

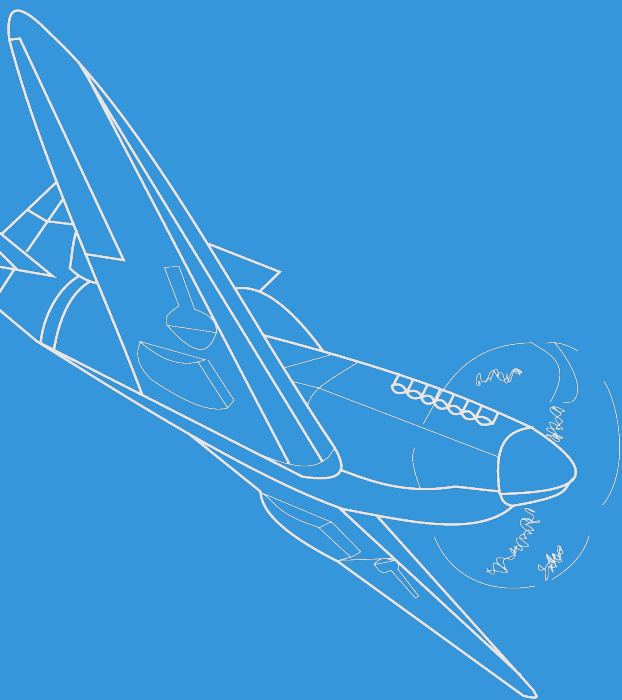
Урок №4

Прыжки по индексу

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

Содержание занятия

1. Прыжки по индексу
2. Цитатные запросы



Прыжки по индексу

Алгоритм пересечения



Brutus →

 →

 →

 →

 →

 →

 →

Calpurnia →

 →

 →

 →

Пересечение ⇒

Алгоритм пересечения



