

# groonga 開発予報

有限会社 未来検索ブラジル

矢田 晋

祝  
本日リリースされた  
groonga 1.2.8 には  
grn\_dat が含まれています

正確には以前のバージョンにも grn\_dat のコードは入っていたのですが、現在の仕様になり、実用に供されるようになったのは groonga 1.2.8 からです。

そこで今日は  
皆さんにちょっと

# ダブル配列の話を 聞いてもらいます

ダブル配列というのは `grn_dat` が用いているデータ構造の名称です.

一部でカルト的な人気を誇っているかもしれません.

# とりあえず grn\_dat とは何か

# grn\_dat とは

- 文字列を ID と関連付けるモジュール
  - grn\_pat, grn\_hash の仲間



ID	文字列
1	Trebor
2	Werdna
3	L'kbreth
4	Gatekeeper
...	...

`dat[1] == "Trebor"`


`dat[2] == "Werdna"`

`dat["L'kbreth"] == 3`

`dat["Gatekeeper"] == 4`

基本的な機能は文字列に ID を割り当てるというものですが、重要なのは ID と文字列で相互に参照できることです。

## grn\_dat と仲間たち

- grn\_pat – パトリシアトライ
  - 前方一致検索をサポートする
- grn\_hash – ハッシュ表
  - 前方一致検索をサポートしない代わりに高速
- grn\_dat – ダブル配列 
  - 前方一致検索をサポートする上に高速

grn\_pat と grn\_dat が前方一致検索をサポートできるのは、どちらもトライの仲間だからです。

ちなみに grn\_dat の「高速」は参照に対する評価です。単純な追加・検索については、ハッシュ表がもっとも優秀です。

# いずれも 参照ロックフリー

groonga といえば参照ロックフリーということで, groonga のコアである grn\_pat/hash/dat はいずれも参照ロックフリーな実装になっています.



## grn\_dat の特徴

- 前方一致検索と参照時間を重視
  - 文字列更新については後述

イマココ

	grn_pat	grn_hash	grn_dat
前方一致検索	○	×	○
参照	○	◎	◎
更新	○	◎	△
サイズ	◎	○	△
文字列更新	×	×	○

それぞれの特徴をまとめた表になっています。

grn\_dat は参照が速い代わりに更新が遅くてサイズが大きいので、速度面で困ったとき、あるいは困ることが予想されるときに利用を検討すれば良いと思います。文字列更新はテーブル・カラムをリネームしたいという要望から生まれた機能です。ダブル配列だから可能というものではありません。


# 前方一致検索とは

- Common Prefix Search
  - クエリの前半に一致する文字列を見つける  
“北海道” ⇒ “北”, “北海”, “北海道”
  - 用途: クエリから索引語への分割
- Predictive Search
  - クエリで始まる文字列を見つける  
“南斗” ⇒ “南斗孤鷺拳(シン)”, “南斗水鳥拳(レイ)”, etc.
  - 用途: クエリの補完・拡張

今回の発表では, 説明を簡単にする目的で, Common Prefix Search と Predictive Search をまとめて前方一致検索ということにしました.

# 文字列更新とは

- ID を残して文字列のみを更新すること
  - 用途: テーブル情報の管理

ID	文字列		ID	文字列
1	Trebor		1	Trebor
2	Werdna		2	Werdna
3	L'kbreth		3	L'kbreth
4	Gatekeeper		4	Sorn
...				...

Update("Gatekeeper", "Sorn")

文字列を差し替えるだけの単純な機能ですが, これのおかげでテーブルやカラムの名前変更が可能になりました.

