Язык SQL

Лекция 10
Полнотекстовый поиск

Е. П. Моргунов
Сибирский государственный университет науки и технологий
имени академика М. Ф. Решетнева
г. Красноярск
Институт информатики и телекоммуникаций
emorgunov@mail.ru

Компания Postgres Professional
г. Москва

Версия от 18.10.2018

Подготовка к работе (1)

На вашем компьютере уже должна быть развернута база данных demo.

- Войдите в систему как пользователь postgres:

```
su - postgres
```

- Должен быть запущен сервер баз данных PostgreSQL:

```
pg_ctl  start  -D /usr/local/pgsql/data
```

- Для проверки запуска сервера выполните команду

```
pg_ctl  status  -D /usr/local/pgsql/data
```

или

```
ps -ax | grep postgres | grep -v grep
```

- Если у вас база данных demo была модифицирована, то для ее восстановления выполните команду

```
psql -f demo_small.sql -U postgres
```
(для ОС Debian)

```
psql -f demo_small.sql
```
(для ОС Xubuntu)

Подготовка к работе (2)

- Запустите утилиту psql и подключитесь к базе данных demo

```
psql -d demo -U postgres
```
(для ОС Debian)

```
psql -d demo
```
(для ОС Xubuntu)

- Назначьте схему bookings в качестве текущей

```
demo=# set search_path = bookings;
```

Введение

- Полнотекстовый поиск (или просто *поиск текста*) — это возможность находить *документы* на естественном языке, соответствующие *запросу*, и, возможно, дополнительно сортировать их по релевантности для этого запроса.
- Наиболее распространённая задача — найти все документы, содержащие *слова запроса*, и выдать их отсортированными по степени *соответствия* запросу.
- Понятия запроса и соответствия довольно расплывчаты и зависят от конкретного приложения.
- В самом простом случае запросом считается набор слов, а соответствие определяется частотой слов в документе.
- В PostgreSQL есть операторы ~, ~*, LIKE и ILIKE, но их возможностей недостаточно.

Полнотекстовая индексация

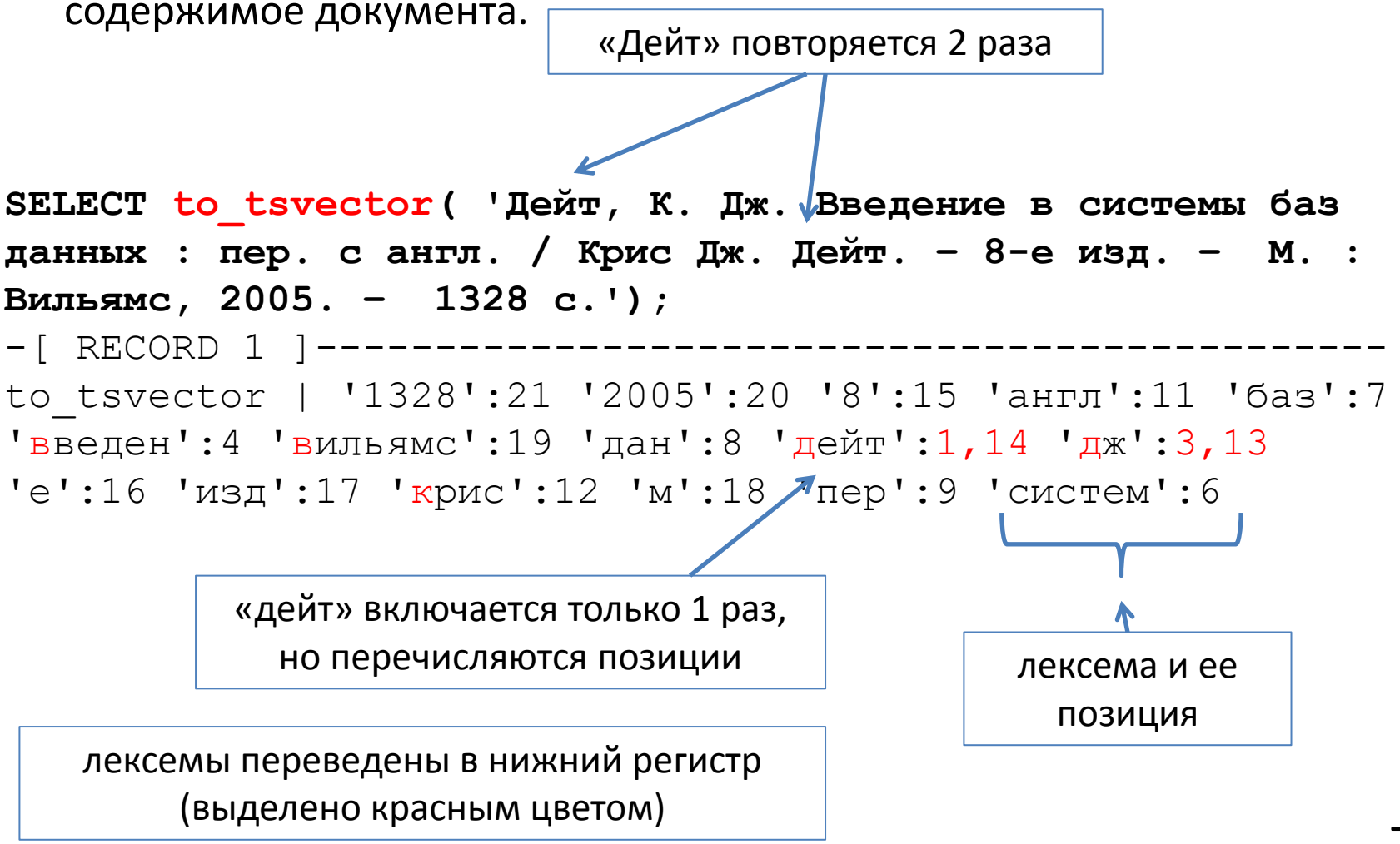
- Полнотекстовая индексация заключается в *предварительной обработке* документов и сохранении индекса для последующего быстрого поиска.
- Предварительная обработка включает следующие операции:
  - ***Разбор документов на фрагменты*** – числа, слова, словосочетания, почтовые адреса и т. д., которые будут обрабатываться по-разному.
  - ***Преобразование фрагментов в лексемы***. Лексема — это *нормализованный* фрагмент, в котором разные словоформы приведены к одной (например, буквы верхнего регистра приводятся к нижнему, а из слов обычно убираются окончания, исключаются *стоп-слова*). Для выполнения этого шага в PostgreSQL используются *словари*.
  - ***Хранение документов в форме, подготовленной для поиска.*** Помимо лексем часто желательно хранить информацию об их положении для *ранжирования по близости*

