

Теория информационного поиска

Лекция 2. Индекс. Булев поиск

Дмитрий Грановский

СПбГУ

17.03.2025

В предыдущей серии

- 1 Рассматриваем поиск по текстовой *коллекции*.
- 2 Коллекция состоит из *документов*.
- 3 Документ состоит из *терм(ин)ов*.
- 4 Пользователь имеет *информационную потребность* и задает *запросы*.
- 5 То, насколько документ подходит к запросу, — *релевантность*.

Как же сделать поиск?

Самый простой вариант (gгер):

- получаем запрос,
- просматриваем всю коллекцию,
- выбираем подходящие документы

Как же сделать поиск?

Самый простой вариант (gгер):

- получаем запрос,
- просматриваем всю коллекцию,
- выбираем подходящие документы
- Чем это плохо?
- Чем это хорошо?
- Что делать, если в запросе >1 термина?

608 Subject Index

- bootstrap test, **70**
- bootstrapping, **70**
 - in IE, 339
- bound pronoun, **418**
- boundary tones, **534**
- BPE, **18**
- BPE, 20
- bracketed notation, **234**
- bridging inference, **420**
- broadcast news
 - speech recognition of, 571
- Brown corpus, **11**
 - original tagging of, 170
- byte-pair encoding, **18**

- CALLHOME, **550**
- Candide*, 228
- canonical form, **307**
- Cantonese, 207
- capture group, **10**
- cardinal number, **239**
 - origin of term, 148
- closed class, **149**
- closed vocabulary, **40**
- closure, stop, **529**
- cluster, **416**
- clustering
 - in word sense
 - disambiguation, 372
- CNF, *see* Chomsky normal form
- coarse senses, **372**
- cochlea, **542**
- Cocke-Kasami-Younger algorithm, *see* CKY
- coda, syllable, **531**
- code switching, **13**
- coherence, **442**
 - entity-based, **451**
 - relations, **444**
- cohesion
 - lexical, 443, 456
- cold languages, 208
- content planning, **515**
- context embedding, **117**
- context-free grammar, 231, **232, 236, 255**
 - Chomsky normal form, 249
- invention of, 257
- non-terminal symbol, 233
- productions, 233
- rules, 233
- terminal symbol, 233
- weak and strong
 - equivalence, 249
- continuation rise, **534**
- conversation, **492**
- conversation analysis, **523**
- conversational agents, **492**
- conversational analysis, **495**
- conversational implicature, 496
- conversational speech, **550**
- counts
 - treating low as zero, 165
- CRF, **162**
 - compared to HMM, 162
 - inference, 166
 - Viterbi inference, 166
- CRFs
 - learning, 167
- cross-brackets, 270
- cross-entropy, **51**
- cross-entropy loss, **82, 137**
- cross-validation, **68**
 - 10-fold, **68**
- crowdsourcing, **398**
- CTC, **557**
- currying, **315**
- cycles in a wave, 534
- cycles per second, 534

- datasheet, **14**
- date
 - fully qualified, **347**

Индекс

словарь	словопозиции
чебоксары	7, 290
чебурашка	62, 290, 304
чебурек	14, 17, 155
чемодан	2, 9, 14, 75, 76, 84, 90, 102, ...

Индекс

словарь	словопозиции
чебоксары	7, 290
чебурашка	62, 290, 304
чебурек	14, 17, 155
чемодан	2, 9, 14, 75, 76, 84, 90, 102, ...

- список словопозиций тж. наз. *постинг* или кишка
- словарь отсортирован (зачем?)
- всё вместе — **обратный** или **инвертированный индекс** (или просто индекс)

Минутка этимологии

Yandex = Yet Another Indexer

Булев поиск

- (boolean retrieval)
- если термин встретился в документе, то документ релевантен
- иначе нерелевантен (бинарная релевантность)
- что делать, если в запросе >1 термина?
- в этой модели запросы должны выглядеть иначе:
 - [компьютерный AND лингвистика]
 - [лингвистика OR языкознание]
 - [прикладной AND (лингвистика OR языкознание)]
 - [морфология AND NOT клетка]

Булев поиск (алгоритм)

