groonga
新年と収穫の祭り

森大二郎

有限会社未来検索ブラジル

全文検索エンジンgroongaを囲む夕べ2

2011/11/29

内容

- 来し方
  - 開発の経緯など
- 現在
  - 最近の動向
- 行く末
  - 今後開発したいもの

来し方

開発の経緯

- 2005年2月にSennaをリリース
- 2chを検索するために開発
- 全板のカキコを即時検索可能に
- 当初から索引の動的構築ひとすじ
- その後groongaに改名
- カラムストア機能を追加

索引の動的構築とは

- 登録した文書を索引に即時反映
- (i.e. リアルタイムサーチ)
- DBMSなら普通だが..
- 検索エンジンとしては当時斜め上
  - 完全転置索引の動的構築は煩雑
- その後世間でも徐々に盛んに
  - twitter, real-time web の影響?

索引の静的構築と動的構築

- 静的構築
  - 構築が完了した時点で検索可能になる
  - 小さい作業領域で高速に構築可能
- 動的構築
  - 検索可能な状態を維持しながら構築
  - ランダムI/Oを抑えるために工夫が必要
  - webで使うならロックは極力抑えたい
    - 大量の検索と更新を同時にこなしたい!

groongaは動的構築一筋

- いろいろ工夫してます
  - メモリ上の索引とディスク上の索引
  - インプレイス更新とマージと併用
  - ロックフリーなデータ構造の採用
- その効果は..?
  →「mroongaのベンチマーク」の枠で

静的構築vs動的構築

技術的な情報については以下を

> 検索エンジンはいかにして動くのか？
> http://gihyo.jp/dev/serial/01/search-engine
> by 山田浩之さん

良記事!オススメ!! 書籍化期待age!

静的構築vs動的構築

!!類似品に注意!!

> 検索エンジンはなぜ見つけるのか
> http://amzn.to/jScDv6
> by 森大二郎

索引構築については触れていません!

groongaに改名してからは

- カラムストア機能を追加
  - 列単位でデータを格納
  - トランザクション処理はやや不得手
  - 集計処理が得意
  - 圧縮効率が追求しやすい
- MySQLやPostgreSQLからも
  これらのメリットを享受可能に!?

現在

最近の動向

- Geographical Searching
  →「Geographical Searching」の枠で
- DBMSとの結合強化
  →「mroonga」「groonga with PostgreSQL」の枠で
- 高速な語彙表
  →「groonga開発予報」の枠で
- 索引の静的構築

え？静的構築？今さら？

- 動的構築ひとすじじゃなかったの?
- オフラインで索引を作る際は
  やはり静的構築が性能的に有利
- 静的構築の機能を後付けで
  入れるのは比較的容易
  (逆はけっこう面倒)
- 静的構築もできた方がいいよね

静的構築の用途

- できると嬉しい場合が増えてきた
- インデックスを後付けする場合
  - 祝! mroonga ALTER TABLE対応
- offlineで索引を作るシステムで
  - 祝! Sedue groonga plugin
    - PFI Seminar 2011/1201 に行こう!

静的構築のアプローチ

- 転置索引構築の2pass化
  メモリが少ないマシンでもサクサク
- 語彙表参照の高速化
  →「groonga開発予報」の枠で
- けっこう速くなりそう
  今日数字が出せずごめんなさい!
- 年末ぐらいにリリースしたい!
  新年ぐらいには収穫の祭りが!(苦しい..)

行く末

今後開発したいもの1

- カラムストアとしての性能強化
  - キャッシュを考慮したデータ構造
  - 高速なファセット検索
- 索引の圧縮方式の拡充
  - 参照頻度に応じて圧縮方式にメリハリ
- 類似文字列検索
  - 編集距離･コサイン類似度 etc

今後開発したいもの2

- 頻出パターン抽出
  - n-gram, 共起語 etc
- ストリーム処理機能
  - 時間窓内での頻出パターン・最大値・最小値計算
- スキーマレス化
  - データの分布に応じて自動的にデータ構造を最適化

最後に。。

「今後開発したいもの」の多くは
まだ開発者が決まってません。

開発者募集!

開発者募集!!

開発者募集!!!