Типы данных tsvector и tsquery

- Для хранения подготовленных документов в PostgreSQL предназначен тип данных **tsvector**, а для представления обработанных запросов — тип **tsquery**.
- Поиск и ранжирование документов выполняется исключительно с представлением документа в формате tsvector.
- Исходный текст потребуется извлечь, только когда документ будет отобран для вывода пользователю.

Простые примеры

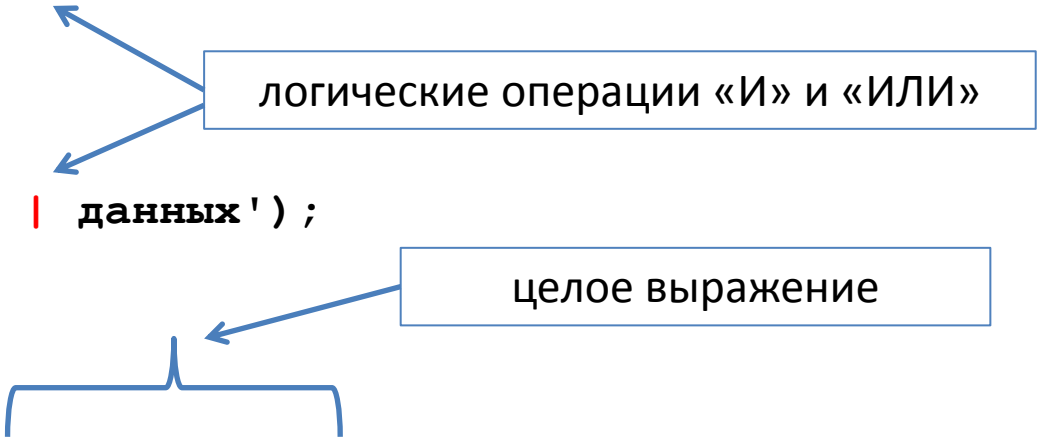
- Функция to_tsvector может разобрать и нормализовать текстовое содержимое документа.



