

## 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 .
```

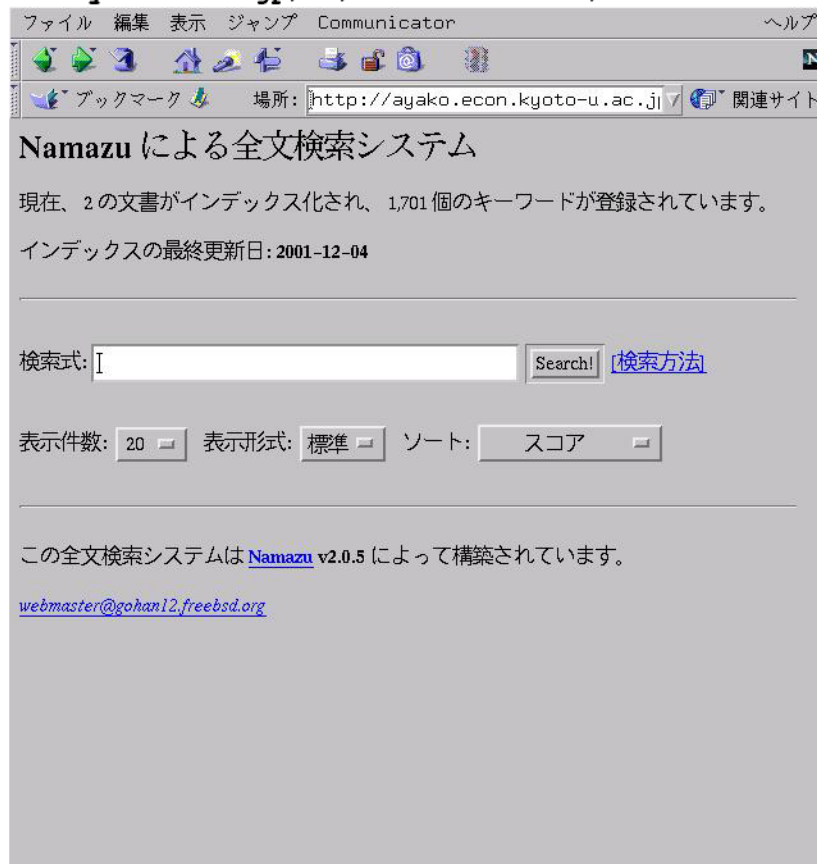
namazu.htmlの作成

```
% 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>

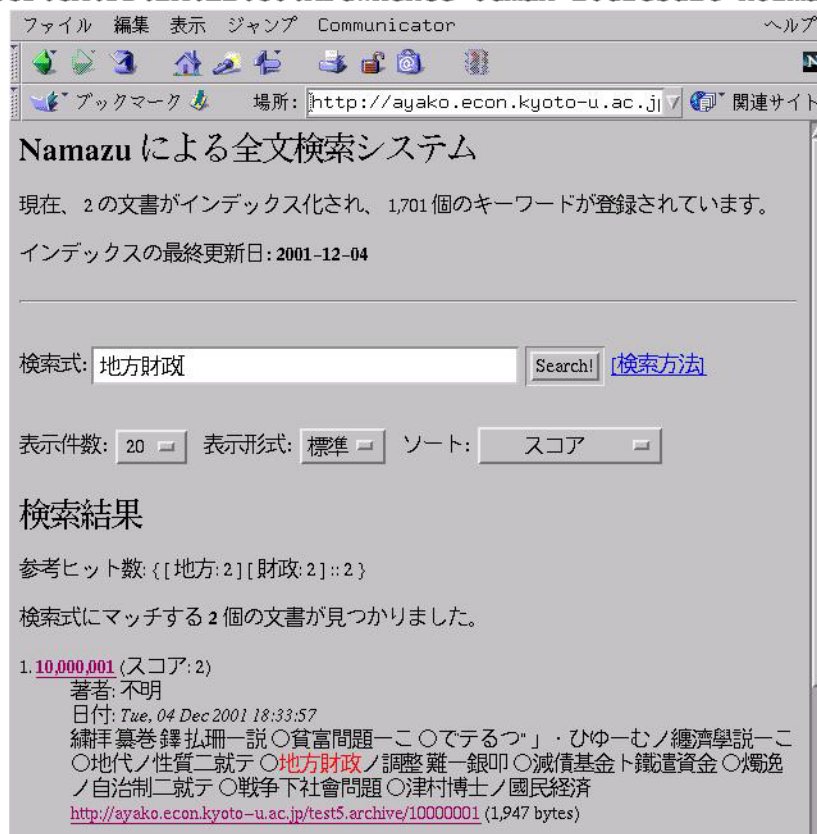


Firefox Communicator browser window showing the Namazu search system homepage. The address bar contains <http://ayako.econ.kyoto-u.ac.jp/>. The page title is "Namazu による全文検索システム". The main content includes: "現在、2の文書がインデックス化され、1,701個のキーワードが登録されています。", "インデックスの最終更新日: 2001-12-04", a search input field with "検索式:" and a "Search!" button, and a link to "検索方法". Below the search field are controls for "表示件数:" (set to 20), "表示形式:" (set to 標準), and "ソート:" (set to スコア). At the bottom, it says "この全文検索システムは [Namazu v2.0.5](#) によって構築されています。" and provides the email [webmaster@gohan12.freebsd.org](mailto:webmaster@gohan12.freebsd.org).

