



MySQLとSennaによる 日本語全文検索

住商情報システム株式会社

プラットフォームソリューション事業部門

IT基盤ソリューション事業部

オープンソースシステム部

池田 徹郎

2006年6月 <http://www.scs.co.jp/mysql>

MySQLとは？

オープンソースのデータベース



- 世界で最も普及しているオープンソースのデータベース (RDBMS)
- LAMPスタックの一つ
- MySQL ABがソースコードを管理。様々な有償サービスも提供中
- 国内にもパートナー企業多数。
- ライセンスはデュアルライセンス方式 (GPLおよび商用ライセンス)
- 最新安定版はver5.0.21
- <http://www.mysql.com>



今日の内容

1. MySQLで全文検索を行う際の問題点
2. なぜ全文検索ができないのか
3. 解決策 MySQL + Senna
4. MySQL + Sennaのデモ紹介
5. Sennaの解説
6. まとめ

1-1. MySQLの全文検索の問題点

- 日本語の全文検索ができない

```
C:\WINDOWS\system32\cmd.exe - mysql -uroot -p

mysql> CREATE TABLE t1 (c1 TEXT CHARSET cp932, FULLTEXT (c1)) ENGINE = MyISAM;
Query OK, 0 rows affected (0.06 sec)

mysql> INSERT INTO t1 VALUES ('今日は良い天気ですね。'),
-> ('明日も晴れるといいですね。'),
-> ('去年の夏は暑かったです。'),
-> ('台風が接近しているらしいですよ。'),
-> ('今年の夏も暑いらしいですよ。');
Query OK, 5 rows affected (0.00 sec)
Records: 5 Duplicates: 0 Warnings: 0

mysql> SELECT c1 FROM t1 WHERE MATCH(c1) AGAINST ('夏');
Empty set (0.02 sec)

mysql>
```

1-2. MySQLの全文検索の問題点

- LIKE演算子を使えば検索できるが遅い

```
C:\WINDOWS\system32\cmd.exe - mysql -uroot -p
mysql>
mysql> SELECT c1 FROM t1 WHERE c1 LIKE '%夏%';
+-----+
| c1                |
+-----+
| 去年の夏は暑かったです。 |
| 今年の夏も暑いらしいですよ。 |
+-----+
2 rows in set (0.00 sec)

mysql> EXPLAIN SELECT c1 FROM t1 WHERE c1 LIKE '%夏%'¥G
***** 1. row *****
      id: 1
  select_type: SIMPLE
        table: t1
         type: ALL
possible_keys: NULL
          key: NULL
         key_len: NULL
          ref: NULL
          rows: 5
   Extra: Using where
1 row in set (0.00 sec)

mysql>
```

2-1. なぜ全文検索ができないのか？

- 英語では全文検索ができる

```
C:\WINDOWS\system32\cmd.exe - mysql -uroot -p

mysql> CREATE TABLE t2 (c1 TEXT, FULLTEXT (c1)) ENGINE = MyISAM;
Query OK, 0 rows affected (0.06 sec)

mysql> INSERT INTO t2 VALUES ('I have a pen.'),
-> ('Please tell me why you went there.'),
-> ('Have a nice day, see you.'),
-> ('May I help you?'),
-> ('I bought a pen yesterday. ');
Query OK, 5 rows affected (0.02 sec)
Records: 5 Duplicates: 0 Warnings: 0

mysql> SELECT c1 FROM t2 WHERE MATCH(c1) AGAINST ('pen');
+-----+
| c1                |
+-----+
| I have a pen.     |
| I bought a pen yesterday. |
+-----+
2 rows in set (0.00 sec)

mysql> █
```

2-2. なぜ全文検索ができないのか？

- 英語では全文検索ができる(EXPLAIN)

```
CA コマンドプロンプト - mysql -uroot -p

mysql> EXPLAIN SELECT c1 FROM t2 WHERE MATCH(c1) AGAINST ('pen')%G
***** 1. row *****
      id: 1
  select_type: SIMPLE
        table: t2
         type: fulltext
possible_keys: c1
          key: c1
         key_len: 0
          ref:
         rows: 1
      Extra: Using where
1 row in set (0.05 sec)

mysql> _
```