Функции to_tsquery и plainto_tsquery

- Можно воспользоваться функциями to_tsquery и plainto_tsquery для преобразования заданного пользователем текста, который содержит слова для поиска, в значение tsquery. Они нормализуют слова в этом тексте.

```
SELECT to_tsquery( 'базы & данных');
 to_tsquery
---------------
 'баз' & 'дан'
(1 строка)
SELECT to_tsquery( 'базы | данных');
  to_tsquery
---------------
 'баз' | 'дан'
(1 строка)
SELECT plainto_tsquery( 'базы данных');
 plainto_tsquery
-----------------
 'баз' & 'дан'
(1 строка)
```



Простая проверка работы поиска

- Выполним проверку работы поиска, не создавая таблицы.

```
SELECT to_tsvector( 'Дейт, К. Дж. Введение в системы баз
       данных : пер. с англ. / Крис Дж. Дейт. – 8-е изд. –
       М. : Вильямс, 2005. –  1328 с.') @@
       to_tsquery( 'базы & данных' );
```

```
 ?column?
----------
 t
(1 строка)
```

обратите внимание

Проверка документа, описываемого значением типа tsvector, на соответствие критерию, заданному значением типа tsquery

Сравните: в описании – «баз данных», а в запросе – «базы данных»

Пример с общим табличным выражением

```
WITH books ( description ) AS
 ( VALUES ( 'Дейт, К. Дж. Введение в системы баз данных :
            пер.  с англ. / Крис Дж. Дейт. – 8-е изд. –
            М. : Вильямс, 2005. –  1328 с.‘ ),
          ( 'Грофф, Дж. SQL. Полное руководство : пер. с
            англ. / Джеймс Р. Грофф,  Пол Н. Вайнберг,
            Эндрю Дж. Оппель. – 3-е изд. – М. : Вильямс,
            2015. – 960 с.' ),
          ( 'Лузанов, П. PostgreSQL для начинающих /
            П. Лузанов, Е. Рогов, И. Лёвшин ; Postgres
            Professional. – М., 2017. – 146 с.' ) ),
  books_2 ( description, ts_description ) AS
  ( SELECT description, to_tsvector( description )
    FROM books )
SELECT * FROM books_2;
```

Выполним выборку

```
-[ RECORD 1 ]--+---------------------------------------------
description    | Дейт, К. Дж. Введение в системы баз данных : пер.  с англ. / Крис Дж. Дейт. – 8-е изд. –  М. : Вильямс, 2005. –  1328 с.
ts_description | '1328':21 '2005':20 '8':15 'англ':11 'баз':7 'введен':4 'вильямс':19 'дан':8 'дейт':1,14 'дж':3,13 'е':16 'изд':17 'крис':12 'м':18 'пер':9 'систем':6
-[ RECORD 2 ]--+---------------------------------------------
description    | Грофф, Дж. SQL. Полное руководство : пер. с англ. / Джеймс Р. Грофф,  Пол Н. Вайнберг, Эндрю Дж. Оппель. – 3-е изд. – М. : Вильямс,  2015. – 960 с.
ts_description | '2015':23 '3':18 '960':24 'с':25 'sql':3 'англ':8 'вайнберг':14 'вильямс':22 'грофф':1,11 'дж':2,16 'джеймс':9 'е':19 'изд':20 'м':21 'н':13 'оппел':17 'пер':6 'пол':12 'полн':4 'р':10 'руководств':5 'эндр':15
-[ RECORD 3 ]--+---------------------------------------------
description    | Лузанов, П. PostgreSQL для начинающих / П. Лузанов, Е. Рогов, И. Лёвшин ; Postgres Professional. – М., 2017. – 146 с.
ts_description | '146':16 '2017':15 'postgr':12 'postgresql':3 'profession':13 'е':8 'лузан':1,7 'лёвшин':11 'м':14 'начина':5 'п':2,6 'рог':9
```

Изменим главный запрос

- Будем показывать только главный запрос, поскольку запрос в CTE не изменяется.

```
WITH books ( description ) AS (...),
  books_2 ( description, ts_description ) AS (...)
SELECT  description
FROM books_2
WHERE ts_description @@ to_tsquery( 'SQL' );
```

```
-[ RECORD 1 ]-----------------------------------
description | Грофф, Дж. SQL. Полное руководство : пер.
с англ. / Джеймс Р. Грофф,  Пол Н. Вайнберг, Эндрю Дж.
Оппель. – 3-е изд. – М. : Вильямс,  2015. – 960 с.
```

Обратите внимание, что книга, в названии которой есть слово «PostgreSQL», не попала в выборку. Почему?

