МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ

Урок №4

Прыжки по индексу

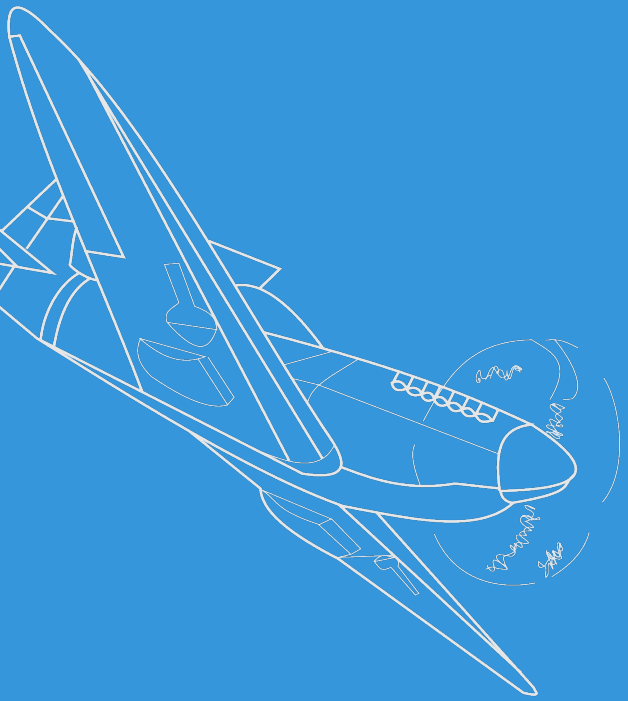
(основано на слайдах Андрея Калинина, Hinrich Schütze, Christina Lioma)