## grn\_dat の役割

- grn\_pat の代替として
  - 前方一致検索が必要なとき
  - 更新より参照の方が多いとき
  - メモリ使用量より参照時間を重視するとき
- テーブル情報の管理に使うと
  - テーブルやカラムの名前変更が可能になる
  - MySQL で ALTER TABLE RENAME が可能になる

# 技術的な情報がほしい方へ

- grn\_dat 開発のポイント
  - ダブル配列の参照ロックフリー化
  - 更新の効率化
  - 前方一致検索の効率化
- 参考資料
  - 参照ロックフリーなダブル配列
    - [http://groonga.org/ja/blog/2011/11/08/grn\\_dat.html](http://groonga.org/ja/blog/2011/11/08/grn_dat.html)

参考資料に書いてある内容は、本来は参照ロックフリーでないダブル配列をどうやって参照ロックフリーにしたのか、参照ロックフリーにすることで悪化した更新効率をどうやって補っているのか、ダブル配列の苦手な Predictive Search をどうやって効率化したのかというものです。

# そろそろ説明は終わりにして

いいにくの日 2011

全文検索エンジン groonga を囲むタブ 2 #groonga

銀河の歴史がまた 14 ページ

# 見せてもらおうか 新しいモジュールの性能とやらを

いいにくの日 2011

全文検索エンジン groonga を囲むタブ 2 #groonga

銀河の歴史がまた 15 ページ

ベンチマークに入りますよの合図です.

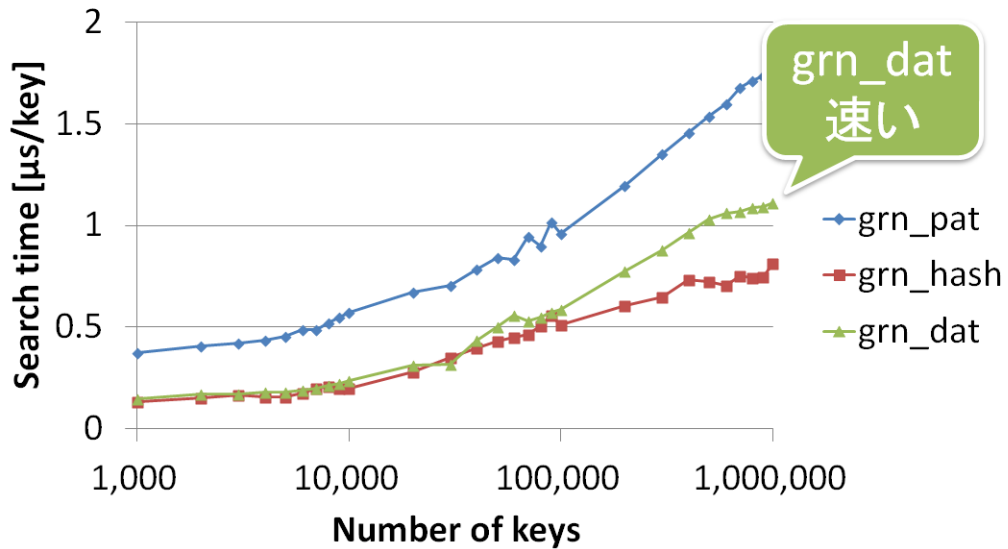
## ベンチマーク(準備)

- データ
  - jawiki-20111111-all-titles-in-ns0
  - 先頭の 100 万件を使用
- 構築・参照方法
  - ランダム順に登録
  - ランダム順に登録文字列を参照
- 計測方法
  - 試行回数 11 で中央値を採用

