全文フルサーチ型データベース「戦後50年 朝日新聞見出しデータベース」について

中村　英
朝日新聞社電子電波メディア局データベースセクション
〒103-11 東京都中央区築地 5-3-2
Tel: 03-5541-8552, Fax: 03-5541-8553, E-Mail: PBB02160@niftyserve.or.jp

概要

「戦後50年　朝日新聞記事データベース」は、1945年1月1日から1995年12月31日までの朝日新聞縮刷版巻頭索引を完全にデータベース化したものだ。

探したい記事の所在情報（掲載日、掲載面、縮刷版での掲載ページ）が瞬時に検索できる。

収容件数は合計307万件。思いついた言葉で検索する自由語検索のほかに、「政治」「経済」といった縮刷版索引が採用している分類語でも検索することができる。

このデータベースを全文フルサーチ型エンジンにのせたものが、朝日新聞社内で実験データベースとして稼働しているほか、検索用キーワードを付加して5枚組のCD-ROMにしたCD-ASAXも商品化されている。

キーワード

戦後50年, 朝日新聞, 見出しデータベース, 縮刷版, 全文フルサーチ型,CD-ROM

Building "Asahi Shimbun headline Database 1945-1995" by fast full-text Search Method

Ei NAKAMURA
Data Base Section of Electronic Media and Broadcasting Division, The Asahi Shimbun
5-3-2,Tsukiji, Chuoku, Tokyo, 104-11, Japan
Phone: +81 3-5541-8552, Fax: +81 3-5541-8553, E-mail: PBB02160@niftyserve.or.jp

Abstract

"Asahi Shimbun Headline Database 1945-1995" is a database of all headlines in The Asahi Shimbun reduced-size collective edition (Shukusatuban),issued monthly. When you search an article carried during the 50 years after World War II, just type a keyword, then you can get the article's publication date, page number, and colum position in the newspaper. This databse covers almost 3,070,000 headlines. We provided this database with fast full-text search method. It is used on-line system in The Asahi Shimbun. CD-ASAX, a set of 5 CD-ROMs , is also available.

Keywords

Asahi Shimbun Headline Database, Shukusatuban, full-text search, CD-ASAX, CD-ROM

1．はじめに

新聞は第一級の歴史資料であるが、それを利用するための索引は、月あるいは年単位であり、数十年を俯瞰できるものはこれまで日本の新聞界にはなかった。また、近年、外国人研究者による日本研究が盛んになってきたが、これらも多くは日本の新聞を研究のとっかかりとする場合が多い。しかし、マイクロフィルムを見ながら適切なテーマを探すというのは至難のわざで、これらが日本研究の進展をさまたげてきた。「戦後50年　朝日新聞見出しデータベース」は、これらの懸案を解決すべく、5年間の歳月を費やし完成した新聞インデックスである。現在はまだ、記事所在情報を示すだけだが、記事紙面イメージとリンクさせれば戦後の新聞ライブラリーも簡単に構築できるであろう。

2．これまでの各種索引

新聞の索引づくりには、これまで数々の試みがなされてきた。朝日新聞社を例にとれば、まず、朝日新聞縮刷版の巻頭索引がある。

朝日新聞縮刷版は大正8年(1919年）の創刊で、第二次大戦中も中断されることなく（1945、1946年の2年分は半年ごとの合本で後に製作）毎月発行されてきている。

各巻の巻頭に紙面の見出しをそのまま整理した索引がついている。「政治」「経済」といった大分類の下に「政党」「金融」などの中分類がつくといった、ツリー構造をもつ分類に従って並べてある。

しかし、この索引は1ヶ月単位であるため、昔の記事を探そうとしても、何年何月かまでを特定しないと使えない、またテーマ別に見出しを編集しているため、分類の体系をのみこまないと、なかなか目的の記事をさがしだせない、アイウエオ順で探せないなどの問題があった。

「朝日ニュース・イヤブック」というのも1973、74、75年の3年間存在した。これは新聞記事をはじめ、写真・図表、そして広告までも、関連主題（テーマ）や固有名詞から多角的に探せることをねらったものだ。1年間分が1冊で、見出しは50音順、ローマ字はアルファベット順に配列した。これは、十人近い専門スタッフをかかえた仕事であったため、オイルショック後の不況も重なって短命なものに終わってしまった。