18 марта 2021 года

Антон Кухтичев

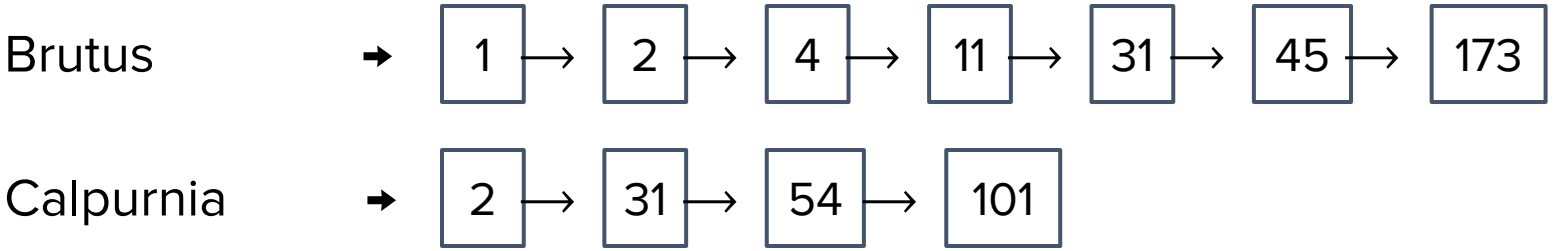
Содержание занятия

1. Алгоритм пересечения
2. Цитатные запросы



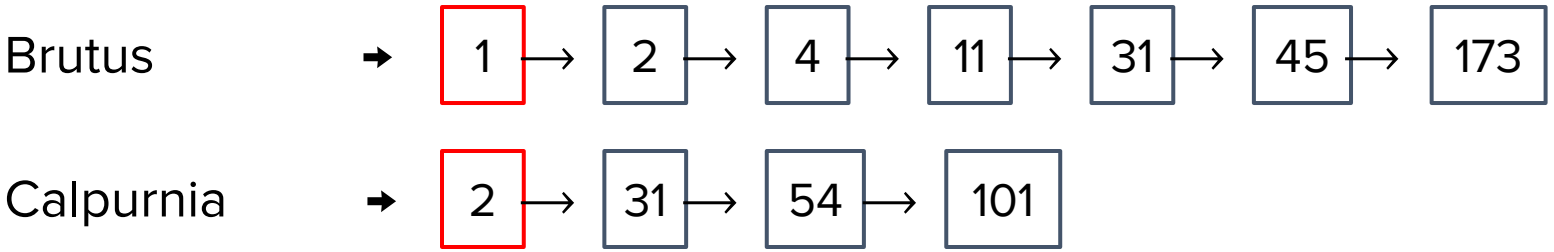
Алгоритм пересечения

Алгоритм пересечения



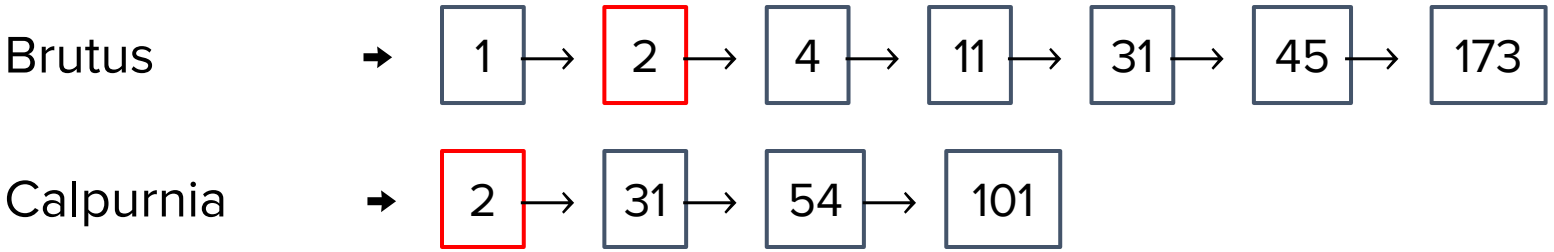