2-3. なぜ全文検索ができないのか？

- 日本語と英語の記述方法の違い

- 英語は単語がスペースで区切られている
 - “I have a pen.”
 - “I bought a pen yesterday.”
- 日本語はスペースで区切られてはいない
 - “去年の夏は暑かったです。”
 - “今年の夏も暑いらしいですよ。”

2-4. なぜ全文検索ができないのか？

- 入力データを分かち書きすることで可能になる

```
mysql> CREATE TABLE t3 (c1 TEXT CHARSET utf8, FULLTEXT (c1)) ENGINE = MyISAM;
Query OK, 0 rows affected (0.13 sec)

mysql> INSERT INTO t3 VALUES ('今日は良い天気ですね。'),
-> ('明日も晴れるといいですね。'),
-> ('去年の夏は暑かったです。'),
-> ('台風が接近しているらしいですよ。'),
-> ('今年の夏も暑いらしいですよ。');
Query OK, 5 rows affected (0.00 sec)
Records: 5 Duplicates: 0 Warnings: 0

mysql> SELECT c1 FROM t3 WHERE MATCH(c1) AGAINST ('夏');
+-----+
| c1                |
+-----+
| 去年の夏は暑かったです。 |
| 今年の夏も暑いらしいですよ。 |
+-----+
2 rows in set (0.00 sec)

mysql>
```



3-1. 解決法 MySQL + Senna

- この組み合わせですべて解決

- MySQL + Sennaを使うことで...
 - 日本語の全文検索を行えるようになります
 - インデックスを使うので検索速度が速いです
 - 文章をスペースで分ける必要もありません
 - アプリケーションから見た場合、これまでMySQLを使ってきたのと同じ感覚で使うことができます
 - 他に全文検索エンジンを用意するといったことが必要ありません