似たような形式の索引に読売新聞社発行の読売ニュース総覧がある。

しかし、書籍の形をとる索引は、収容期間が限られ、十年とか二十年といった長期間にわたる検索ができないという制約がある。

3．戦後見出しデータベースの構想

「戦後50年　見出しデータベース」の構想に取りかかったのは、1990年の秋ごろである。朝日新聞社は1984年8月から、新聞記事データベースを構築しているが、それ以前の記事については、記事のデータベース化はもちろん索引データベースも存在していなかった。社内レファレンスを担当する旧編集局調査部（現在は電子電波メディア局データベースセクション）からも、見出しのデータベース化の要望があり、筆者をキャップとして検討を始めた。

戦後だけでも数百万件もあろう新聞記事だから、いちいち紙面をみて見出しをカードに書き抜いて整理するなどといった方式は手間からみても費用から見てもとても不可能である。

そこで着目したのが先述の縮刷版の巻頭見出し索引であった。

この索引は数人の専任編集者でつくっている。

まず「政治」「経済」といった大分類があり、その下には「政党」「内閣」といった中分類がつく。その下にも下位分類がいくつかあり、現在では6階層の分類となっている。

編集者は毎日発行される紙面から、見出しをカードに書き写し、該当する分類をつけて整理ボックスにいれておく。ある程度量がまとまると、印刷工場に送られ、活字化（文字通りの鉛活字！）され、棒組みされる。

月替わりになると、棒組みした活字を分類テーマに沿って組み直し、印刷する。電子データがなく、活字印刷というのは非常に大きな負担であった。データベース構築にあたっての作業量の半分以上はこのテキストおこしに費やされたといってもよい。

それはともかく、諸先輩に感謝すべきなのであろう。この分類は大分類、中分類、小分類くらいまでは50年たっても大きな変化はない。

少々のズレは承知ということで、この分類はそのまま採用した（「社説」などのように統一性を保つため、一部手直ししたものもある）。

またデータベース化することで、この巻頭索引の弱点であった、長期間を対象にした検索や、アイウエオ順での検索もクリアーした。

4．データベースの構造

図1は縮刷版巻頭索引の例である。ここから、どれだけのデータを人手をかけずに取り出すことができるかが、ポイントであった。

見出しデータベースは、検索対象が見出し部分だけだから、当然検索キーワードが少ない。一方、「首相」といった表現は日常茶飯事に登場するが、長い期間を対象にすると、どの首相か判別がむつかしいということになる。

少ない手がかりをもとに、どれだけ幅広く対象をすくい上げることができるか、またその逆にあいまいな多数の候補のなかからどれだけ正確にめざすものに絞り込むことができるか、二律背反の性格をもっている。

そこで大いに役立ったのが「分類」であった。

例えば「首相、年頭の記者会見（1970年1月5日付朝刊1面）」と「首相の年頭あいさつ　要旨（1975年1月1日付朝刊2面）」という見出しはは一見しただけでは、同一人物なのか別人物なのかわからない。

ところが、「分類」を使えば、この見出しははっきりと区別することができる。

1970年1月5日付朝刊は［大分類］政治［中分類］内閣［小分類］なし［サブ1分類］佐藤首相という分類がふられており、一方、1975年1月1日付の方は［大分類］政治［中分類］内閣［小分類］なし［サブ1分類］三木首相となっている。

前者では「政治、内閣、佐藤首相、首相、年頭の記者会見」

後者では「政治、内閣、三木首相、首相の年頭あいさつ」

を、検索対象とすれば

検索語　首相　and　佐藤

で後者の見出しは排除することができる。

「分類」も「見出し」も同格の検索対象として同一フィールドにおけばよい。

他方、厳密な絞り込みを要求する利用者に対してはシソーラス的な使い方もできるように、分類だけの検索領域ももうけた。

図2はCD-ROM版での検索画面だが、「全検索」欄は見出しも分類もすべを共通に検索するためのフィールドだ。「見出し」、「大分類」～「サブ3」分類までは、見出しだけ、あるいは分類だけで検索したい利用者のためのフィールドである。

