自然言語処理　114－2
（1996. 7. 18）

固有名詞に着目し記事群を整理分類し提供するシステム

増田恵子† 梅村恭司‡

† masuda@avenue.tutics.tut.ac.jp
‡ umemura@tutics.tut.ac.jp
豊橋技術科学大学 情報工学系 梅村研究室
〒441 豊橋市天伯町雲雀ヶ丘 1-1

あらまし

近年，電子化された多量のテキストを処理する必要性が高まっている．テキストの処理にはいろいろな手法があるが，我々はまず取掛りとして，固有名詞に着目して新聞記事を整理分類するシステムを作成した．本稿では，固有名詞のデータベースを用いて，記事に固有名詞が含まれるかどうかで，記事群の整理を行なった．扱っている固有名詞は人名と地名である．分類整理した記事を提供する部分では，WWW ブラウザを利用し，少ない工程で実現することができた．

キーワード　情報検索，整理分類，固有名詞，新聞記事

A browser for newspaper articles based on proper nouns

Keiko Masuda† and Kyoji Umemura‡

† masuda@avenue.tutics.tut.ac.jp
‡ umemura@tutics.tut.ac.jp
Umemura Laboratory, Department of Information and Computer Sciences
Toyohashi University of Technology
1-1, Tempaku, Toyohashi, Aichi 441, Japan

Abstract

Recently, we need to process a lot of online texts. There are various methods in the text processing. We are interested in the usefulness of proper nouns. We have developed a system that provides and lays out articles based on proper nouns. We have selected the articles which contain the proper nouns on our database and laid out the articles. The proper nouns that we used are person's names and place's names. We have used the WWW browser and realized the system quickly.

key words　Information Retrieval, Browser, Proper Noun, Newspaper

1 はじめに

World Wide Web(WWW)には，たくさんの電子化されたテキストが存在し，今後ますます増えていくだろう．ユーザーはそれを利用することができる．中には，新聞社のWWWサービスとして新聞が読めるページがある．そこでは毎日の情報を手に入れることができる．ページに提供されるニュースは現在の所，その都度更新され蓄積されない所が多い．しかし，最近では過去の新聞を検索できる所も出てきた[1]．ユーザーはその中から自分が欲しい情報を得ようとする．検索して得た情報が多くなれば，分類したり，整理することが必要となってくるだろう．

この問題から最近では，テキストを自動分類する手法が提案されている．テキストの自動分類にはいろいろな手法があるが，それらの多くはキーワードやシソーラスを用い，テキストの解析が必要とされる[2]-[6]．

本稿ではキーワードを用いずに，テキストに出現する固有名詞に着目して分類整理するシステムを試作した．固有名詞には，物事を識別する機能があり，場所・時間・ラベル(名前)という基本的な属性が備わっている．我々は，固有名詞の物事を識別する機能を用いることで，多くの情報を分類できると考えた．電子化されたテキストには，固有名詞の出現が多い新聞記事を用いた．固有名詞によって分類整理された記事を提供する部分では，WWWブラウザを最大限に利用し，少ない工程で実現することができた．この記事を提供するツールを，News Organizerと呼ぶ．

本稿では，第2節で分類・整理方法を，第3節では記事の提供法とNews Organizerの使用例を，第4節では本システムの課題について議論し，第5節でまとめる．

2 分類・整理方法

我々は，新聞記事を分類するために固有名詞を用いる．記事は，固有名詞をインデックスとして分類される．分類するまでの流れは次の通りである．

1. 記事インデックスを作成．
2. 固有名詞のデータベースを作成．
3. 記事に出現する固有名詞を探索．
   固有名詞毎の記事インデックスのリストを作成．
4. 固有名詞のサブ・インデックスを作成．

新聞記事は，毎日新聞全文記事データベースCD-毎日新聞91-94年度版を用いた．固有名詞は，今回人名と地名のみを対象にした．人名は，政治・行政・社会の各分野で活躍する日本人約19000人分に関するデータベースを用いた．地名は，日本の47都道府県名を使用した．

1. 記事インデックスは，記事名・日付・記事のテキストファイルの先頭からの位置・記事の長さで構成され，個々の記事は，記事のテキストファイルの先頭からの位置によって識別される．