3-2. 解決法 MySQL + Senna

- Sennaとはどんなソフトウェアなのか

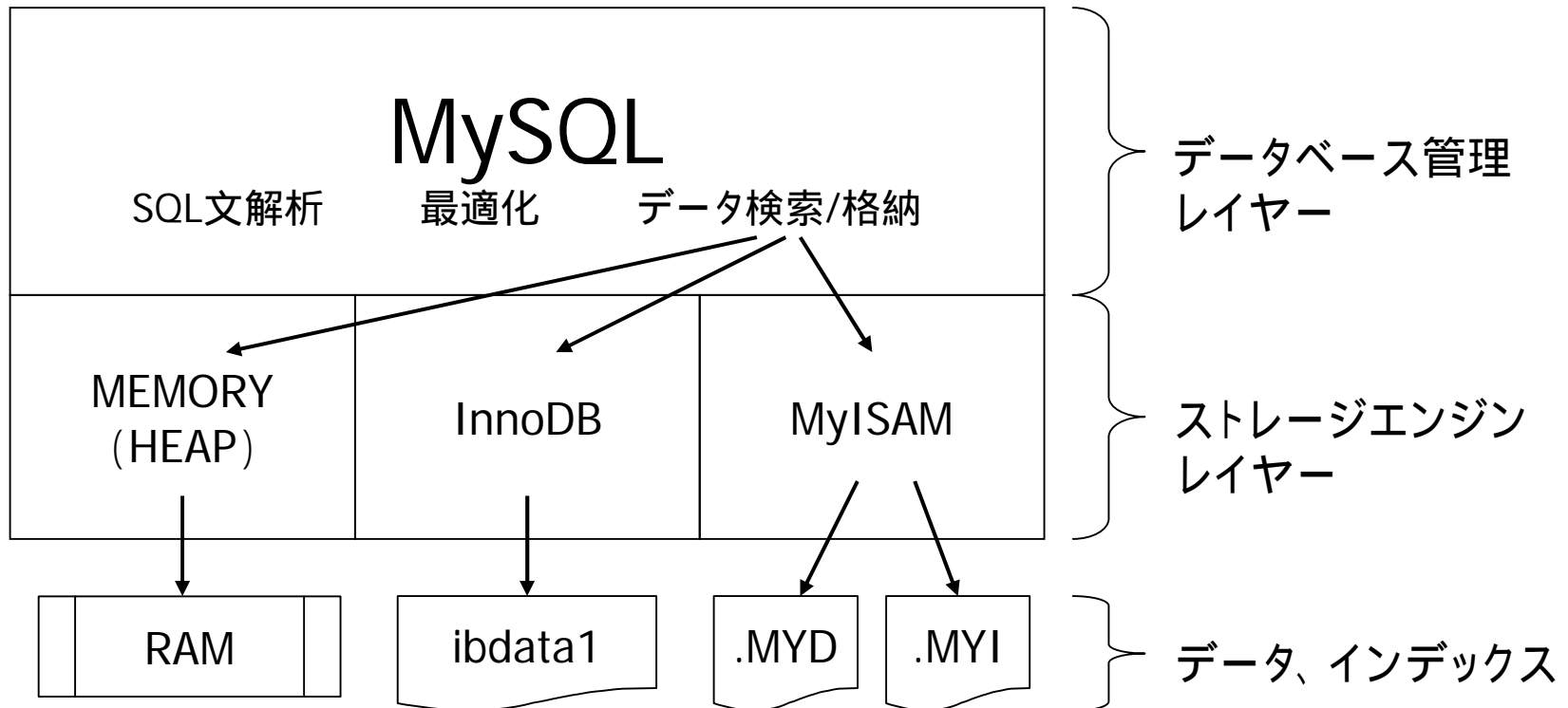
- 日本発のオープンソースソフトウェア
- 組み込み用の高速全文検索エンジン
- MySQLで日本語で全文検索を行う場合の問題を解決してくれる
- (有)未来検索ブラジル様により開発中
- 配布ライセンスはLGPL