Пересечение ⇒

Алгоритм пересечения



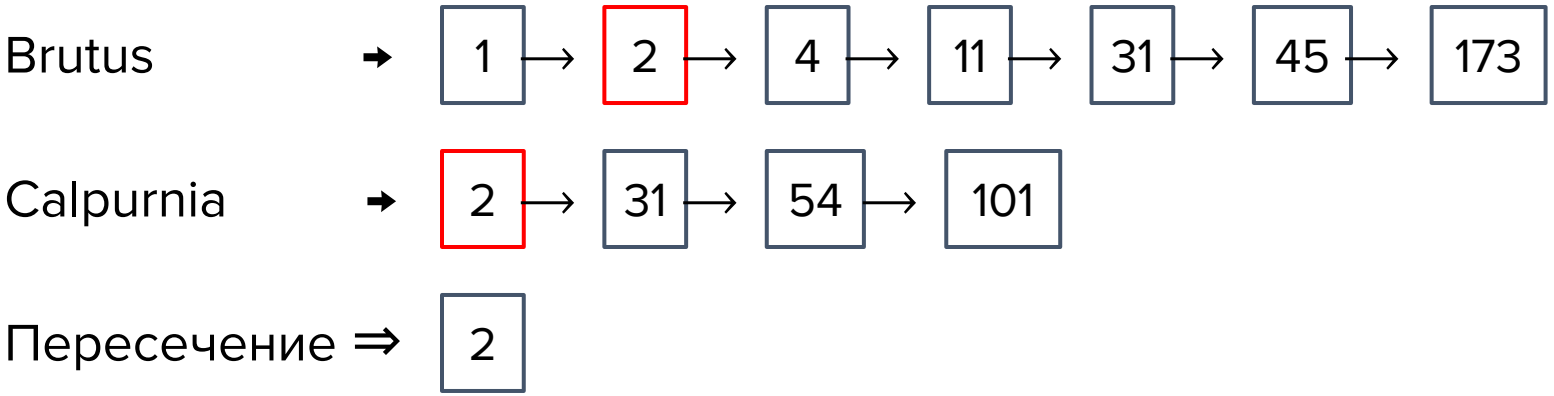
Пересечение ⇒

Алгоритм пересечения

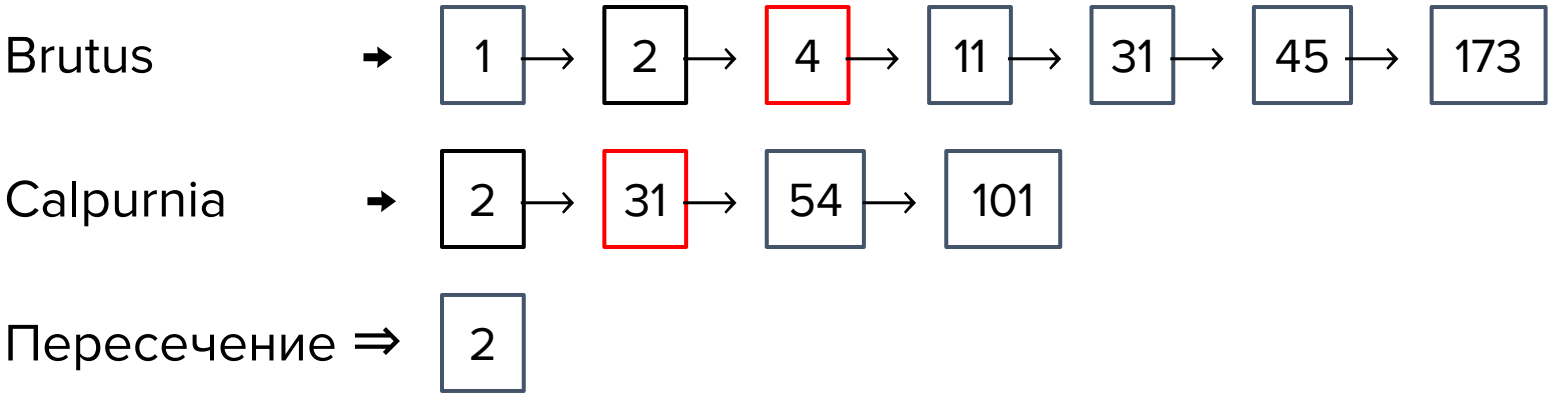


Пересечение ⇒

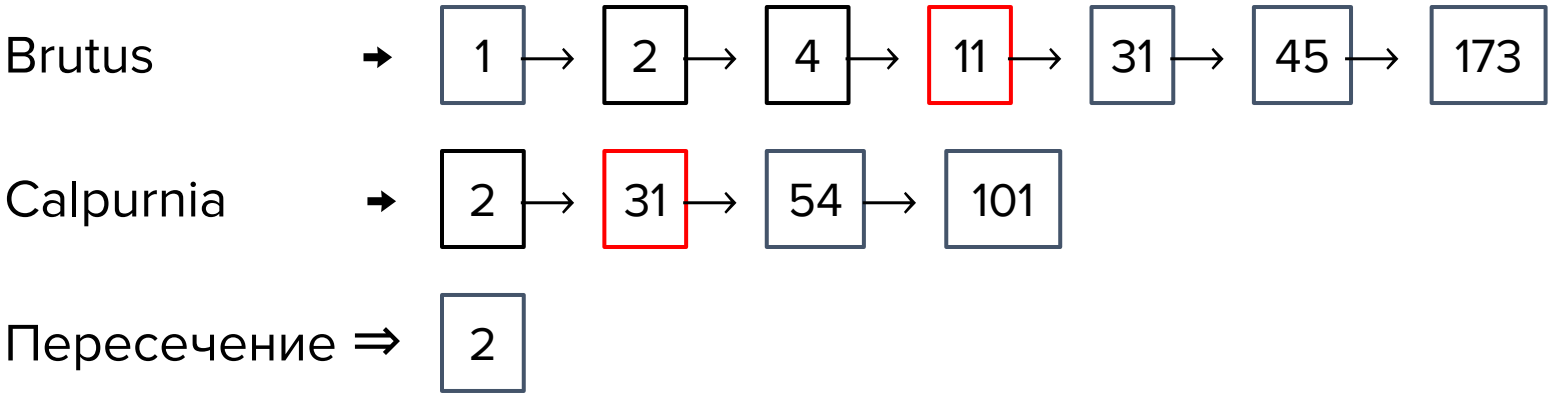
Алгоритм пересечения