А что сделает регулярное выражение ?

```
WITH books ( description ) AS (...),
  books_2 ( description, ts_description ) AS (...)
SELECT description
FROM books_2 WHERE description ~ 'SQL';

-[ RECORD 1 ]-----------------------------------------
description | Грофф, Дж. SQL. Полное руководство : пер. с
англ. / Джеймс Р. Грофф,  Пол Н. Вайнберг, Эндрю Дж.
Оппель. – 3-е изд. – М. : Вильямс,  2015. – 960 с.
-[ RECORD 2 ]-----------------------------------------
description | Лузанов, П. PostgreSQL для начинающих / П.
Лузанов, Е. Рогов, И. Лёвшин ; Postgres Professional. –
М., 2017. – 146 с.
```

Логическая операция «ИЛИ»

```
WITH books ( description ) AS (...),
  books_2 ( description, ts_description ) AS (...)
SELECT description
FROM books_2
WHERE ts_description @@
      ( to_tsquery( 'SQL' ) ||
        to_tsquery( 'PostgreSQL' ) );
```

```
to_tsquery( 'SQL | PostgreSQL' )
```

```
-[ RECORD 1 ]-----------------------------------
description | Грофф, Дж. SQL. Полное руководство : пер.
с англ. / Джеймс Р. Грофф,  Пол Н. Вайнберг, Эндрю Дж.
Оппель. – 3-е изд. – М. : Вильямс,  2015. – 960 с.
-[ RECORD 2 ]-----------------------------------
description | Лузанов, П. PostgreSQL для начинающих /
П. Лузанов, Е. Рогов, И. Лёвшин ; Postgres
Professional. – М., 2017. – 146 с.
```

Функция plainto_tsquery

```
WITH books ( description ) AS (...),
  books_2 ( description, ts_description ) AS (...)
SELECT description
FROM books_2
WHERE ts_description @@  plainto_tsquery( 'базы данных' );
```

```
-[ RECORD 1 ]-----------------------------------------
description | Дейт, К. Дж. Введение в системы баз данных :
пер.  с англ. / Крис Дж. Дейт. – 8-е изд. –  М. : Вильямс,
2005. –  1328 с.
```

Выборка из библиотечной базы данных

Государственная универсальная научная библиотека Красноярского края

https://irbis.kraslib.ru/cgi-bin/irbis64r/irbis64r_91/cgiirbis_64.exe

Сделаем выборку по критерию:

«Ключевое слово= ПРОГРАММИРОВАНИЕ»

1. Любанович, Билл. Простой Python [Текст] : современный стиль программирования / Билл Любанович. - Санкт-Петербург ; Москва ; Екатеринбург : Питер, 2018. - 476 с.

2. Робсон, Элизабет. Изучаем HTML, XHTML и CSS [Текст] / Элизабет Робсон, Эрик Фримен ; [пер. с англ. В. Черник]. - Москва ; Санкт-Петербург ; Нижний Новгород : Питер, 2017. - 718 с.

3. Бхаргава, Адитья. Грокаем алгоритмы [Текст] : иллюстрированное пособие для программистов и любопытствующих / Адитья Бхаргава ; [пер. с англ. Е. Матвеев]. - Санкт-Петербург : Питер, 2017. - 288 с.

. . .

(1000 записей)

Подготовка данных для занесения в таблицу БД

1. Открыть файл в формате doc и сохранить в виде простого текста в файле books.txt.

2. Удалить лишние символы '\r':

```
sed -e 's/^M//' books.txt > books2.txt
```

3. Удалить точки после номеров книг и сделать разделителем полей символ табуляции. Это текст программы books.pl:

```
!/usr/bin/perl -w
while (<STDIN>)
{
  chomp;
  my ( $num, $descr ) = $_ =~ /^(\d+)\.\s*(.+)$/;
  print $num . "\t" . $descr . "\n";
}
exit( 0 );
```

```
./books.pl < books2.txt > books3.txt
cp books3.txt  /home/postgres
```

Занесение данных в таблицу БД

```
CREATE TABLE books
( book_id integer PRIMARY KEY,
  book_description text
);
CREATE TABLE

COPY books FROM '/home/postgres/books3.txt';
COPY 1000

ALTER TABLE books ADD COLUMN ts_description tsvector;
ALTER TABLE

UPDATE books
SET ts_description = to_tsvector( 'russian',
                                  book_description );
UPDATE 1000
```

