддержки. К ним относятся информация о стоп-словах, исключенных из запроса, результат применения стемминга к словам запроса, количество совпадений для каждого термина или фразы, а также информация о том, какие последовательности слов запроса обрабатывались как фразы. Информационно-поисковая система может также предложить поисковые термины с помощью тезауруса или контролируемого словаря. Пользователю можно предоставить возможность просмотра словаря инвертированного индекса и таким образом помочь найти хорошие термины, встречающиеся в коллекции.

9.2.2. Расширение запроса

В методе обратной связи по релевантности пользователи вводят дополнительную информацию о документах (помечая релевантные и нерелевантные документы в списке ре-

зультатов), и эта информация используется для изменения весов терминов в запросе. С другой стороны, при *расширении запроса* (query expansion) пользователи вводят дополнительную информацию о словах запроса или фразах, возможно, предлагая дополнительные термины запроса. Некоторые поисковые системы (особенно в вебе) предлагают в ответ на запрос ряд связанных запросов; пользователи могут выбрать один из этих альтернативных запросов. На рис. 9.6 приведен пример вариантов запроса, предлагаемых поисковой веб-системой Yahoo!. Основной вопрос, связанный с этой формой расширения запроса, — как сгенерировать альтернативные или расширенные запросы для пользователя. Чаще всего для расширения запроса применяется метод глобального анализа, использующий определенные форму тезауруса. Для каждого термина t запрос можно автоматически расширить с помощью синонимов или близких слов из тезауруса. Использование тезауруса можно сочетать с идеями взвешивания термина; например, можно приписать добавленным терминам веса, которые меньше весов исходных терминов запроса.

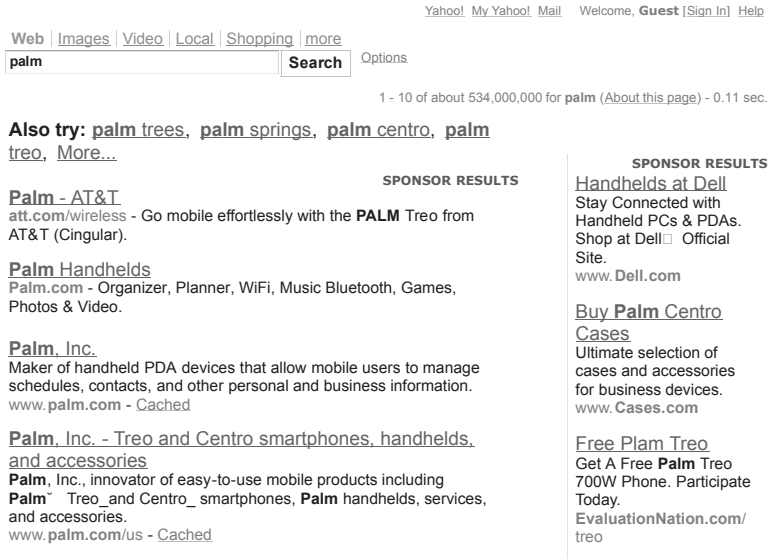

Рис. 9.6. Пример расширения запроса в интерфейсе поисковой веб-системы Yahoo! в 2008 году. Варианты расширенных запросов появляются непосредственно после панели "Search Results" (результаты поиска)

Существуют различные методы построения тезауруса, предназначенного для расширения запросов.

- *Использование контролируемого словаря, поддерживаемого редакторами*. Для каждого понятия в этом словаре есть канонический термин. Примерами контролируемых словарей являются тематические рубрики в каталогах традиционных библиотек, такие как тематические рубрики Библиотеки конгресса США (Library of Congress Subject Headings) или десятичная система классификации Дьюи. Использование контролируемых словарей характерно для предметных областей с большим количеством источников информации. Ярким примером является Unified Medical Language System (UMLS), которая используется службой Medline для поиска статей по биологии и медицине. Например, на рис. 9.7

к поиску слова `cancer` (рак) добавляется слово `neoplasms` (новообразования). Это расширение запроса к системе Medline контрастирует с примером расширения запроса к поисковой системе Yahoo! Интерфейс Yahoo! представляет собой пример интерактивного расширения запроса, а система PubMed расширяет запрос автоматически. Если пользователь не захочет проверить посланный запрос, он может даже не узнать, что произошло расширение запроса.