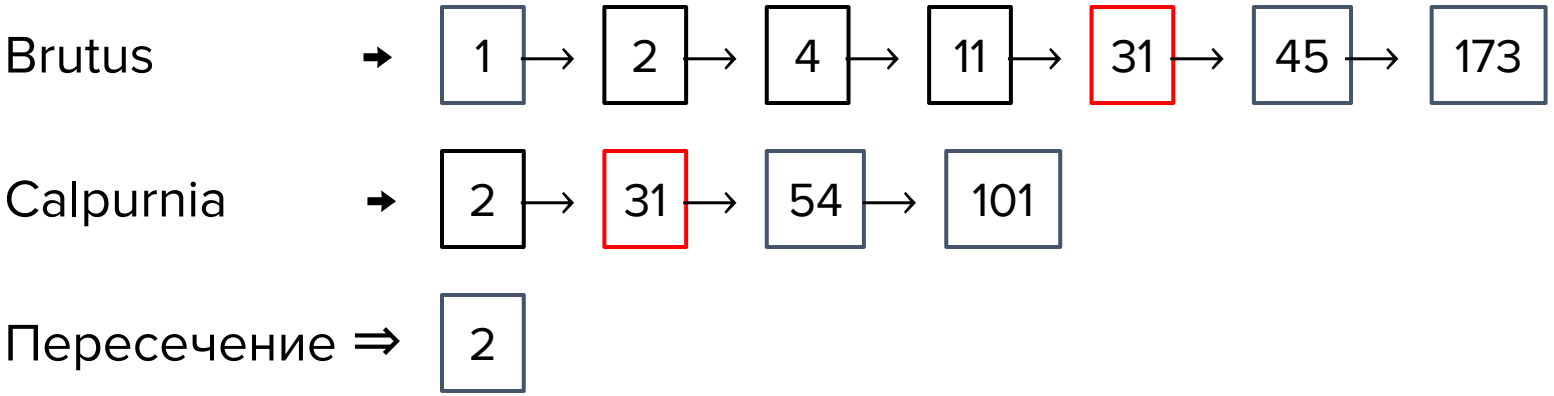
Алгоритм пересечения



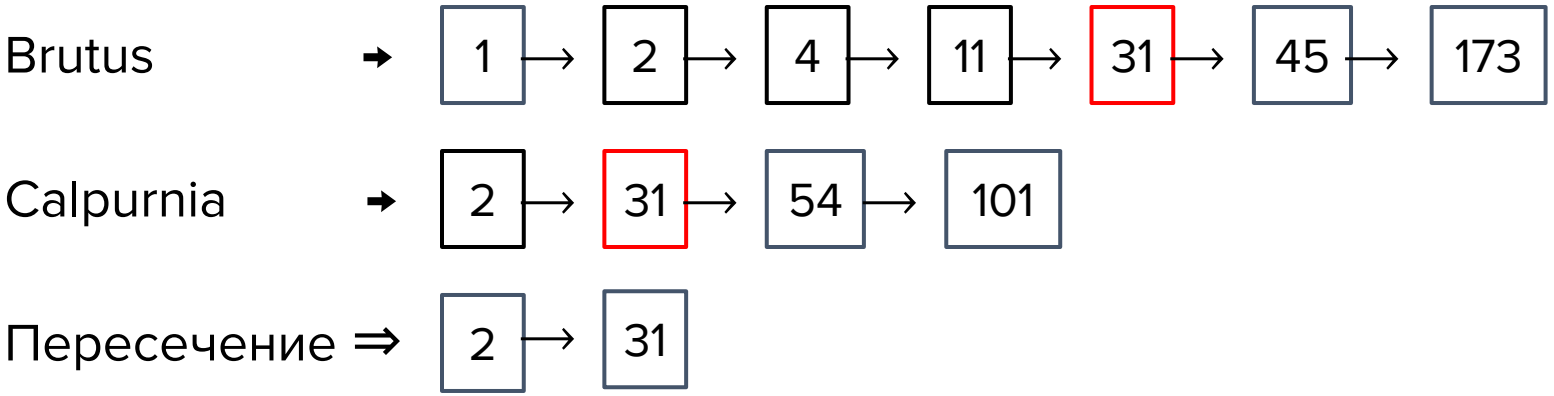
Алгоритм пересечения



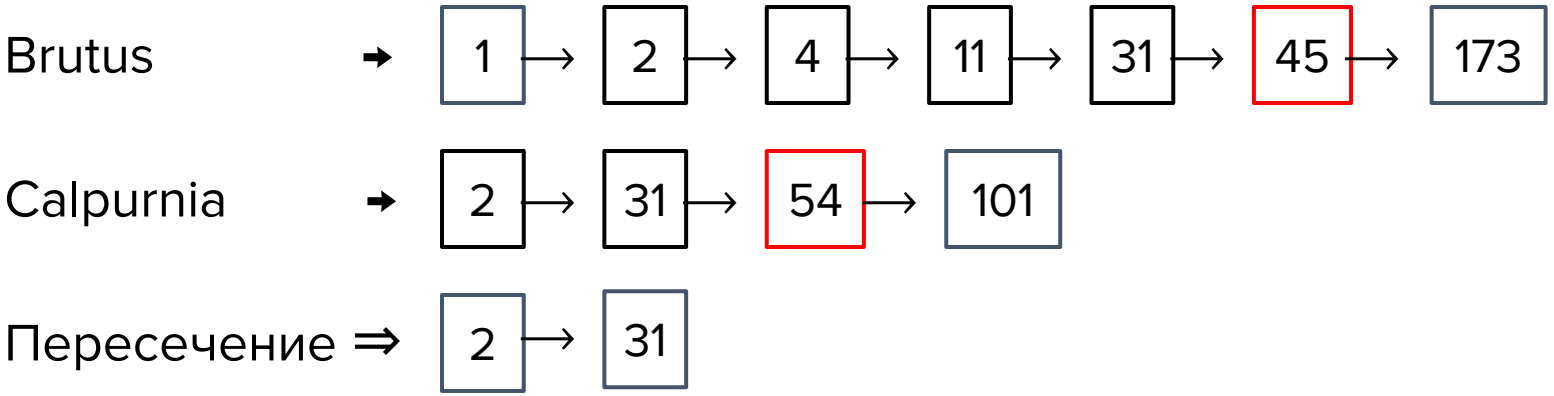
Алгоритм пересечения



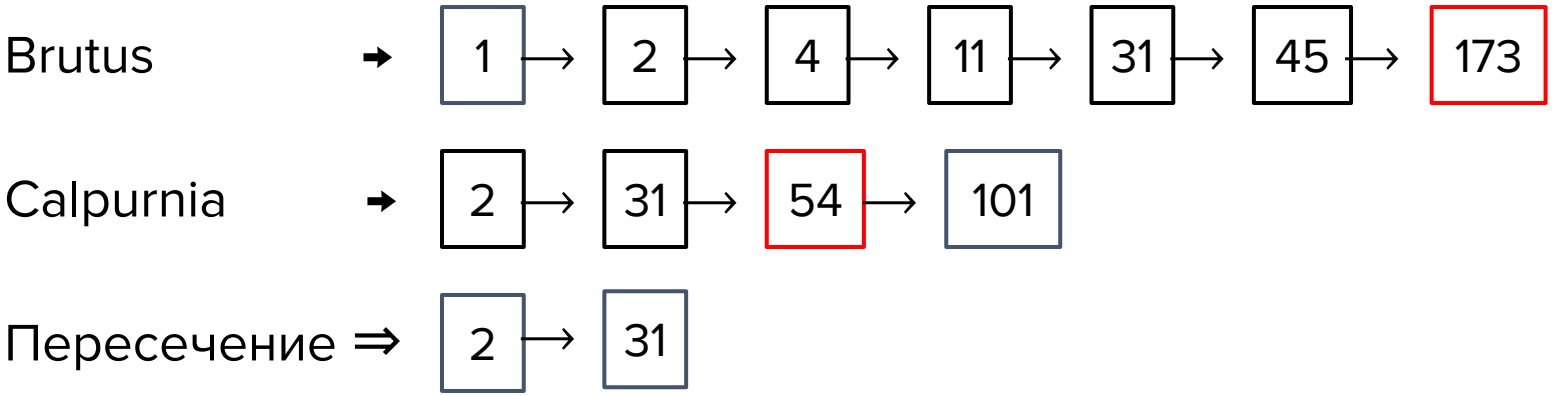
Алгоритм пересечения