Выборка

```
EXPLAIN SELECT * FROM books
WHERE ts_description @@ to_tsquery( 'SQL' );
                          QUERY PLAN
---------------------------------------------------------------
 Seq Scan on books  (cost=0.00..402.50 rows=2 width=721)
   Filter: (ts_description @@ to_tsquery('SQL'::text))
(2 строки)
-[ RECORD 1 ]----+-----------------------------------------
book_id          | 807
book_description | Маркин, Александр Васильевич. Построение
запросов и программирование на SQL [Текст] : учебное пособие
: [для студентов высших учебных заведений, обучающихся по
специальности 230201 "Информационные системы и технологии"] /
А. В. Маркин. - Москва : Диалог-МИФИ, 2011. - 344 с.
ts_description   | '2011':33 '230201':21 '344':34 'sql':9
'александр':2 'васильевич':3 'высш':15 'диалог':31 'диалог-
миф':30 'заведен':17 'запрос':5 'информацион':22
'маркин':1,28 'миф':32 'москв':29 'обуча':18 'пособ':12
'построен':4 'программирован':7 'систем':23 'специальн':20
'студент':14 'текст':10 'технолог':25 'учебн':11,16
```

Каковы затраты ресурсов?

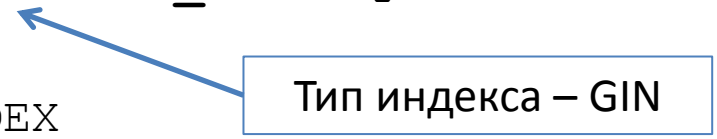
```
EXPLAIN ANALYZE SELECT * FROM books
WHERE ts_description @@ to_tsquery( 'SQL' );
                                 QUERY PLAN
-------------------------------------------------------------------------
Seq Scan on books  (cost=0.00..402.50 rows=2 width=721)
                   (actual time=29.380..33.913 rows=1 loops=1)
   Filter: (ts_description @@ to_tsquery('SQL'::text))
   Rows Removed by Filter: 999
 Planning time: 0.985 ms
 Execution time: 34.000 ms
(5 строк)
```

Создадим индекс

- Создадим индекс по столбцу, содержащему нормализованное описание.

```
CREATE INDEX books_idx ON books
  USING GIN ( ts_description );

CREATE INDEX
```




```
\d books
...
Индексы:
    "books_pkey" PRIMARY KEY, btree (book_id)
    "books_idx" gin (ts_description)
```

Повторим выборку

```
EXPLAIN ANALYZE SELECT * FROM books
WHERE ts_description @@ to_tsquery( 'SQL' );
                                QUERY PLAN
-----------------------------------------------------------------------
 Bitmap Heap Scan on books  (cost=12.26..20.07 rows=2
                                width=721)
                  (actual time=0.100..0.103 rows=1 loops=1)
   Recheck Cond: (ts_description @@
                  to_tsquery('SQL'::text))
   Heap Blocks: exact=1
   ->  Bitmap Index Scan on books_idx  (cost=0.00..12.26
                                        rows=2 width=0)
                  (actual time=0.081..0.082 rows=1 loops=1)
         Index Cond: (ts_description @@
                      to_tsquery('SQL'::text))
 Planning time: 1.261 ms
 Execution time: 0.218 ms
(7 строк)
```

Литература

1. Лузанов, П. PostgreSQL для начинающих / П. Лузанов, Е. Рогов, И. Лёвшин ; Postgres Professional. – М., 2017. – 146 с.
2. Моргунов, Е. П. Язык SQL. Базовый курс : учеб.-практ. пособие. / Е. П. Моргунов ; под ред. Е. В. Рогова, П. В. Лузанова ; Postgres Professional. – М., 2017. – 257 с.
3. PostgreSQL [Электронный ресурс] : официальный сайт / The PostgreSQL Global Development Group. – http://www.postgresql.org.
4. Postgres Professional [Электронный ресурс] : российский производитель СУБД Postgres Pro : официальный сайт / Postgres Professional. – http://postgrespro.ru.

Задание

1. Изучить материал лекции. За дополнительной информацией следует обратиться к главе 12 документации – «Полнотекстовый поиск».
2. Провести эксперименты с таблицей books, содержащей описания книг и нормализованные представления этих описаний, выполняя различные операции, описанные в разделе документации 9.13 «Функции и операторы текстового поиска».