- 2006年7月にver1.0をリリース予定
- <http://qwik.jp/senna/>



3-3. 解決法 MySQL + Senna

- 確認: MySQLのアーキテクチャ

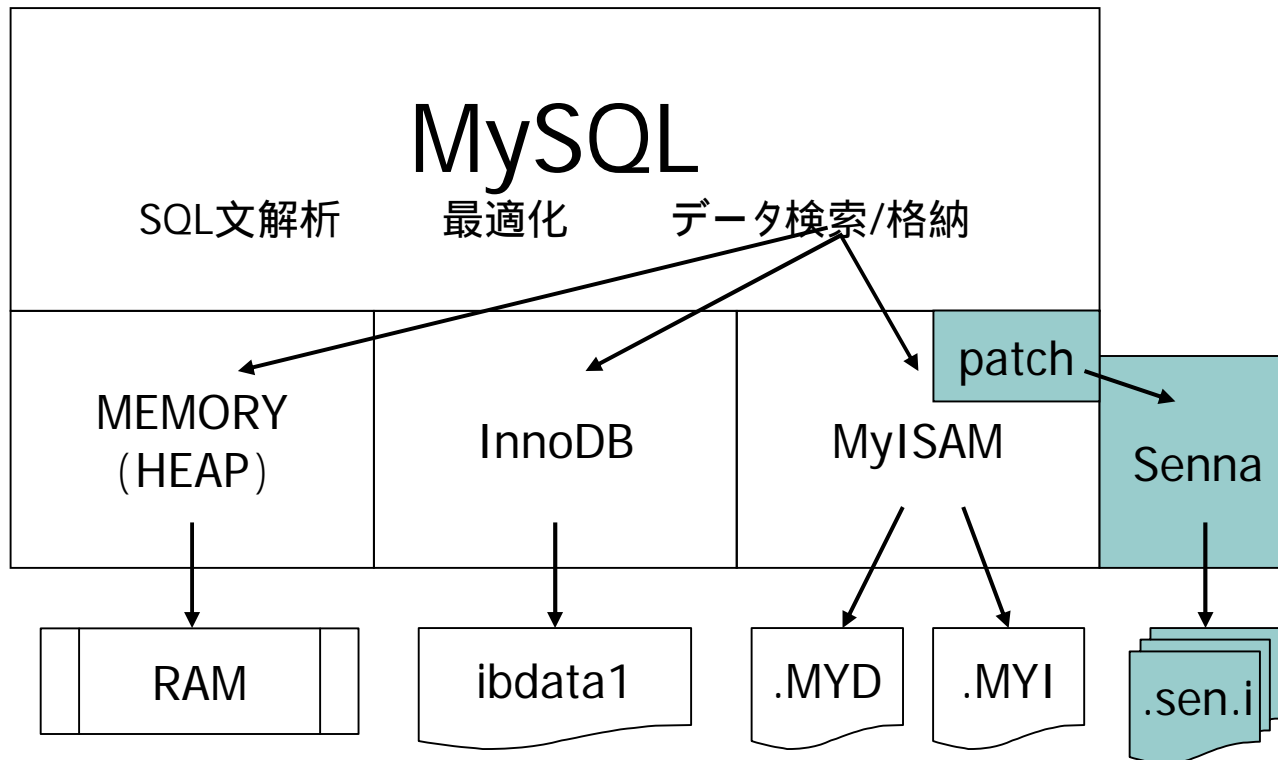
- MySQLはマルチストレージエンジン・アーキテクチャを採用
- MySQLでFULLTEXTインデックスをサポートしているのはMyISAM



