

データベース構築促進及び技術開発に関する報告書

OCRを利用したキーワード自動抽出 に関する調査研究

平成6年3月

財団法人 データベース振興センター

委託先 (株) エレクトロニック・ライブラリー

KEIRIN



この事業は、競輪の補助金を受けて実施したものです。

序

データベースは、わが国の情報化の進展上、重要な役割を果たすものと期待されている。今後、データベースの普及により、わが国において健全な高度情報化社会の形成が期待される。さらに海外に対して提供可能なデータベースの整備は、国際的な情報化への貢献および自由な情報流通の確保の観点からも必要である。しかしながら、現在わが国で流通しているデータベースの中でわが国独自のものは1/3にすぎないのが現状であり、わが国データベースサービスひいてはバランスある情報産業の健全な発展を図るためには、わが国独自のデータベースの構築およびデータベース関連技術の研究開発を強力に推進し、データベースの拡充を図る必要がある。

このような要請に応えるため、(財)データベース振興センターでは日本自転車振興会から機械工業振興資金の交付を受けて、データベースの構築および技術開発について民間企業、団体等に対して委託事業を実施している。委託事業の内容は、社会的、経済的、国際的に重要で、また地域および産業の発展の促進に寄与すると考えられているデータベースの構築とデータベース作成の効率化、流通の促進、利用の円滑化・容易化などに関係したソフトウェア技術・ハードウェア技術である。

本事業の推進に当って、当財団に学識経験者の方々と構成されるデータベース構築・技術開発促進委員会(委員長 山梨学院大学教授 蓼沼良一氏)を設置している。

この「OCRを利用したキーワード自動抽出に関する調査研究」は平成5年度のデータベースの構築促進および技術開発促進事業として、当財団が(株)エレクトロニック・ライブラリーに対して委託実施した課題の一つである。この成果が、データベースに興味をお持ちの方々や諸分野の皆様方のお役に立てば幸いである。

なお、平成5年度データベースの構築促進および技術開発促進事業で実施した課題は次表のとおりである。

平成6年3月

平成5年度 データベース構築・技術開発促進委託課題一覧

分野	課題名	委託先
社 会	1 CD-ROMによるテレビ視聴率データベースの構築	(株)ビデオ・リサーチ
	2 輸入畜肉貨物の規格・重量等の検証用データベースの構築	五十嵐冷蔵(株)
	3 知的資源型データベースの調査研究	(株)ジャパンコミュニケーションズインスティテュート
	4 ビジネス雑誌記事データベースの共同構築とその利用に関する調査研究	経済文献研究会
	5 新しい電子情報サービスに関する調査研究	(株)日本経済新聞社/ (株)日経データ社
	6 研修用教材データベースのプロトタイプ作成	(財)大阪科学技術センター
	7 マイクロマシン技術情報データベースの構築調査	(財)マイクロマシンセンター
中小企業 振興地域 活性化	8 異分野研究のための知的オリエンテーション・データベースシステムの構築	(株)けいはんな
	9 関西広域データベースセンター設立のための調査研究	関西データベース協議会
	10 地域活性化のための産・学交流支援データベースのプロトタイプ作成	東北インテリジェント・ コスモス学術機構
	11 中小企業技術情報データベースの構築	(株)オーネット
	12 地域情報を対象にした分散協調型データベースシステムの開発	(株)エマーズ
海 外	13 電子デバイス情報の海外提供サービスに関する調査	電子デバイス情報 サービス(株)
	14 英日キーワード変換機能をもつデータベース検索システムの開発	カテナ(株)
	15 CD-ROMによる5カ国対訳特許用語辞典及び関連諸制度一覧の構築	丸善(株)
技 術	16 人体形状画像データ合成のための技術開発	(社)人間生活工学研究センター
	17 OCRを利用したキーワード自動抽出に関する調査研究	(株)エレクトロニック・ ライブラリー
	18 既存画像データのフォーマット変換システムのプロトタイプの作成	(株)ジー・サーチ
	19 安全研究における多重センサー・システム構築のための基本安全用語データベースの開発	(株)紀伊國屋書店
	20 データベース検索サポートシステムのプロトタイプ作成	セントラル開発(株)
		情報図書館RUKIT
	21 グループワーク支援のための分散型トランザクション管理方式の調査研究	(株)新世代システムセンター

【目 次】

1. 調査研究の概要

1.1 背景	3
1.2 目的	4
1.3 実施内容と結果の概略	4

2. 調査研究の実施結果

2.1 実施方法の概略	7
2.2 キーワード自動抽出と人手作業の比較【テスト1】	9
2.2.1 テストデータの選定	9
2.2.2 OCRによる文字データの作成	9
2.2.3 キーワード候補の抽出とキーワードの決定	9
2.2.4 キーワードの評価づけと分布	12
2.2.5 評価キーワードによる検索とヒット結果の分布	14
2.2.6 人手と自動の比較評価	17
2.3 キーワード数の調整とヒット結果の関係【テスト2】	20
2.3.1 テストの目的	20
2.3.2 テストの方法	20
2.3.3 テストの結果	21
2.4 作業効率化への寄与	38
2.4.1 効率化の測定	38
2.4.2 測定の結果	38
2.4.3 測定の限界	40

3. 成果と今後の課題

3.1 成果	41
3.1.1 精度・効率の良いキーワード自動抽出システムへの展望	41
3.1.2 作業の効率化と省力効果の確認	42
3.1.3 現実的な導入の可能性確認	42
3.2 課題と展望	43
3.2.1 品質向上対策	43
3.2.2 システム化への課題	44
3.2.3 実用化から汎用システムへ	44

4. 資料

4.1 ELNETのサービス概要	45
4.2 「MOR」登録検索式サンプル	49
4.3 テストサンプル記事の見出し一覧	51
4.4 原文イメージ、OCR読み取り原文、採用キーワードサンプル	57
4.5.1 OCR誤読の統計データ	61
4.5.2 OCR誤読補正の結果	75
4.6 自動抽出キーワードの「漏れ」「ノイズ」分析	79
4.7 インデクサー処理後の追加処理	91
4.8 チームメンバーとスケジュール	93

【図表目次】

●図2—1 テストの流れ	8
●図2—2 抽出キーワード群の分布	13
●図2—3 4種類の検索	14
●図2—4 ヒット記事群の分布	16
●図2—5 文字数当たりの抽出キーワード制限値	20
●表2—1 キーワード制限値と平均抽出数	21
●図2—6 制限値の変化に伴うヒット記事数の変化	22
●表2—2 各群のヒット記事数の変化	23
●図2—7 制限値の変化に伴うヒット記事群の構成比の変化	24
●図2—8 同上ヒット記事数の変化	25
●図2—9 同上必須記事の「漏れ」の変化	26
●図2—10 「(L) 15」と「(L) 10」のサンプル比較	27
●図2—11 同上ヒット記事数の関係図	28
●図2—12 キーワード数(増加)の変化	29
●図2—13 ヒット記事数(増加)の変化	30
●図2—14 現行作業の分担	38
●図2—15 「MOR」作業の進行パターン	39

1. 調査研究の概要

1.1 背景

株式会社エレクトロニック・ライブラリー（略称EL）は、わが国で発行されている新聞、雑誌の記事をイメージのまま網羅的・横断的に蓄積し、検索サービスに供することを目的として、1986年に設立されたデータベース・サービス会社である。新聞、雑誌社から著作権の使用許諾を得て、94年2月末現在で新聞45紙、雑誌約120誌の記事約350余万件を蓄積し、1500を超す会員に1日平均4～5万件の記事をファクシミリを通じてサービスしている。

データベースは切り抜きスタイルの記事原文イメージと、見出しやキーワードからなる書誌的事項で構成されている。

作業対象となる記事は1日平均約2,100件。記事に対するキーワード付与は人手で行っており、100人近くの人々が1日2交代で作業に当たっている。このため、データベース構築コストが高く、コンピュータによるキーワード自動抽出を導入して索引・入力作業の省力化を図ることが課題になっていた。

キーワードの自動抽出は、新聞記事データベースの領域ではすでに常識的な手法である。各新聞社が紙面製作をコンピュータ化して、記事本文を文字コード化しているからこそ可能だった。これに対して、ELの記事本文は、切り抜きを光ディスクに蓄積したものであり、単なるイメージ・データであって文字コード化していない。コンピュータ可読型でないのがキーワード自動抽出を阻む壁になっていた。

ELでは90年以降、株主社である日本電信電話株式会社技術陣の協力を得て、OCR（光学的文字読み取り装置）と日本語処理ソフトによるキーワード自動抽出の検討を行ってきた。ここ2、3年、OCRの文字認識精度と処理速度の向上には目覚ましいものがあり、新聞、雑誌記事の切り抜きを文字コード化することが現実的になった。

キーワード自動抽出導入に当たって、ELにとってはもう一つ解決しなければならない問題があった。自動抽出したキーワードによる検索結果の品質と量とである。

ELの主力サービスは、「MOR」（モーニングレビュー）と呼ぶものであって、ユーザーが検索式を登録しておく、毎朝、当日の朝刊から検索結果が自動的にファクシミリで送られる仕組みである。多くの新聞記事データベースでは、検索をすると、まず画面に見出し一覧が表示され、ユーザーはそこから必要とする記事を選ぶことになる。「MOR」でも見出しをファクシミリ出力して、そこから記事を選ぶことができるが、多くのユーザーはこの選択の過程を飛ばして、いきなり記事を受け取っている。ELの場合、検索結果の品質評価と量の測定がとくに問題になるのはこのためである。

本調査研究『OCRを利用したキーワードの自動抽出』は、直接的には以上のような背景の下で行われたものである。

1.2 目的

以上の背景から、本調査研究では、商用サービスを行っているE L社のシステム、作業態勢、サービス内容との比較検証を基本に、①現在商品化されている汎用機器とソフトの組み合わせに一定の工夫を加えることによって、商用に耐え得る水準のキーワード自動抽出が可能か②その際、どの程度の効率化が見込め、省力効果が期待できるか③実用化に当たっての課題は何か、の3点を見極めることを直接の目的とした。

最近、イメージ処理技術の進展やファクシミリの普及などによって、データベースの世界でも「イメージ志向」が高まっている。「マルチメディア時代」を控え、最も豊富な素材といえる印刷物のイメージデータベースは、その大きなすそ野として期待される。

本調査研究では、これらの状況も視野に入れつつ、効率的で高性能な印刷物のイメージデータベース構築の可能性を提示することもまた、重要な目的とした。

1.3 実施内容と結果の概略

「MOR」でサービスした記事を使ってキーワードを自動抽出し、実サービスの登録検索式を使って検索を行い、その結果と人手作業との比較を行った。

OCRで読み出された記事から抽出されたキーワード候補語は、パソコン上で自動的に「重みづけ」がなされ、やはり自動的に「絞り込み」が行われた後、キーワードとして抽出された。1記事当たり平均キーワード数10.5個のサンプルに始まって、平均キーワード数を少しずつ増やしてゆき、15.2個まで8群のテスト用サンプルを用意した。キーワード数を制限したのは、主題に絞ってキーワードを付与しているE Lデータベースの特徴を生かしながら、システムへの負荷を抑え、かつ、ユーザーに届く記事件数が急激に増えないようにするための配慮からである。

キーワードが記事内容にふさわしいかどうかは人が判断し、まず「必須」「不要」に分け、必須とも不要とも決めかねるものを「関連」とする3段階の評価を行った。

キーワード付与の後は検索の段階に進む。検索は実サービスに登録されていた1,990の検索式をそのまま使った。検索結果に対する評価は、ヒットした付与キーワードの評価群の組み合わせによって、やはり「必須」「不要」「関連」の3段階に自動的に分類した。

なお、自動抽出と人手作業を比較した相対評価に当たっては、以上の3段階に「漏れ」を加えた。「必須」のうち、人手作業だけによって発生したものは自動抽出にとって「漏れ」であり、人手作業の「漏れ」はその逆のケースである。

また、これらの品質テストとは別に、人手作業と自動抽出方式による作業量の増減を比

較し、自動抽出が全体としてどの程度の作業効率化をもたらすかを試算した。

本調査研究の結果、以下の諸点が確認でき、比較的少量のキーワードで効率が良く、かつ商用に耐えうる精度を持つ、印刷物を対象としたキーワード自動抽出システム構築への展望を得ることができた。

- ①すでに商品化されているOCR、日本語処理ソフトと既存の各種辞書をベースに、一定の補正機能と絞り込みルールを付加することによって、1記事平均8～15個前後に絞ったキーワードの自動抽出が可能である。
- ②1記事12個以上のキーワードを抽出すれば、EL社が現に商用サービスしている人手付与によるキーワードと比較しても質的に劣らず、付与量、均質性などを加味すれば、より優れた結果が得られると判断できる。
- ③これらのキーワードを使った検索結果も、商用サービスと比較して遜色がない。
- ④キーワード絞り込みのルールを操作することによって、キーワード、検索結果双方で一定の質、量の調整ができる。
- ⑤調査で使用したOCR、日本語処理ソフトなど、既存ツールの改良をはじめとして、各種システム辞書の整備・開発、キーワード絞り込みルール・修正機能の改善など、品質と処理速度にかかわる今後の開発、改善、工夫の余地が多々あり、その成果はいずれも今回のテスト結果に上積みされる。
- ⑥現行作業量で20～30%の作業効率化がはかられ、作業量が増えるにつれ省力効果はさらに拡大していくことが見込まれる。
- ⑦自動抽出方式導入に当たっては、抽出キーワード数を制限することによって検索結果の出力量の調整が可能であり、現行商用サービスの激変を避けながら緩やかに移行することができる。

2.調査研究の実施結果

2.1 実施方法の概略

調査研究では、OCR読み取りと、便宜上2段階に分けて行ったキーワード自動抽出および検索テスト、自動抽出方式による作業効率化への寄与度の試算を行い、それぞれの結果を分析・検証した。各段階の実施方法の概略は以下の通りである。

【テスト1】

EL社の商用サービス「ELNET」のうち「MOR」の実データをサンプルとし、まず以下①～④手順で1記事平均15.2個のキーワード自動抽出した後、⑤⑥の方法でテスト検索を行い、人手作業によるサービスとの全般的な比較分析を行った。

- ①OCRによる文字コード・データの作成
- ②日本語処理による名詞の自動抽出
- ③OCR誤読文字の自動修正
- ④独自の「絞り込みルール」によるキーワードの決定
- ⑤抽出キーワードの評価づけ
- ⑥評価づけキーワードを使った実サービス用登録検索式によるテスト検索

<図2-1>

【テスト2】

「キーワード抽出制限値」を変化させることによって、1記事平均15.2から10.5個まで計8段階のキーワードの抽出と検索テストを行い、キーワードと検索結果の質、量にどのような変化が生じるかを分析・検証した。