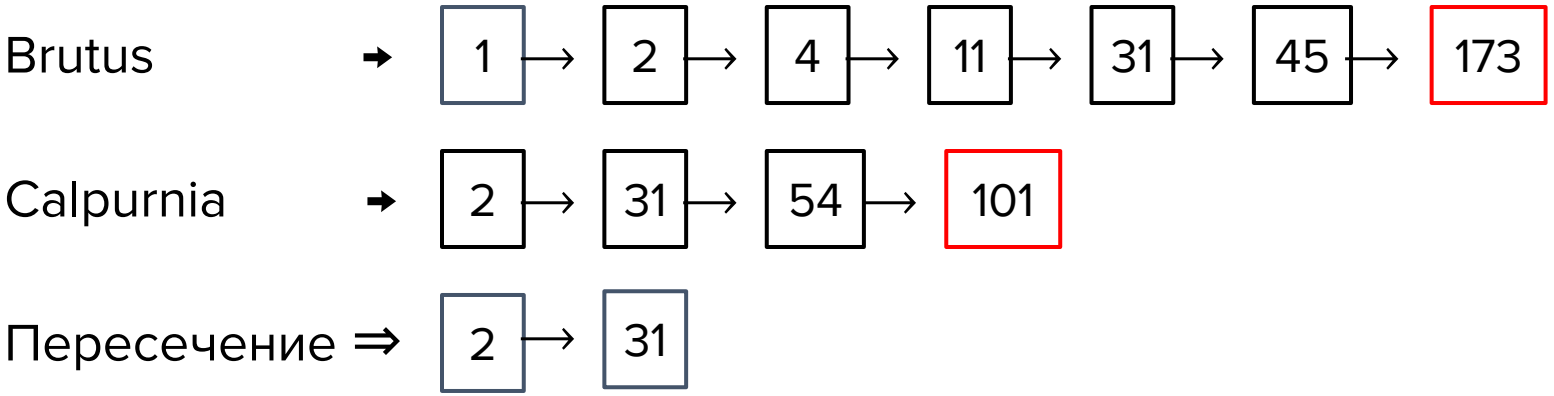
Алгоритм пересечения



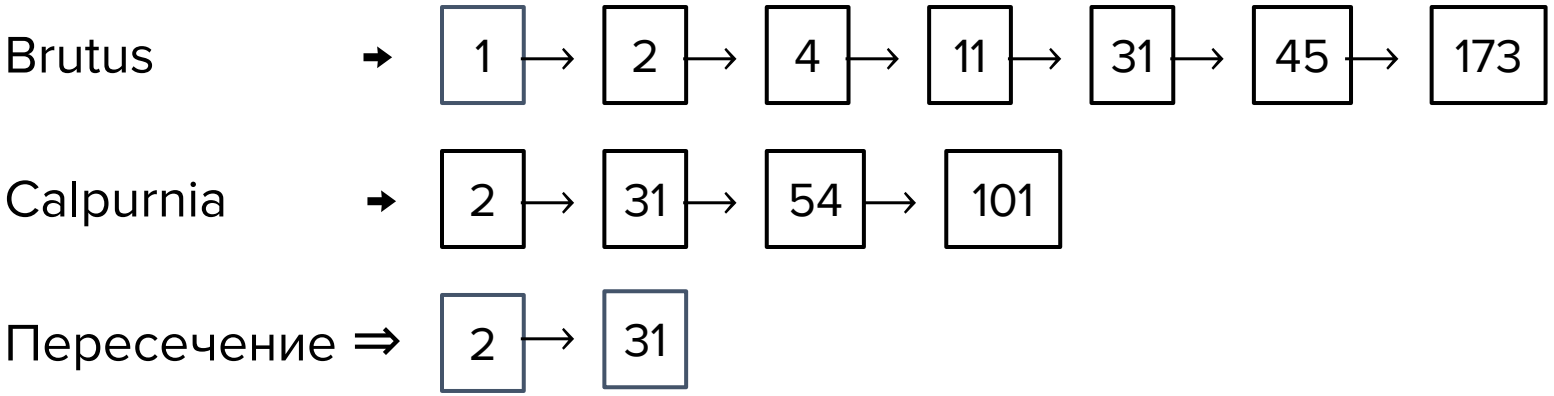
Алгоритм пересечения



Алгоритм пересечения



Алгоритм пересечения

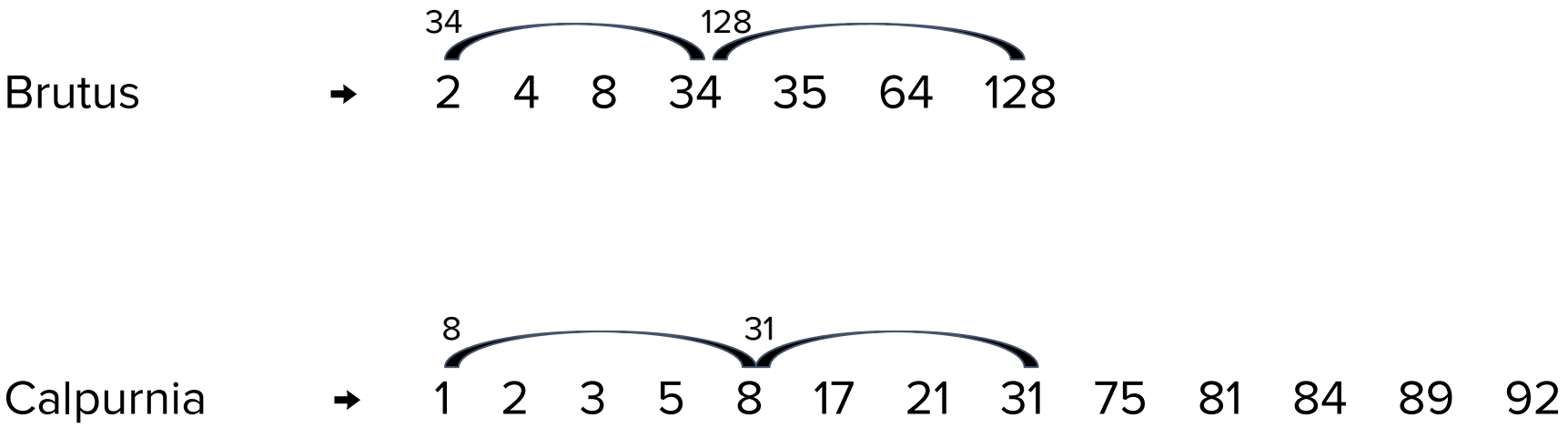


- Линеен по длине списков координат
- Можно ли улучшить?