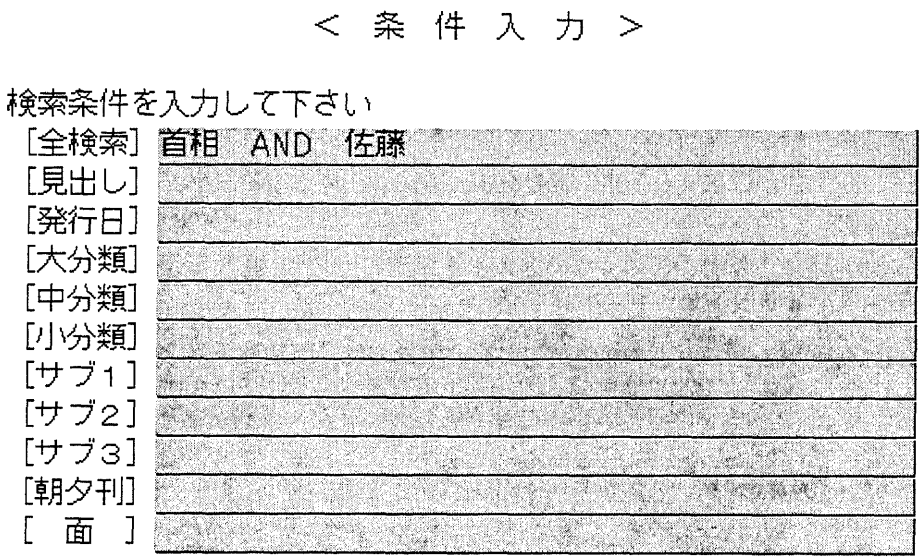

図2

フィールド同士の掛け合わせもできるようにした。

分類をテキストにするにあたっては、「大分類」は｛ ｝で囲む、「中分類」は＜ ＞で囲むというように符号で区別し、それを手がかりにBASICプログラムで見出し本文に自動付加させた。

発行日は縮刷版の掲載ページをにあたる和数字を手がかりに自動発生させた。

縮刷版は1ヶ月に1冊の発行だから、年、月までは特定できる。次に、1日発行の朝刊は○ページから、夕刊は×ページからというように図3のような変換テーブルを各月ごとに作成し、和数字の読み替えをおこない，発行日を自動作成させた。1951年10月からは掲載段も情報として表示するようになったので、これも取り入れた。

図4は検索結果の表示画面だ。上段は見出し本文、下段は掲載日、朝夕刊の別、掲載面、縮刷版における掲載ページ、記事の掲載されている段になっている。

テキストおこしは光学読みとり機(OCR)を使いMS-DOSファイルに変換、それを外注で校閲、修正してもらった。多大のエネルギーをさいたのが、このテキストおこしで、電子データが存在すれば、1年もあればこのデータベースは完成したと思う。

OCRの泣き所は判断の難しい文字でも、それらしい字を打ち出してしまうことである（警告機能のあるOCRもあるが、いちいちそれをチェックしていると入力のスピードが大幅に低下する)。例えば「ベトナム」と「ベトナム」。前者は「べ」がひらかなの「べ」、「ト」は漢字の「ト」なのだが、一見したわけでは見分けがつきにくく、校閲漏れがきわめて多かった。外注は家庭婦人をパソコン通信で組織している電

```
' 1M   1E   2M   2E   3M   3E   4M   4E   5M   5E          頁データ   01M-05E
'-------------------------------------------------
   1,   0,   0,   0,  65,   0,  89, 113, 123, 143
'
' 6M   6E   7M   7E   8M   8E   9M   9E  10M  10E          頁データ   06M-10E
'-------------------------------------------------
 153,   0, 173, 193, 203, 227, 237, 261, 271, 295
'
'11M  11E  12M  12E  13M  13E  14M  14E  15M  15E          頁データ   11M-15E
'-------------------------------------------------
 311, 341, 353, 381, 389,   0, 421, 445, 457,   0
'
'16M  16E  17M  17E  18M  18E  19M  19E  20M  20E          頁データ   16M-20E
'-------------------------------------------------
 485, 509, 519, 543, 559, 589, 601, 625, 633,   0
```

図3

鉄系の会社に頼んだのだが、当初は見落としが多く、仕事に関するパソコン通信のボードを設けるなどコミュニケーションの確立につとめた結果、次第に満足できるレベルに仕上がっていった。

このようにして出来上がったテキストをBASICでつくったパソコンプログラムにかけて出来上がったのが、図5のようなデータである。

これを全文フルサーチ型検索システムに読み込めば、データベースとなる。

5．データベースの構築

テキスト入力は1991年3月に開始したのだが、そのころは検索システムについて確としたあてがあるわけではなかった。最悪の場合は、KW切り出し方式の朝日新聞記事データベース(HIASK)の汎用機システムを使うしかないだろうが、そのうち何らかのシステムがあらわれるだろうという楽天的考えであった。