Brutus → 1 → 2 → 4 → 11 → 31 → 45 → 173

Calpurnia → 2 → 31 → 54 → 101

Пересечение ⇒

Алгоритм пересечения

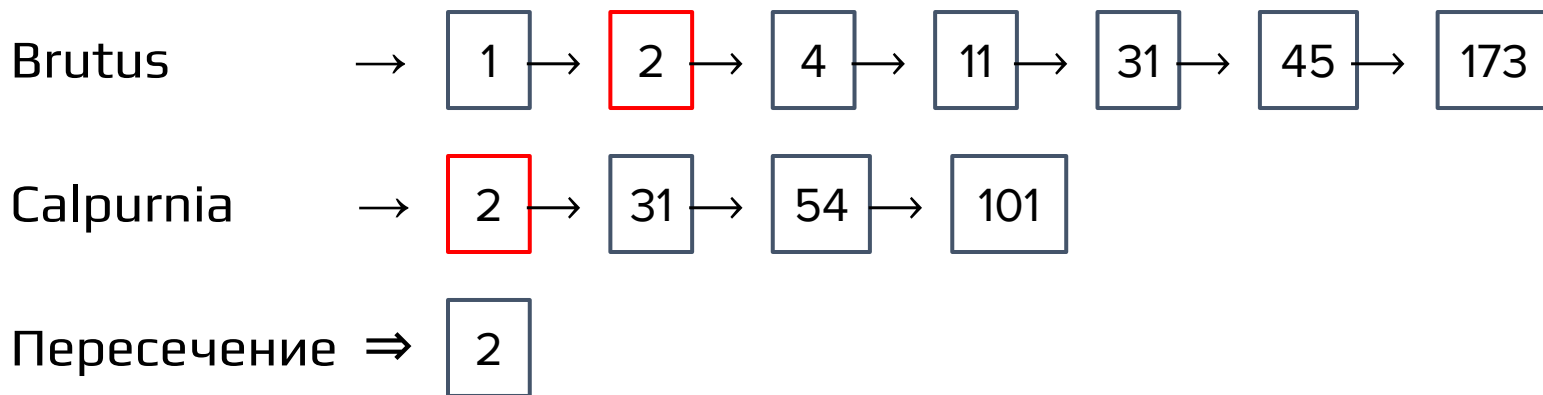


Brutus → 1 → 2 → 4 → 11 → 31 → 45 → 173

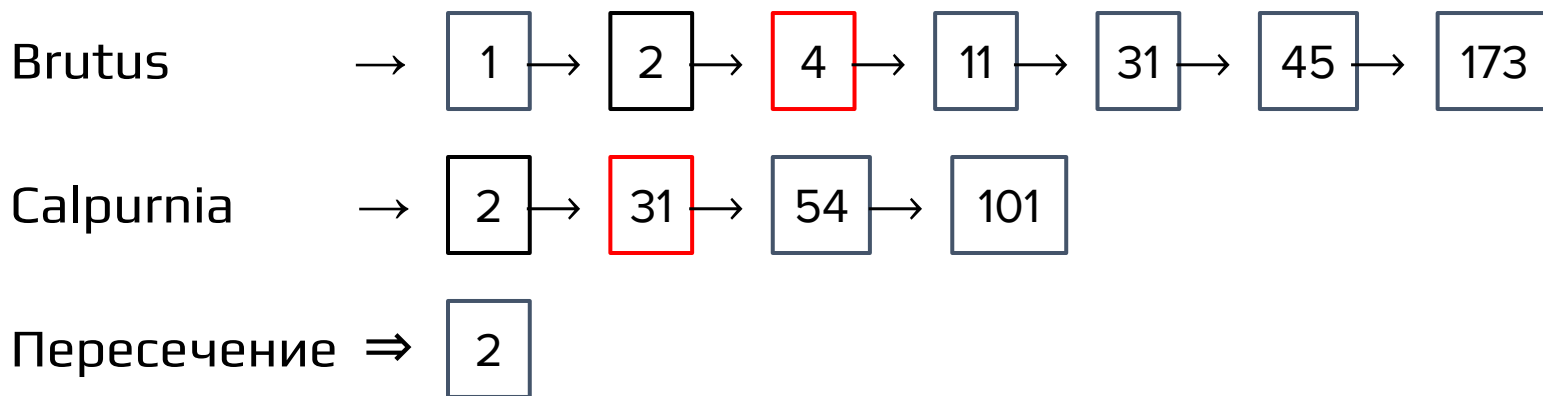
Calpurnia → 2 → 31 → 54 → 101

Пересечение ⇒

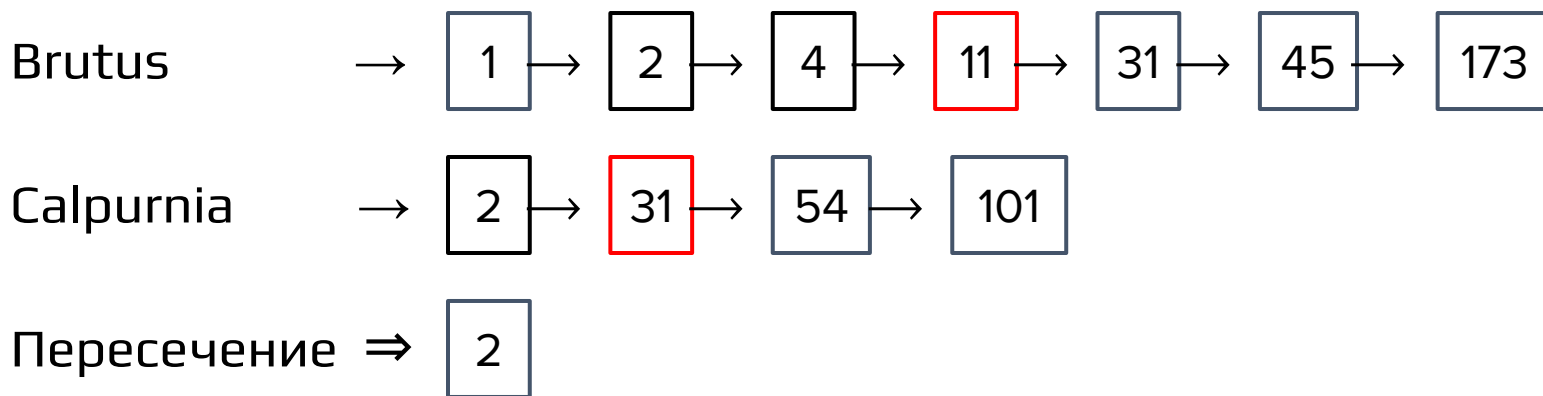
Алгоритм пересечения