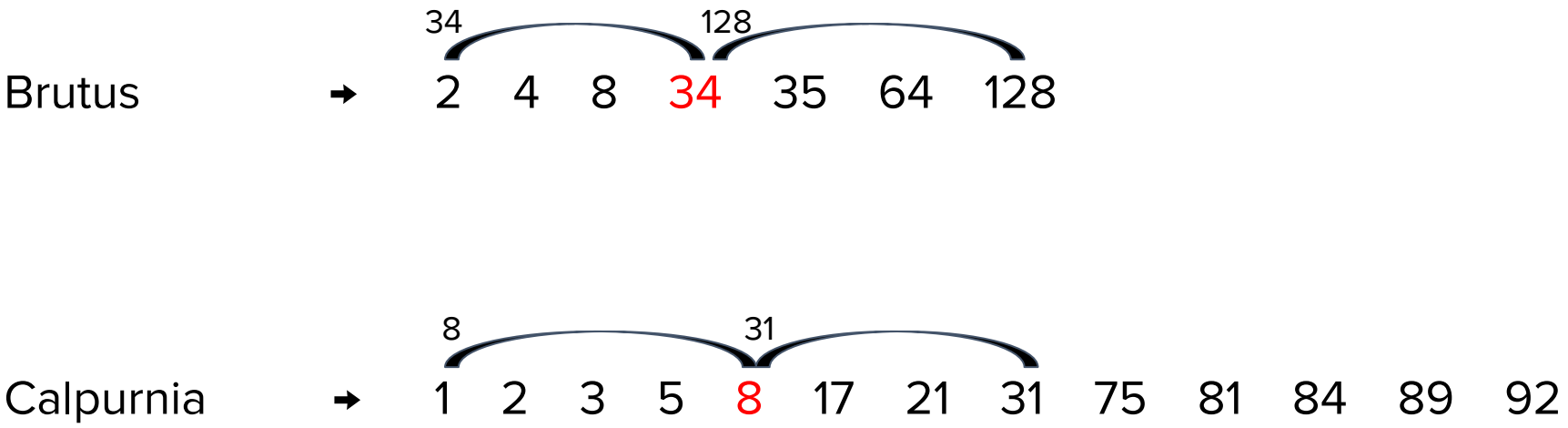
Пропуск координат

- Специальные указатели внутри индекса позволяют пропускать те постинги, которые не появятся в результатах поиска.
- Тем самым пересечение списков становится более эффективным.
- Некоторые списки могут содержать в себе миллионы вхождений, так что вопрос производительности встаёт несмотря на линейность алгоритма.
- Куда поместить эти указатели?
- И как добиться, что результаты пересечения будут правильными?

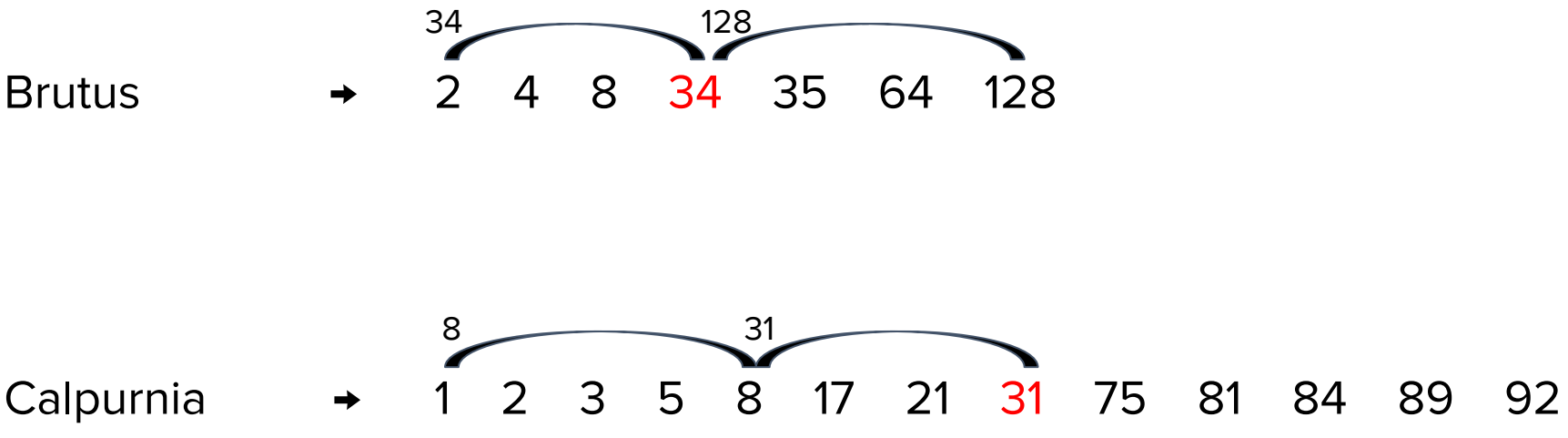
Основная идея



Основная идея



Основная идея



Алгоритм пересечения с прыжками по индексу

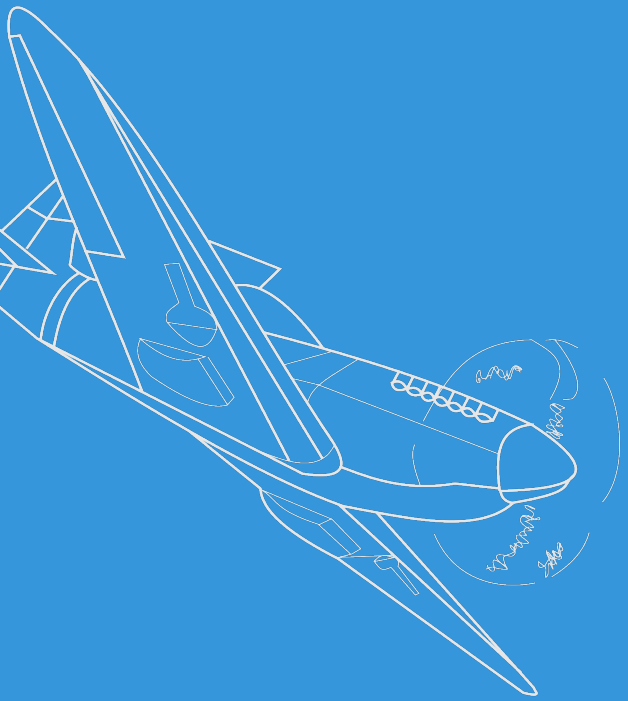

```
IntersectWithSkips(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then Add(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then if hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
 9                        then while hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
10                                do p1 ← skip(p1)
11                        else p1 ← next(p1)
12                else if hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
13                        then while hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
14                                do p2 ← skip(p2)
15                        else p2 ← next(p2)
16  return answer
```

Где размещать указатели?