【作業効率化への寄与度の試算】

作業効率化への寄与度は、現行人手作業の分析と自動抽出方式採用による作業の増減を比較し、全体としてどの程度の効率化が見込まれるかを試算した。

【総合分析と課題の整理】

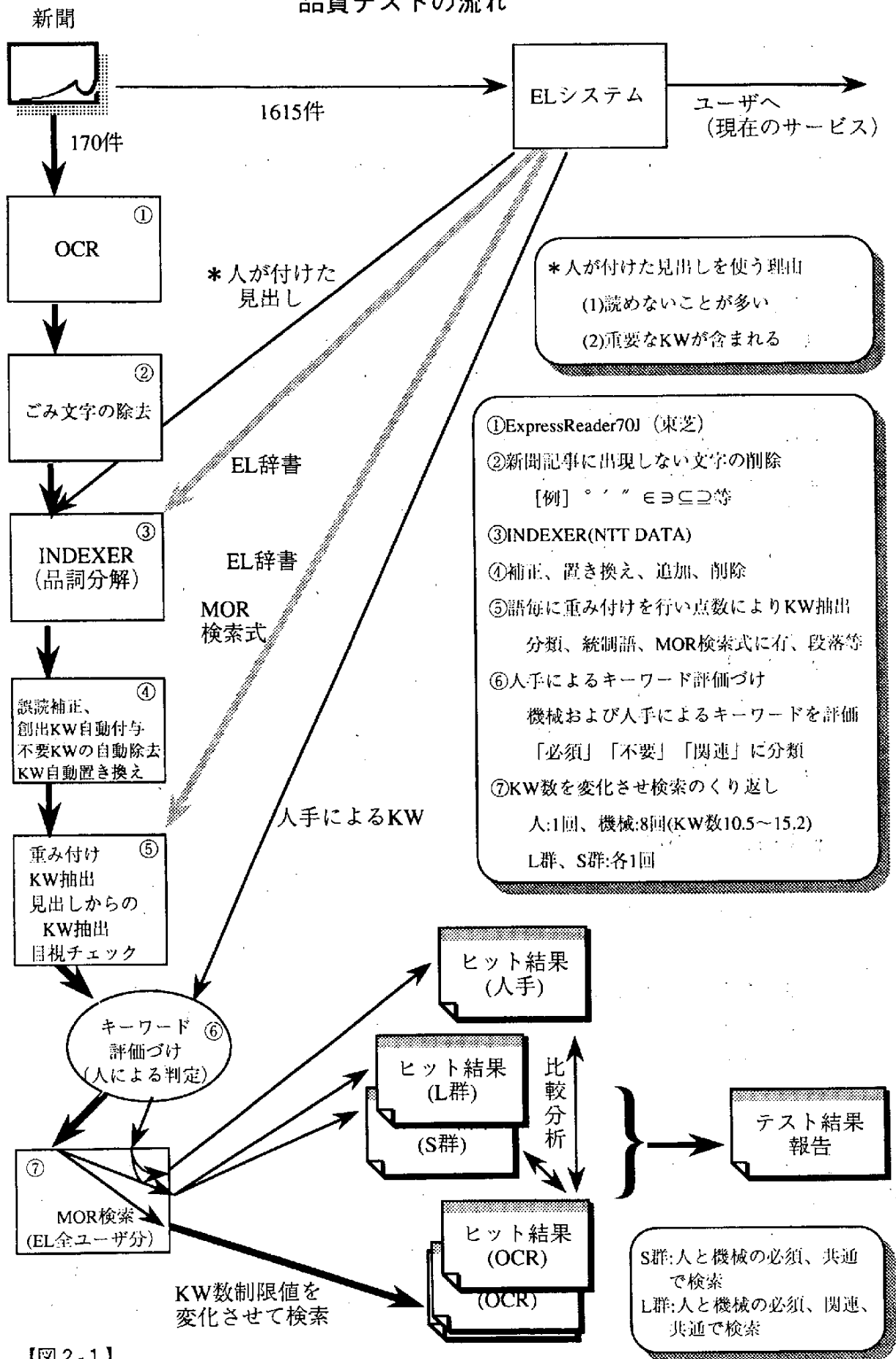
両テストの分析結果をもとに、①現行商用サービスとの品質の比較②自動抽出方式による作業効率化の度合い③今後の改善のポイントとシステム化の可能性の検討、などを行った。

【調査研究の態勢と日程】

EL社内に副社長を長とし、技術、情報、営業各担当役員を含む14名からなるチームを編成、93年7月から94年3月までの間、各種作業を行った。

(資料4.8に担当者名と実施スケジュール)

品質テストの流れ



【図2-1】

2.2 キーワード自動抽出と人手作業との比較……………【テスト1】

2.2.1 テストデータの選定

テスト用データの選定に当たっては①現行サービスとの比較検証を行うため、ユーザーの検索式があらかじめ登録されている「MOR」を対象とし②データ量は分析・検証作業の負荷を考慮して商用サービスの1日分の約1割にあたる170記事とし③一般紙、専門紙などのバランス、記事のジャンル、形式、分量などの分布に配慮してサンプル抽出した。抽出の対象は、93年5月18日（火）にサービスした朝刊25紙＜注1＞の1,615記事。
（資料4.3にサンプル記事の見出し一覧）

＜注1＞サンプル25紙 朝日新聞 読売新聞 毎日新聞 産経新聞 日本経済新聞
日経産業新聞 日経流通新聞 電波新聞 日刊工業新聞 日本工業新聞 株式新聞
ニッキン 電気新聞 化学工業日報 薬事日報 日刊建設工業新聞 建設通信新聞
日刊自動車新聞 織研新聞 日本繊維新聞 日本食糧新聞 日本農業新聞
日刊水産経済新聞 住宅新報 日本生産性新聞

2.2.2 OCRによる文字データの作成

OCRによる文字の読み取りは、東芝製の「EXPRESS READER 70 J」で行った。レイアウト自動解析のエラーは人手で修正し、記号など記事に出現しない文字は誤読として自動削除を行った。

2.2.3 キーワード候補の抽出とキーワードの決定

上記の方法で読み取った文字データを、NTTデータ通信株式会社の日本語処理ソフト「INDEXER」（NTT情報通信網研究所で開発）にかけて単語単位に分解した後、誤読文字の補正、単語の連結・分解、置き換え、文中に現われない語の追加を行った。さらに「重みづけ」による独自の絞り込みルールを使ってキーワードを決定した。

それぞれの段階で行った処理の内容は以下のようなものである。

（1）辞書

キーワード候補の抽出に当たっては、INDEXERに組み込まれている汎用辞書とEL辞書（約60万語）を併用した。

(2) OCR誤読文字の修正

単語分解の結果切り出された名詞がE L辞書にない場合、「誤読の統計データ」をもとに一部を別の文字に置き換えて再度E L辞書と照合し、辞書にあれば自動修正するという方法をとった。

(資料4.5.1に「誤読の統計データ」の一覧、
資料4.5.2に「誤読の補正の結果」)

(3) 自動または人手による一部修正

文字データからキーワード候補となる語を切り出した後、最低限必要と思われる以下の「後処理」を、自動または人手で加えた。

(資料4.7に「インデクサー処理後の追加処理」)

①置き換え・追加

- ・「米」→「米国」、「工場を建設」→「工場建設」、「発売」→「新製品」、「開発・販売」→「新製品」など、一定のルールに従い置き換えを行った(自動)。

②名詞分類などの修正

- ・「タイ」(一般名詞と地域名詞)など違った名詞分類を持つ語、INDEXERのシステム辞書とE L辞書の間で名詞分類が食い違ったキーワードなど計117語をチェックし、一部を修正した(人手)。

③「限定句」の追加

- ・「中国」のように表記も名詞分類も同じ(いずれも地域名詞)で、違った意味を持つ語12個に「限定句」<注2>を追加した(人手)。

<注2>限定句E L N E Tでは同形異義語に対し、内容を識別するためのキーワード種別(一般、地域、企業、人名)を設け、検索時に指定できるようにしている。さらに、種別が同じでも中国(地方、国名)、C D(オーディオ、支払機、預金)、ロータス(コンピュータ、自動車)のように異なった複数の意味を持つ語には「限定句」を付け、区別できるようにしている。

(4) 「重みづけ」と抽出制限式による絞り込み

INDEXERで切り出された名詞のうち、EL辞書にあるものに対し(a)のような配点基準(サンプル)で「重みづけ」をし、総合点の高い順から(b)の制限式に当てはめてキーワードとした。

(a) 重みづけ(配点例)

①名詞分類別基礎点

- ・企業名、人名、団体名……………50点
- ・一般名詞、地域名……………20点
- ・サ変名詞……………10点

②統制語加算

- ・EL辞書の統制語……………5点

③MOR加算

- ・MORのユーザー検索式にあるもの……………10点

④初出位置加算

- ・見出し……………20点
- ・第1段落……………10点
- ・第2段落(地域名、サ変名詞を除く)……………5点
- ・第3段落(同上)……………5点
- ・第4・第5段落(一般名詞、地域名、サ変名詞を除く)……………5点

⑤強調語加算

- ・「」で囲まれた強調語など……………10点

⑥分割語減算

- ・INDEXERが切り出した複合語から分割した単位単語……5点

(b) キーワード抽出数制限式

1記事当たりの抽出総数の上限は、現行システムとの関係で40個とし、具体的な抽出数は①企業名、人名、団体名は合わせて35個まで②その他の語は下の制限式から導かれる数を上限とし、双方を合わせた中から総合得点の高いもの順に40個まで採用するというルールによって決定した。

$$\text{制限式: } 2 + (\text{制限値} - 2) \times \{1 - e^{(-0.003 \times \text{記事の文字総数})}\}$$

* e: 自然対数

(式中の「制限値(L)」は【テスト1】では15に仮設定、【テスト2】では14から8まで変化させた)

以上の処理を加えた結果、対象の170記事から2,595個のキーワードが抽出された。

1記事当たりの平均キーワード数は15.2個だった。

一方、比較対象とした同じ記事に対する人手による付与キーワード総数は1,060個。

1記事当たりの平均キーワード数は6.2個だった。

(資料4.4に記事原文、読み取り原文、採用キーワードのサンプル)

2.2.4 キーワードの評価づけと分布

(1) 評価づけの方法

人手で付与された1,060個のキーワードと、自動抽出による2,595個を合わせた2,988個(うち667個は共通)のキーワードを以下の3ランクに評価、分類した。

- ・「必須」……………必要不可欠なキーワード
- ・「不要」……………検索時にノイズとなるキーワード
- ・「関連」……………「必要」とも「不要」ともいえないキーワード

評価は人によって違いが出るため、商用サービスのデータ作成に従事している検索作業当者、情報制作管理者、営業担当者の7人による評点を合算し、4人以上の判断が一致したものを「必須」または「不要」とし、それ以外を「関連」〈注3〉とした。

〈注3〉「関連」あるキーワードが必要かどうかの判断は、ユーザーの検索方法や目的によってかなりの幅があり、すべてのキーワードをすっきり「必要」「不要」に二分することは困難である。本テストでは、明らかに必要と判断できるものだけを「必須」、ノイズとなるものだけを「不要」と特定し、それ以外のグレーゾーンにあるキーワードを「関連」(「記事主題に何らかの関連を持つ語」の意味)とくくった。

(2) 評価キーワードの分布

人手によるキーワードと、自動抽出キーワードの評価づけ結果の分布は〈図2-2〉のようになり、以下のような特徴がみられた。なお、人手、自動双方に共通したキーワード(「共通」)は、評価の際、「必須」と同じ扱いにした。

(a) キーワード総数

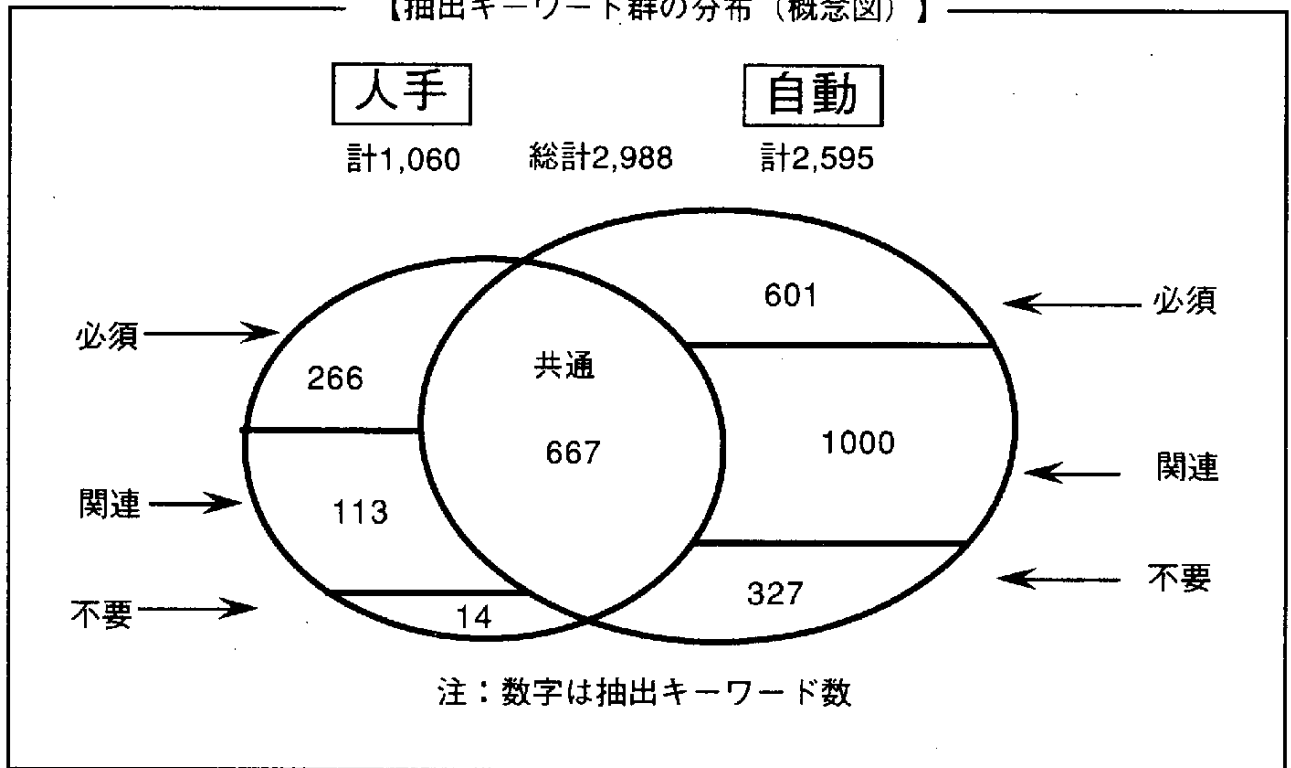
キーワード総数は人手付与の1,060(1記事平均6.2)に対し、自動抽出では2,595(15.2)と約2.5倍になった。

