MySQLと PostgreSQLと 日本語全文検索

Azure Databaseで Mroonga・PGroongaを使いたいですよね！？

須藤功平　クリアコード

OSS on Azure非公式コミュニティ #5 『Azure Database』勉強会
2017-06-26

アンケート

Azure Databaseで高速な日本語全文検索をしたい人！

最後にもう一度似たようなことを聞くよ！

日本語全文検索：LIKE

LIKE

- 😄SQL標準
  - MySQLでもPostgreSQLでも使える
- 😄少ないデータなら速度は十分
  - 400文字×20万件くらいなら1秒とか

LIKEのパフォーマンス

- ☹少なくないデータ
  - レスポンスが遅い
- ☹多くの同時アクセス
  - スループットがでない
  - 1回のLIKE毎にCPUが専有されるため

パフォーマンスの考え方

- 少ないデータ &&
  多くない同時アクセス
  - LIKEで十分
- 少なくないデータ ||
  多くの同時アクセス
  - 高速化が必要

高速日本語全文検索

インデックスで高速化できる

- MySQL
  - 5.7から標準対応
- PostgreSQL
  - GIN（組込）＋pg_trgm（標準添付）

高速？

ベンチマーク！

- 対象：Wikipedia日本語版
- レコード数：約185万件
- データサイズ：約7GB
- メモリー4GB・SSD250GB (ConoHa)

https://github.com/groonga/wikipedia-search/issues/4

（他人のベンチマークは参考程度）
（検討時はちゃんと実際の環境でベンチマークをとろう！）

注意

- pg_trgmではなくpg_bigmを使用
  - pg_bigm：外部プラグイン
  - 性能の傾向はだいたい同じ
    1,2文字のときはpg_bigmの方が速い
    3文字以上はpg_trgmの方が速い

検索1

キーワード：テレビアニメ

（ヒット数：約2万3千件）

InnoDB ngram	3m2s
InnoDB MeCab	6m20s
pg_bigm	4s

検索2

キーワード：データベース

（ヒット数：約1万7千件）

InnoDB ngram	36s
InnoDB MeCab	0.03s
pg_bigm	2s

検索3

キーワード：PostgreSQL OR MySQL

（ヒット数：約400件）

InnoDB ngram	N/A（エラー）
InnoDB MeCab	0.005s
pg_bigm	0.185s

検索4

キーワード：日本

（ヒット数：約63万件）

InnoDB ngram	1.3s
InnoDB MeCab	1.3s
pg_bigm	0.84s

高速…？

- InnoDB FTS MeCab
  - ハマれば速い
  - クエリーが複数語だと遅い
- pg_bigm
  - ハマれば速い
  - ヒット数が多いと遅い
- InnoDB FTS ngram：安定して遅い

Mroonga・PGroonga

- Mroonga（むるんが）
  - MySQLに
    高速日本語全文検索機能を追加する
    プロダクト
- PGroonga（ぴーじーるんが）
  - PostgreSQLに
    高速日本語全文検索機能を追加する
    プロダクト

検索1

キーワード：テレビアニメ

（ヒット数：約2万3千件）

InnoDB ngram	3m2s
InnoDB MeCab	6m20s
Mroonga:1	0.11s
pg_bigm	4s
PGroonga:2	0.29s

検索2

キーワード：データベース

（ヒット数：約1万7千件）

InnoDB ngram	36s
InnoDB MeCab:1	0.03s
Mroonga:2	0.09s
pg_bigm	2s
PGroonga:3	0.17s

検索3

キーワード：PostgreSQL OR MySQL

（ヒット数：約400件）

InnoDB ngram	N/A（エラー）
InnoDB MeCab:1	0.005s
Mroonga:2	0.028s
pg_bigm	0.185s
PGroonga:3	0.063s

検索4

キーワード：日本

（ヒット数：約63万件）

InnoDB ngram	1.3s
InnoDB MeCab	1.3s
Mroonga:1	0.21s
pg_bigm:2	0.84s
PGroonga	1s

検索速度まとめ

- Mroonga・PGroonga
  - 安定して速い
- InnoDB FTS MeCab・pg_bigm
  - ハマれば速い
- InnoDB FTS ngram
  - 安定して遅い

Mroonga（むるんが）

MySQLに
高速日本語
全文検索機能を
追加

Mroonga：インデックス作成

普通のMySQLの使い方

```
CREATE TABLE ... (
  ...,
  FULLTEXT INDEX (column)
) ENGINE=Mroonga;
```