- Компромисс между тем, сколько пропускать и как часто.
- Больше пропусков: каждый указатель пропускает несколько элементов, но можно чаще использоват.
- Меньше пропусков: каждый пропуск переносит дальше, но использовать можно реже.

Где размещать указатели? (продолжение)

- Простая эвристика: при длине списка P использовать $\sqrt{P}$ равномерно распределённых указателей.
- Но это правило игнорирует распределение терминов запроса.
- Проще делать, если индекс статический. Для динамических индексов — сложнее из-за обновлоений.
- Насколько указатели помогают?
- Раньше помогали сильно.
- Сейчас их использование уже не настолько значимо (скорость процессоров)



Цитатные запросы

Цитатные запросы

- Желательно выполнять запросы вида [“stanford university”] — поиск цитаты.
- Поэтому предложение *The inventor Stanford Ovshinsky never went to university* не должно быть в списке результатов.
- Концепция цитатных запросов легко воспринимается пользователями.
- Примерно 10-15% запросов к веб-поиску — цитатные.
- Следовательно, уже недостаточно просто хранить идентификаторы документов в постингах.
- Два способа расширения обратного индекса:
  - индекс биграмм.
  - координатный индекс.

Индекс биграмм

- Добавляем в индекс любую последовательную пару терминов.
- Например, по *Friends, Romans, Countrymen* сгенерируется две биграммы: “*friends romans*” и “*romans countrymen*”
- Каждая из биграмм добавляется в словарь.
- Теперь легко выполнять двусловные запросы.

Более длинные запросы

- Запрос вида [“*stanford university palo alto*”] может быть представлен в виде булевского запроса “*stanford university*” AND “*university palo*” AND “*palo alto*”
- Кроме того, нужно будет отфильтровать выдачу, чтобы выбрать только те документы, которые содержат всю 4-х словную цитату.

Расширенные биграммы

- Разбираем документ и выполняем разметку частей речи.
- Далее объединяем, например, существительные (N) с
- предлогами (X)
- Теперь считаем любую последовательность терминов вида NX*N расширенной биграммой.
- Примеры: [catcher in the rye], [king of Denmark]
- Включаем расширенные биграммы в словарь.
- Выполняем запросы так же.

Проблемы с биграммами

Почему биграммные индексы редко используются?

- Ложные вхождения.
- Индекс «взрывается» из-за огромного количества новых терминов.

Координатные индексы

- Координатные индексы — более эффективная альтернатива биграммам.
- Постинги в индексе без координат: каждый постинг всего лишь идентификатор документа.
- Постинги в координатном индексе: каждый постинг — идентификатор документа и список координат

Пример координатного индекса

Запрос: [$to_1\ be_2\ or_3\ not_4\ to_5\ be_6$]

Пример координатного индекса

Запрос: [$to_1\ be_2\ or_3\ not_4\ to_5\ be_6$]

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
  2: ⟨1, 17, 74, 222, 255⟩;
  4: ⟨8, 16, 190, 429, 433⟩;
  5: ⟨363, 367⟩;...⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
  4: ⟨17, 191, 291, 430, 434⟩;...⟩

Пример координатного индекса

Запрос: [$to_1\ be_2\ or_3\ not_4\ to_5\ be_6$]

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
  2: ⟨1, 17, 74, 222, 255⟩;
  4: ⟨8, 16, 190, 429, 433⟩;
  5: ⟨363, 367⟩;...⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
  4: ⟨17, 191, 291, 430, 434⟩;...⟩

Пример координатного индекса

Запрос: [$to_1\ be_2\ or_3\ not_4\ to_5\ be_6$]

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
  2: ⟨1, 17, 74, 222, 255⟩;
  4: ⟨8, 16, 190, 429, 433⟩;
  5: ⟨363, 367⟩;...⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
  4: ⟨17, 191, 291, 430, 434⟩;...⟩

Пример координатного индекса

Запрос: [$to_1\ be_2\ or_3\ not_4\ to_5\ be_6$]

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
  2: ⟨1, 17, 74, 222, 255⟩;
  4: ⟨8, 16, 190, 429, 433⟩;
  5: ⟨363, 367⟩;...⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
  4: ⟨17, 191, 291, 430, 434⟩;...⟩

Пример координатного индекса

Запрос: [$to_1\ be_2\ or_3\ not_4\ to_5\ be_6$]

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
2: ⟨1, 17, 74, 222, 255⟩;
4: ⟨8, 16, 190, 429, 433⟩;
5: ⟨363, 367⟩;...⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
4: ⟨17, 191, 291, 430, 434⟩;...⟩

Пример координатного индекса

Запрос: [$to_1\ be_2\ or_3\ not_4\ to_5\ be_6$]

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
  2: ⟨1, 17, 74, 222, 255⟩;
  4: ⟨8, 16, 190, 429, 433⟩;
  5: ⟨363, 367⟩;...⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
  4: ⟨17, 191, 291, 430, 434⟩;...⟩

Пример координатного индекса

Запрос: [$to_1\ be_2\ or_3\ not_4\ to_5\ be_6$]

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
2: ⟨1, 17, 74, 222, 255⟩;
4: ⟨8, 16, 190, 429, 433⟩;
5: ⟨363, 367⟩;...⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
4: ⟨17, 191, 291, 430, 434⟩;...⟩

Пример координатного индекса

Запрос: [$to_1\ be_2\ or_3\ not_4\ to_5\ be_6$]