日本語 Wikipedia のタイトル一覧を利用しました。  
ウィキペディア創設者ジミー・ウェールズからのお願いを  
無下にすることはできません。



## ベンチマーク(参照時間)



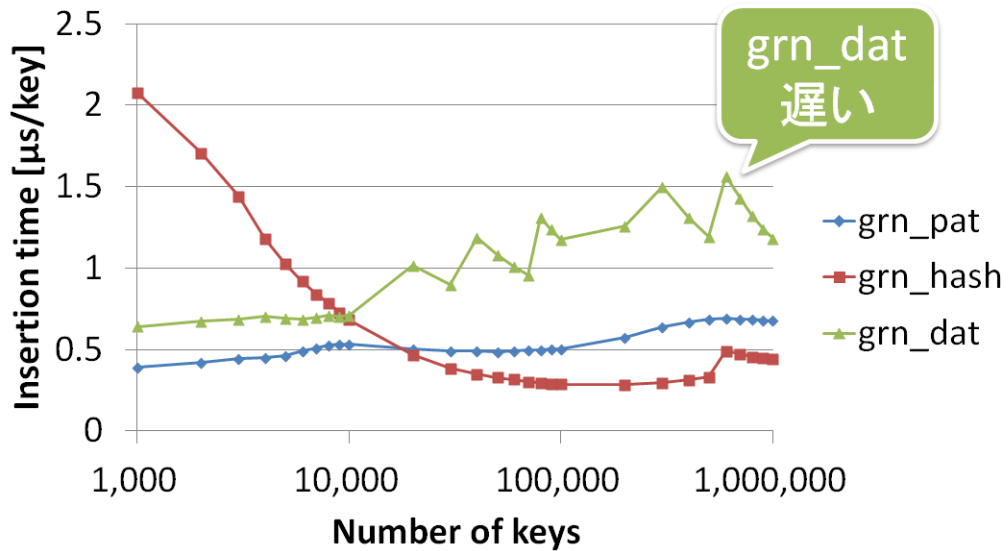
いいにくの日 2011

全文検索エンジン groonga を囲むタブ 2 #groonga

銀河の歴史がまた 17 ページ

grn\_dat は grn\_pat の代替としての役割を持つので、grn\_pat(青)とgrn\_dat(緑)を比較してください。下にあるほど高速です。

## ベンチマーク(構築時間)



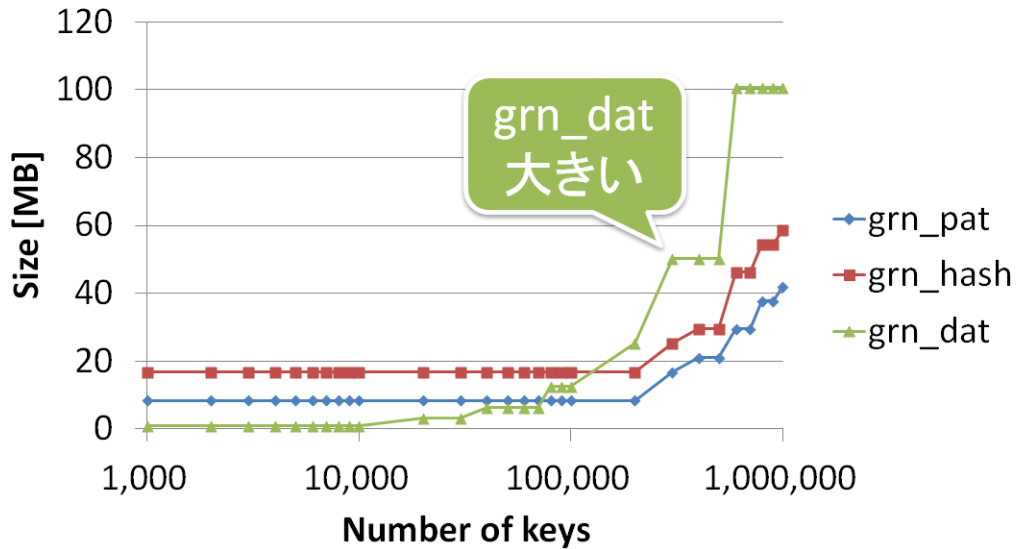
いいにくの日 2011

全文検索エンジン groonga を囲むタブ 2 #groonga

銀河の歴史がまた 18 ページ

grn\_dat は参照の割合が多い用途に使いましょうということを示す実験結果です.