Mroonga：全文検索

普通のMySQLの使い方

```
SELECT * FROM ...
  WHERE
    MATCH(column)
    AGAINST('キーワード'
              IN BOOLEAN MODE);
```

Mroonga：クエリー言語

デフォルトOR→AND

```
-- ↓AまたはBが含まれていればマッチ
AGAINST('A B' IN BOOLEAN MODE);
AGAINST('+A +B' IN BOOLEAN MODE);
-- ↑↓AとBが含まれていればマッチ
-- ↓Mroongaの拡張
AGAINST('*D+ A B' IN BOOLEAN MODE);
```

Mroonga：Windows

Windows用バイナリーあり

- MariaDBとセット
- ダウンロードして展開するとすぐに使える

PGroonga

PostgreSQLに高速日本語全文検索機能を追加

PGroonga：インデックス

普通のPostgreSQLの使い方

```
CREATE INDEX name ON texts
  USING pgroonga (content);
```

PGroonga：全文検索

専用演算子を使用

```
SELECT * FROM ...
  WHERE
    column &? 'キーワード';
```

PGroonga：JSON

JSON内の全テキストを全文検索

```
CREATE TABLE logs (record jsonb);
CREATE INDEX i ON logs
  USING pgroonga (record);
-- ログのどこかに「error」があればマッチ
SELECT * FROM logs
  WHERE record &? 'error';
```

PGroonga：JSON全文検索例

以下は全部マッチ

```
{"message": "Error!"}
{"tags": ["web", "error"]}
{"syslog": {"message": "error!"}}
```

PGroonga：入力補完1

検索ボックスで便利なアレ

```
CREATE TABLE terms
  (term text,          -- 候補単語
   readings text[]); -- ヨミガナ
-- インデックス
CREATE INDEX i ON terms USING pgroonga
  (term pgroonga.text_term_search_ops_v2,
   readings pgroogna.text_array_term_search_ops_v2);
```

PGroonga：入力補完2

用意するデータ：
候補とカタカナのヨミガナだけ

```
INSERT INTO terms
  VALUES
    ('牛乳',                          -- 補完候補
     ARRAY['ギュウニュウ', -- ヨミガナ1
           'ミルク']);     -- ヨミガナ2
```

PGroonga：入力補完3

ローマ字で検索

```
SELECT term FROM terms
  -- 「ギュウニュウ」にヒット
  WHERE readings &^~ 'gy';
--  term
-- ------
--  牛乳
-- (1 row)
```

PGroonga：入力補完4

ひらがなで検索

```
SELECT term FROM terms
  -- 「ギュウニュウ」にヒット
  WHERE readings &^~ 'ぎゅう';
--  term
-- ------
--  牛乳
-- (1 row)
```

PGroonga：入力補完5

カタカナで検索

```
SELECT term FROM terms
  -- 「ギュウニュウ」にヒット
  WHERE readings &^~ 'ギュウ';
--  term
-- ------
--  牛乳
-- (1 row)
```

PGroonga：入力補完6

別のヨミガナでもヒット

```
SELECT term FROM terms
  -- 「ミルク」にヒット
  WHERE readings &^~ 'mi';
--  term
-- ------
--  牛乳
-- (1 row)
```

PGroonga：入力補完7

漢字でもヒット

```
SELECT term FROM terms
  -- 「牛乳」にヒット
  WHERE readings &^ '牛';
--  term
-- ------
--  牛乳
-- (1 row)
```

PGroonga：Windows

Windows用バイナリーあり

- 商用ログ管理製品
  「VVAULT AUDIT」が採用
  http://vvault.jp/product/vvault-audit/
  - アクセスログに対して
    ユーザー名・パスを全文検索
- 決め手：高速・省スペース

まとめ

- Mroonga（むるんが）
  - MySQLで高速日本語全文検索！
  - しかも使いやすいし便利！
- PGroonga（ぴーじーるんが）
  - PostgreSQLで高速日本語全文検索！
  - しかも使いやすいし便利！

アンケート1

Azure Databaseで 高速な日本語全文検索を したい人！

最初より増えているといいな

アンケート2

Azure DatabaseでMroonga・PGroongaを使いたい人！

Azure Database開発者にアピールして！
どういう風に使いたいか教えて！