2. ここでは人名・地名のうち，人名のデータベースを説明する．人名データベースは，個人の名前・読み・仕事・出身地の組から成る．

3. 固有名詞を含む記事群を取り出すためには形態素解析を用いるのが一般的だが，我々はそれを使用せず，固有名詞のトライ構造を作り，メモリ上に配置した固有名詞のデータと，記事のテキストデータとのマッチングを行なった．なぜなら，たくさんのデータベースを用いれば，テキストとのマッチングでも有効な結果が得られると考えたからである．この結果は，固有名詞とその固有名詞が記事の内容に含まれた記事群のリスト(これを記事リストと呼ぶ)とをペアとした一覧表となる．ここでの記事群のリストは，記事インデックスと同様の形式をとる．一覧表は，一覧表へのインデックスを持ち，固有名詞・一覧表へのポインタ・一覧表におけるデータの長さという形式をとる．また，人名と地名の一覧表は別々に作成する．

4. 人名の元データベースから，記事中に含まれた人名のみを抽出し，新たな人名リストを作成した．形式は元データベースと同じく，名前・読み・職業・出身地の組から成る．このリストを元に，名前の読みを五十音順に並べた人名リストと出身地で並べた人名リストを作成した．

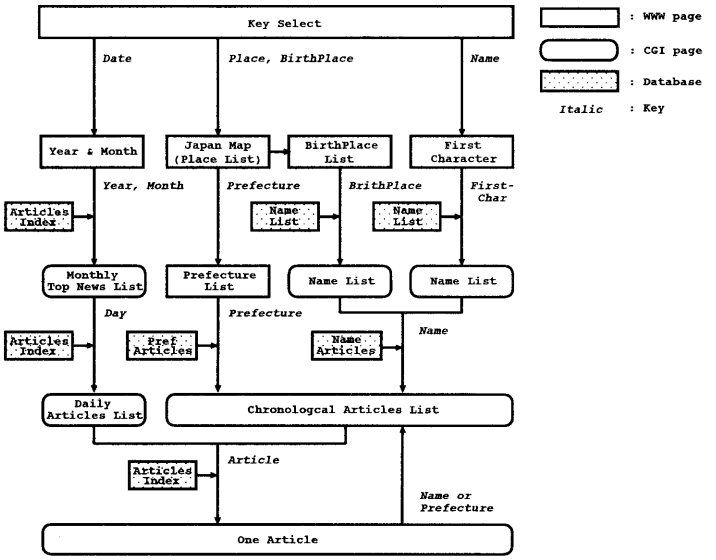

図 1: News Organizer の構成

3 システムの構造

News Organizer は新聞記事を，時間・場所・人という視点で提供する．時間は記事の発行された日付であり，全記事を対象にする．場所は地名の都道府県で，作成した地名の一覧表から提供する．人は，人名の一覧表から提供する．また人は，五十音順で並べられた名前一覧と出身地で配置した名前の 2 つの方法で提供する．システムの構成を図 1に示す.

システムは現在 WWW 上に展開しており，一部 CGI を通して提供している．CGI (Common Gateway Interface) は WWW サーバとサーバ上で動作するスクリプト / プログラムとのインタフェースで，ページを表示する際に予め何らかの処理を行ないたい場合や，他のデータベース等に処理を依頼する場合などに使用される [7]．図 1の背景が白の四角部分が，ユーザーに提供されるページである．角のある四角はいつも存在するページで，角のまるい四角は CGI を通して提供される (その都度作られる) ページである．四角の下の斜体文字は，上のページにおいて選ばれたキーを表しており，背景がドッドの四角は，システムが CGI を通して使用するデータを示している.

News Organizer は，5 段階のステップで記事を提供する．第 1 ステップは，どの視点で記事を見るかを提示する．これには日付・場所・人 (名前，出身地) があり，1 つ選択する.

第 2 ステップは，各視点でのパラメータを選ぶ時の指標を提示する．指標は，日付では年と月，場所と出身地では地方，名前では五十音を提示する．選択された指標は，引数 (キー) として次に渡される.

第 3 ステップは，各視点でのパラメータを提示する．日付では，年月から記事インデックスの日付の部分を参照し，その年月の日を全て提示する．記事を見る時の参考としてその日のトップニュースも提示している．場所では，選んだ地方の都道府県名を提示する．名前では，五十音から読みで並べた人名リストを参照し，選んだ五十音の名前を提示する．人の参考として職業と出身地も提示している．出身地では，地方から都道府県を選び，その都道府県出身者のリストを出身地で並べた人名リストから提示する．名前と同様に職業と出身地も提示している．これらからパラメータとして，日付は日，場所は都道府県，人は名前を選ぶ.