(b) 「必須」キーワード数

「必須」キーワード数(「共通」を含む)は、人手の933個に対し、自動抽出は1,268個で約1.4倍。

「関連」も広い意味で「必要キーワード」と考えると、その総計は人手1,046個に対し自動抽出は2,268個で、約2.2倍となった。

【抽出キーワード群の分布（概念図）】



【図2-2】

(c) 精度

人手と自動を合わせた全抽出キーワードに占める「必須」キーワードの比率をみると、人手31%に対し自動は42%で約11ポイント上回った。「必要キーワード」の比率では35%対76%となり、自動の方が41ポイント上回った。

(d) 漏れ

「必須」キーワードの「漏れ率」＜注4＞は、人手39%に対し自動は17%で、自動は人手の半分以下だった。「必要キーワード」に広げてみると、人手が60%強にのぼるのに対し、自動は14%弱で4分の1以下になった。

＜注4＞漏れ率ここでは人手と自動で付与された「必須」キーワード（「共通」を含む）の和を「必須」キーワードの総数と仮定し、これに対する人手、自動それぞれの「必須」の欠落の率をみた。

(e) ノイズ

ノイズとなる「不要」キーワードは、人手では14個（全抽出キーワードの0.5%）にとどまったのに対し、自動抽出は327個（同約11%）にのぼった。

（資料4.6に「漏れ」「ノイズ」の分析データ）

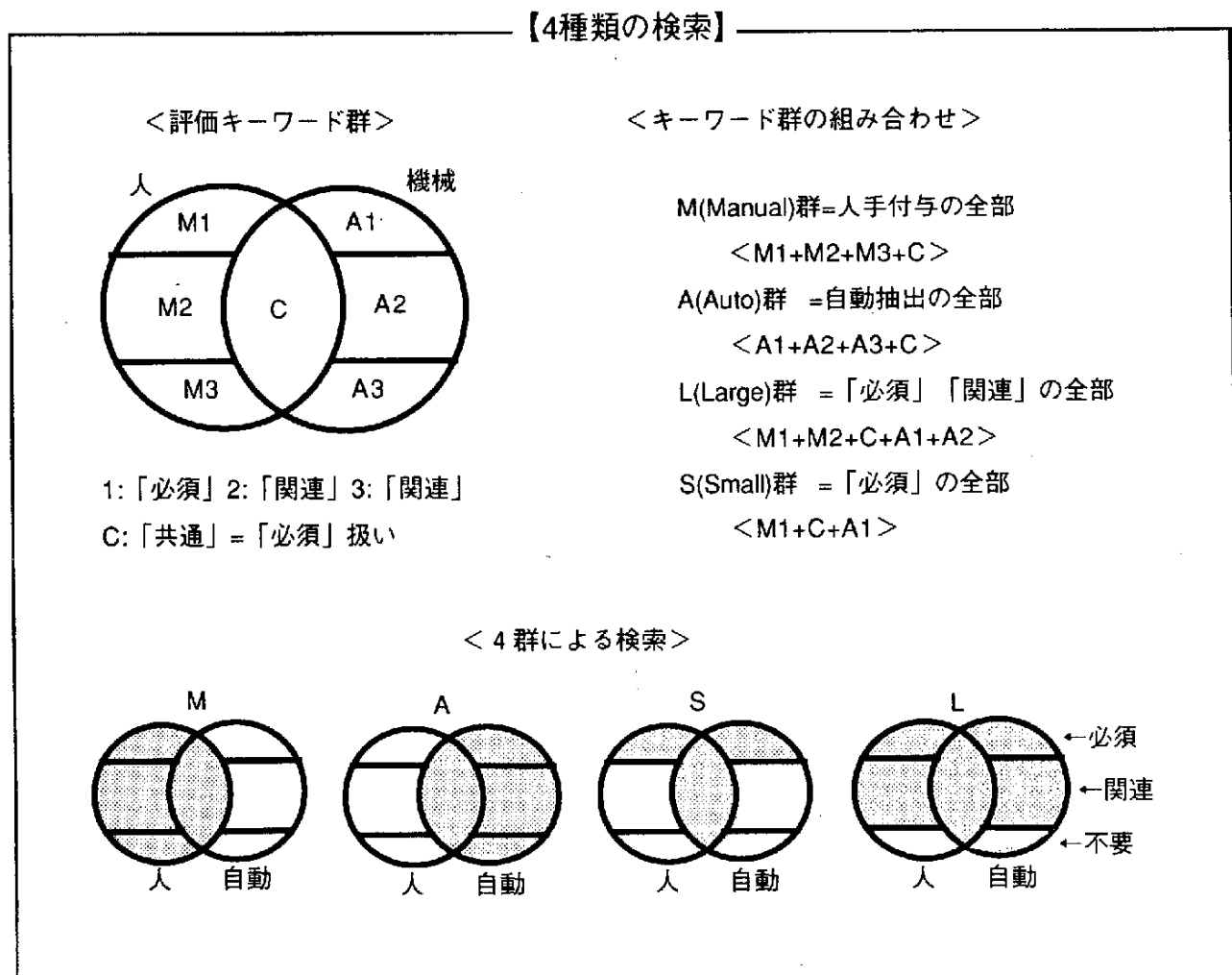
2.2.5 評価キーワードによる検索とヒット結果の分布

(1) 評価づけキーワードによる検索

人手付与と自動抽出のキーワードが実際の検索結果にどのような影響を与えるかをみるために、それぞれのキーワードを商用サービス「MOR」に登録されている1,990の検索式に当ててテスト検索を行った。

(資料4.2に検索式のサンプル)

検索は、「必須」「不要」「関連」の3ランクに評価づけされたキーワード群を、さらに<図2—3>に示す4グループにくくって行い、それぞれの群でヒットした記事のリストを照合して同じ3ランクに自動分類・集計した。



【図2—3】

(2) ヒット結果の分布

ヒット結果は、キーワードの評価づけに従って以下の方法で自動分類した。検索式に含まれる論理演算の影響を自動分類に正しく反映させるため、「S」群と「L」群の検索結果の照合を行った<注5>。

- ・「必須」……………「S」群と「L」群の検索で共にヒットした記事
- ・「不要」……………「S」群と「L」群で共にヒットしなかった記事
- ・「関連」……………「S」群または「L」群の一方のみにヒットした記事

<注5>「関連」キーワードは「必要とも不要ともいえないもの」と定義づけたため、検索結果の分類に当たっては「関連」キーワードの有無によって結果が異なる記事を「グレーゾーンにある記事」として、「関連」に入れた。

「データベース」(必須)と「エレクトロニクス」(関連)というキーワードを持つ記事を例にとれば、検索式とヒット結果、ヒット結果の分類は以下のようになる。

《検索式1》データベースORエレクトロニクス

- ヒット結果……………「S」「L」双方ともヒット
- 結果の分類……………「必須」

《検索式2》データベースANDエレクトロニクス

- ヒット結果……………「S」ではヒットせず、「L」でヒット
- 結果の分類……………「関連」

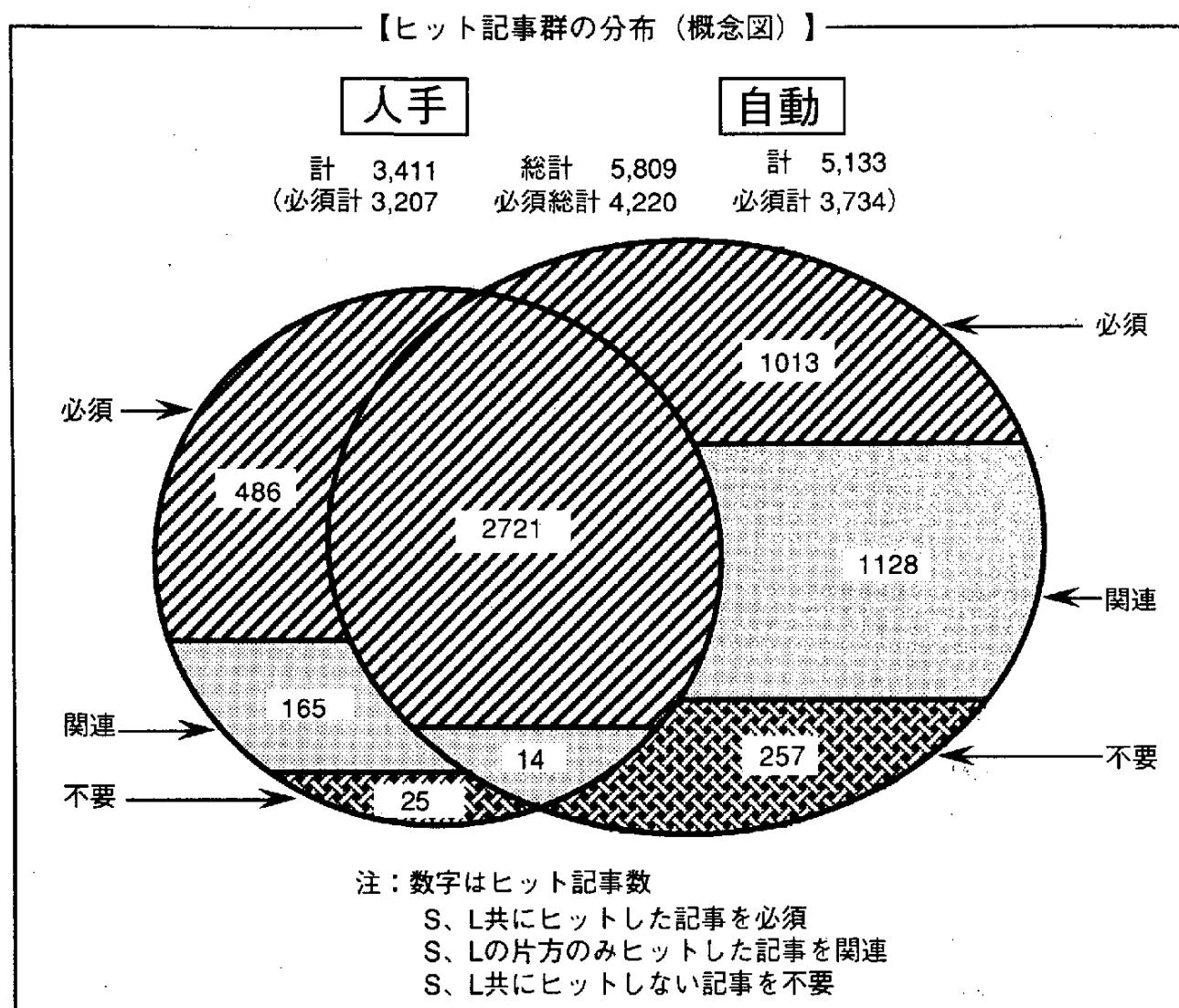
《検索式3》データベースNOTエレクトロニクス

- ヒット結果……………「S」ではヒット、「L」ではヒットせず
- 結果の分類……………「関連」

《検索式4》エレクトロニクスNOTデータベース

- ヒット結果……………「M」または「A」にはヒットしたが、「S」「L」
いずれにもヒットせず
- 結果の分類……………「不要」

ヒット記事群の分布は<図2-4>のようになり、以下に述べる特徴がみられた。



【図2-4】

(a) 総数

ヒット記事総数では人手3,411に対し自動抽出は5,133記事と約1.5倍になった。

(b) 「必要」記事のヒット数

「必須」記事は人手3,207記事に対し、自動抽出は約1.2倍の3,734記事。「関連」を含めた「必要記事」数では3,386対4,876で、自動の方が約1.4倍強になった。

(c) 精度

人手と自動を合わせた全ヒット記事に占める「必須」記事の比率（精度）は、人手 55% に対し自動は 64% で、自動の方が約 9 ポイント、527 記事多くヒットした。

「関連」も含めた「必要」記事では、人手 58% 強に対し自動は 84% 弱で、自動の方が約 26 ポイント上回った。

(d) 再現率、漏れ

人手と自動双方でヒットした「必須」の和を「必須」記事の全量とすると、再現率は人手 76.0% に対し自動は 88.5% となった。

これと裏腹の関係にある「必須」の漏れは、人手 24.0% に対し自動は 11.5%。

「関連」を含めた広い意味での「必要記事」では、人手約 39% に対し自動は約 12% だった。

(e) ノイズ

キーワードレベルでは 11% だった自動抽出のノイズは、検索結果では 257 記事で、全ヒット記事の 4.5% となった。これに対し人手の方は 25 記事、0.4% で、自動抽出の 10 分の 1 弱にとどまった。

自動の場合は「不要」キーワードがあったためにノイズとなったものがほとんどであるのに対し、人手の方は「必須」キーワードがついていなかったため、論理演算の「NOT」で除けないためノイズとなったものが大部分を占めた。

2.2.6 人手と自動の比較評価

(1) 評価の前提

本テストの場合、人手によるキーワード付与は 1 記事平均 6.2 個、自動抽出は 15.2 個であり、比較の基礎となるキーワード数に 1 対 2.5 の差があるため、単純な結果比較は客観性に欠けるといえる。

しかし、人手の平均 6.2 個は、主題キーワードのみを付与するという ELNET の基本方針の他、一定時間にこなせる作業量の物理的制約による「限界値」の側面も持つ。自動抽出の場合、このような制約は少ない。人手と自動抽出の優劣を総合的に判断するためには、キーワード付与量に対する柔軟性への視点も欠かせない。【テスト 1】ではまず、自動抽出キーワードの数を 15 個前後に仮設定し、総体的な優劣比較をみるを目的とした。

キーワード数と品質の関係の客観的な分析は、【テスト 2】の結果にもとづいて行った。

(2) キーワードレベルの評価

(a) 付与量への柔軟性

現行の人手による作業は、限られた時間内に一定の要員で一定の記事を処理するため、処理記事数が増えればキーワード数が減るという反比例の関係が生じ、キーワード数を増やすには作業時間か要員数を増やさなければならない。

これに対し自動抽出は、システムサイドからの制約を除けば付与量に制限はなく、付与キーワード数の増加にも柔軟に対応できる利点がある。

(b) 均質性の確保

人手によるキーワード付与の悩みの一つは、インデクサーの個人差による付与キーワードの「ぶれ」である。本テストに先だって行った予備調査で得られたデータによると、「MOR」の実作業と同じ環境で2組のインデクサーが33記事に対して付与したキーワードを比較した結果、双方が共通して付与したキーワードは全体の51%強にとどまり、ヒット記事数でも60余の差が生じていた。

自動抽出では本来このような事態は起こり得ず、付与キーワードの均質性の点で優れている。

(c) 「精度」と「漏れ」

前項「キーワード分布」でみたように、人手と自動では「必須」キーワードレベルの「精度」で約11ポイント、「漏れ」で約22ポイントの差があり、いずれも自動の方が優れた結果を示している。

(d) ノイズ

人手によるキーワード付与では、ノイズはほとんど出ない。この点は人手作業の最大の強み。逆に自動抽出ではノイズの率が高まることが最大の悩みである。本テストでは自動抽出のノイズ率は全体の約11%だった。

