embarcadero
Developer Camp

【G6】ライトニングトーク

ngram方式を採用した全文検索VCLコンポーネント

株式会社ビズ
田付　幸一

embarcadero
Developer Camp

概要

開発の目的

- ヘルプデスク用のインシデント管理システムの開発に伴い大量の受付インシデントやメールを検索したい。

  色んな検索エンジンがあるけど
  フリーのはいまいち使えないし
  売ってるのは高いし
  管理システムとの連携をどの様に行うか？
  きめ細やかな制御をしたいので有れば自作でしょ
  （車輪の再発明と言わないで）

VCLなら配布も楽だしいいんじゃない

ngram

- 文字列をn文字（本コンポーネントでは2文字）ごと区切って文書内の位置情報を付与して索引とする方式

  ＜長所＞

  - 辞書が不要で原理が簡単
  - 検索漏れが無い（「東京都」を「京都」で検索OK）

  ＜短所＞

  - 検索ノイズ有り（「東京都」を「京都」で検索しちゃう）
  - です、ます等余り意味の無いフレーズで索引が埋まる
  - 頻出語句の検索に時間がかかりがち
  - 索引が大きめ

現状

- 現状こんなもんです。
  - Unicode非対応 （思いっきり文字コード依存）
  - 単純なAND検索だけ
    （正規表現ってどうやるの？）
  - 索引作成に掛かる時間がまだまだ不満
  - 索引ファイルの大きさも不満
  - 既存索引に対する追加、削除が弱い
  - 頻出単語が含まれると結構遅いかも
  - HDDのランダムアクセス性能に大きく左右される
    →SSDを使うことで性能UP

索引作成　サンプルコード

```
uses ・・・,FULLTextSearch;

  Indexer : TFullTextSearchMakeIndex;   // 索引作成VCLコンポーネント

implementation

  Indexer.FilePath := 'c:¥demo';     // 索引保存先ファイルパス
  Indexer.FileName := 'wikipedia';   // 索引ファイル名
  Indexer.Clear;                     // 初期化
  Indexer.SaveText := False;         // 索引にオリジナルテキストを含まない
  while データ有り do
    Indexer.AddData(                 // テキスト追加
                     Text,             // 検索対象文字列
                     key,              // 文書を特定するユニーク文字列(文書キー)
                    'text',            // フィールド名(オリジナルテキスト保存時)
                    TimeStamp,         // 文書日付（検索範囲を指定する場合に使用)
                    1                  // フィールド番号（検索時に指定)
                     );
  Indexer.Save(True);                // 索引を保存して終了
```

検索　サンプルコード

```
uses ・・・,FULLTextSearch;

  Search : TFullTextSearch;        // 検索VCLコンポーネント

implementation

var Kekka : TStrings;

  Search.FilePath := 'c:¥demo';        // 索引保存先ファイルパス
  Search.FileName := 'wikipedia';      // 索引ファイル名
  Kekka := Search.Search(              // 検索実行(文書キーの文字列リストが返る)
                         SearchWord,     // 検索キーワード
                         [1],            // 検索対象フィールド番号リスト
                         Start_date,     // 検索範囲開始日付（省略可）
                         End_date        // 検索範囲終了日付（省略可）
                         );
```

embarcadero
Developer Camp

デモンストレーション

デモに使用するデータについて

Wikipediaの一部をダウンロードしたものからテキスト部分を抽出

- テキストファイル件数　16．3万ファイル
- ファイルあたりの平均サイズ　7．0KB
- 総データ量　1．08GB
- 索引作成に掛かった時間　32分
- 索引サイズ　1．62GB
- 索引サイズ比率　150%

※ちなみに全テキストファイルを秀丸でgrep検索したら19分かかりました

マシンスペック

＜索引作成＞

- Athlon64 x2 4000+（Dual Core 2.1GHz）
- メモリ　2GB
- 2.5inch SSD(40GB)　+　2.5inch SSD(120GB)　SATA2
- Windows　2003Server

＜検索デモ＞

- Core2Duo P8600（Dual Core 2.4GHz）
- メモリ　2GB
- 2.5inch HDD(250GB)　+　2.5inch SSD(32GB USB接続)
- Windows XP sp3

開発ツール

開発バージョン
　Delphi7 Enterprise

本日のデモ用に移植
　Delphi 2009 Professional

64bit化？　（メモリいっぱい使えるからいいかも）
　Delphi XE2　　未購入^^;

embarcadero
Developer Camp

ご静聴ありがとうございました