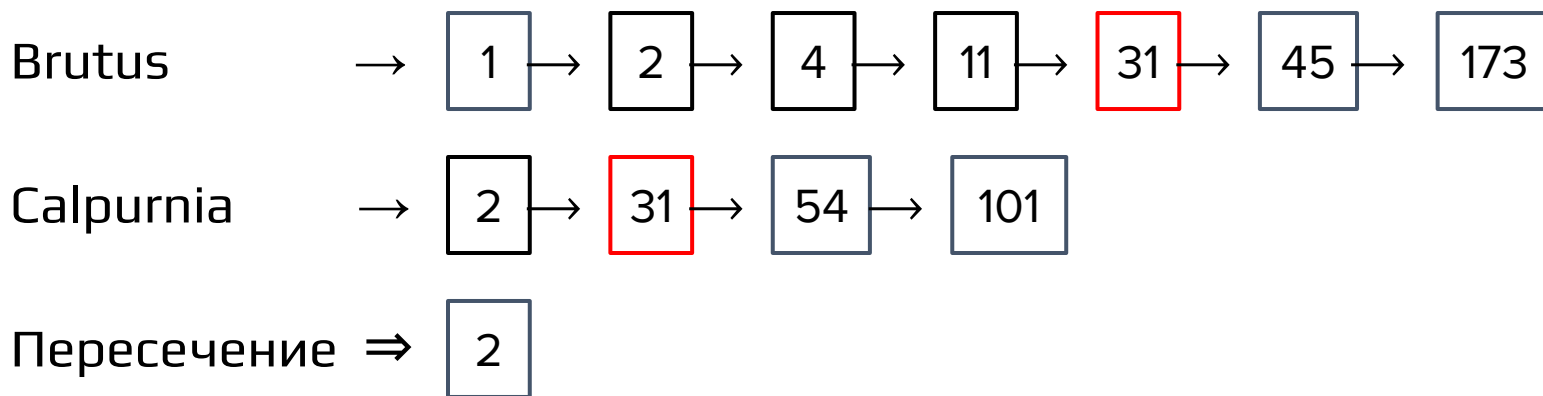
Алгоритм пересечения



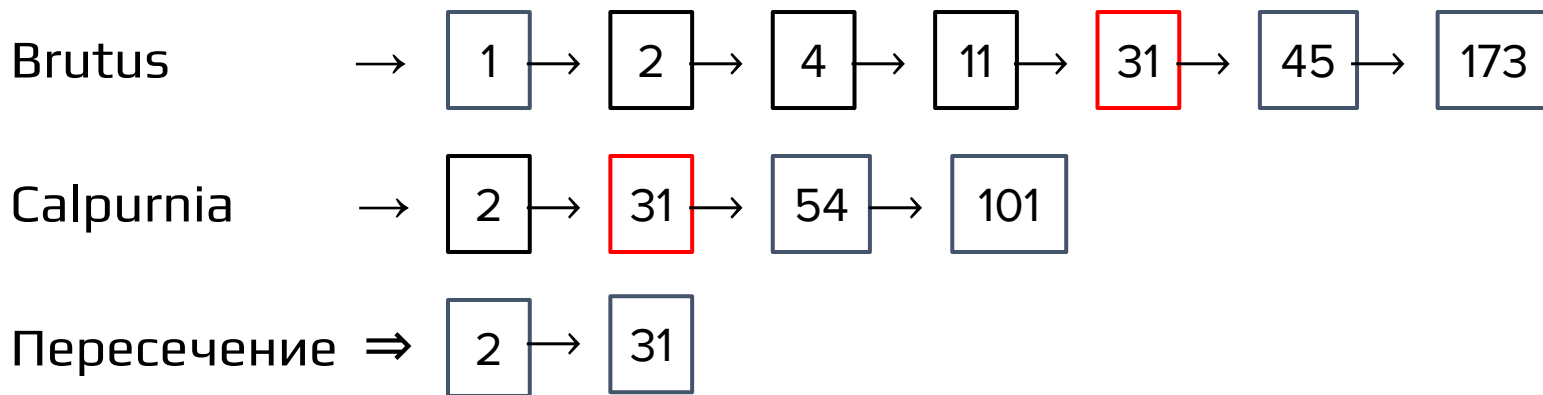
Алгоритм пересечения



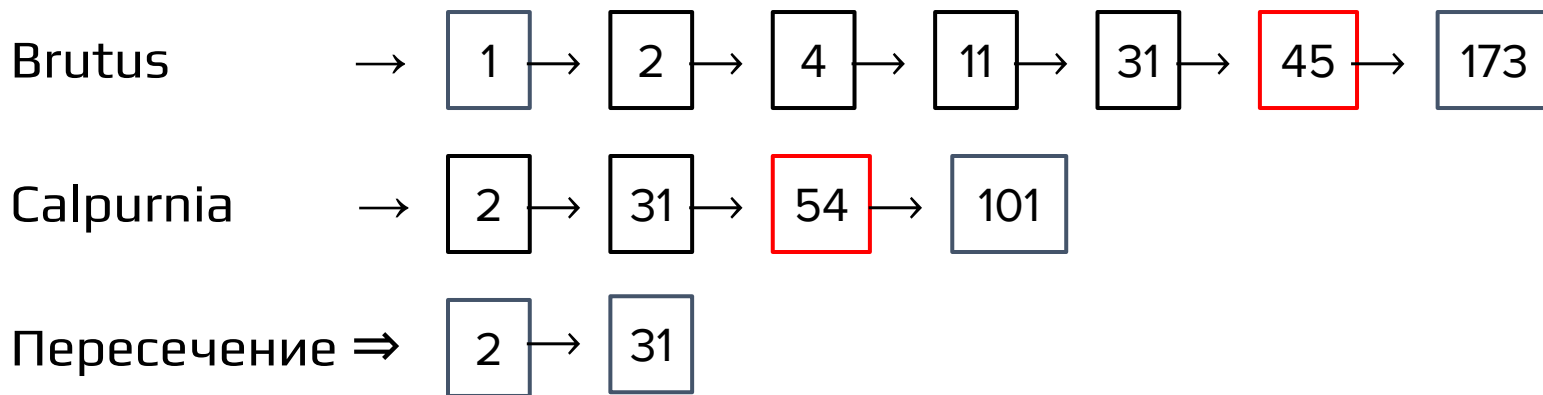
Алгоритм пересечения



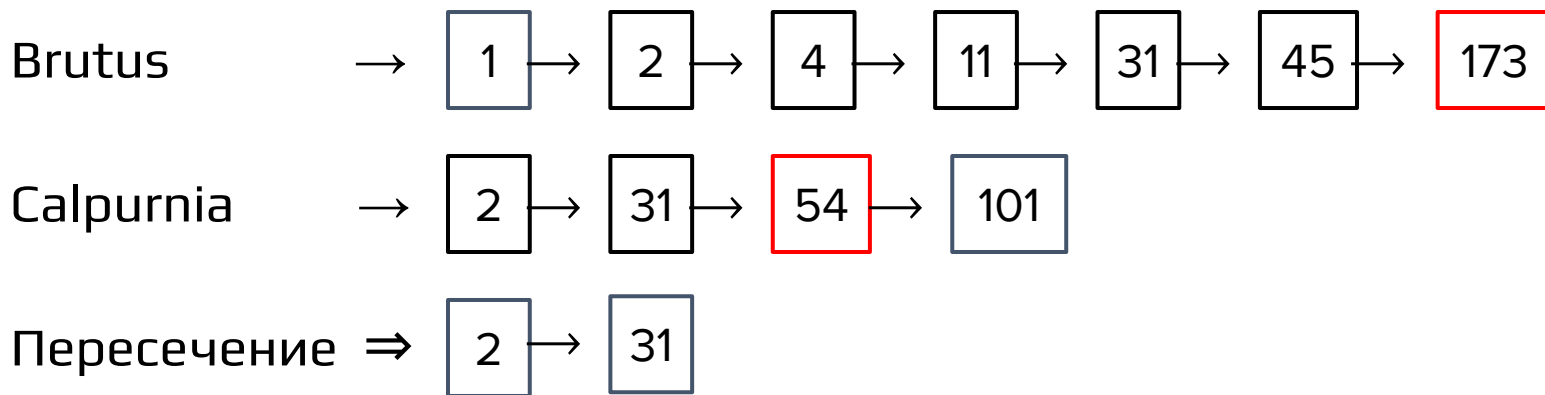
Алгоритм пересечения



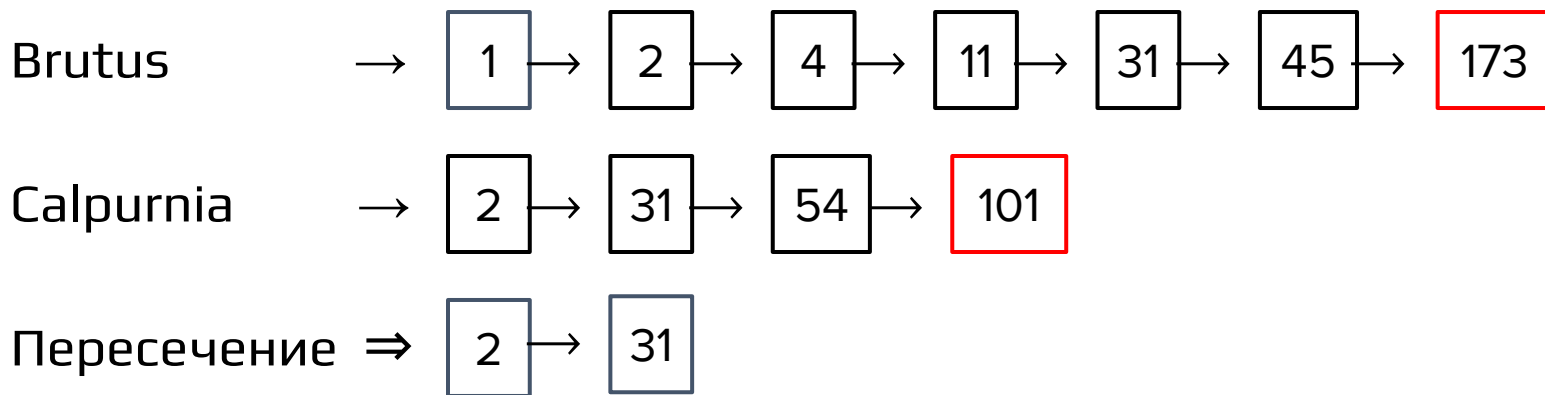
Алгоритм пересечения



Алгоритм пересечения



Алгоритм пересечения



Алгоритм пересечения



Brutus → 1 → 2 → 4 → 11 → 31 → 45 → 173

Calpurnia → 2 → 31 → 54 → 101

Пересечение ⇒ 2 → 31

- Линеен по длине списков координат
- Можно ли улучшить?



- 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



Brutus → 2 4 8 34 35 64 128

Diagram illustrating the sequence of numbers for Brutus. The sequence is 2, 4, 8, 34, 35, 64, 128. Arcs connect 2 to 34 (labeled 34) and 34 to 128 (labeled 128).

Calpurnia → 1 2 5 8 17 21 31 75 81 84 89 92

Diagram illustrating the sequence of numbers for Calpurnia. The sequence is 1, 2, 5, 8, 17, 21, 31, 75, 81, 84, 89, 92. Arcs connect 1 to 8 (labeled 8) and 8 to 31 (labeled 31).