テストのサンプルデータのうち、A社の9記事に付与された商用サービス用のフリータームをテストと同じ尺度で検証してみたところ、「必須」は約10%、ノイズは50%を超えていた。

1記事に対する付与キーワード数が100個を超えるケースが多いフリーターム方式と、ELNETの主題に絞ったキーワード付与とを単純に比較はできないが、本テストの「ノイズ率11%」はかなり低い数字とはいえよう。

(3) 検索結果の評価

(a) 精度アップ

「必須」記事が全ヒット記事に占める比率は、人手55%対自動64%で自動の方が約9ポイント高く、優れた結果となった。

「関連」も含めると58%対84%となり、自動の方が26ポイント高くなった。

(b) 「漏れ」の大幅減少

自動の「漏れ」は「必須」記事群で人手に比べ半減、「関連」を合わせると3分の1以下になり、相対的にはかなりの品質向上が期待される。

ただ、人手の場合にヒットしながら自動では漏れた486記事の内容分析と、この数を可能な限り減らす対策が必要となる。

(c) 比較的少ないノイズ

ノイズについても「漏れ」と同様、単なる数字の比較だけでは済まない性格の問題だが、自動抽出方式によるノイズが、少なくとも検索結果全量の4.5%にとどまっていることは明るい展望を抱かせる。

(4) 総合評価

以上、「精度」「もれ」「ノイズ」の3つの側面からの分析で、人手の方が優れていたのは「ノイズ」が少ない点だけであり、「精度」「漏れ」での差をみれば、キーワードを1記事平均15個程度に絞っても、自動抽出の方が相対的に優れているという判断ができた。また、自動抽出が本来的に持っているキーワードの付与量に対する柔軟性、付与キーワードの均質性を併せて評価すれば、その差はさらに大きくなる、といえる。

2.3 キーワード数の調整とヒット結果の関係………【テスト2】

2.3.1 テストの目的

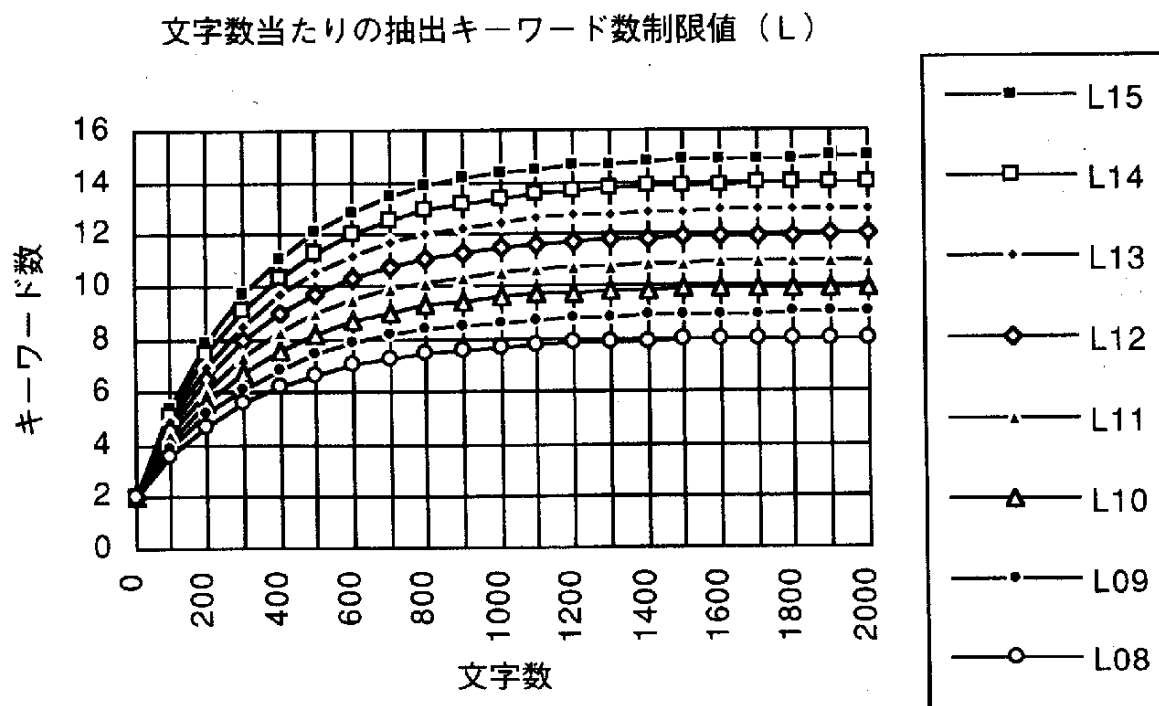
【テスト1】では、現行人手作業の約2.5倍に当たる1記事平均15.2個のキーワードを自動抽出、人手作業によるサービスとの総体的な品質の差をみた。

本テストでは、自動抽出キーワードの「制限値」を変えることによって8段階のキーワード抽出レベルを設定、1記事当たりの平均抽出キーワード数の変化が品質に与える影響を探るとともに、人手によるキーワード付与との関係を検証することを目的とした。

2.3.2 テストの方法

テストデータは【テスト1】と同じ93年5月18日付「MOR」サービスの170記事と1,990のユーザー登録検索式。キーワードの評価づけも、同じものを使った。

本テストではまず、【テスト1】で15に設定していた「キーワード抽出制限値」(L)を1つずつ8まで減らし<図2-5>、それぞれのレベルでキーワードを抽出した。



【図2-5】

次にこのキーワードを、【テスト1】で行った評価づけデータに従って「必要」「関連」「不要」の3つにグループ分けし、【テスト1】と同様に「Manual」「Auto」「Large」「Small」の4グループにくくって検索を行い、ヒット記事を「必須」「関連」「不要」「共通」の4群に自動分類した。

＜【テスト1】を行った後、本テストに当たって検索処理ソフトを一部を手直ししたため、両テスト間でヒット記事数に若干の差が生じた。今後示す数字は手直し後のものであり、(L) 15段階の数字が【図2-4】に示したものと一部食い違っているのはこのためである＞

2.3.3 テストの結果

(1) キーワード群の変化とヒット群の関係

キーワード抽出数調整の仕組みは、①企業名、人名、団体名は得点順に上限35個まで②その他のキーワードは「制限式」によって導かれる数を上限とし、③双方を合わせた中から総合得点の高い順に最大40個まで採用する、というもの。

従って抽出キーワード総数は各記事に含まれる企業、人名、団体名の数と総文字数によって変化することになるが、実際に抽出された平均キーワード数は、制限式内の「制限値」に近い数字になった＜表2-1＞。

【表2-1】 キーワード制限値と平均抽出数

制限値	L15	L14	L13	L12	L11	L10	L9	L8
	↓	↓	↓	↓	↓	↓	↓	↓
抽出数	15.2	15.0	14.4	13.7	12.9	12.1	11.3	10.5

(「L」はキーワード制限値)

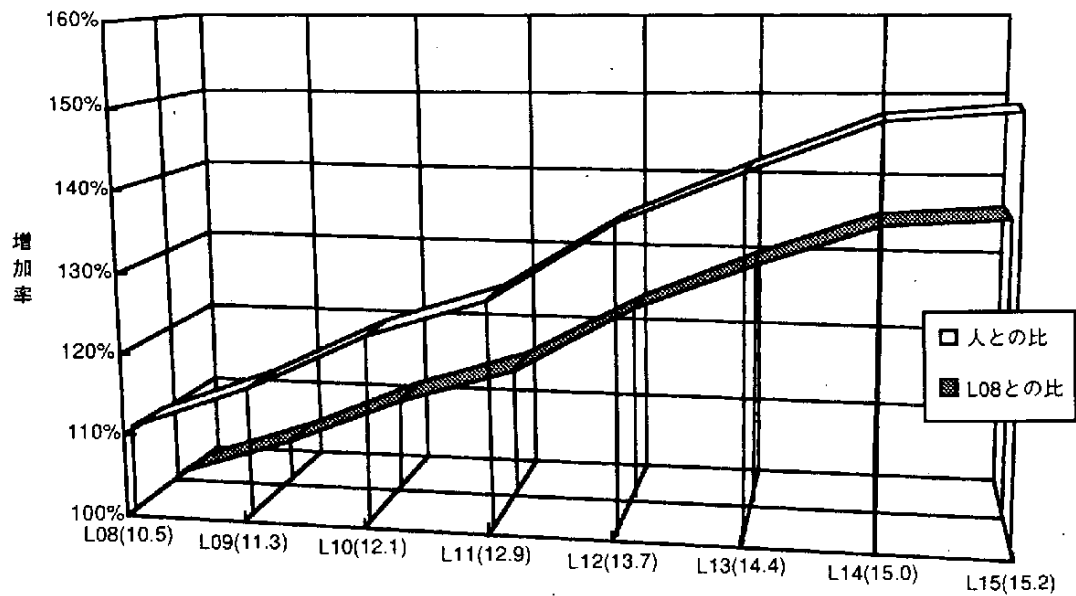
(2) ヒット総量の変化

キーワード制限値(L)を15から8まで1刻みで減らしていくと、抽出キーワード数、総ヒット記事数とも、ほぼ相似の形で減っていくことがわかった。

現行人手の場合と較べると、制限値15＝【テスト1】＝ではヒット記事総数が50%強増えたのに対し、制限値10では23%弱の増加にとどまり、制限値8では10%強まで下がることわかった＜図2-6＞。

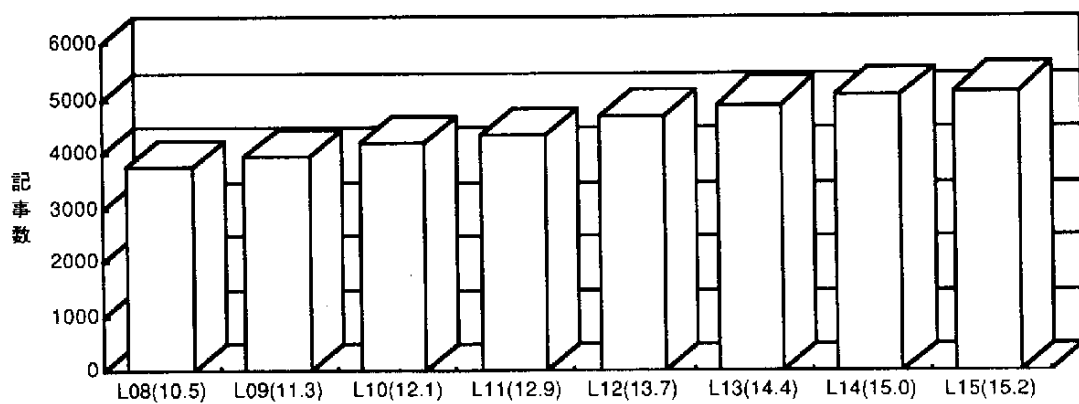
制限値の変化に伴うヒット記事数の変化< (L 8) → (L 1 5) >

増加率(人との比、L08との比)



	L08(10.5)	L09(11.3)	L10(12.1)	L11(12.9)	L12(13.7)	L13(14.4)	L14(15.0)	L15(15.2)
人との比	110%	116%	123%	128%	137%	143%	149%	150%
L08との比	100%	105%	111%	116%	124%	130%	135%	136%

ヒット記事数



	L08(10.5)	L09(11.3)	L10(12.1)	L11(12.9)	L12(13.7)	L13(14.4)	L14(15.0)	L15(15.2)
ヒット記事数	3766	3949	4183	4353	4669	4890	5089	5124

【図2—6】

(3) 「必須」「関連」「不要」群の変化

ヒット記事数が減るに従って「必須」「関連」「不要」「共通」の記事群はいずれも漸減した。

ヒット記事の実数でみると、減少カーブは「関連」が最も大きい。制限値 (L) 15～(L) 12 では「関連」の方が「必須」を上回っているが、(L) 12 と (L) 11 の間で交差し、(L) 11 以下では逆転して「関連」は「必須」を下回る<表2-2>。

【表2-2】各群のヒット記事数の変化

制限値	L15	L14	L13	L12	L11	L10	L9	L8
必須	1,015	1,010	945	885	782	736	698	675
	(100.0)	(99.5)	(93.1)	(87.2)	(77.0)	(72.5)	(68.8)	(66.5)
関連	1,117	1,088	1,028	917	765	676	590	490
	(100.0)	(97.4)	(92.0)	(82.1)	(68.5)	(60.5)	(52.8)	(43.9)
不要	259	257	220	192	167	165	143	132
	(100.0)	(99.2)	(84.9)	(74.1)	(64.5)	(63.7)	(55.2)	(51.0)

<カッコ内は制限値 (L) 15 のヒット記事数を 100 とした比率>

また、(L) 15 から (L) 8 までの構成比の変化をみても、「関連」の減少率が最も大きく、次いで「不要」「必須」の順になっていることがわかる<図2-7>。

これは、「必須」キーワードが制限値の低い段階で採用される比率が高く、制限値が緩和されるに従って「関連」キーワードの採用率が相対的に高まり、これがヒット記事数に反映されることを物語っている。

制限値の変化に伴うヒット記事群の構成比の変化< (L) 8 → (L 15) >

	共通	必須	関連	不要	合計
L08(10.5)	2469	675	490	132	3766
	65.6%	17.9%	13.0%	3.5%	

L09(11.3)	2518	698	590	143	3949
構成比	63.8%	17.7%	14.9%	3.6%	
増加率	-1.8%	-0.2%	1.9%	0.1%	

L10(12.1)	2606	736	676	165	4183
構成比	62.3%	17.6%	16.2%	3.9%	
増加率	-1.5%	-0.1%	1.2%	0.3%	

L11(12.9)	2639	782	765	167	4353
構成比	60.6%	18.0%	17.6%	3.8%	
増加率	-1.7%	0.4%	1.4%	-0.1%	

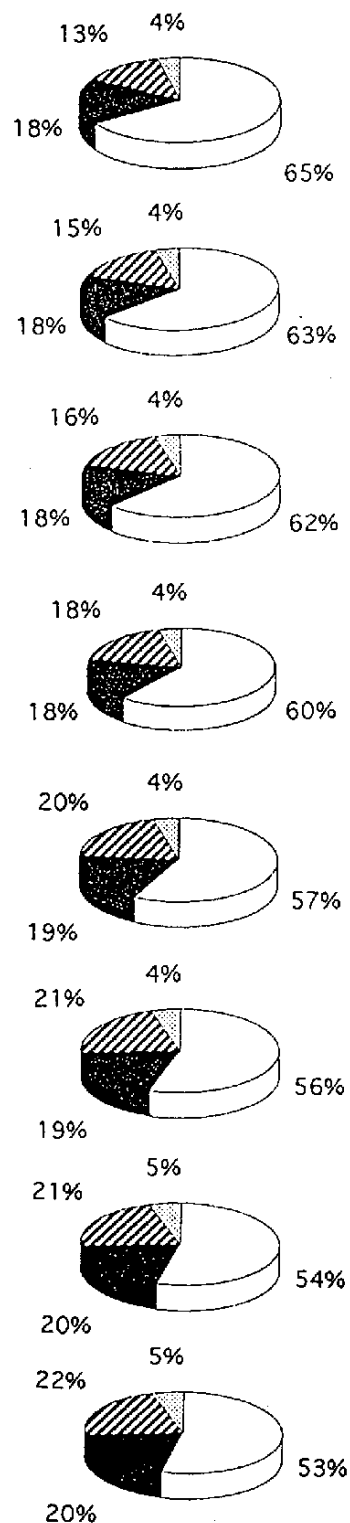
L12(13.7)	2675	885	917	192	4669
構成比	57.3%	19.0%	19.6%	4.1%	
増加率	-3.3%	1.0%	2.1%	0.3%	

