МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ

Урок №5

Словари и нечёткий поиск

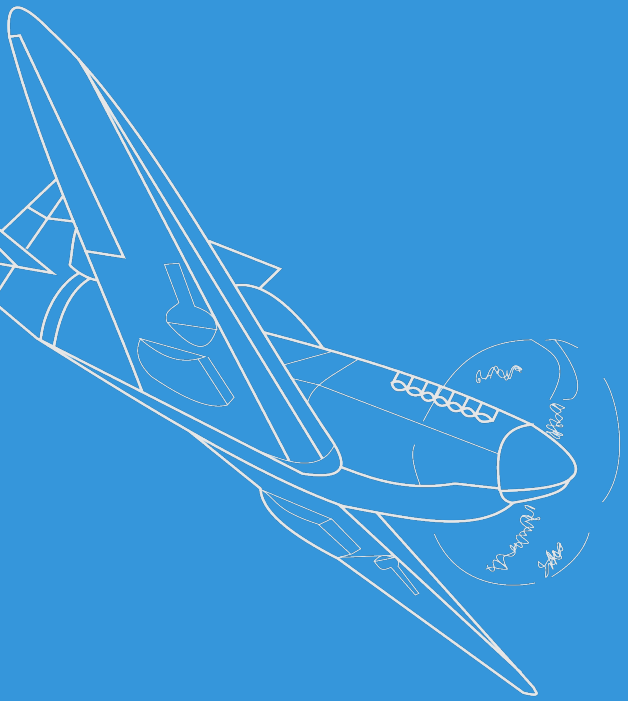
(основано на слайдах Андрея Калинина, Hinrich Schütze, Christina Lioma)

18 марта 2021 года

Антон Кухтичев

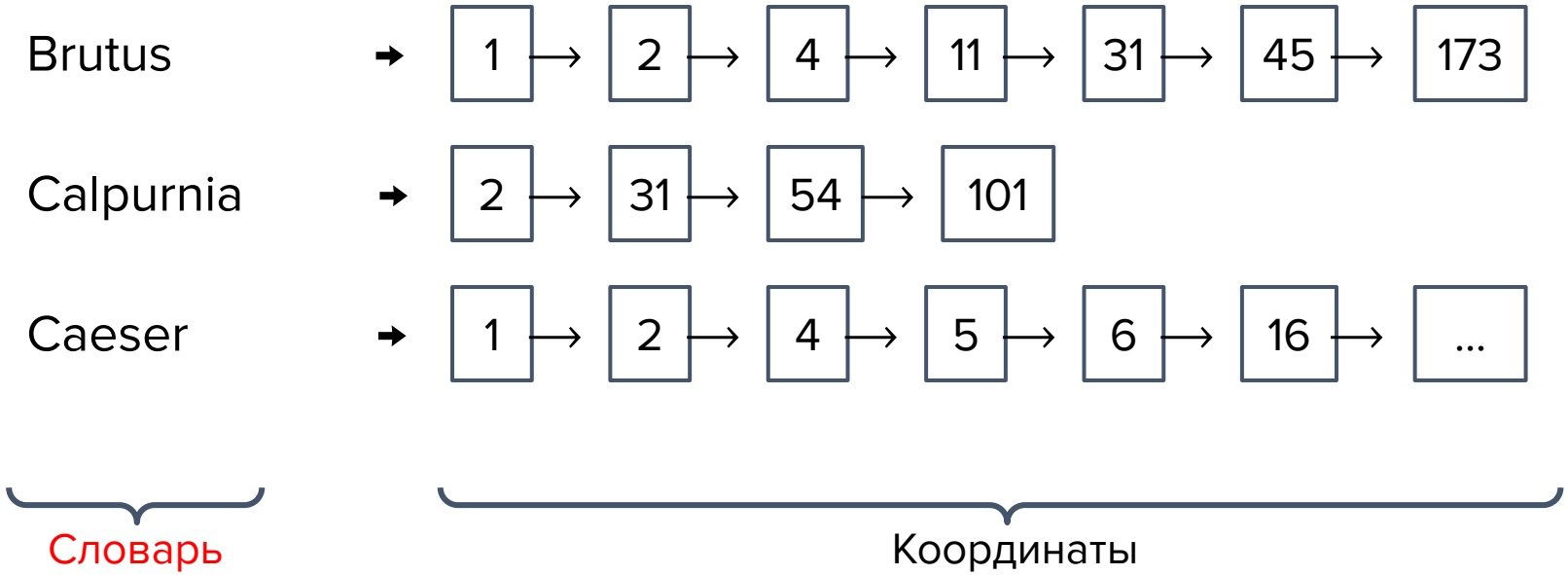
Содержание занятия

1. Словарь
2. Запросы с мета-символами
3. Проверка правописания
4. Soundex
5. Исправление запросов