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
  2: ⟨1, 17, 74, 222, 255⟩;
  4: ⟨8, 16, 190, 429, 433⟩;
  5: ⟨363, 367⟩;...⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
  4: ⟨17, 191, 291, 430, 434⟩;...⟩

Пример координатного индекса

Запрос: [$to_1\ be_2\ or_3\ not_4\ to_5\ be_6$]

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;

2: ⟨1, 17, 74, 222, 255⟩;

4: ⟨8, 16, 190, 429, 433⟩;

5: ⟨363, 367⟩;...⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;

4: ⟨17, 191, 291, 430, 434⟩;...⟩

Пример координатного индекса

Запрос: [$to_1\ be_2\ or_3\ not_4\ to_5\ be_6$]

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;

2: ⟨1, 17, 74, 222, 255⟩;

4: ⟨8, 16, 190, 429, 433⟩;

5: ⟨363, 367⟩;...⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;

4: ⟨17, 191, 291, 430, 434⟩;...⟩

Пример координатного индекса

Запрос: [$to_1\ be_2\ or_3\ not_4\ to_5\ be_6$]

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;

2: ⟨1, 17, 74, 222, 255⟩;

4: ⟨8, 16, 190, 429, 433⟩;

5: ⟨363, 367⟩;...⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;

4: ⟨17, 191, 291, 430, 434⟩;...⟩

Пример координатного индекса

Запрос: [$to_1\ be_2\ or_3\ not_4\ to_5\ be_6$]

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
  2: ⟨1, 17, 74, 222, 255⟩;
  4: ⟨8, 16, 190, 429, 433⟩;
  5: ⟨363, 367⟩;...⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
  4: ⟨17, 191, 291, 430, 434⟩;...⟩

Пример координатного индекса

Запрос: [$to_1\ be_2\ or_3\ not_4\ to_5\ be_6$]

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;

2: ⟨1, 17, 74, 222, 255⟩;

4: ⟨8, 16, 190, 429, 433⟩;

5: ⟨363, 367⟩;...⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;

4: ⟨17, 191, 291, 430, 434⟩;...⟩

Пример координатного индекса

Запрос: [$to_1\ be_2\ or_3\ not_4\ to_5\ be_6$]

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;

2: ⟨1, 17, 74, 222, 255⟩;

4: ⟨8, 16, 190, 429, 433⟩;

5: ⟨363, 367⟩;...⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;

4: ⟨17, 191, 291, 430, 434⟩;...⟩

Пример координатного индекса

Запрос: [$to_1\ be_2\ or_3\ not_4\ to_5\ be_6$]

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;

2: ⟨1, 17, 74, 222, 255⟩;

4: ⟨8, 16, 190, 429, 433⟩;

5: ⟨363, 367⟩;...⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;

4: ⟨17, 191, 291, 430, 434⟩;...⟩

Пример координатного индекса

Запрос: [$to_1\ be_2\ or_3\ not_4\ to_5\ be_6$]

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
  2: ⟨1, 17, 74, 222, 255⟩;
  4: ⟨8, 16, 190, 429, 433⟩;
  5: ⟨363, 367⟩;...⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
  4: ⟨17, 191, 291, 430, 434⟩;...⟩

Пример координатного индекса

Запрос: [$to_1\ be_2\ or_3\ not_4\ to_5\ be_6$]

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
2: ⟨1, 17, 74, 222, 255⟩;
4: ⟨8, 16, 190, 429, 433⟩;
5: ⟨363, 367⟩;...⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
4: ⟨17, 191, 291, 430, 434⟩;...⟩

В 4-м документе — вхождение!

Учёт близости

- Координатный индекс можно использовать для цитатного поиска.
- Так же его можно использовать для поиска с учётом расстояний.
- Например: [employment /4 place]
- Найти все документы, содержащие термины employment и place на расстоянии 4-х слов друг от друга.
- Employment agencies that place healthcare workers are seeing growth — есть вхождение.
- Employment agencies that have learned to adapt now place healthcare workers нет вхождения.

Учёт близости

- Используем координатный индекс.
- Простейший алгоритм: смотрим на пересечение координат (i) `employment` в документе и (ii) `place` в том же документе.
- Неэффективно для частых слов.
- Мы хотим вернуть вхождения, подходящие под запрос.
- Это важно для построения сниппетов.

Пересечение с учётом близости

```
POSITIONALINTERSECT(p1, p2, k)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then l ← ⟨ ⟩
 5             pp1 ← positions(p1)
 6             pp2 ← positions(p2)
 7             while pp1 ≠ NIL
 8             do while pp2 ≠ NIL
 9                do if |pos(pp1) − pos(pp2)| ≤ k
10                      then ADD(l, pos(pp2))
11                      else if pos(pp2) > pos(pp1)
12                              then break
13                   pp2 ← next(pp2)
14                while l ≠ ⟨ ⟩ and |l[0] − pos(pp1)| > k
15                do DELETE(l[0])
16                for each ps ∈ l
17                do ADD(answer, ⟨docID(p1), pos(pp1), ps⟩)
18                pp1 ← next(pp1)
19             p1 ← next(p1)
20             p2 ← next(p2)
21        else if docID(p1) < docID(p2)
22                then p1 ← next(p1)
23                else p2 ← next(p2)
24  return answer
```

Комбинированный подход

- Биграммы и координатные индексы могут использоваться вместе.
- Многие биграммы часто используются: Michael Jackson, Britney Spears и т.п.
- Для этих биграмм большая скорость обработки существенна по сравнению с пересечением координат.
- Комбинированный подход: включить все частые биграммы в словарь. Остальные пересекать «координатным» способом.

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