L13(14.4)	2697	945	1028	220	4890
構成比	55.2%	19.3%	21.0%	4.5%	
増加率	-2.1%	0.4%	1.4%	0.4%	

L14(15.0)	2734	1010	1088	257	5089
構成比	53.7%	19.8%	21.4%	5.1%	
増加率	-1.4%	0.5%	0.4%	0.6%	

L15(15.2)	2733	1015	1117	259	5124
構成比	53.3%	19.8%	21.8%	5.1%	
増加率	-0.4%	0.0%	0.4%	0.0%	

構成比



【図2-7】

□ 共通 ■ 必須 ▨ 関連 ▩ 不要

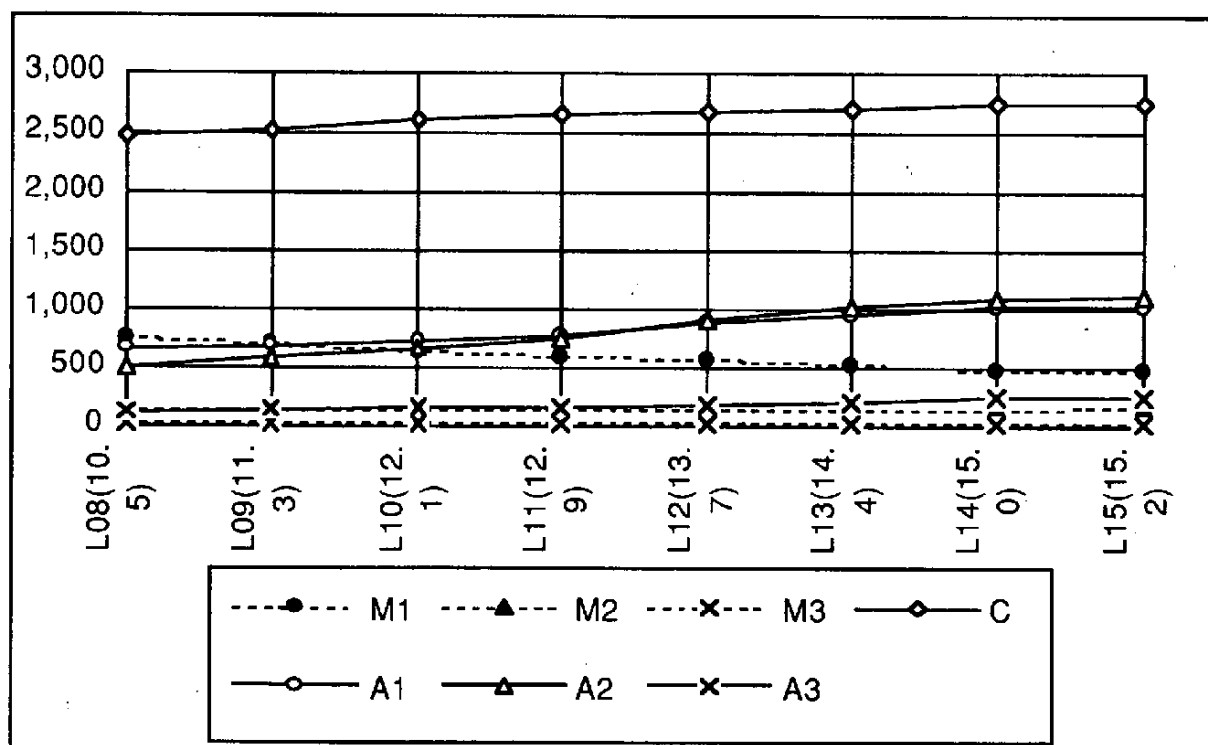
【図 2—7】

人手付与キーワードによるヒット結果との比較でみると、「必須」は平均抽出キーワード数が減るに従って人手との差が狭まり、平均 11.3 個 (L 9) 付近で交差し、平均 10.5 個 (L 8) では逆転し人手の方が上回る。

一方、自動抽出の「不要」部分も、緩やかなカーブではあるが、キーワード数とともに減少していくことがわかる<図 2—8>。

以上の点から、キーワード抽出制限値を変化させることによって、「関連」「不要」記事の抑制が一定程度可能であることがわかったが、キーワード抽出数が 11 個以下では自動抽出の品質上のメリットが失われることも明らかになった。

制限値の変化に伴う「必須」「関連」「不要」「共通」各群のヒット記事数の変化



M:人、A:自動、1:必須、2:関連、3:不要、C:共通

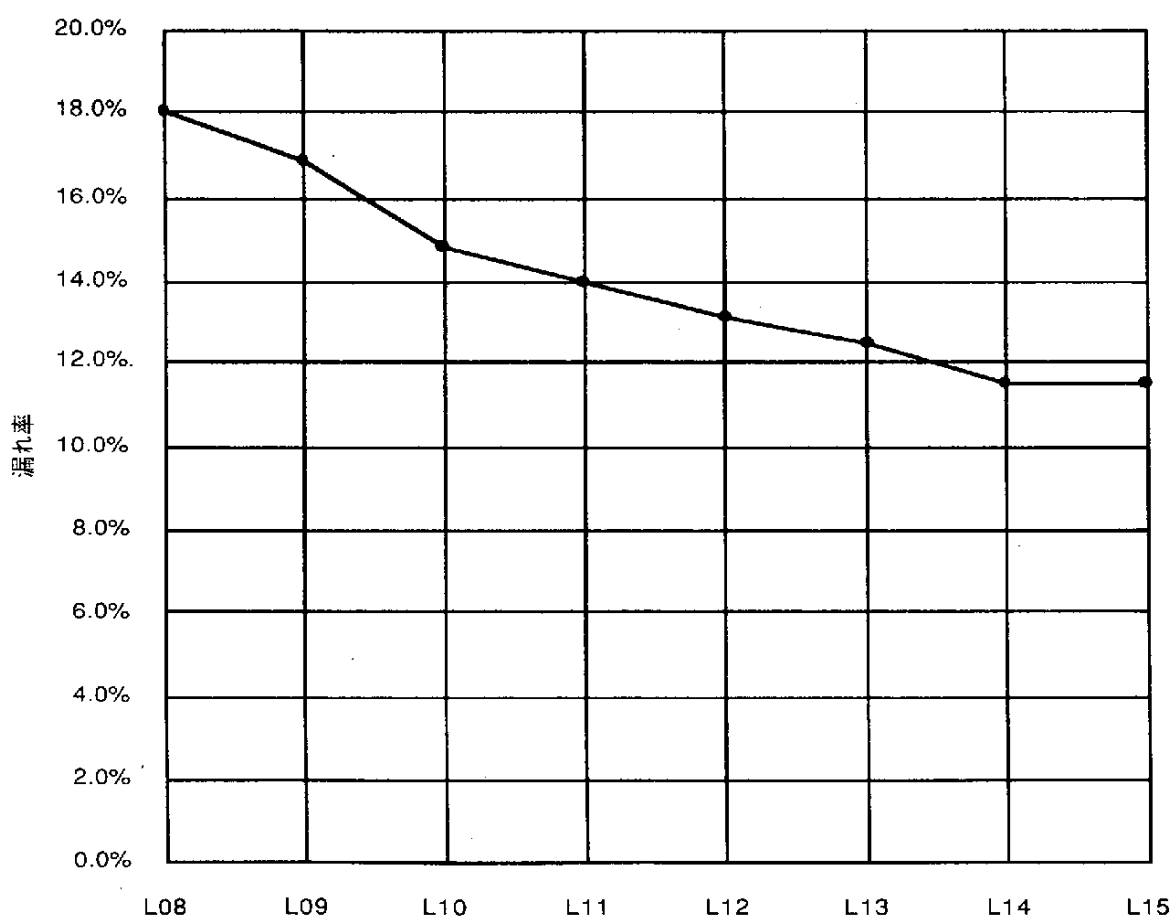
【図 2—8】

(4) 「漏れ」の変化

必要記事の「漏れ」は、キーワード抽出数が減るのに反比例する形で増える。「必須」に限定して漏れの変化をみると、平均キーワード15.2（L15）では488記事（自動、人手を合わせた「必須」記事全体に占める比率は11.5%）だったが、キーワード12.1レベル（L10）では626記事（同14.8%）、10.5レベル（L8）では763記事（同18.1%）に達する＜図2－9＞。

「漏れ防止」の観点からとらえると、キーワード制限値の絞り込みは自動抽出の利点を減殺することがわかる。

制限値の変化に伴う「必須記事」の漏れの変化



	L08	L09	L10	L11	L12	L13	L14	L15	Smallのヒット 総数
M1	763	714	626	591	555	530	489	488	4236
漏れ率	18.0%	16.9%	14.8%	14.0%	13.1%	12.5%	11.5%	11.5%	

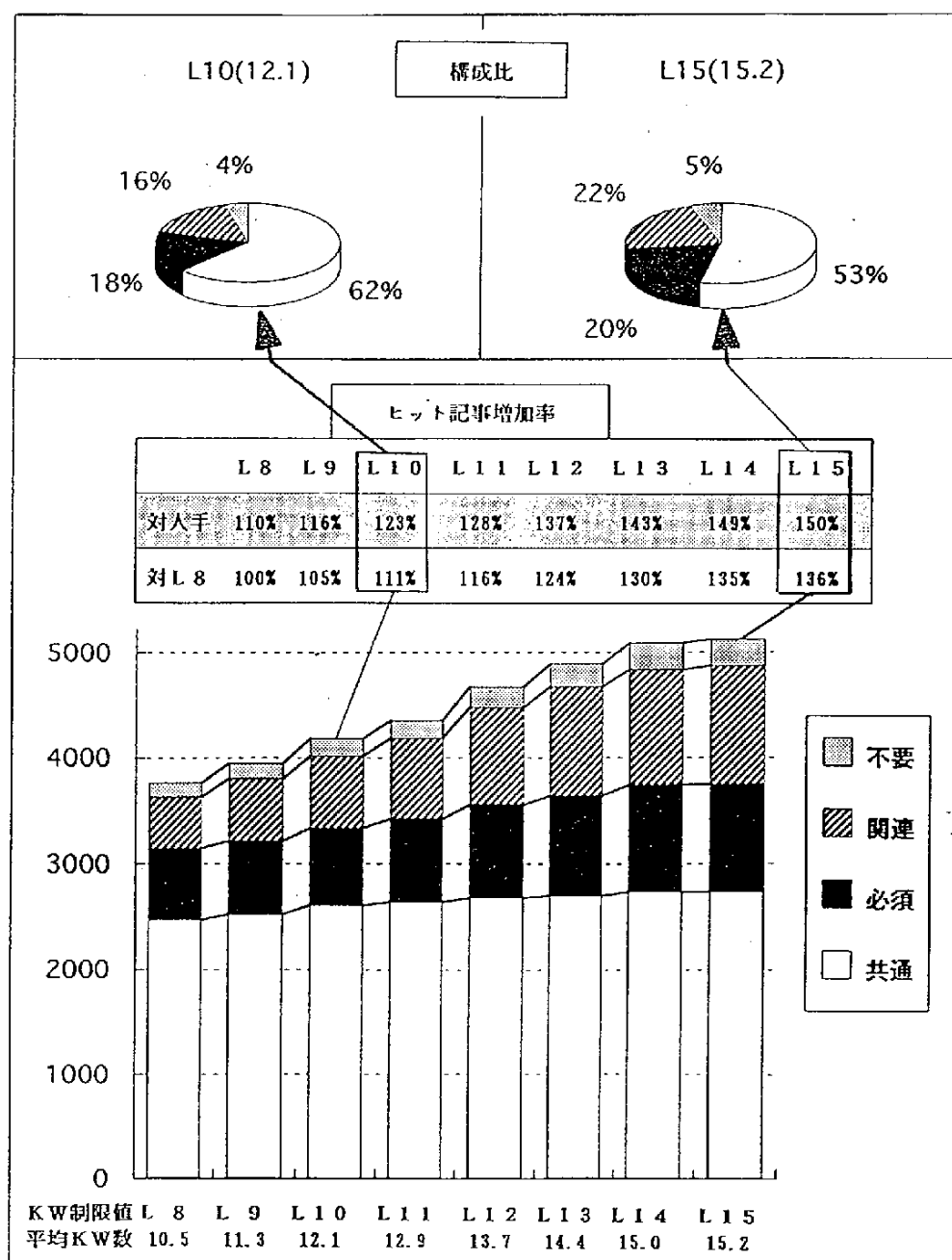
【図2－9】

(5) 「L15」と「L10」の比較

モデルケースとしてキーワード制限値 (L) 15と (L) 10の場合を比較してみると、おおむね以下のような点があげられる。

- ①人手と比較したヒット記事数増加は、(L) 15では約50%だったが (L) 10では約23%と、ほぼ半減した<図2-10>。

制限値 (L) 15と (L) 10のサンプル比較

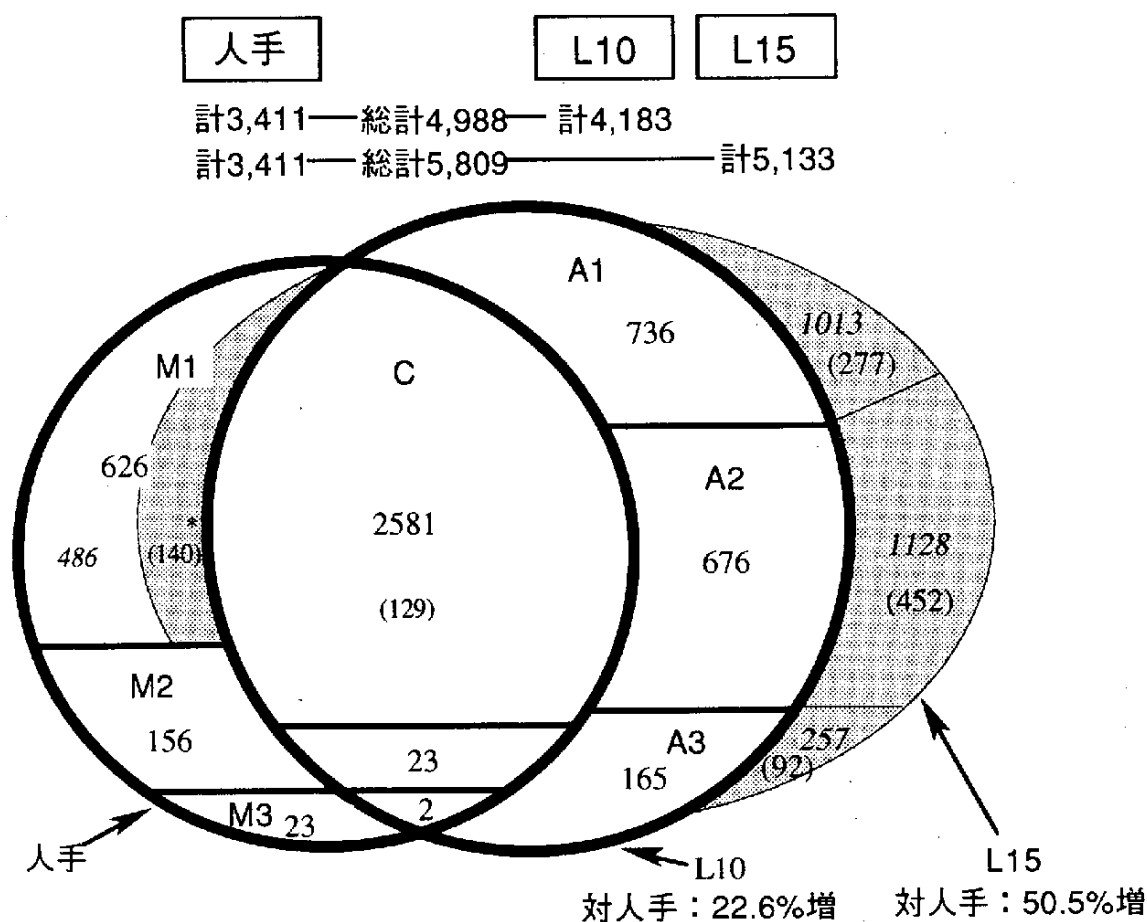


【図2-10】

② (L) 10は全体としてみると (L) 15の円が縮んだ形で、質的に大きな変化はみられないが、各部分の減少率をみると「必須」約27%、「関連」約40%、「不要」約36%となっており、「関連」部分と「ノイズ」の抑制に一定の効果がみられる。