3-4. 解決法 MySQL + Senna

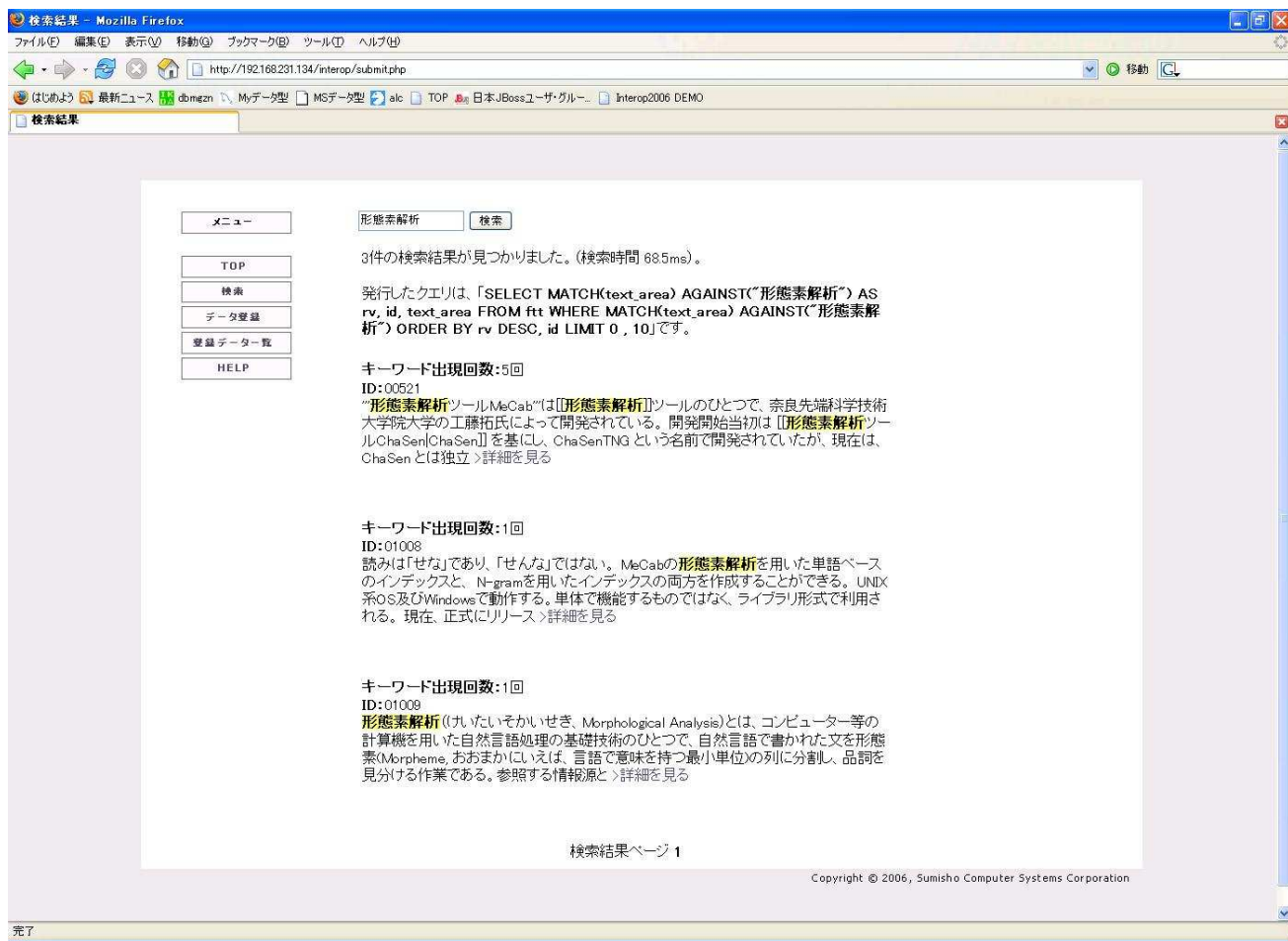
- MySQL+Sennaのアーキテクチャ

- MyISAMにパッチを当てて、FULLTEXTインデックスを使うときにMySQL経由でSennaが使われるようにする



4. MySQL+Sennaのデモ紹介

- Apache+PHP+MySQL+Sennaのデモ



The screenshot shows a Mozilla Firefox browser window displaying search results for the keyword '形態素解析'. The browser's address bar shows the URL 'http://192.168.231.134/interop/submit.php'. The search results page includes a navigation menu on the left with buttons for 'メニュー', 'TOP', '検索', 'データ登録', '登録データ一覧', and 'HELP'. The main content area displays the search results for '形態素解析', indicating that 3 items were found (search time: 68.5ms). The results are sorted by relevance (rv) in descending order, limited to 10 items. The first result is for ID:00521, with a keyword frequency of 5. The second result is for ID:01008, with a keyword frequency of 1. The third result is for ID:01009, with a keyword frequency of 1. The page footer shows '検索結果ページ 1' and 'Copyright © 2006, Sumisho Computer Systems Corporation'.

検索結果 - Mozilla Firefox

ファイル(F) 編集(E) 表示(V) 移動(O) ブックマーク(B) ツール(T) ヘルプ(H)

http://192.168.231.134/interop/submit.php

検索結果

メニュー

TOP

検索

データ登録

登録データ一覧

HELP

形態素解析 検索

3件の検索結果が見つかりました。(検索時間 68.5ms)。

発行したクエリは、「SELECT MATCH(text_area) AGAINST("形態素解析") AS rv, id, text_area FROM ftt WHERE MATCH(text_area) AGAINST("形態素解析") ORDER BY rv DESC, id LIMIT 0 , 10」です。

キーワード出現回数:5回
ID:00521
"形態素解析"ツールMeCab™は[[形態素解析]]ツールのひとつで、奈良先端科学技術大学院大学の工藤拓氏によって開発されている。開発開始当初は[[形態素解析]ツールChaSen[ChaSen]]を基にし、ChaSenTNG という名前で開発されていたが、現在は、ChaSenとは独立 >詳細を見る

キーワード出現回数:1回
ID:01008
読みは「せな」であり、「せんな」ではない。MeCabの形態素解析を用いた単語ベースのインデックスと、N-gramを用いたインデックスの両方を作成することができる。UNIX系OS及びWindowsで動作する。単体で機能するものではなく、ライブラリ形式で利用される。現在、正式にリリース >詳細を見る