そのうち、当社はフルテキストサーチ型データベースシステム「検蔵君」を松下電器産業と共同開発することになり、その一環として見出しデータベースを構築してみることにした。

「検蔵君」はUNIXをOSに採用しており、SUNのSPARC互換機で作動する。

当初は検索エンジンを搭載した3枚のVMEバス仕様の基盤で構成されており、メモリーいっぱいにテキストを吸い込んで、一気にそれを吐き出し、その中から検索文字列と一致した文字列を含む文書を拾い上げていく、ストリーマーと名付けられた方式であった。

使用したシステムは、ワークステーションとしてSolbourne5/600(SPARC互換機、メモリー32MB、ハードディスク660MB）とイーサーネットで接続したNECの9800型パソコン1台。

それ以前に朝日新聞記事データベース構築の実験をしていたので、1ヶ月でプロトタイプが完成した。

ユーザー側としては、データベース構築は容易であった。パソコンで見出しに日付、分類を付け加え、CSV形式にしたデータをワークステーションに転送し、シフトJISコードからEUCコードに変換する。

＜ 一 覧 表 示 ＞

検索結果　595 件

1．中国には柔軟姿勢　首相、七〇年代の抱負を語る＿佐藤首相
　19700101　M　1　　1　　1
2．佐藤首相の記者会見　要旨＿佐藤首相
　19700101　M　2　　2　　3
3．きょう伊勢神宮へ　佐藤首相＿佐藤首相
　19700104　M　1　　73　　6
4．（写真）伊勢神宮での佐藤首相　豆力メラマンに、にっこり＿佐藤首相
　19700105　M　1　　89　　1
5．首相、年頭の記者会見　伊勢　国会、自民が責任　手を貸し後継者育成
　19700105　M　1　　89　　1
6．新年のメッセージ交換　佐藤首相　米大統領＿佐藤首相
　19700106　M　1　　117　　7
7．日中改善は実行で　古井氏が首相発言など批判＿佐藤内閣の対中接触政策
　19700114　M　2　　350　　1
8．社会開発を推進　首相談話＿第三次佐藤内閣発足
　19700115　M　2　　378　　1
9．佐藤首相帰京＿佐藤首相
　19700126　E　1　　709　　12
10．日中改善望む姿勢　一般的意味で発言　首相、石井氏に明かす
　19700128　M　1　　761　　1

図 4

「検蔵君」はユーザーが変換可能な項目テーブルをもっており、このテーブル書き換えを行えば、読み込んだデータがそのままデータベース化できる。ストリーマー方式ではデータを転送した段階で即座に検索ができた。インデックス方式を採用したのちでも、10 年分 (50～75MB) のデータなら数時間で構築できた。

この見出しデータベースは 1993 年 9 月、東京・池袋で開かれた「DATABASE'93 TOKYO」で 1972 年から 1992 年分を参考出展として発表した。

6．PanaSearch による 50 年分の構築

ストリーマー方式による検索は、検索エンジンの容量を超えるデータを処理するときは、大幅に検索スピードが下がってしまう傾向があった。また、特殊なハードウエアを使うため、使用できる機種が限定される、また進歩の激しいハードウエアの世界においてたえず能率アップを迫られるなどの問題があるように思われた。

ところで、松下電器では「検蔵君」のほかに、複数の全文検索型エンジンの開発がすすんでいた。

全文検索型エンジンは当時からいくつかのメーカーが開発を手がけており、日立の成分表方式など、いろいろな方法が模索されていた。

現在 PanaSearch という名前で商品化されている松下のこのシステムは、「テキストに出現する文字、あるいは文字列を照合単位とし、あらかじめ、その文字位置情報を照合単位種別にグループ化して配列した検索ファイルを用い、低出現照合単位から順に文字位置照合を行うもの」である [1]。

特殊なハードは必要とせず、すべてソフトで処理する。

1945 年から 1993 年までのデータ入力は、1994 年 8 月には完了する予定であったので、今度は「DATABASE'94 TOKYO」で 49 年分の見出しデータベースを共同出展しようということで話をすすめた。

