

メール全文検索アプリケーション Sylph-Searcher のご紹介

SRA OSS, Inc. 日本支社
技術部 チーフエンジニア・Sylpheed開発者

山本 博之
yamamoto@sraoss.co.jp

Sylph-Searcherとは

- Sylpheed向け電子メール全文検索アプリケーション
- PostgreSQL 8.2の全文検索機能を利用
- Linux/Unix、Windows 2000以降で動作
- オープンソース(修正BSDライセンス)で配布

Sylph-Searcherの特長

- PostgreSQL の全文検索エンジン tsearch2 を利用
- メールの検索に特化
- Sylphedで管理しているメールボックス(MH, IMAP4, News)、MH(Mew, Wanderlust等)をインポート可能
- 高速なインポート・検索
- クライアントとデータベースを別マシンに分けることが可能
 - 複数台のマシンからもアクセス可能
- 内部文字コードはすべてUTF-8
 - 多言語に対応

Sylph-Searcherが使用するライブラリ

- GLib (基本的なデータ構造、アルゴリズムを提供)
- GTK+ (GUIを提供)
- LibSylph (Sylpheedのメールデータの操作)
- MeCab (テキストのわかち書き)
- libpq (PostgreSQLとの通信)

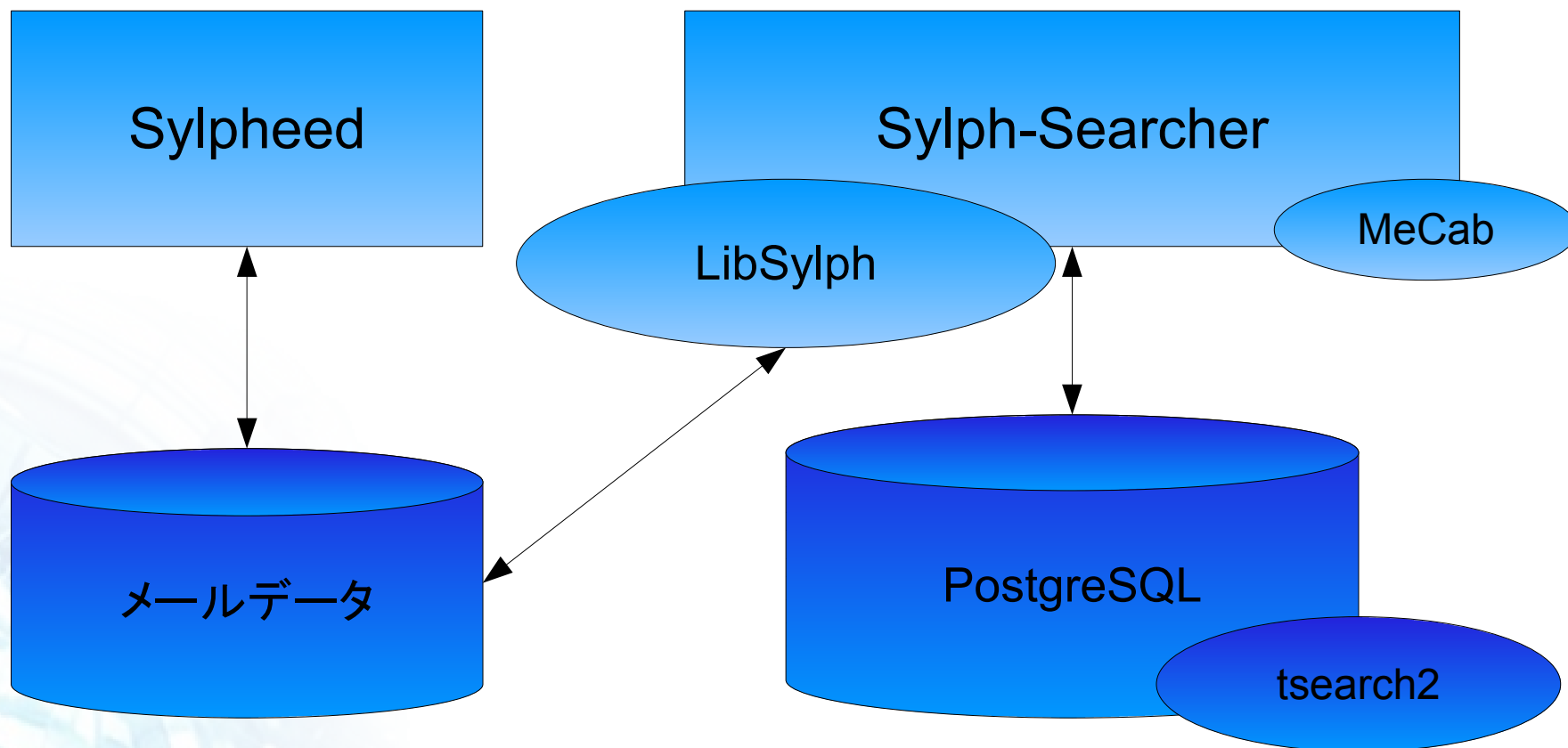
tsearch2

- PostgreSQLに全文検索機能を追加する拡張モジュール
- PostgreSQL 7.4以降のcontribディレクトリに付属
- PostgreSQL 8.3から本体に組み込まれる予定
- マルチバイト文字(UTF-8)対応
- 8.2で追加されたGIN(汎用転置インデックス)を使用
 - 転置インデックス: 単語がどの文書に含まれるかを表すデータ構造
- 単語ごとに区切られた言語が前提
 - 日本語ではわかち書きが必須