Словарь

Обратный индекс



Словарь как массив

- Для каждого термина нужно сохранить:
  - количество документов (частотность)
  - указатель на координаты
  - ...
- На время допустим, что можно представить эту информацию в виде структуры фиксированной длины.
- Тогда можно использовать массив для хранения словаря.

Словарь как массив

	Термин	Частотность	Координатный блок
	a	656256	→
	aachen	65	→
	...	...	...
	zulu	221	→
объём:	20 байт	4 байта	4 байта

Как искать термин запроса q_i в этом массиве? То есть: какую структуру данных можно использовать, чтобы найти строку, в котрой находится q_i?

Структуры данных поиска терминов

- Два основных класса: хеши и деревья.
- Некоторые ИСП используют хеши, некоторые — деревья.
- Основные вопросы выбора:
  - Количество терминов фиксировано, или растёт?
  - Какие относительные частоты доступа к разным ключам?
  - Сколько разных ключей имеется?

Хеши

- Каждый термин хешируется в целое число.
- Боремся с коллизиями.
- Во время запроса: хешируем термин запроса, разрешаем коллиции, находим нужную строку в массиве.
- Плюсы:
  - Поиск в хеш-таблице быстрее, чем поиск в дереве.
  - Время поиска — константа.

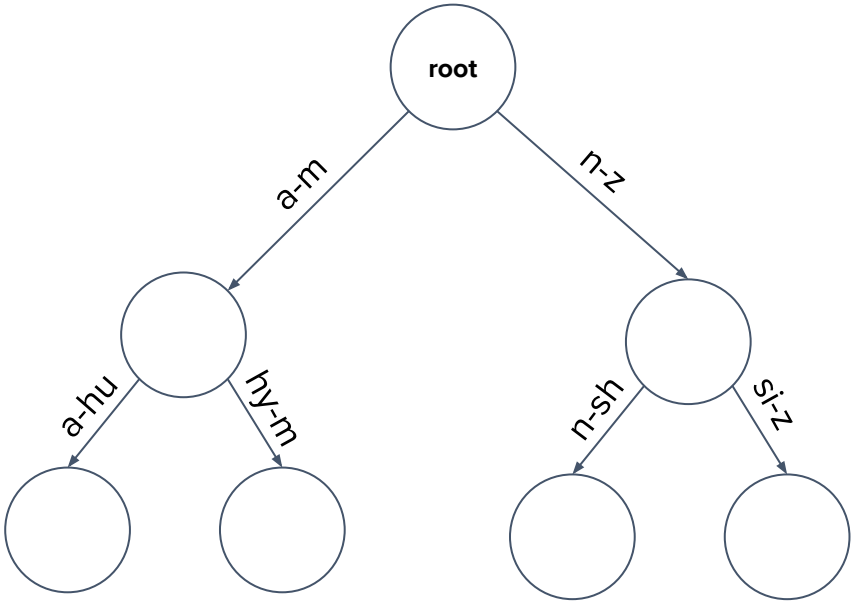
Хеши

- Минусы
  - Нельзя найти небольшие различия (resume и résumé)
  - Нельзя искать по префиксу (все термины, начинающиеся с
  - automat)
  - Для растущего словаря придётся время от времени всё рехешировать.
- Теоретически, можно сделать «морфологическую» хеш-функцию.