- User query: `cancer`
- PubMed query: ("neoplasms"[TIAB] NOT Medline[SB]) OR "neoplasms"[MeSH Terms] OR cancer[Text Word]
- User query: `skin itch`
- PubMed query: ("skin"[MeSH Terms] OR "integumentary system"[TIAB] NOT Medline[SB]) OR "integumentary system"[MeSH Terms] OR skin[Text Word] AND (("pruritus"[TIAB] NOT Medline[SB]) OR "pruritus"[MeSH] OR itch[TextWord])

Рис. 9.7. Примеры расширения запроса с помощью тезауруса PubMed. Запрос пользователя, заданный через интерфейс PubMed системе Medline по адресу `www.ncbi.nim.nih.gov/entrez/`, отображается в словарь Medline, как показано на рисунке

- *Тезаурус, создаваемый вручную*. Здесь редакторы создают множества синонимов для понятий без назначения канонического термина. Одним из таких тезаурусов является метатезаурус UMLS. Система Statistics Canada поддерживает тезаурус предпочтительных терминов, синонимов, а также более широких и узких терминов по отраслям, по которым правительство собирает статистические данные, например товары и услуги. К тому же этот тезаурус является двуязычным (на английском и французском языках).
- Автоматически *создаваемый тезаурус*. Для автоматического создания тезауруса используются статистические данные о совместной встречаемости слов в документах предметной области (см. раздел 9.2.3).[5]
- *Переформулирование запроса на основе анализа лога запросов*. Здесь используется предыдущее переформулирование запросов, сделанных вручную пользователями, чтобы предложить их в качестве подсказки новому пользователю. Для этого требуется огромное количество запросов, поэтому метод подходит для веб-поиска.

Расширение запроса с помощью тезауруса имеет одно преимущество: оно не требует дополнительного ввода информации от пользователя. Использование расширения запроса обычно увеличивает полноту поиска. Этот метод широко применяется во многих научных и технических областях. Кроме глобального анализа, для расширения запроса можно применять методы локального анализа, например, анализируя документы в списке результатов. В этом случае пользователь должен ввести дополнительную информацию, причем сохраняется различие: обратная связь может относиться к терминам запроса или к документам.

[5] В последнее время в контексте веб-поиска чаще используется совместная встречаемость не столько в одном документе, сколько в множестве ссылок на один документ или в множестве запросов пользователя внутри одной поисковой сессии. — *Примеч. ред.*

9.2.3. Автоматическая генерация тезауруса

В качестве альтернативы затратному ручному созданию тезауруса можно попытаться сгенерировать тезаурус автоматически путем анализа коллекции документов. Существует два основных метода. Один из них просто использует совместную встречаемость слов. Считается, что слова, встречающиеся в одном документе или абзаце, являются близкими по смыслу или связанными между собой, поэтому для поиска таких слов используется простой подсчет текстовых статистических показателей. Другой подход основан на использовании поверхностного грамматического анализа текста, а также грамматических отношений или зависимостей. Например, если нечто выращивается, готовится, съедается и переваривается, то, скорее всего, речь идет о еде. Простое использование смежных слов более надежно (на него не влияют ошибки грамматического анализа), но поиск с помощью грамматических отношений является более точным.

Для того чтобы составить тезаурус на основе совместной встречаемости, проще всего опираться на попарное сходство терминов. Начнем с матрицы "термин–документ" A, в которой каждая ячейка $A_{t,d}$ содержит взвешенное количество вхождений $w_{t,d}$ термина t в документ d, в котором вес выбирается так, чтобы строки матрицы A были нормализованы по длине. Затем вычисляется матрица $C = AA^T$, где $C_{u,v}$ — мера сходства между терминами u и v (чем больше эта мера, тем лучше). На рис. 9.8 приведен пример тезауруса, созданного с помощью описанного выше метода за исключением того, что для снижения размерности в нем предусмотрен еще один этап — латентное семантическое индексирование (latent semantic indexing), которое будет рассмотрено в главе 18. Одни термины тезауруса хороши или по крайней мере могут служить подсказкой, другие — несущественны и плохи. Качество ассоциаций между словами обычно представляет собой проблему. Неоднозначность терминов легко создает нерелевантные статистически коррелированные термины. Например, запрос `Apple computer` может быть расширен до `Apple red fruit computer`[6]. Недостатки таких тезаурусов — как ложноположительные, так

word	nearest neighbors
absolutely	absurd, whatsoever, totally, exactly, nothing
bottomed	dip, copper, drops, topped, slide, trimmed
captivating	shimmer, stunningly, superbly, plucky, witty
doghouse	dog, porch, crawling, beside, downstairs
makeup	repellent, lotion, glossy, sunscreen, skin, gel
mediating	reconciliation, negotiate, case, conciliation
keeping	hoping, bring, wiping, could, some, would
lithographs	drawings, Picasso, Dali, sculptures, Gauguin
pathogens	toxins, bacteria, organisms, bacterial, parasite
senses	grasp, psyche, truly, clumsy, naive, innate