使い方

- データベースの設定
 - データベースの作成
 - 付属のスクリプトでテーブルと関数の登録
 - tsearch2の組み込み
- メールのインポート
 - syldbimportコマンドを使用
 - 例: `symbdbimport -d dbname "#mh/メール箱/inbox"`
 - 定期的に(cron等で)メールを差分インポートすると便利
- 検索
 - sylph-searcher (GUIフロントエンド) を実行

Sylpheed全文検索アプリ構成図



処理の流れ

- LibSylph経由でメールデータにアクセス
- MeCabで本文をわかち書き(日本語を単語に分解)
- libpq経由でPostgreSQLにデータを投入
- PostgreSQLはtsearch2により全文検索インデックスを生成
- 検索時は検索語を投入時と同様にMeCabで分かち書きしてデータベースに問い合わせる
 - 全文検索インデックスを使用して検索されるので高速

テーブル定義

msginfo

```

msg_sid BIGSERIAL PRIMARY KEY
file_size INTEGER
file_mtime TIMESTAMP
msg_date TIMESTAMP
flags INTEGER
hdr_from TEXT
hdr_to TEXT
hdr_cc TEXT
hdr_newsgroups TEXT
hdr_subject TEXT
hdr_msgid TEXT UNIQUE NOT NULL
hdr_inreplyto TEXT
hdr_references TEXT
body_text TEXT
body_index TSVECTOR
    
```

msg_folderinfo

```

msg_sid BIGINT NOT NULL
folder_id TEXT
msgnum INTEGER

PRIMARY KEY (folder_id, msgnum)
    
```

msginfo: メッセージの情報

msg_folderinfo: メッセージを含む
フォルダの情報

msg_sid: DB内部で使用する
メッセージ識別ID

hdr_msgid: [Message-Id] メッセージの
一意性の判別に使用

```
CREATE INDEX msginfo_body_index ON msginfo USING gin (body_index);
```

実際に発行されるSQL文(インポート時)

- INSERT INTO msginfo(file_size, file_mtime, msg_date, flags, <各種ヘッダ...>, body_text, body_index)
VALUES(..., to_tsvector('わかち 書き テキスト'));
- to_tsvector() 関数によりわかち書きされたテキストを tsvector型に変換
 - tsvector: キーワードのテキスト上の位置情報
- body_indexからGINインデックスを自動生成

実際に発行されるSQL文(検索時)

- ```
SELECT msg_sid, hdr_from, hdr_to,
hdr_subject, msg_date, folder_id FROM msginfo
LEFT JOIN msg_folderinfo USING (msg_sid)
WHERE query('検索 文字列') @@ body_index;
```

  - ※ query(): 

```
SELECT to_tsquery(replace($1, ' ', '&'))
```
- to\_tsquery()関数により検索文字列をtsearch2のクエリに変換
- 「WHERE <tsquery> @@ 対象カラム」という特別な書式を使用

## ベンチマーク

- 測定に使用したマシンのスペック
  - CPU: Core 2 Duo E6600 (2.4GHz, dual-core)
  - RAM: DDR2-667 1GB (512MB x2)
- 変更したPostgreSQLの設定
  - fsync=off
  - maintenance\_work\_mem=128MB

## ベンチマーク結果

- 32598通のメッセージ(448MB)のインポートにかかった時間
  - 8分56秒 (2回目: 14.2秒) (60.8通/秒)
- 上記のメッセージを以下のキーワードで検索した場合にかかった時間
  - 「全文検索」: 4.006 (msec) (ヒット件数: 215)
  - 「Sylpheed」: 63.590 (msec) (ヒット件数: 6890)
- データベース全体のサイズ
  - 199MB

## 実装済みの機能(1.0)

- インポート用コマンド (syldbimport)
- 簡易検索用コマンド (syldbquery)
- GUIフロントエンド (sylph-searcher)
- フォルダ単位のインポート
- フォルダの再帰的インポート、差分インポート
- Sylpheed管理外のMHフォルダのインポート
  - Mew, Wanderlustなどのメールもインポート可能
- 本文、From、To、Subject、日付による検索
- Sylpheedとの連携
  - 検索結果のメッセージをSylpheedで開く

## 実装予定の機能

- GUI上でのインポート操作
- 添付ファイルの検索
- 検索条件の追加
- フレーズ検索
- 検索ヒット率によるランキング
- 他フォーマット(mbox等)への対応
- Webインタフェースなどの提供
- その他の全文検索エンジンへの対応
- Sylpheedへの統合(将来的に)
  - 受信時に自動的にインポート、Sylpheed上での全文検索

## 関連サイト

- Sylpheed ホームページ
  - <http://sylpheed.sraoss.jp/>
- Sylph-Searcher のダウンロード
  - <http://sylpheed.sraoss.jp/ja/download.html#searcher>