第 4 ステップは，選んだパラメータに関する記事リストを提示する．日付では，選ばれた日の記

事リストを，記事インデックスの日付の部分を参照して提示する．場所では，都道府県から地名の一覧表のインデックスを参照して，一覧表のキーにアクセスし記事リストを得る．名前と出身地では，名前から人名の一覧表のインデックスを参照して，一覧表のキーにアクセスし記事リストを得る．これらから記事を選ぶ．

最後のステップは，記事の内容を提示する．これは，記事リストに含まれる記事データのファイルの先頭からの位置と記事の長さによって，記事の内容を提示している．

図1に記事の内容から上に戻るパスがあるが，これは選んだ固有名詞の他に同じ記事に含まれた固有名詞を選ぶことができるようにしているためである．各固有名詞を選んだ時には，選んだ固有名詞の記事リストが提示される．すでに見た記事かどうかの判断は，一度見たインデックスや記事は，アンカーの色が変化するために容易にできる．ここで名前で記事を見る時の使用例を示す．

1. 3つの視点の中からnameを選択する．(図2)
2. 名前の頭文字を選択する．
   ここでは”み”を選んでいる．(図3)
3. 名前リストが提示される．
   ここでは”宮沢喜一”を選ぶ．(図4)
4. 記事リストが提示される．
   ここでは”1991年1月3日の記事を選ぶ．(図5)
5. 記事の内容を見る．(図6)

日付，場所，出身地も同様の手順で見ることができる．

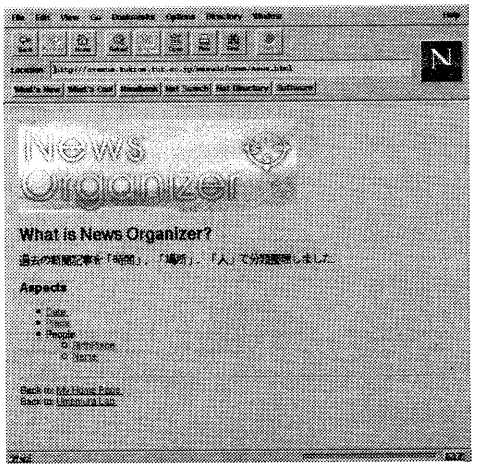

図 2: News Organizer Home Page

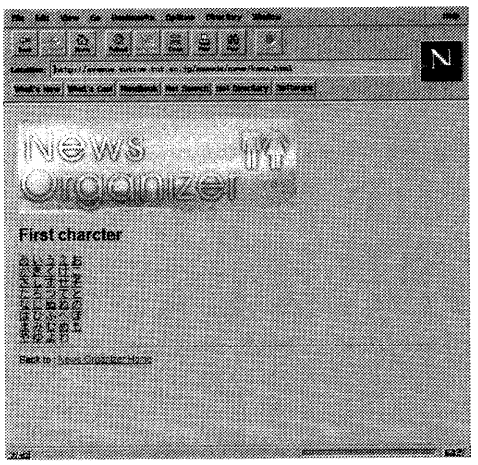

図 3: News Organizer(name) 頭文字の選択

4 課題

News Organizerには，3つの課題が残されている．1つは名前の重なりによる間違いがある．2つめは同姓同名の区別である．3つめはインデックスの追加である．