- 1 Разбираем запрос на термины и операторы
- 2 Определяем порядок выполнения запроса
- 3 Достаем из индекса постинг для каждого термина
- 4 Пересекаем/объединяем постинги
 - если они отсортированы, пересечение можно сделать за один проход: $O(N + M)$

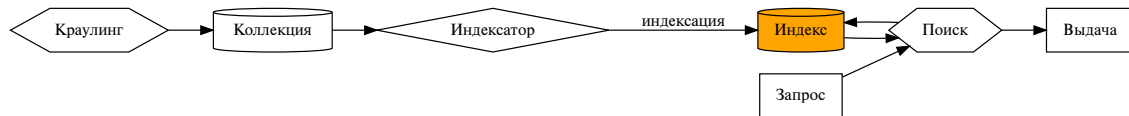
Пересечение списков

1 → 3 → 4 → 7

1 → 2 → 3 → 5

1 → 3

Общая схема



Оптимизация

[ремонт AND газового AND водонагревателя]

Оптимизация

[ремонт AND газового AND водонагревателя]

- предположим, термины встречаются с такой частотой:
 - ремонт: 15000 документов
 - газовый: 2500 документов
 - водонагреватель: 20 документов
- сколько в худшем случае операций требует пересечение?
 - $(15000 + 2500) + \min(15000, 2500) + 20 = 20020$
- можно ли улучшить?
 - [водонагревателя AND газового AND ремонт]
 - $(2500 + 20) + \min(2500, 20) + 15000 = 17540$

Проблемы 1

[ремонт электронных часов casio]

Проблемы 1

[ремонт электронных часов casio]

Получаются одинаково нерелевантны документы:

- “ремонт электронных часов”
- “ремонт наручных часов casio”
- “ремонт мебели”
- ...

Проблемы 1

[ремонт электронных часов casio]

Получаются одинаково нерелевантны документы:

- “ремонт электронных часов”
- “ремонт наручных часов casio”
- “ремонт мебели”
- ...

Обычно хотим не так радикально.

Проблемы 2

[московский кремль]

Проблемы 2

[московский кремль]

Получаются одинаково релевантны документы:

- “**московский кремль** построен в конце XV века”
- “экспозиция «псковский **кремль**» по адресу: **московский** проспект...”

Проблемы 2

[московский кремль]

Получаются одинаково релевантны документы:

- “**московский кремль** построен в конце XV века”
- “экспозиция «псковский **кремль**» по адресу: **московский** проспект...”

Это проблема модели мешка слов (*bag-of-words*).

Проблемы 3

Некоторые термы слишком часто встречаются и:

- занимают много места в индексе
- тратим на них много времени при пересечении/объединении
- как правило, выполняют грамматическую функцию
- не добавляют смысла ни в документ, ни в запрос

Проблемы 3

Простое решение:

- объявляем такие термы **СТОП-словами**
- выбрасываем из индекса
- выбрасываем из запросов

Проблемы 3

Простое решение:

- объявляем такие термины **СТОП-словами**
- выбрасываем из индекса
- выбрасываем из запросов

Но теперь есть другие проблемы:

- поиск по точной цитате: [песня я и ты], [быть или не быть]
- неоднозначность этих форм:
 - [история 19 в]
 - [по для обработки видео]

Проблемы

- если ищем через AND — слишком мало результатов
 - высокая точность, низкая полнота
- если ищем через OR — слишком много результатов
 - низкая точность, высокая полнота
- не учитывается частота термина
- хотим более тонкую настройку — **ранжирование**
- в текущем виде это скорее *фильтрация*

Проблемы 4

Требуется квалифицированный пользователь:

- тщательно формулирует поисковый запрос
- способен задавать запросы вида (A AND B) OR C
- готов просмотреть все найденные документы
- хорошо представляет себе коллекцию

Резюме

- ❶ Не можем просматривать всю коллекцию на каждый запрос.
- ❷ *Инвертированный индекс* задает соответствие термин \rightarrow документ.
- ❸ Такой индекс требует запросов специального вида (термины + *операторы*).
- ❹ Всё вместе это *булев поиск*, имеет целый ряд проблем.