## ベンチマーク(サイズ)



いいにくの日 2011

全文検索エンジン groonga を囲むタベ 2 #groonga

銀河の歴史がまた 19 ページ

メモリ消費が気になるときは grn\_pat を使った方がいいですよということを示す実験結果です。

# まとめると 前方一致検索ができて 参照時間に優れる

# そういえば

いいにくの日 2011

全文検索エンジン groonga を囲むタベ 2 #groonga

銀河の歴史がまた 21 ページ

# 「groonga 開発予報」 というタイトルでした

いいにくの日 2011

全文検索エンジン groonga を囲むタブ 2 #groonga

銀河の歴史がまた 22 ページ

# 検討中の内容を紹介します

いいにくの日 2011

全文検索エンジン groonga を囲むタベ 2 #groonga

銀河の歴史がまた 23 ページ

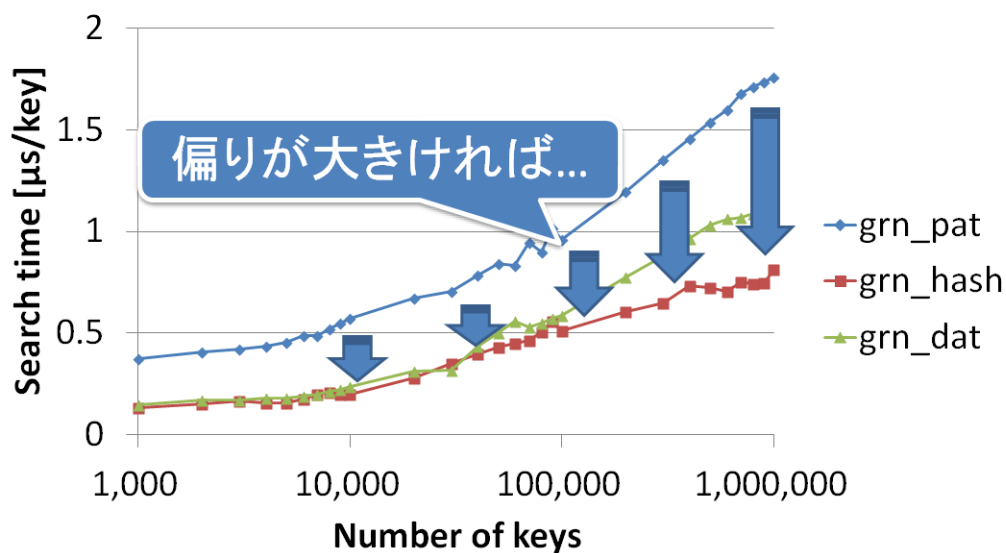
ひとつ

## 頻出する索引語をキャッシュ

<http://groonga.org/ja/blog/2011/07/13/lexicon-cache.html>



## ベンチマーク(参照時間)



いいにくの日 2011

全文検索エンジン groonga を囲むタベ 2 #groonga

銀河の歴史がまた 25 ページ

全文検索に用いる索引語を登録する場合, どうしても偏りが大きくなるので, それを利用すれば grn\_pat とキャッシュの組み合わせで grn\_hash 並みの性能が出せるはずという案です.