③ (L) 10に対する (L) 15の膨らみ部分を人手と重ねてみると、人手との共通部分はわずかで、自動抽出分が大きいことがわかる。これは、抽出キーワード数を増やせば増やすほど、人手ではとらえきれない部分を大幅にカバーできるが、反面「関連」「不要」部分のヒット記事も増大することを示している<図2-11>。

制限値 (L) 10と (L) 15のヒット記事数の関係<概念図>



*()内の数字はL10とL15の差
 M1の140はM1,M2,M3の合計 129+9+2=140
 Cの129はC1,C2,C3をそれぞれ合計した差 (2721+14)-(2581+23+2)=129

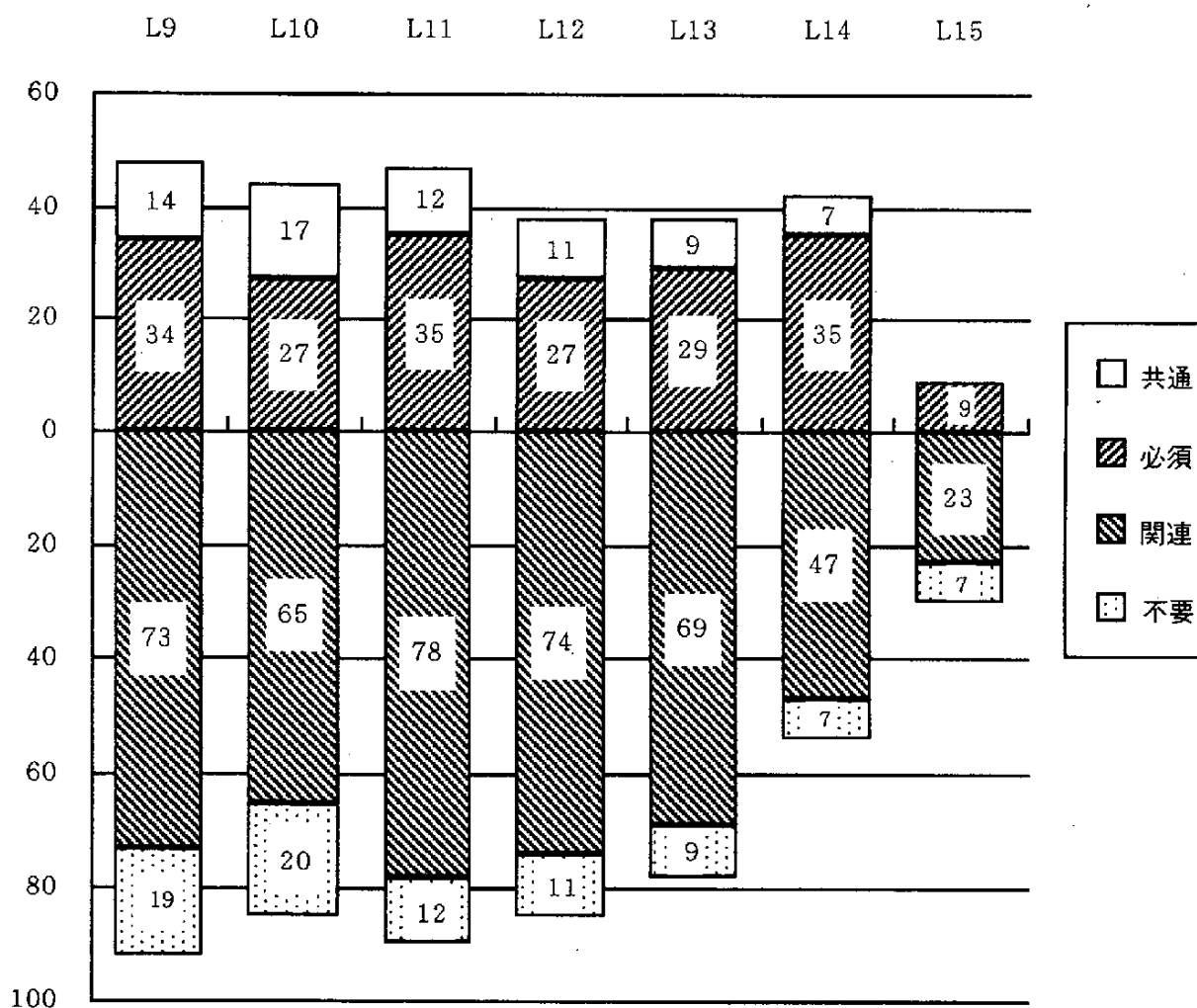
【図2-11】

(6) キーワード群の変化とヒット記事との関係

本テストで抽出されたキーワード群とヒット記事との関係を見ると、おおむね以下のよう傾向がわかった。

- ①「必須」キーワードは、キーワード抽出制限値が1増えるごとに30前後ずつ増えていく<図2-12>。

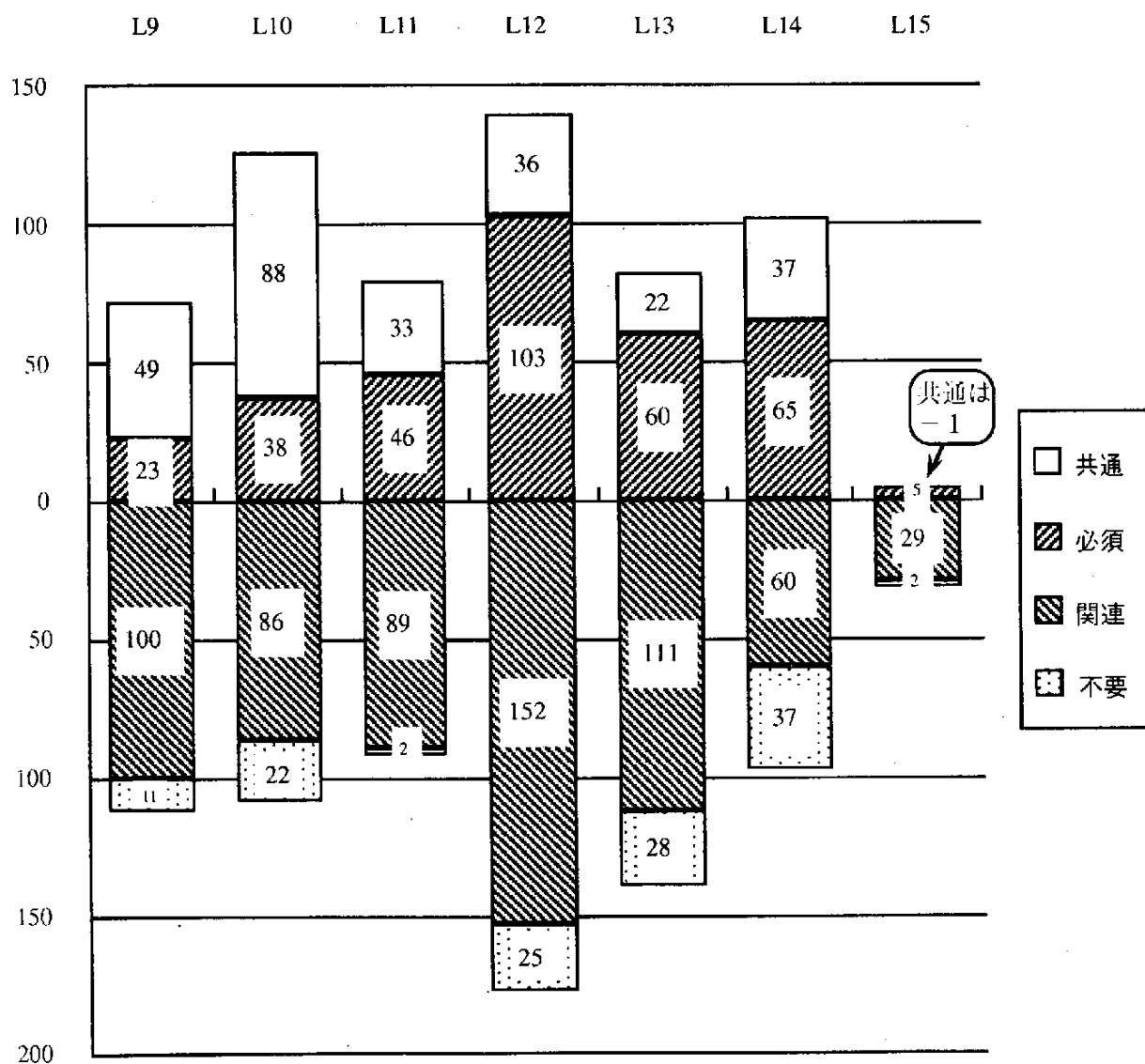
KW数（増加）の変化



【図2-12】

- ②「必須」キーワードによるヒット記事数は、制限値が増えるに従って23→38→46→60→65…と右肩上がりに増えていく<図2—13>。
- ③「関連」部分のキーワードも、「必須」のほぼ倍に当たる60～70ずつ増え、ヒットも「必須」とほぼ同様の傾向を示す。

ヒット記事数（増加）の変化



【図2—13】

(7) 個別記事別にみたキーワードとヒットの関係

キーワード制限値の調整の結果が、採用キーワードとヒット記事件数にどのような変化をもたらしたかを、具体的なサンプル記事に沿って検証すると、①制限値(L)を8から15まで1ずつ上げていくと、170記事のほとんどでキーワードが1個ずつ増えた②新たに付与されたキーワードが検索式内でNOT演算の対象となってヒット件数を減らすケースもあるが、ほとんどの記事ではキーワード増がヒット増につながった、などの点がわかり、さらにヒット件数増と具体的なキーワードの関係がかなりクリアに把握できた。

《例1》記事番号11読売新聞コラム「編集手帳」

(L) 8段階で26件ヒットしたが、これに関与したキーワードは「ダイエー」(18件)、「編集手帳」(3件)、「脱税」(2件)、「自民党」(2件)。1件はキーワード以外の新聞コードによるヒットだった。

(L) 9で「全国調査」が増え4件、(L) 13で「所得税」で2件、(L) 14では「ゴルフ会員権」で7件のヒットが増えて、最終的には39件となった。(L) 8段階に較べれば、1.5倍のヒット増に当たり、うち26件、70%近くが必須キーワード(新聞コードを含む)でヒットしていた。

「関連」と評価づけされた「全国調査」「ゴルフ会員権」でヒットした11件が、要求に合ったものかどうかは、それぞれのユーザーの判断によろう。

《例2》記事番号22毎日新聞社説「消費者PL制導入は年内に決着を」

(L) 8段階で30件、(L) 9から(L) 15までで31件のヒットがあり、ヒット件数は倍になった。キーワードが増えていく中で、「消費者保護」「製造物責任」という社説のテーマにぴったりの必須キーワードが付与され、いいヒットにつながった反面、「ゆとり」という不要キーワードが(L) 13段階で採用されたことによって、5件のノイズが発生した。

《例3》記事番号156日経流通新聞「開発トピックス」

「液体手袋は作業後、水で洗い流せばよい」という記事の「水」が(L) 12段階で採用され、8件ヒットした。登録検索式の「水」は、ほとんど飲料水や環境問題に関連するもので、採用されたキーワードとは意味するところが違っている。自動索引の限界を示す例といえる。

【OCRで読み出した原文】

編集手帳 = 高額納税者と脱税番付

「長者番付」の・トップは江副浩正・前リクルート会長だった。ダイエーへの自社株売却益が主という。他にもバブル崩壊後の土地・株の処分¹で番付入りした人が目立つ「宴のあと」の感がある。一方で政治家たちはどうか。「所得税額一千万円超」のこの番付に名を連ねたのは全国国会議員の一割強、いくら探しても金丸信・前自民党副総裁の名は見つからない・かつて番付の基準は一定以上の所得額だった。国会議員の所得もほぼ全員が公示されて社会的なチェックができたのだが、昭和五十八年から所得税額を基準とすることになり、高額納税者「顕彰」の色彩だけが強まったそれではと、小紙が独自にまとめている「脱税番付」も今年で十年目になった。前年中に脱税容疑で刑事訴追された事件を全国調査、一覧表にしている・今回のそのワーストはゴルフ会員権乱売の「ケン・インターナショナル」、法人脱税額五十七億円は史上最悪記録だ。その他、バブル期の絵画取引に絡む脱税など、こちらは「宴の裏側」「虚業のツケ」を映ず各界の「長者番付」を眺め、他人様の懐具合をあれこれ推察するよりは、幾分か胸のつかえがおりる。今年起訴された金丸被告は次回、来年の「脱税番付」で上位にランクされるだろう。