キーワード出現回数:1回
ID:01009
形態素解析 (丸たいそかいせき、Morphological Analysis)とは、コンピューター等のコンピューターを用いた自然言語処理の基礎技術のひとつで、自然言語で書かれた文を形態素(Morpheme、おもかにいえば、言語で意味を持つ最小単位)の列に分割し、品詞を見分ける作業である。参照する情報源 >詳細を見る

検索結果ページ 1

Copyright © 2006, Sumisho Computer Systems Corporation

完了

5-1. Sennaの解説

- 利用方法はMySQLを使う方法と同じ

```
10.101.160.96 - PuTTY
mysql>
mysql> CREATE TABLE t1 (c1 TEXT CHARSET utf8, FULLTEXT (c1)) ENGINE = MyISAM;
Query OK, 0 rows affected (0.00 sec)

mysql> INSERT INTO t1 VALUES ('今日は良い天気ですね。'),
-> ('明日も晴れるといいですね。'),
-> ('去年の夏は暑かったです。'),
-> ('台風が接近しているらしいですよ。'),
-> ('今年の夏も暑いらしいですよ。');
Query OK, 5 rows affected (0.02 sec)
Records: 5 Duplicates: 0 Warnings: 0

mysql> SELECT c1 FROM t1 WHERE MATCH(c1) AGAINST ('夏');
+-----+
| c1                |
+-----+
| 去年の夏は暑かったです。 |
| 今年の夏も暑いらしいですよ。 |
+-----+
2 rows in set (0.00 sec)

mysql>
mysql>
mysql>
```



5-2. Sennaの解説

- Senna独自のインデックスが作成される

```
-rw-rw---- 1 mysql mysql 8.1M May 31 15:30 t1.000.SEN
-rw-rw---- 1 mysql mysql 2.6M May 31 15:30 t1.000.SEN.i
-rw-rw---- 1 mysql mysql 4.0K May 31 15:30 t1.000.SEN.i.c
-rw-rw---- 1 mysql mysql 13M May 31 15:30 t1.000.SEN.I
-rw-rw---- 1 mysql mysql 240 May 31 15:30 t1.MYD
-rw-rw---- 1 mysql mysql 1.0K May 31 15:30 t1.MYI
-rw-rw---- 1 mysql mysql 8.4K May 31 15:30 t1.frm
```

t1.000.SEN.i はバッファ用に使われる

t1.000.SEN.i.cがインデックスファイルの実体

t1.000.SENとt1.000.SEN.Iはインデックス管理用

t1.frmはMySQLのテーブル定義ファイル

t1.MYIはMyISAMのインデックスファイル

t1.MYDはMyISAMのデータファイル



5-3. Sennaの解説

- 2種類のインデックス方式から選択

例: 「今日は良い天気ですね」

- 分かち書き方式
 - 日本語なら日本語の辞書を使って文字列を分ける
 - 「今日/は/良い/天気/です/ね」のようにインデックスを作成する
- N-gram方式 (N=2の場合)
 - 機械的に2文字ずつ分ける
 - 「今日/日は/は良/良い/いい天/天気/気で/です/すね」と識別してインデックスを作成する