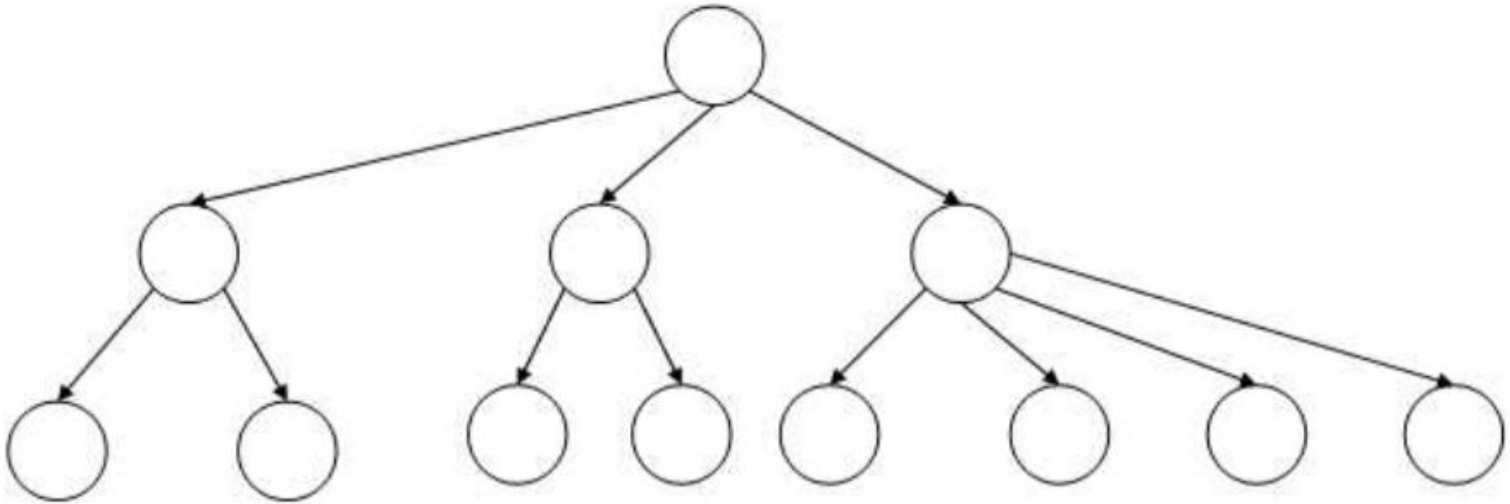
Деревья

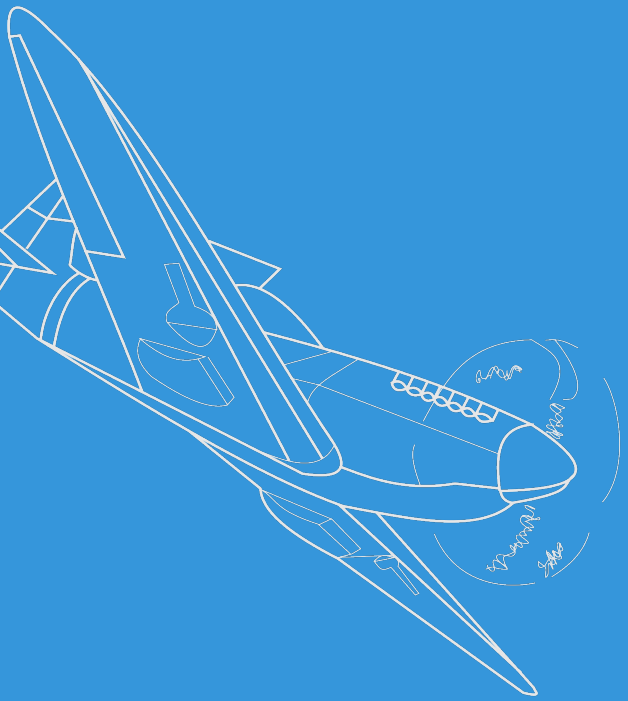
- Деревья позволяют искать термины с общим префиксом.
- Простейшее дерево — бинарное.
- Поиск медленее хешей, *O(log M)*, где M — размер словаря.
- *O(log M)* соблюдается для сбалансированных деревьев.
- Так же можно использовать B-деревья.

Бинарное дерево



B-дерево





Запросы с мета-символами

Запросы с мета-символами

- mon*: найти все документы, содержащие термин, начинающийся с mon
- Просто для B-дерева: найти все термины t, находящиеся в диапазоне mon ≤ t < moo
- *mon: найти все термины, заканчивающиеся на mon
  - Создаём дополнительное дерево, для терминов, записанных задом наперёд.
  - Теперь по этому дереву получаем термины t в диапазоне nom ≤ t < non
- Результат: множество терминов, подходящих под маску.
- Теперь нужно найти документы, содержащие любой из этих терминов.