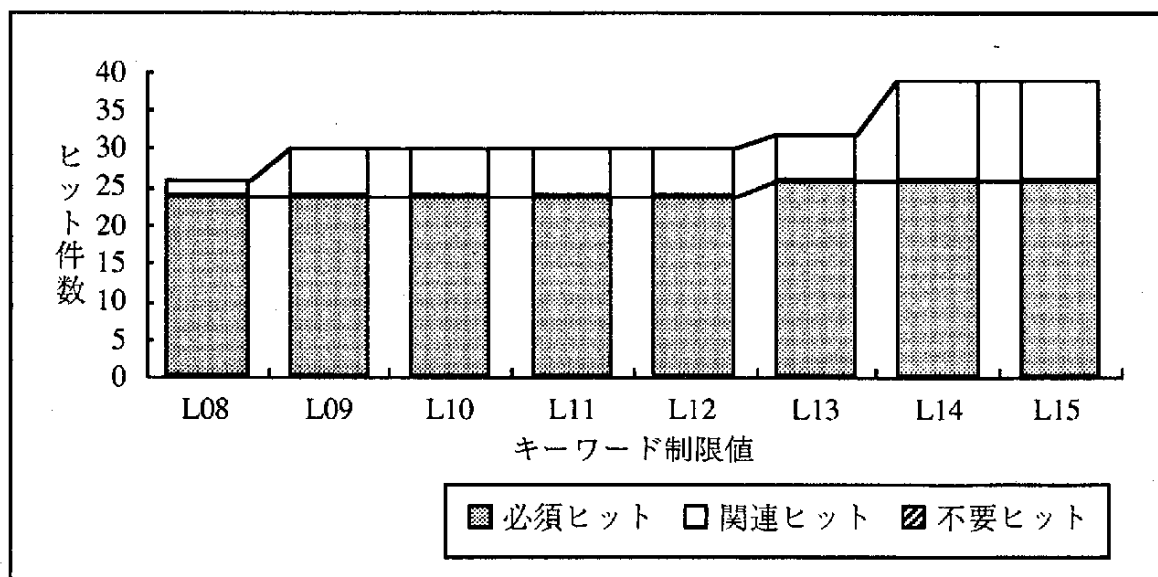
【各レベルの採用キーワードとヒット件数】

KW制限	採用キーワード	評価点	ヒット件数
L08	脱税	1	2
L08	「編集手帳」	1	3
L08	ダイエ	1	18
L08	長者番付	1	
L08	高額納税者	1	
L08	ケン・インターナショナル	1	
L08	自民党	2	2
L08	昭和	3	
L08	基準	3	
L08	手帳	3	
			小計 26*
L09	全国調査	2	4
L10	人	3	
L11	処分	3	
L13	所得税	1	2
L14	ゴルフ会員権	2	7
			小計 13
15 KW			合計 39

(注) 評価点 1、2、3 は、それぞれ必須、関連、不要 KW を示す

*印は KW 以外の新聞コードで 1 件ヒットしたため、計に加えた。

キーワードとヒットの関係



【OCRで読み出した原文】

社説＝消費者 P L制導入は年内に決着を

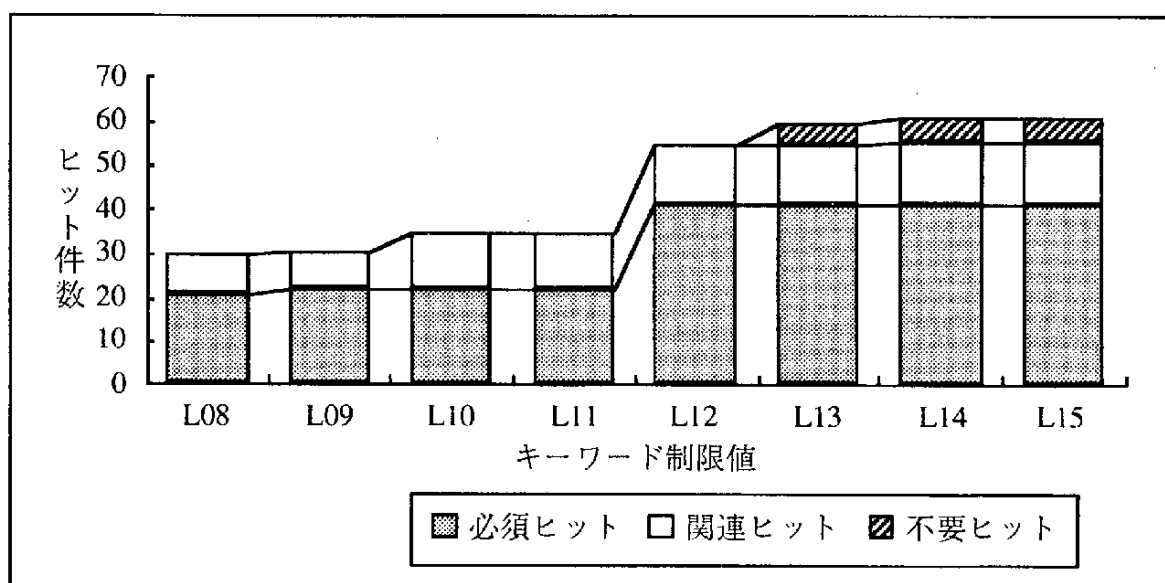
今月が「消費者月間」であることは意外に知られていない。四半世紀前の一九六八年五月に、消費者保護の憲法といわれる「消費者保護基本法」が、自民、社会、公明、民社四党提案で成立、施行された。その後、消費者の利益を守り、増進させる意味から、経済企画庁が中心になり、五月には全国各地でさまざまな行事が展開されている。今年のテーマは「くらしの安全、心のゆとり」で、先週、東京では中央集会が開かれた。消費者側から提起された論点は円高差益の還元による内外価格差の是正と「製造物責任（プロダクト・ライアビリティ＝P L）」制の早期導入の二つだった。いずれも消費者利益の確保、権利保護の基本になる主題だ。皮肉をいえば、「生活大国」を目指そうという日本で、消費者保護がいかに立ち遅れているかのあかしである。この日、開会のあいさつに立った船田元・経済企画庁長官は、差益還元については強い意欲を示したものの、P L制導入についてはいまひとつ態度がはっきりしなかった。昨年十月、同庁が所管する国民生活審議会消費者政策部会が、P L制導入の一年先送りを決めたこともあって、沈黙は金を決め込んだのかもしれない。しかし、消費者の権利と利益を守る生活者官庁を自任する以上、歯切れのいい発言を聞きたかった。残念なことだが、日本は製造物責任制の導入では大きな後れを取っている。いま、日本の消費者が置かれている立場は、欠陥商品で被害を受けた場合、自らの手で企業の過失を立証しない限り損害賠償金をとれない状況にある。個人が企業相手に訴訟を起こすとなると、大変な手間と費用がかかる。欧米などP L先進国の場合、国によって差異はあるものの、消費者は無過失責任を企業に問うことができる。つまり、消費者が過失を立証することなく企業の責任を問い、賠償を求めることができるわけだ。国民生活審議会がこの問題と取り組み始めたのは、七五年に民法学者の我妻栄氏らが「製造物責任法要綱試案」を提出した時からだから、ざっと十八年前になる。この間、自動車、家電、製薬業など消費財メーカーの反対に遭い、何度も議論が中断したり、棚上げされるなど経緯を経て、今日に至っている。政府は昨年十二月の消費者保護会議（会長・宮沢喜一首相）そして今年一月の宮沢首相の施政方針演説の中で消費者被害の救済策について「年内にP L法立法について結論を出す」と公約している。通産省所管の各業界団体は、今春以降繰り返し勉強会を開いて対応策の詰めを急いでいるが、なお消極論が多く、予断を許さない。P L法制定は、消費者保護基本法の考え方からいって当然の道筋である。乱訴社会といわれるアメリカでは、訴訟多発で産業界が悲鳴をあげ、国内の生産活動に支障をきたしているとも伝えられている。だが、日本が独自の知恵を出す方法はあるはずだ。企業の賠償限度額を決めるとか、商品開発時点で予期できなかった技術的問題については、一定の範囲で責任を免除するといった弾力的な対応策もありうる。現に欧州共同体（E C）加盟国の中では、国伯に応じて工夫をこらし、P L法を制定している。消費者月間を機に、いささか盛り上がり欠けるP L論議を活性化させ、年内決着を実現させたい。

【各レベルの採用キーワードとヒット件数】

K W制限	採用キーワード	評価点	ヒット件数
L08	「社説」	1	4
L08	消費者保護基本法	1	10
L08	民社党	1	
L08	消費者	1	1
L08	自民党	1	2
L08	我妻榮	1	
L08	国民生活審議会	1	
L08	公明党	1	
L08	経済企画庁	1	2
L08	通産省	1	2
L08	欧州共同体	2	
L08	利益	2	
L08	E C	2	3
L08	消費者月間	2	
L08	導入	2	
L08	P L	2	6
L08	宮沢首相	2	
			小計 30
L09	消費者保護	1	1
L10	生活大国	2	4
L11	行事	3	
L12	製造物責任	1	20
L13	ゆとり	3	5
L14	円高差益	2	1
			小計 31
1 5 K W			合計 61

(注) 評価点 1、2、3 はそれぞれ必須、関連、不要 K W を示す

キーワードとヒットの関係



【OCRで読み出した原文】

開発トピックス＝富士製薬の液体手袋「ノングローブ」手に塗り乾燥、汚れ防ぐ細かい手作業もOK

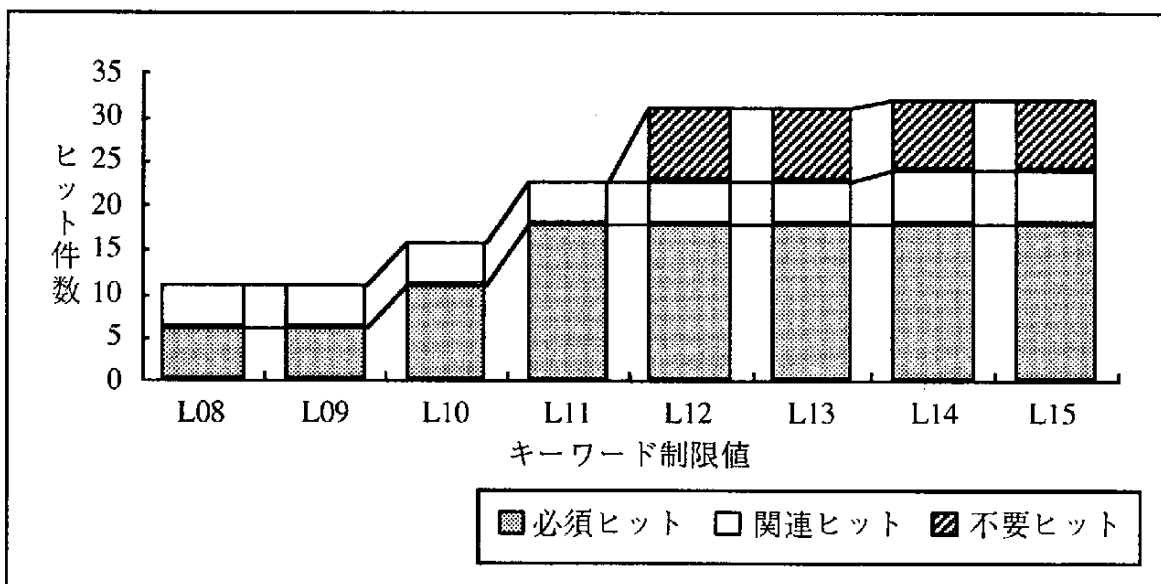
ペンキ塗りなど汚れ作業をする前に手のひらに擦り込み、一分程度、自然乾燥させて透明な薄い被膜を作っておく。すると石油製品系統の汚れから守ることができる製品「ノングローブ」が好評だ。富士製薬（東京・目黒、清水礼子社長）が九二年六月から発売しているが、家庭向けは三士、T 篇入り九百円（税別）で、約四十回使用できる。手袋をはめると微妙な指先での作業がやりにくい、が、「ノングローブ」はローションタイプの液体を塗るだけで済む。界面活性剤などの働きで手に汚れをつけない仕組みだ。使用後は水で洗い流せばよい。「ノングローブ」は石油化学製品には反応しないため溶けず、水に溶ける特徴を持つ界面活性剤の働きを取り入れている。工場などでは研磨剤入りの工業用せっけんで汚れを落とすケースも少なくないが、液体手袋はそもそも手を汚さない、という発想によって開発された。米国のある化学メーカーが販売していた同類製品に、羽根田武邦取締役が着目したのがきっかけ。早速メーカー側と直接交渉し、共同開発を開始したのは発売の一年ほど前のこと。以前からある他社の同類製品について、羽根田氏は「効果の持続時間が限られたり、完全に乾燥せずに使用中にべとつく製品もあった。滑りやすいと作業する上で危険なため、改善しようと考えた」と話す。そこで製品の差別化を図るため、技術スタッフの畑井政信氏は「主成分の界面活性剤の種類は豊富にある。その中から何と何を選択して組み合わせるか、徹底的に検討した。それらをどのような比率で混ぜ合わせるかについても実験を重ね、二十回ほど試作を繰り返しながら追求した」と、振り返る。製品検査をすると、少なくとも十五時間の持続効果があることがわかった。またペースト状やクリーム状にせず、ローションタイプにしたのは乾燥しやすくべとべとした使用感をなくすためだった。使用にあたっての肌の安全性については大学の研究室に依頼して確認した。もともと、やや刺激性があるにおい対策として、「アフタシェービングローションに似た香り」（羽根田氏）でマスキングし、液体の色は清潔感を表現するためブルーを選んだ。商品名は手袋（グローブ）が不要で汚れ作業ができるという意味から「ノングローブ」とつけた。全国のホームセンターなどで販売しており、日曜大工、換気扇やガスレンジなどの掃除やペンキ塗り、ワックスがけなどの際によく使われている。業務用には五百、T 居入りがあり、機械工具を扱う商社を通じて販売している。「当初は年間で千二百壇の販売を見込んでいたが、三千沼売れる見通しになった」という羽根田氏は、今後の展開について、「一般家庭に一本備えるのが常識になるようコンビニエンスストアにも普及させたい。業務用は3K対策に向けて展開を強化していきたい」と語っている。

【各レベルの採用キーワードとヒット件数】

KW制限	採用キーワード	評価点	ヒット件数
L08	「開発トピックス」	1	1
L08	富士製薬	1	5
L08	液体手袋	1	
L08	グローブ	2	
L08	汚れ	2	
L08	トピックス	2	2
L08	手作業	2	
L08	乾燥	2	3
L08	羽根田武邦	2	
L08	開発	2	
			小計 11
L09	製品	2	
L10	新製品	1	5
L11	界面活性剤	1	7
L12	水	3	8
L13	共同開発	1	
L14	社長	2	1
L15	米国	2	
			小計 21
15 KW			合計 32

