データベース構築論2

20章「全文検索の高速化」

(Namazu)

中島康彦

§20. 1 検索の手法

「手作業で作成したキーワード」による検索
- キーワード作成が繁雑，類擬語が膨大／不統一
- 利用時には高速，キーワードの体系が不明

「文書そのもの」に対する全文検索
- システムを作るのは簡単
- 利用時には低速，キーワード作成者による擾乱がない

「自動生成したインデックス」による検索
- 一度，インデックス生成の仕組みを作れば，あとは簡単
- 利用時には高速，インデックスと類擬語のバランスが必要

§20. 2 インデックス検索までの流れ

インデックスを作成する.

- 文書の内容から種類を推測する.

  File-MMagic

- 種類に応じて文書をテキストファイルに変換する.

  word,excel,pdfなど ⇒ text形式

- 文書を「わかち書き」にする. (形態素解析)

```
% echo "本日は晴天なり" | kakasi -w
本日 は 晴天 なり

% cd ~/Html/ml00-x.archive
% mknmz -k [0-9]*
% ls
```

検索する.

```
% namazu 検索式 .
```

§20. 3 WEBとの連携

.namazurcの作成

```
% vi .namazurc
Index      .
Template   .
Replace    /usr0/majordomo/lists/ml00-x.archive/
   http://i.econ.kyoto-u.ac.jp/~xxx/ml00-x.archive/
Logging    on
Lang       ja
```

namazu.cgiの作成

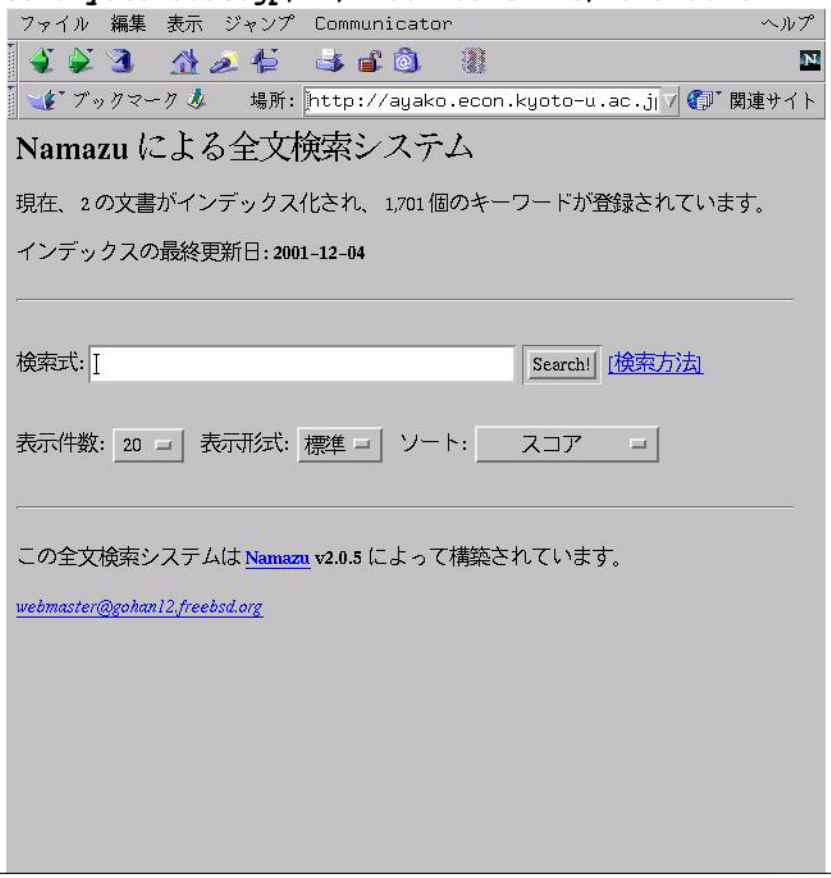
```
% cp /usr/local/libexec/namazu.cgi .
```

namazu.htmlの作成

```
% cat NMZ.head.ja NMZ.foot.ja
     | sed -e "s/{cgi}/namazu.cgi/" > namazu.html
```