5-4. Sennaの解説

- 推奨は分かち書き方式

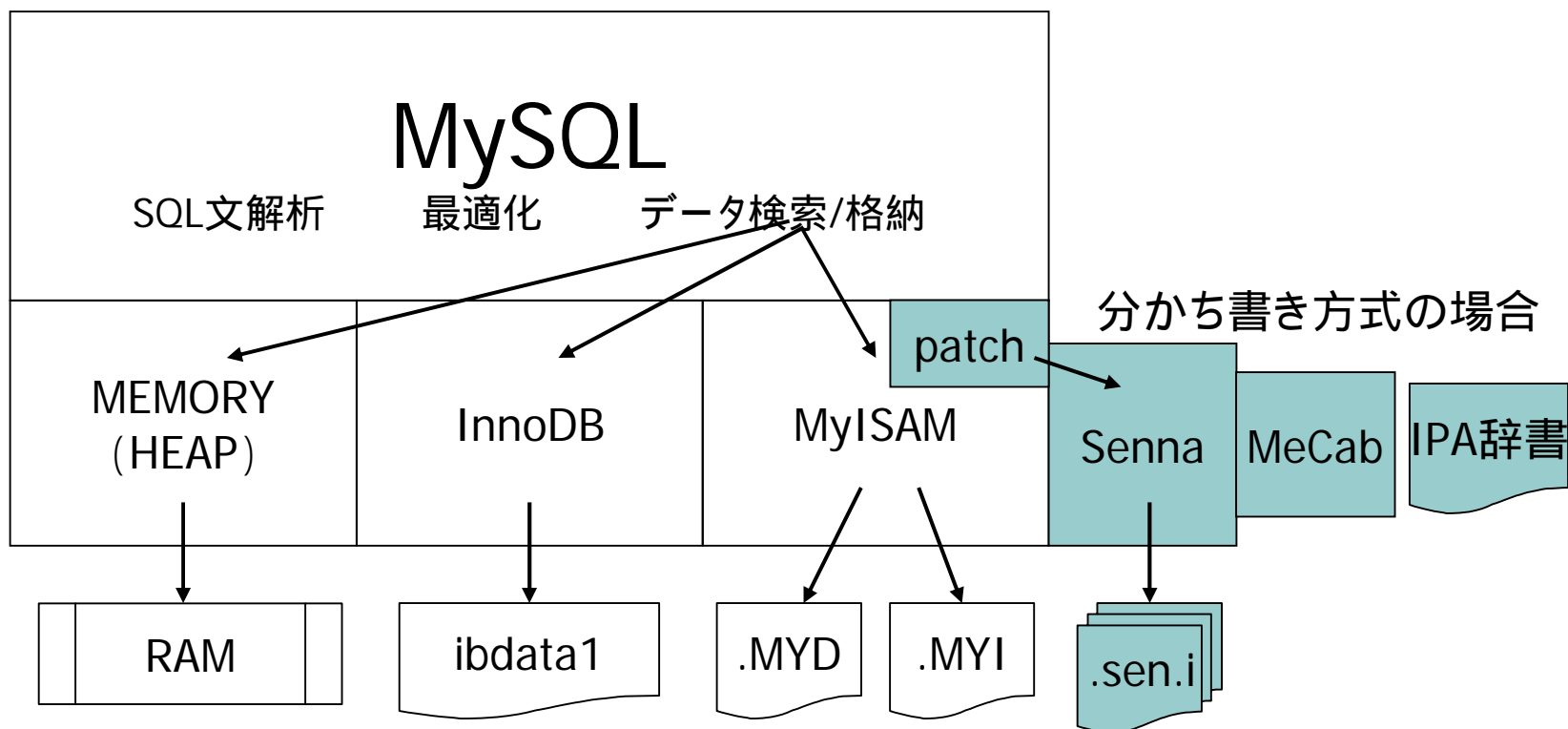
- 分かち書き方式の方がインデックスファイルが小さくてすむ
 - 分かち書き方式では約1.3倍、N-gram方式では約2.5倍
 - 100GBのデータに分かち書き方式でFULLTEXTインデックスを作成するとインデックスファイルは合計130GB
 - 100GBのデータにN-gram方式でFULLTEXTインデックスを作成するとインデックスファイルは合計250GB
- 分かち書き方式では辞書を使うため適合率が高い
 - 分かち書き方式は欲しいデータを効率よく検索できる
 - 形態素解析エンジンMeCab+IPA辞書などを利用

オープンソースの形態素解析エンジン MeCab
<http://mecab.sourceforge.jp/>

5-5. Sennaの解説

- MySQL+Sennaのアーキテクチャ

- 分かち書き方式の場合には形態素解析ソフトMeCabを使う
- MeCabを使った分かち書き方式が推奨されている





5-6. Sennaの解説

- MySQL+Sennaのビルド方法

MeCabのビルド

```
./configure --prefix=/usr --with-charset=utf8  
make  
make install
```