4.1 名前の重なり

例として，「千葉一」(ちば・はじめ)という名前で検出された記事があった．しかし，記事の内

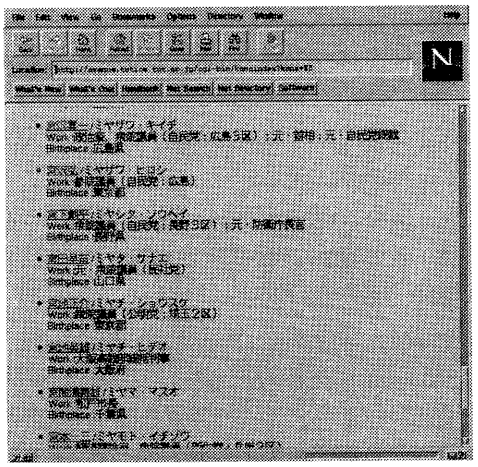

図 4: News Organizer(name) 人名リストの提示

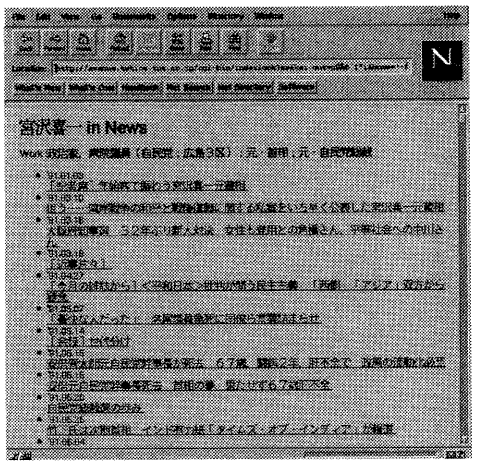

図 5: News Organizer(name) 記事リストの提示

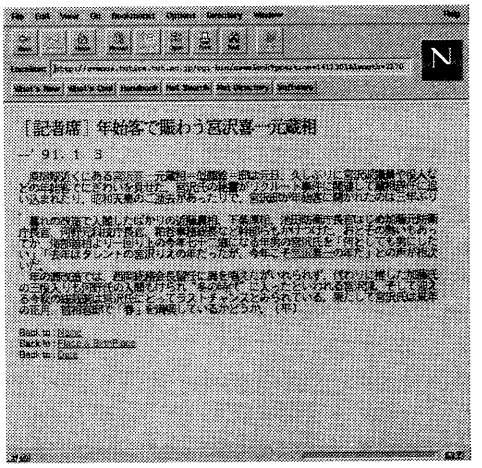

図 6: News Organizer(name) 記事の提示

容を見てみると，選挙の「千葉一区」と重なっている．

現在，新聞記事から固有名詞を含む記事を検出する時に，あえてパターンマッチのみで実行している．4 文字以上の固有名詞ならば，パターンマッチでも役立つインデックスになるが，3 文字以下の固有名詞に対しては，形態素解析を用いるのがよいと思われる [8].

4.2 同姓同名の区別

例として，「伊藤みどり」という名前で検出された記事は，スケートの「伊藤みどり」に関する記事であった．しかし，実際に使用した人名データでの「伊藤みどり」は台湾出身の人の日本名で，スケートの「伊藤みどり」ではなかった．同姓同名の区別はかなり難しい本質的な問題であり，ここをアプローチするのが，このシステムを作成した動機の一つにもなっている．

4.3 インデックスの追加

人名のインデックスとして，名前・出身地・職業を用いた．今回は付けられなかったがインデックスはその他にも，勤務地 (住所)・生年月日などがある．これらのインデックスを追加する方がより細かい整理を行なうことができると考える．また，インデックスを and や or 条件で引ける機能が

あってもよいだろう.

5 おわりに

我々は，新聞記事を固有名詞に着目して整理分類を行なった．新聞記事は人や場所に関する内容が多い，つまり固有名詞が多いためにシステムを使用して有効であるという感触を得た．しかし，客観的な評価は行なっていない．記事の提供ツールは，WWW ブラウザの利用により簡単に作成することができた.

我々は News Organizer を，WWW のレファレンスシステムの構成要素とする計画である．我々の研究室において作成したレファレンスシステム：Autoref[9] は，WWW 上のテキストを辞書引きするシステムである．ユーザーは，英文で書かれたページを Autoref を通して開くことにより，Autoref によってアンカーを加えられたドキュメントを利用する．News Organizer は，将来 Autoref のアンカー先とする予定である．これによりユーザーは，新聞記事をレファレンスに使用できるだろう.

なお，このシステムは毎日新聞社との契約により，研究室内でのみ使用している.

謝辞

本研究で使用した新聞記事データベースの使用許可を頂いた (株) 毎日新聞社に感謝します.

参考文献

[1] 朝日新聞：Digital News Archives, ⟨http://kensaku.asahi.com:1080/dnahead.html⟩

[2] 河合敦夫：意味属性の学習結果にもとづく文書自動分類方式, 情報処理学会論文誌, Vol.33, No.9, pp.1114-1122, 1992.

[3] 湯浅夏樹, 上田徹, 外川文雄：大量の文書データから自動抽出した名詞間共起関係による文書の自動分類, 情報処理学会研究報告, NL98-11, 1993.

[4] 亀田弘之, 藤崎博也：テーマ・キー概念・キーワード間の階層構造を利用する新聞記事情報の分類・検索システム, 情報処理学会論文誌, Vol.28, No.11, pp.1103-1111, 1995.

[5] 山本和英, 増山繁, 内藤昭三：分類体系相互の関係を利用したテキストの自動分類, 情報処理学会研究報告, NL106-2, 1995.

[6] 西野文人：日本語テキスト分類における特徴素抽出, 情報処理学会研究報告, NL112-14, 1996.

[7] ⟨http://w3.lab.kdd.co.jp/technotes/WWW/-CGI/aboutcgi.html⟩

[8] 田中康仁：自然言語の解析による知識獲得と拡張 - 四文字漢字列を用いて - 情報処理学会研究報告, NL94-9, 1993.

[9] 伊藤修一, 梅村恭司：WWW での辞書引き方法の比較検討, 情報処理学会研究報告, HI64-9, 1996.