## §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?](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)

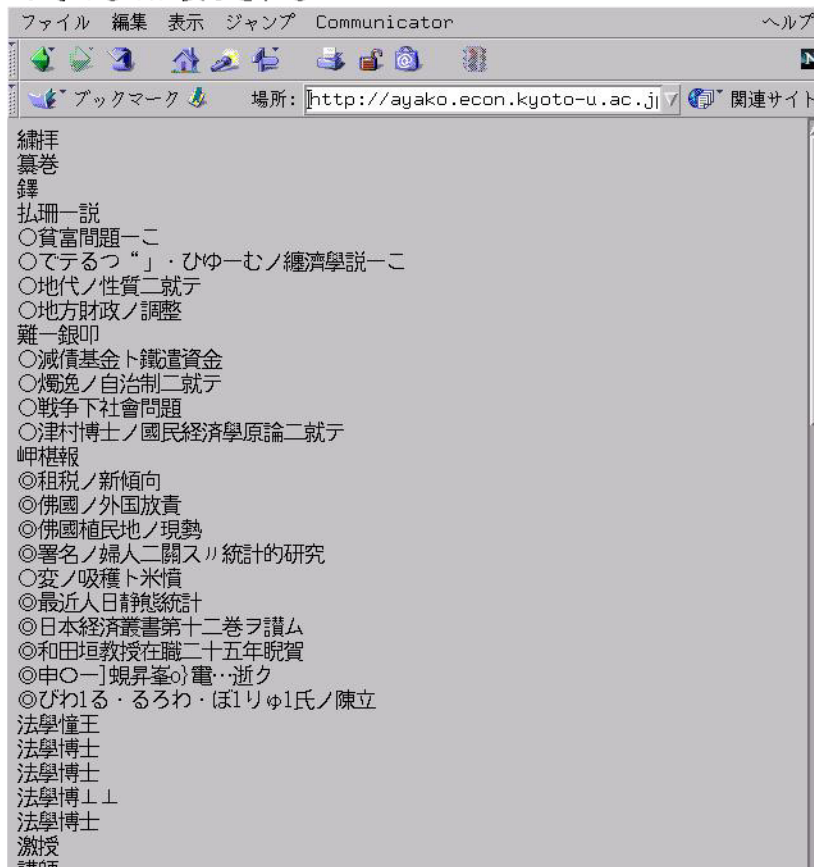
[query=%C3%CF%CA%FD%BA%E2%C0%AF&whence=0&max=20&result=normal&sort=score](http://i.econ.kyoto-u.ac.jp/test4.archive/namazu.cgi?query=%C3%CF%CA%FD%BA%E2%C0%AF&whence=0&max=20&result=normal&sort=score)



Firefox Communicator browser window showing the search results page. The address bar contains <http://ayako.econ.kyoto-u.ac.jp/>. The page title is "Namazu による全文検索システム". The main content includes: "現在、2の文書がインデックス化され、1,701個のキーワードが登録されています。", "インデックスの最終更新日: 2001-12-04", a search input field containing "地方財政" and a "Search!" button, and a link to "検索方法". Below the search field are controls for "表示件数:" (set to 20), "表示形式:" (set to 標準), and "ソート:" (set to スコア). The section "検索結果" shows "参考ヒット数: {[地方:2][財政:2]}:2". Below this, it says "検索式にマッチする2個の文書が見つかりました。" and lists one result: "1. [10,000,001](#) (スコア: 2)", with details: "著者: 不明", "日付: Tue, 04 Dec 2001 18:33:57", and a long title: "繡洋纂卷録抄一説○貧富問題一こ○でてるつ・ひゆ一むノ經濟學説一こ○地代ノ性質二就テ○地方財政ノ調整難一銀印○減債基金ト鐵道資金○燭逸ノ自治制二就テ○戦争下社會問題○津村博士ノ國民經濟". The URL <http://ayako.econ.kyoto-u.ac.jp/test5.archive/10000001> (1,947 bytes) is provided at the bottom.

## §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

### Namazul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/

The screenshot shows a web browser window displaying the Kyoto University Economics Database. The page title is "経済学研究科 経済論叢データベース". It indicates there are 7,433 documents and 1,149,415 keywords. A search bar is present with a search button. Below the search bar, there are dropdown menus for "件数" (20), "形式" (標準), and "ソート" (スコア). A table of search results is visible, listing document IDs, author names, and titles. A large text box on the right side of the page contains a notice in Japanese regarding PDF display issues.

ここにPDFが表示されます。

画像から作成したPDFについては、文字認識誤りを含むテキスト情報が埋め込まれていることがあります。検索結果中に不適切な文字が混入し得る点は御了承下さい。

50年を経過しない本文については表示できません。原本の入手方法については、経済学会に直接お問い合わせ下さい。

10171209	赤岡功	〈追憶文〉冬嶺孤松秀
10171208	岸田民樹	〈追憶文〉降旗武彦先
10171207	村木正義	予防原則と費用効果か
10171206	杉浦勉	イギリス行財政改革に
10171205	陳力陽	リスク回避、契約から
10171204	松山一紀	会社人間の閉塞感
10171203	近藤文男	シャープの対米輸出マ
10171202		故 降旗武彦名誉教授遺
10171201		哀辞
10171105	伊藤宣広	D. H. ロバートソン
10171104	劉吟衡	保護関税政策の国際政
10171103	中村隆之	ハロッド "An Essay in
10171102	菅原歩	ユーロ債市場の形成と
10171101	近藤文男	シャープの対米輸出マ
10170508		経済論叢 第169巻・第
10170507	田中秀夫 川名雄	〈研究ノート〉アダム
10170506	坂本雅則	門鎖的所有構造下にお

今日はここまで