Основная идея



Brutus → 2 4 8 **34** 35 64 128

Diagram illustrating the sequence of numbers for Brutus. The sequence is 2, 4, 8, 34, 35, 64, 128. The number 34 is highlighted in red. Arcs above the sequence indicate intervals: an arc from 2 to 8 is labeled 34, and an arc from 34 to 128 is labeled 128.

Calpurnia → 1 2 5 **8** 17 21 31 75 81 84 89 92

Diagram illustrating the sequence of numbers for Calpurnia. The sequence is 1, 2, 5, 8, 17, 21, 31, 75, 81, 84, 89, 92. The number 8 is highlighted in red. Arcs above the sequence indicate intervals: an arc from 1 to 5 is labeled 8, and an arc from 8 to 31 is labeled 31.

Основная идея



Brutus → 2 4 8 34 35 64 128

Diagram illustrating the sequence of numbers for Brutus. The sequence is 2, 4, 8, 34, 35, 64, 128. The number 34 is highlighted in red. Arcs above the sequence indicate the following values: 34 above the arc from 2 to 8, and 128 above the arc from 34 to 128.

Calpurnia → 1 2 5 8 17 21 31 75 81 84 89 92

Diagram illustrating the sequence of numbers for Calpurnia. The sequence is 1, 2, 5, 8, 17, 21, 31, 75, 81, 84, 89, 92. The number 31 is highlighted in red. Arcs above the sequence indicate the following values: 8 above the arc from 1 to 5, and 31 above the arc from 8 to 21.

Алгоритм пересечения с прыжками по индексу



INTERSECTWITHSKIPS(p_1, p_2)

```
1  answer  $\leftarrow \langle \rangle$ 
2  while  $p_1 \neq \text{NIL}$  and  $p_2 \neq \text{NIL}$ 
3  do if  $\text{docID}(p_1) = \text{docID}(p_2)$ 
4      then  $\text{ADD}(\text{answer}, \text{docID}(p_1))$ 
5           $p_1 \leftarrow \text{next}(p_1)$ 
6           $p_2 \leftarrow \text{next}(p_2)$ 
7  else if  $\text{docID}(p_1) < \text{docID}(p_2)$ 
8      then if  $\text{hasSkip}(p_1)$  and  $(\text{docID}(\text{skip}(p_1)) \leq \text{docID}(p_2))$ 
9          then while  $\text{hasSkip}(p_1)$  and  $(\text{docID}(\text{skip}(p_1)) \leq \text{docID}(p_2))$ 
10             do  $p_1 \leftarrow \text{skip}(p_1)$ 
11             else  $p_1 \leftarrow \text{next}(p_1)$ 
12      else if  $\text{hasSkip}(p_2)$  and  $(\text{docID}(\text{skip}(p_2)) \leq \text{docID}(p_1))$ 
13          then while  $\text{hasSkip}(p_2)$  and  $(\text{docID}(\text{skip}(p_2)) \leq \text{docID}(p_1))$ 
14             do  $p_2 \leftarrow \text{skip}(p_2)$ 
15             else  $p_2 \leftarrow \text{next}(p_2)$ 
16  return answer
```

Где размещать указатели?



- Компромисс между тем, сколько пропускать и как часто.
- Больше пропусков: каждый указатель пропускает несколько элементов, но можно чаще использовать.
- 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 обновлений.
- Насколько указатели помогают?
- Раньше помогали сильно.
- Сейчас их использование уже не настолько значимо (скорость процессоров)



Цитатные запросы

Цитатные запросы (1)



- 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.

Цитатные запросы (2)



- 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 биграммami



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: $[to_1 \text{ } be_2 \text{ } or_3 \text{ } not_4 \text{ } to_5 \text{ } be_6]$

T0, 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: $[to_1 \text{ } be_2 \text{ } or_3 \text{ } not_4 \text{ } to_5 \text{ } be_6]$

T0, 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: $[to_1 \text{ } be_2 \text{ } or_3 \text{ } not_4 \text{ } to_5 \text{ } be_6]$

T0, 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: $[to_1 \text{ } be_2 \text{ } or_3 \text{ } not_4 \text{ } to_5 \text{ } be_6]$