Рис. 9.8. Пример автоматически сгенерированного тезауруса. Этот пример основан на работе Шютце (Schütze, 1998), использующей латентное семантическое индексирование (см. главу 18)

[6] Из этого примера следует, что расширение слов запроса через тезаурус является контекстнозависимым, т.е. расширение слова, пригодное для одного запроса, непригодно для другого. — *Примеч. ред.*

и ложноотрицательные связи. Более того, поскольку термины, представленные в автоматическом тезаурусе, и так сильно коррелированны в документе (причем часто для создания тезауруса и индексирования используется одна и та же коллекция), эта форма расширения запроса не позволяет найти много новых документов.

Расширение запроса часто эффективно повышает полноту. Однако ручное создание тезауруса и его дальнейшее обновление с учетом научного развития области и изменения терминологии связано с большими затратами. В принципе, тезаурус предметной области необходим: универсальные тезаурусы и словари плохо покрывают богатую и специфичную терминологию научных дисциплин. В то же время расширение запроса может сильно снизить точность, особенно если запрос содержит неоднозначные термины. Например, если пользователь формулирует запрос `interest rate`, расширение запроса до `interest rate fascinate evaluate` вряд ли целесообразно. Итак, расширение запроса менее полезно, чем применение метода обратной связи по релевантности, хотя оно может оказаться так же эффективно, как метод обратной связи по псевдорелевантности. Преимущество расширения запроса в том, что оно в большей степени понятно пользователю.

? ***Упражнение 9.7.*** Допустим, что матрица A является двоичной матрицей встречаемости термина в документе. Какой смысл имеют элементы матрицы C?

9.3. Библиография и рекомендации для дальнейшего чтения

Информационно-поисковые системы рано столкнулись с проблемой вариантных выражений, когда слова запроса не содержатся в документе, несмотря на то что сам документ является релевантным этому запросу. Ранний эксперимент, проведенный примерно в 1960-м году, на который ссылается Свансон (Swanson, 1988), выяснил, что только одиннадцать из двадцати трех документов, индексированных по теме *toxicity*, содержали слово с корнем *toxi*. Кроме того, существует проблема перевода, когда пользователи не знают, какие термины использует документ. Блэр и Марон (Blair and Maron, 1985) пришли к выводу, что “пользователю слишком сложно предсказать точные слова, словосочетания и фразы, которые будут использованы во всех (или в большинстве) релевантных документах и только (или в основном) в этих документах”.

Метод обратной связи по релевантности с помощью модели векторного пространства впервые описан в статье Солтона (Salton, 1971*b*), алгоритм Роккио — в его работе (Rocchio, 1971), а вариант Ide dec–hi вместе с оценкой нескольких вариантов — в статье Иде (Ide, 1971). Другой вариант обратной связи сводится к рассмотрению *всех* документов в коллекции как нерелевантных, кроме тех, которые были оценены как релевантные, вместо документов, явно оцененных как нерелевантные. Однако Шютце и др. (Schütze et al., 1995) и Сингал и др. (Singhal et al., 1997) показали, что лучшие результаты можно получить, используя только документы, близкие к запросу, а не все документы. Более поздние исследования описаны в статьях Солтона и Бакли (Salton and Buckley, 1990), Ризлера и др. (Riezler et al., 2007), посвященных статистическому NLP-подходу к методу RF, а также в недавнем обзоре Рутвена и Лалмас (Ruthven and Lalmas, 2003).

Качество интерактивных систем RF обсуждается в работах Солтона, Харман, Бакли и др. (Salton, 1989; Harman, 1992; Buckley et al., 1994*b*). Эксперименты с участием поль-

зователей по исследованию эффективности обратной связи по релевантности описаны в работе Конеманн и Белкина (Koenemann and Belkin, 1996).

Традиционно наиболее известным тезаурусом английского языка считается *тезаурус Роже* (Roget's thesausus), описанный в работе (Roget, 1946). В настоящее время исследователи практически всегда используют тезаурус WordNet не только потому, что он свободно распространяется, но и благодаря его богатой структуре связей. Этот тезаурус доступен по адресу `http://wordnet.princeton.edu`.

Автоматическая генерация тезаурусов обсуждалась в работах Кью и Фрая (Qui and Frei, 1993) и Шютце (Schütze, 1998). Использование локальных и глобальных методов расширения запросов исследовано в работе Ксу и Крофта (Сюй and Croft, 1996).