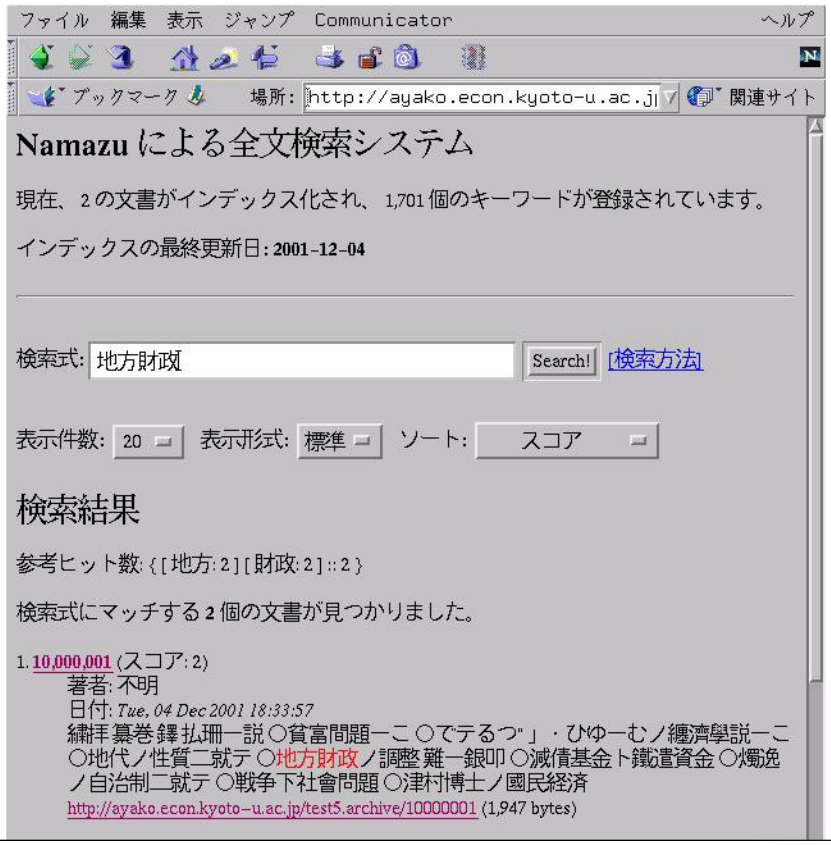
§20. 4 各自のホームページを表示してみる

http://i.econ.kyoto-u.ac.jp/~x/ml00-x.archive/namazu.html



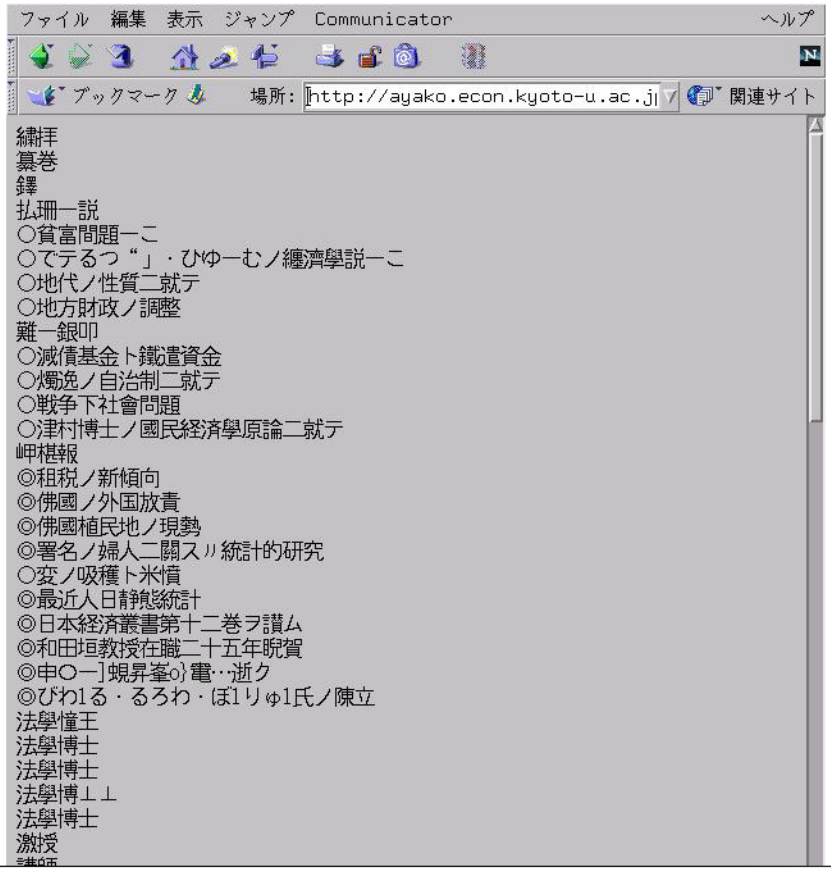
§20. 5 前回の経済論叢データベースのOCR結果を使用した例

http://i.econ.kyoto-u.ac.jp/test4.archive/namazu.cgi?query=%C3%CF%CA%FD%BA%E2%C0%AF&whence=0&max=20&result=normal&sort=score