19700104,M,15,87,10,社会,事件,交通事故,交通事故死,,,神奈川県下で十人　三が日の交通事故死__交通事故死
19700104,M,15,87,11,社会,人事,死去,訃報,,,永野護氏、矢野仁一氏__訃報
19700105,M,1,89,1,政治,内閣,,佐藤首相,,,（写真）伊勢神宮での佐藤首相　豆カメラマンに、にっこり__佐藤首相
19700105,M,1,89,1,政治,政党,日本社会党,社党再建問題,,,社党再建策練る　きょう、新年初の中執委__社党再建問題
19700105,M,1,89,1,政治,内閣,,佐藤首相,,,首相、年頭の記者会見　伊勢　国会、自民が責任　手を貸し後継者育成
19700105,M,1,89,3,外交,,,,,,招待外交　今年は、にぎやか　万国博への来日を柱に__外交
19700105,M,1,89,5,文化,教育,宗教,,,,創価学会、制度改革へ　副会長新設　政党活動を分離強化__宗教
19700105,M,1,89,6,外交,渉外,沖縄,基地労働者大量解雇と全軍労スト,,,８日に４８時間スト　きょう団交不調なら__基地労働者大量解雇と全軍労スト
19700105,M,1,89,8,本社,,社告,,,,創刊　朝日アジア・レビュー　３月１日発売__社告
19700105,M,1,89,9,政治,政党,公明党,,,,公明、きょう中央幹部会　国会対策など検討__公明党
19700105,M,2,90,1,政治,,安保問題,自衛の現実（「７０年安保」第４部）,,,（２）日本の空__自衛の現実（「７０年安保」第４部）

図５

これは、UNIX のワークステーションを使い、イーサーネットでクライアントの DOS/V パソコンとつなぎ、WINDOWS3.1 で画面を作成というものであった。

49 年分、291 万件というデータは松下電器としては初めての体験で、インデクシングにはかなり苦労したようだったが、約 300MB のデータを約 2 秒で検索するという性能を示した。

現在、朝日新聞社内で実験データベースとして稼働しているものは、このシステムを無手順のパソコン通信に対応させたものだ。普通のパソコンやワープロから市販の通信ソフトで記者であれば誰でもアクセスできる。

7．全文検索型の長所・短所

全文検索型データベースの利点は、データベースの構築が容易と言うことだ。

ワープロ文書を追加していくだけの感覚でデータベースができる。

従来のキーワード切り出し型のデータベースでは、キーワード切り出し段階で切り出しミスが発生し、その手直しが大きな負担になってきた。朝日新聞の場合、カタカナの読みもキーワードとして付加しているから（というよりカタカナ KW からスタートしたというのが本当は正しい）、正しい読み調べがさらに負荷として加わる。

シソーラスを付け加える場合も、人手が必要で、しかも個人的ばらつきが必ずしも避けられない。

全文検索型は機械的に KW をつくるので、この点大幅な省力化がはかれる。ニュースのように急ぎの場合はデータ入手とほぼリアルタイムにデータベース構築といった離れ業も可能であろう。

KW 切り出し型では、KW 増加をさけるためカットしていた「初めて」「珍しい」などの名詞以外のことばでの検索が可能である。

古典などのテキストの電子化が進めば、「あはれ」の用法にはどんなものがあるか、などは即座にデータがそろってしまう。これまで、コツコツとためてきた業績を一気にむなしいものにしてしまうおそれはあるが、それをふまえた上でまた新しい学問のあり方を切り開くであろう。

弱点はよくいわれることだが、雑音が多いということである。

「京都」を検索すると「東京都」もヒットしてしまう。「EC」をひけば「OPEC」もヒットしてしまうなどである。これはNOT検索を行えばある程度は排除できる。

異体字にも弱い。故渋沢竜彦氏は澁澤龍彦、澁澤竜彦など多くの異体字表記がありデータベース屋泣かせであるが、これはカナ検索ならば「シブサワタツヒコ」ですむ。

一つの方法は「龍」という字があれば「竜」という異体字も自動的にインデックスに発生させることだが、それでは仕組みが複雑になりすぎるかもしれない。

もう一つの方法は、こういった特殊ケースはそんなにあるものではないから、同義語辞書をつくって、入力段階で候補を提示して利用者に選択してもらう、あるいはシステムがある段階で利用者に質問を出す、といったことであろう。