T0, 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: $[to_1 \ be_2 \ or_3 \ not_4 \ to_5 \ be_6]$

T0, 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: $[to_1 \text{ } be_2 \text{ } or_3 \text{ } not_4 \text{ } to_5 \text{ } be_6]$

T0, 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: $[to_1 \text{ } be_2 \text{ } or_3 \text{ } not_4 \text{ } to_5 \text{ } be_6]$

T0, 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: $[to_1 \text{ } be_2 \text{ } or_3 \text{ } not_4 \text{ } to_5 \text{ } be_6]$

T0, 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: $[to_1 \text{ } be_2 \text{ } or_3 \text{ } not_4 \text{ } to_5 \text{ } be_6]$

T0, 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: $[to_1 \text{ } be_2 \text{ } or_3 \text{ } not_4 \text{ } to_5 \text{ } be_6]$

T0, 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: $[to_1 \text{ } be_2 \text{ } or_3 \text{ } not_4 \text{ } to_5 \text{ } be_6]$

T0, 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: $[to_1 \text{ } be_2 \text{ } or_3 \text{ } not_4 \text{ } to_5 \text{ } be_6]$

T0, 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: $[to_1 \text{ } be_2 \text{ } or_3 \text{ } not_4 \text{ } to_5 \text{ } be_6]$

T0, 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: $[to_1 \text{ } be_2 \text{ } or_3 \text{ } not_4 \text{ } to_5 \text{ } be_6]$

T0, 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: $[to_1 \text{ } be_2 \text{ } or_3 \text{ } not_4 \text{ } to_5 \text{ } be_6]$

T0, 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: $[to_1 \text{ } be_2 \text{ } or_3 \text{ } not_4 \text{ } to_5 \text{ } be_6]$

T0, 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: $[to_1 \text{ } be_2 \text{ } or_3 \text{ } not_4 \text{ } to_5 \text{ } be_6]$

T0, 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!



- 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**.



- 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(p_1, p_2, k)

```
1  answer  $\leftarrow \langle \rangle$ 
2  while  $p_1 \neq \text{NIL}$  and  $p_2 \neq \text{NIL}$ 
3  do if  $\text{docID}(p_1) = \text{docID}(p_2)$ 
4      then  $I \leftarrow \langle \rangle$ 
5           $pp_1 \leftarrow \text{positions}(p_1)$ 
6           $pp_2 \leftarrow \text{positions}(p_2)$ 
7          while  $pp_1 \neq \text{NIL}$ 
8          do while  $pp_2 \neq \text{NIL}$ 
9              do if  $|\text{pos}(pp_1) - \text{pos}(pp_2)| \leq k$ 
10                 then  $\text{ADD}(I, \text{pos}(pp_2))$ 
11                 else if  $\text{pos}(pp_2) > \text{pos}(pp_1)$ 
12                     then break
13                  $pp_2 \leftarrow \text{next}(pp_2)$ 
14                 while  $I \neq \langle \rangle$  and  $|I[0] - \text{pos}(pp_1)| > k$ 
15                     do  $\text{DELETE}(I[0])$ 
16                 for each  $ps \in I$ 
17                     do  $\text{ADD}(\text{answer}, \langle \text{docID}(p_1), \text{pos}(pp_1), ps \rangle)$ 
18                  $pp_1 \leftarrow \text{next}(pp_1)$ 
19              $p_1 \leftarrow \text{next}(p_1)$ 
20              $p_2 \leftarrow \text{next}(p_2)$ 
21         else if  $\text{docID}(p_1) < \text{docID}(p_2)$ 
22             then  $p_1 \leftarrow \text{next}(p_1)$ 
23         else  $p_2 \leftarrow \text{next}(p_2)$ 
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.

Введение в информационный поиск
I Маннинг Кристофер Д., Шютце
Хайнрих

Рекомендуемая
литература

Для саморазвития (опционально)
Чтобы не набирать двумя
пальчиками



Спасибо за
внимание!

Антон Кухтичев



a.kukhtichev@mail.ru



[@toshunster](https://t.me/toshunster)