§20. 6 検索結果の表示

OCR変換テキストそのものが表示される



§20. 7 速度の比較

Pentium-III 700MHz／RAID5 40MB/s

6898個のテキストファイル(合計3M行，168Mバイト)から，「景気循環」を検索した場合

逐次全文検索

- 検索 … 12分

Namazuによるインデックス検索

- インデックス作成 … 220分
- 検索 … 1秒未満

§20. 8 OCR結果との正しい連携

インデックス生成対象 … 認識誤りを含むテキスト

表示対象 … TIFF画像ファイル

この先は宿題. 成果をレポートに替えてもよい.

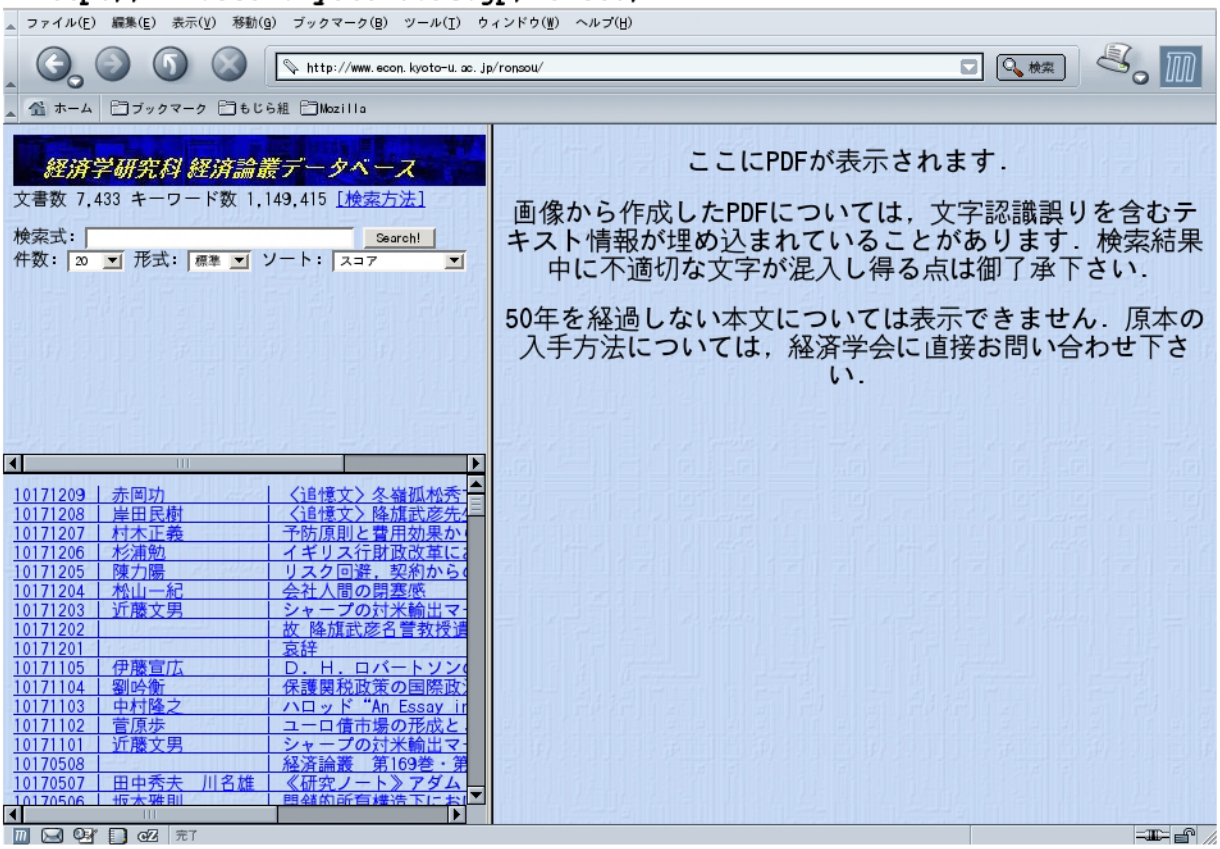
§20. 9 最近できること

画像から透明テキスト付きPDFを生成

- そのままNamazuに入力可
- ただし，それだけではタイトル･著者名がわからない
- `pdftotext -enc EUC-JP "xxx.pdf"` により，`xxx.txt`を生成し，インデックスファイルと組み合わせる.

§20. 10 経済論叢データベース(現在)

http://www.econ.kyoto-u.ac.jp/ronsou/



今日はここまで