8．CD-ROMの製作

一応全文検索型による見出しデータベースは完成したが、世の中に全文検索はまだ認知されておらず、ディストリビューターは依然、KW主体の検索システムを使っている。また、パソコン単体で利用しようとすればパッケージ商品のCD-ROMがやはり便利である。

そこで、さらに一歩進めてCD-ROMも製作することにした。おりしも、時代は戦後50年の節目にさしかかっており、タイミング的にも戦後50年全見出しというのは絶好のキャッチフレーズのように思えた。

1994年には、まだパソコン用CD-ROM検索ソフトでは全文検索型のものは見あたらず、結局従来型のデータベースと同じくKW切り出しを行う必要があった。図6はそのデータ例である。

検索ソフトは米国Dataware Technologies社のCD-Answerというソフトを使った。MS-DOSでもWINDOWSでもMacでも動くというのがセールスポイントであった。

予想通りというか、KW切り出し方式は多大の労力を必要とし、筆者一人でKWチェックのデスクワークを担当したが、まるまる1年半かかってしまった。発行スケジュールを守るため、当初予定していたカタカナのキーワードづけを省略し、漢字のみになってしまったのが、いささか心残りである。これをやっていたらとても1996年3月に全巻完結ということは不可能だったろうが。

CD-ROMの世界も最近は全文検索型のデータベースソフトがあらわれてきたようだ。DVDという新しい媒体もほぼ全容があきらかになってきた。

著作権などの問題が残っているが、2000年に朝日新聞20世紀全紙面電子ライブラリーができれば、というのが筆者の夢である。

参考文献

[1] 菊池忠一．日本語文書用高速全文検索の一手法．電子情報通信学会論文誌．VolJ75-D-I,No.9,p836-846(1992)

……

中村　英．朝日新聞縮刷版見出しデータベースの構築．情報管理．Vol.37,No.2(1994)

中村　英．戦後50年 朝日新聞見出しデータベースの構築．書誌索引展望．Vol.19,No.4(1995)

【一月の記事索引】

図1

7001-3604,19700104,M,15,87,10,社会,事件,交通事故,交通事故死,,,神奈川県下で十人　三が日の交通事故死__交通事故死,030,神奈川,神奈川県,神奈川県下,県,県下,下,十人,三が日,交通事故,交通事故死,死|041,社会|042,事件|043,交通事故|044,交通事故,交通事故死,死||

7001-3364,19700104,M,15,87,11,社会,人事,死去,訃報,,,永野護氏、矢野仁一氏__訃報,030,永野,永野護,護,矢野,矢野仁一,仁一|041,社会|042,人事|043,死去|044,訃報||

7001-0112,19700105,M,1,89,1,政治,政党,日本社会党,社党再建問題,,,社党再建策練る　きょう、新年初の中執委__社党再建問題,030,社党,社党再建,社党再建策,再建,再建策,策,きょう,新年,新年初,初,中執委|041,政治|042,政党|043,日本,日本社会党,社会党|044,社党,社党再建,社党再建問題,再建,再建問題,問題||

7001-0323,19700105,M,1,89,1,政治,内閣,,佐藤首相,,,(写真) 伊勢神宮での佐藤首相　豆カメラマンに、にっこり__佐藤首相,030,写真,伊勢神宮,佐藤,佐藤首相,首相,豆,豆カメラマン,カメラマン|041,政治|042,内閣||044,佐藤,佐藤首相,首相||

7001-0322,19700105,M,1,89,1,政治,内閣,,佐藤首相,,,首相、年頭の記者会見　伊勢　国会、自民が責任　手を貸し後継者育成,030,首相,年頭,記者会見,伊勢,国会,自民,責任,手,貸し,貸し後継,貸し後継者,貸し後継者育成,後継,後継者,後継者育成,者,者育成,育成|041,政治|042,内閣||044,佐藤,佐藤首相,首相||

7001-0723,19700105,M,1,89,3,外交,,,,,,招待外交　今年は、にぎやか　万国博への来日を柱に__外交,030,招待,招待外交,外交,万国博,来日,柱|041,外交|||||

7001-2837,19700105,M,1,89,5,文化,教育,宗教,,,,創価学会、制度改革へ　副会長新設　政党活動を分離強化__宗教,030,創価学会,制度,制度改革,改革,副会長,副会長新設,新設,政党,政党活動,活動,分離,分離強化,強化|041,文化|042,教育|043,宗教|||

図6