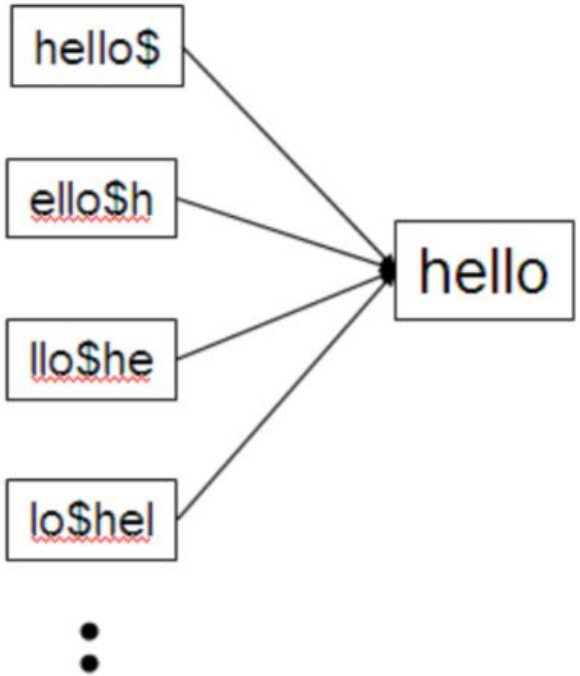
Как обработать * внутри термина

- Например: m*nchen
- Можно поискать m* и *nchen в B-деревьях и пересечь два полученных множества.
- Довольно расточительно.
- Альтернатива: индекс перестановок
- Основная идея: «переворачивать» каждый запрос с маской таким образом, чтобы * оказалась в конце.
- Хранить каждый переворот каждого термина в словаре, в том же B-дереве.

Индекс перестановок

- Для термина hello: добавим *hello$, ello$h, llohe, lohel, и o$hell* в B-дерево, где $ — специальный символ.

Отображение перестановок в термины



Индекс перестановок

- Итак, для hello храним: hello$, ello$h, llohe, lohel и o$hell
- Тогда запросы
  - X, ищем X$
  - X*, ищем X*$
  - *X, ищем X$*
  - *X*, ищем X*
  - X*Y, ищем Y$X*
  - Например: для hel*o ищем o$hel*
  - Как обработать запрос X*Y*Z?

Поиск в индексе перестановок

- Прокрутить запрос так, чтобы * была справа.
- Искать как обычно.
- Однако: такой индекс как минимум учетверяет размер словаря (для английского языка, для русского — увеличит в 7-8 раз).

Индексы k-грамм

- Занимает меньше места, чем индекс перестановок.
- Индексируем все символьные k-граммы (последовательности из k символов) термина.
- 2-граммы часто называют биграммами.
- Например: из April is the cruelest month получим следующие биграммы: $a ap pr ri il l$ $i is s$ $t th he e$ $c cr ru ue el le es st t$ $m mo on nt h$
- $ — специальный символ, обозначающий границу слова.
- Добавляем в новый индекс не термины, а биграммы.

3-граммный обратный индекс

k-граммные индексы

- Теперь у нас два разных вида обратных индексов.
- Есть индекс терминов-документов.
- И есть индекс k-грамм, чтобы находить термины по запросам, состоящим из k-грамм.

Выполнение запроса с метасимволами для биграмм

- Запрос mon* можно обработать так:

  $m and mo and on
- Так получим все термины с префиксом mon…
- … но и много «ложных срабатываний», таких как moon.
- Их нужно отфильтровать, напрямую сравнивая термины с запросом.
- Оставшиеся термины нужно искать в индексе терминов-документов.
- k-граммный индекс и индекс перестановок
  - k-граммный индекс занимает меньше места.
  - Индекс перестановок не требует пост-фильтрации.

Упражнение

Почему у больших веб-поисков нет поддержки запросов с масками?

Упражнение

Почему у больших веб-поисков нет поддержки запросов с масками?

- Много слов.
- Увеличивается количество обрабатываемых терминов.
- Люди будут вводить меньше символов в словах.

Рекомендуемая литература

[Введение в информационный поиск | Маннинг Кристофер Д., Шютце Хайнрих]

Для саморазвития (опционально)

[Чтобы не набирать двумя пальчиками]

Спасибо за внимание!

Антон Кухтичев

a.kukhtichev@mail.ru

@toshunster