MeCabの辞書 (IPA辞書) のビルド

```
./configure --prefix=/usr --with-charset=utf8  
make  
make install
```

Sennaのビルド

```
./configure --prefix=/usr  
make  
make install
```



5-7. Sennaの解説

- MySQL+Sennaのビルド方法(続き)

MySQLのビルド

```
patch -p1 < ../senna/bindings/mysql/mysql-5.0.21.senna.diff
patch -p1 < ../senna/bindings/mysql/mysql-5.0.21.senna.2ind.diff

autoconf

CFLAGS="-O2 -march=pentium4" CXX=gcc ¥
CXXFLAGS="-O2 -march=pentium4 -felide-constructors" ¥
./configure --prefix=/usr/local/mysql --with-charset=utf8 ¥
--with-extra-charsets=complex --with-named-curses-libs=/usr/lib/libncurses.a ¥
--enable-thread-safe-client --enable-local-infile --enable-asm-asm --without-innodb ¥
--disable-shared --with-client-ldflags=-all-static --with-mysqld-ldflags=-all-static ¥
--with-big-tables --without-readline

make

make install
```



5-8. Sennaの解説

- Sennaの設定ファイルとログファイル

senna.confで文字コードの設定を行う例

```
mkdir /var/senna  
vi /var/senna/senna.conf  
DEFAULT_ENCODING utf8
```

senna.logは/var/senna/logディレクトリの有無でON/OFFされる

```
mkdir /var/senna/log
```

5-9. Sennaの解説

- ログファイル senna.log の例

```
senna:/var/senna/log # tail -n20 senna.log
05/31:15:30:22.536860|671d|_mi_ft_add(0x868e790,0,0x868f3c8,0x8694850,40)
05/31:15:30:22.536972|671d|inv=63c9d008 new seg=8
05/31:15:30:22.537485|671d|inv=63c9d008 new seg=10
05/31:15:30:22.538027|671d|inv=63c9d008 new seg=11
05/31:15:30:22.538590|671d|_mi_ft_add(0x868e790,0,0x868f3c8,0x8694850,88)
05/31:15:30:22.538673|671d|inv=63c9d008 new seg=3
05/31:15:30:22.539196|671d|_mi_ft_add(0x868e790,0,0x868f3c8,0x8694850,132)
05/31:15:30:22.539319|671d|_mi_ft_add(0x868e790,0,0x868f3c8,0x8694850,188)
05/31:15:30:22.539377|671d|inv=63c9d008 new seg=13
05/31:15:30:22.539904|671d|inv=63c9d008 new seg=21
05/31:15:30:22.540430|671d|inv=63c9d008 new seg=22
05/31:15:30:22.540947|671d|inv=63c9d008 new seg=35
05/31:15:30:22.541480|671d|inv=63c9d008 new seg=37
05/31:15:30:23.385917|671d|sen_index_sel > (夏)
05/31:15:30:23.385975|671d|n=1 (夏)
05/31:15:30:23.385990|671d|exact: 2
05/31:15:30:23.386000|671d|hits=2
05/31:15:30:23.386022|671d|ft_nlq_reinit_search(0x8670c18)
05/31:15:30:23.386069|671d|ft_nlq_close_search(0x8670c18)
05/31:15:30:23.386082|671d|sen_records_close < (8664b70:3:0)
```



6. MySQL+Sennaまとめ

- 便利、手軽、余計な手間がかからない

- MySQL+Sennaでオープンソースだけでデータベースの日本語全文検索が利用できる
- アプリから見るとSennaはMySQLによって完全に隠蔽されており、MySQLを使う側からするとMySQL単独で使うのと変わらない



ご静聴ありがとうございました。

- 住商情報システムはMySQLに関する技術サポート、ライセンス販売、トレーニング等を行っています。ぜひ一度ご相談下さい。
- <http://www.scs.co.jp/mysql>
- oss@scs.co.jp