# ふたつ

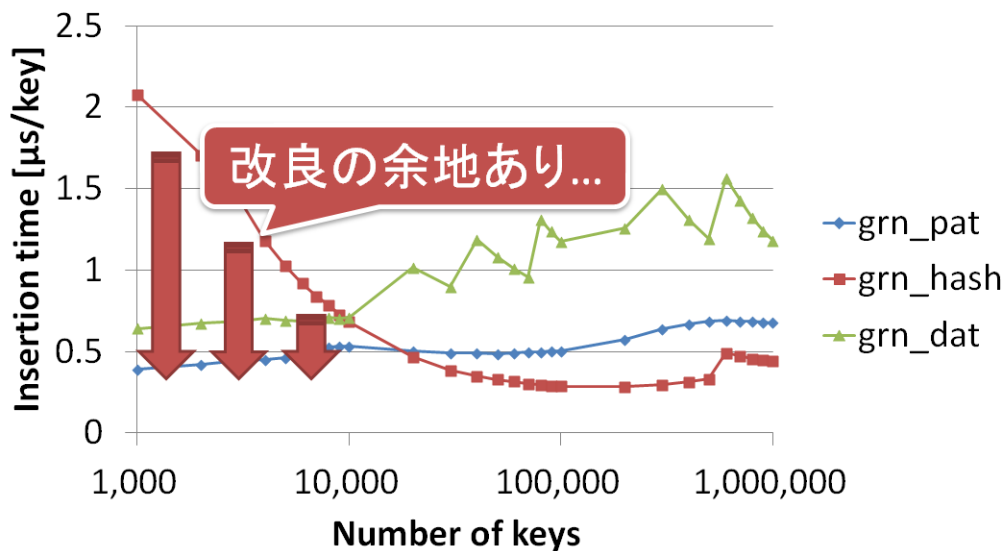
## 不安定なハッシュ表を調整

いいにくの日 2011

全文検索エンジン groonga を囲むタベ 2 #groonga

銀河の歴史がまた 26 ページ

## ベンチマーク(構築時間)



いいにくの日 2011

全文検索エンジン groonga を囲むタブ 2 #groonga

銀河の歴史がまた 27 ページ

小規模な grn\_hash の構築時間が長くなっている部分は原因がほぼ特定されているので、改良の余地がありそうだという案です。

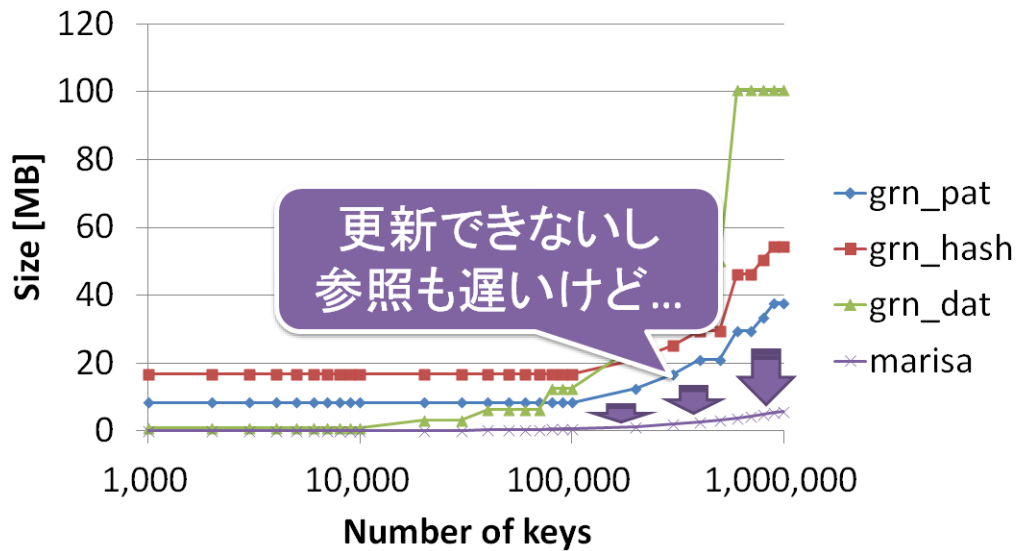
検索の途中経過なんかも grn\_hash に保存しているので、全体的な性能に影響するのではないかと考えています。

みつ

## 見る機会の少ないデータを圧縮

<http://code.google.com/p/marisa-trie/>

## ベンチマーク(サイズ)



いいにくの日 2011

全文検索エンジン groonga を囲むタベ 2 #groonga

銀河の歴史がまた 29 ページ

更新不可で参照も遅い代わりに桁違いにコンパクトなデータ構造があるので、それらを上手く利用できれば無駄をなくすことができるのではないかという案です。

他にもありますが  
このくらいで勘弁してください

# grn\_dat と愉快的仲間たちに 絞って紹介しました

今回は grn\_dat をメインに持ってきたので,  
grn\_pat/hash/dat に関する内容に絞りました.

次回  
「groonga の野望」  
お楽しみに