(注) 評価点 1、2、3 は、それぞれ必須、関連、不要 KW を示す

キーワードとヒットの関係



2.4 作業効率化への寄与

2.4.1 効率化の測定

本調査のキーワード自動抽出方式が作業効率向上に寄与する度合いは、以下の方法で測定した。

- ①現行作業の分析
- ②キーワード付与関係作業の占める割合の測定
- ③自動抽出にともない新たに発生する作業量の測定
- ④品質確保のために要する補完作業量の測定
- ⑤総合判定

2.4.2 測定の結果

(1) 現行作業の分析

EL社の「MOR」およびデータベース構築作業は、「前処理」「入力処理」「後処理」の3ブロックに分けられる。

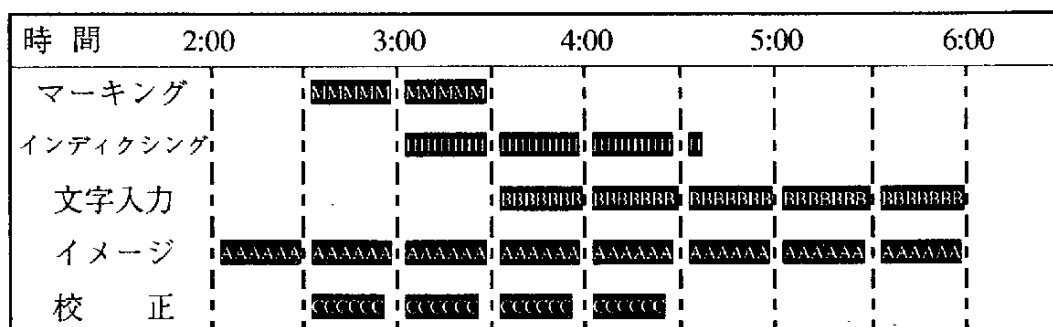
前処理は、採録する記事の選択と記事イメージの切り出し範囲を指定する「マーキング」、キーワード、分類、書誌事項、見出しづけをする「インデイクシング」。入力作業は、キーワード、見出し、書誌データなどの「文字入力」と、イメージを入力し記事単位に切り出す「イメージ入力」。後処理は主として入力文字の校正、「MOR」作業の点検・修正・追加に大別できる<図2—14>。

現行の作業分担		
前処理	入力処理	後処理
◇マーキング	◇文字入力	◇校正
◇インデイクシング	◇イメージ入力	◇MOR点検

【図2—14】

作業の流れは、基本的にはマーキング→インデクシング→文字／イメージ入力→校正／KW修正という順序になるが、現実の作業はほぼ並列状態で進む。特に「MOR」作業では時間の制約から、マーキング、インデクシング、文字入力担当者間で相互応援が行われ、常時ピーク状態が保たれている。

「MOR」作業の進行パターン



【図2-15】

(2) キーワード付与関係作業の占める割合の測定

上述のように、現実には相互応援で作業をこなしているため、キーワード付与に関する作業が占める割合は、作業にかかっている人数と実働時間を個別に測定して集計、全体の実作業時間との比率をみた。

この結果、インデクシングと入力を合わせた作業が全体に占める割合は、40～50%程度と想定された。

(3) 自動抽出にともない新たに発生する作業量の測定

自動抽出方式によって新たに発生する作業は、「OCRのレイアウト補正」である。新聞紙面は複雑な記事の流し方やレイアウトが多いほか、見出しと写真との組み合わせや段間のスペースのふぞろいなど、OCRの自動読み取りにとっての泣き所が多い。OCRが解析できなかったり、読み方の順序を間違えた部分を人手で直してやる作業である。

OCRのレイアウト補正オペレーションの改善テストを行った結果、1記事を1分以内で処理できる見通しがつき、作業量は現行作業量全体の10～20%前後にあたると推定された。

(4) 品質確保のために要する補完作業量の測定

自動抽出にともない、キーワード辞書のメンテナンスの強化など、補完作業が必要になる。本調査では、既述の品質テスト結果を前提に、原則として辞書のメンテナンス以外に大きな負担になる新たな作業は発生しない、と仮定した。

この結果、補完作業に要する作業は、全体の5～6%と想定された。

(5) 総合判定

以上の結果に、作業の流れの変化による影響、作業改善の見通しなどを加え総合的に判断した結果、現行作業に比べ20～30%の作業効率化が期待できると想定された。

2.4.3 測定の限界

キーワード自動抽出システム導入がもたらす作業の効率化の度合いは、品質上の許容限度、システムの構成、システム各パーツの性能と精度、人手の関与の範囲、作業手順の設定など多くの要素に左右される。

本調査では、これらの細部にわたる検討は不可能だった。従って、上にあげた数字はあくまでも計算上の「想定値」ととどまる。

3.成果と今後の課題

3.1 成果

本調査研究の直接的な成果は①すでに商品化されているOCRと日本語処理ソフトを組み合わせ、これに独自の修正および絞り込み方式を加味することによって、効率が良く、かつ精度の高い印刷物を対象としたキーワード自動抽出システムを構築できる②一定の作業効率化が見込め、現行の人手によるデータベース構築の硬直性をやわらげることができる③新システムを採用する場合、現行の商用サービスの激変を避けながら緩やかな移行が可能である、との展望を持ち得たことである。これらの展望を発展させ、印刷物を素材としたイメージデータベース構築のすそ野の拡大に資することができれば、さらに大きな成果となりえよう。

3.1.1 精度・効率の良いキーワード自動抽出システムへの展望

本調査研究で行った二つのテストの結果のポイントは、概ね以下の諸点に要約できる。

- ①すでに商品化されているOCR、INDEXERと既存の各種辞書をベースに、一定の補正機能と絞り込みルールを付加することによって、1記事平均8～15個前後に絞ったキーワードの自動抽出が可能である。
- ②1記事12個以上のキーワードを抽出すれば、EL社が現に商用サービスしている人手付与によるキーワードと比較しても質的に劣らず、付与量、均質性などを加味すれば、より優れた結果が得られると判断できる。
- ③これらのキーワードを使った検索結果も、商用サービスと比較して遜色がない。
- ④キーワード絞り込みのルールを操作することによって、キーワード、検索結果双方で一定の質、量の調整ができる。
- ⑤OCR、INDEXERなど既存ツールの改良をはじめとして、各種システム辞書の整備・開発、キーワードの絞り込みルール・修正機能の改善など、品質と処理スピードにかかわる今後の開発、改善、工夫の余地が多々あり、その成果はいずれも今回のテスト結果に上積みされる。

現在、わが国で商用サービスされている新聞、雑誌記事などのテキスト・データベースでは、コンピュータ編集の文字データからキーワードを自動抽出する「フリーターム方式」が主流である。これらのデータベースでは、1記事当たり数十から百を超えるキーワードが付与されている。文中の名詞の大部分をキーワードとするため、網羅性では優れているが、精度の高い検索をするためには相当の熟練が要求され、膨大なキーワードによるシステムへの負荷が大きい、などの問題がある。

一方、EL社の商用サービス・データベースは、主題に絞ったキーワードを人手で付与しているため、キーワードは1記事平均10個未満にとどまっている。索引者の判断があるため、不要なキーワードが混入する余地は少なく、「MOR」や「SDI」のような自動出力サービスでは優れているが、キーワードの絶対数、均質性、構築コストなどの点で問題が残っている。

本調査では、双方の利点を活かし、弱点をカバーできるシステムの可能性を追求した。

上記の結論を総合すると、比較的少量のキーワードで、実用可能な精度を持つキーワード自動抽出システム構築への展望を得ることができた。

3.1.2 作業の効率化と省力効果の確認

データベース構築にとって最大の問題点は、構築に高いコストを要する点である。本調査では、自動抽出が品質に与える影響を分析するとともに、作業の効率化にどの程度寄与するかを探った。

その結果、やや不確定な要素はあるものの、直接的な効果だけで20～30%の効率化と省力効果が見込まれた。

これは、現在の作業量を前提としたものであり、作業量が増えるに従って省力効果も拡大していくことが想定される。

3.1.3 現実的な導入の可能性確認

EL社の場合、88年1月の「MOR」開始から足掛け7年にわたって商用サービスを続けており、94年2月末現在で約1,500余のユーザーに1日4～5万記事をサービスしている。

また、ユーザーがあらかじめ検索式を登録、これにヒットした記事を直接ファクシミリで送信する「MOR」「SDI」サービスをおこなっている。これらのサービスは、端末で検索し、必要な記事を選択する対話型のデータベースと違い、ユーザーに事前チェックの機会が少ない。

これらの事情から、検索方法やヒット量に極端な変化をもたらす制作工程の変更は現実的ではない。

【テスト2】の結果は、新たな方式を導入する際、データ処理の連続性を維持しながらサービス面でも激変を避け、なだらかに移行する方法を探る手がかりが得られた、と評価される。

3.2 課題と展望

本調査研究の実用化に当たっては今後、品質と入力作業・システムの両面からの詰めが必要になる。品質面では①「漏れ」「ノイズ」などの対策②キーワード抽出ロジックと各種辞書の改良③より適切なキーワード抽出数の設定など、入力作業・システム面からは①投資効果の精査②効率的な作業の流れの構築③OCR、INDEXERの機能改善と補完機能の開発、などが主な課題となろう。

3.2.1 品質向上対策

本調査では、自動抽出による必須キーワードの漏れは、1記事平均15.2個のキーワード抽出のケースで17%程度だった。現行人手作業の39%に較べればかなり低い数字だが、内容を分析すると「環境保護」「調査結果」など文中にない重要なキーワードの漏れが目立った。人手作業の利点を自動抽出にも反映させる工夫が必要となる。

また、辞書にない語（新出キーワード）は、人のチェックに頼らざるをえず、いかに早く必要な新出キーワードを付与するかが、大きなポイントになる。

一方、人手作業ではほとんどなかったノイズが、自動抽出では全体の11%程度発生した。このうち、OCRの誤読によるものは、活字のフォント登録によって防ぐことが期待できる。

また、「重みづけ」を中心とする独自の絞り込みルールは、かなりの効果をあげることが確認された。また、「キーワード抽出制限値」を変え、抽出数を変化させると一定の傾向をもって品質に影響が現われることが実証された。

抽出の上限をさらにあげてみるなどの追加テストを重ねることによって、相対的に優れた品質を得るための、より適切な抽出数のレベルを探る必要がある。

検索式をあらかじめ登録しておき、ヒットした記事を自動配信する「MOR」「SDI」と、対話型のデータベース検索のように、異なったサービスに対応するキーワード抽出のあり方も検討の対象となろう。

3.2.2 システム化への課題

具体的なシステムを設計する場合、最も問題になるのは作業効率アップによるコスト軽減とシステム構築に要する投資の比較検討である。

本調査では、キーワード付与の自動化による省力効果は現行人手作業20～30%、との推定結果が得られた。これがシステム導入に見合うのか、より効果的な作業の流れをどう組み立てるか、がポイントとなろう。

品質だけでなく、作業効率にも影響を与えるOCRのレイアウト補正、キーワード候補切り出し機能の改善や、補完機能の開発などもまた見逃せない課題となる。

3.2.3 実用化から汎用システムへ

『データベース白書1993』は、92年の日本のデータベースの動向を分析する中で、新聞や雑誌記事のイメージデータベース・サービスが伸びていることを指摘し、「データベース・サービスもようやく、出版物に近い表現力を身に付けてきたといえる」と述べている。

マルチメディアやオンライン・ジャーナル時代の到来が脚光を浴びる一方、成熟した情報伝達媒体である印刷物を素材とするイメージ・データベースの構築は緒についたばかりである。

本調査研究の本来の課題は、イメージ・データベース構築のノウハウを開発・改良することによって、今後もおお相当長期にわたって有効性を発揮し続けるであろう印刷物の分野のデータベースのすそ野の拡大に資することにある。その意味では、今回の調査研究は「小さな一歩」にすぎない。

ここで得られた成果を、具体的なシステム化につなげ、さらに汎用的なシステムへと発展させることが究極の課題といえよう。

4.資料

4.1 ELNETのサービス概要

4.1 ELNETのサービス概要

約350万件もの新聞・雑誌記事からパソコンなどで必要な記事を探し、ファクシミリに記事原文（見出し・写真・図表もそのまま）を出力する画期的なデータベースサービスである。

①ELNETのアウトプット

↓記事データ一覧

(パソコン又はファクシミリで出力)

キーワードをもとに選び出した「記事データリスト」。必要な情報をすぐに発見できます。各新聞のとりあげ方や関連記事の掲載頻度などにより、記事の重要度も把握できます。

97777704 93年04月15日18:38 [44564] 1/19 p}

宮部 管理課

■ELNET■

「野生動物関係記事」

ユーザーID番号

検索日

配信番号

対象ファイル

☐ 検索年月日
☐ ファイル名
☐ レンジ
☐ 検索式

93年04月15日

FLDBサービスファイル

1992

S1 19658件 北海道+ロシア

S2 468件 野毛?

S3 15件 S1+S2

FAX/ONLINE S3/ロシア 野生動物保護記事

記事データ 15件 記事原文 17枚 (注) 原文ページは「日」指定の為出力されません

☐ 出力形式
☐ 出力枚数

検索キーワード

ヒット件数

[No.]

●北海道下巻中のオオワシ 捨てられた鳥がエサ43% 「完全自活」わずか11% ロシア研究者が調査

◎読売新聞 93年04月14日 夕刊◇34面 見出し3段 写真

えん;博士:ワシ;調査;野生鳥獣調査;調査結果;北海道;ロシア(生態:オオワシ;越冬:ロブロク、エフゲニー)

(470) 93041400297

2 ●環境庁 鳥獣害 自然保護アビール 飼育保護モデルに活用

◎日本経済新聞 93年03月18日 朝刊◇34面 見出し4段 写真

自然保護;生態系;野生動物;環境庁;鳥獣害;北海道【飼育保護】

(550) (460) 93031800686

キーワード

アクセッション番号

紙誌名/発行日他

3 ●朝日と冬 水鳥(上)一獣医の憂と希望

◎読売新聞 93年02月23日 朝刊◇27面 見出し3段 写真

星油;ボラチニア;治療;渡り鳥;獣医;船舶事故;野生動物;動物保護;水鳥;北海道【冬小牧場;

野生動物保護観察協会;「便ひと冬」】

(510) (470) (連載) 93022300257

4 ●石井井川手負いガモ 骨中の矢毒々しく「宮小牧 油まれの水鳥救いの手」

◎朝日新聞 93年02月04日 朝刊◇31面 見出し3段 写真

検査;賞物税;汚染;渡り鳥;被害;救出;動物保護;カモ;水鳥;鳥獣保護法;北海道;東京;宮小牧市;保護区【月天;

野生動物保護観察協会;石井井川】

(470) 93020400176

記事見出し

5 ●東京南北一週 道徳大災 強盗にタバコの投げ捨て

◎エコノミスト 70巻51号 p.27 93年12月01日 写真

マナー;心証;観光客;火災;特別天然記念物;生態系

北海道;国語市【「東西南北」;ラムサール条約;鶴

(560) 93120900007

6 ●日本テレビ系 北海道の大自然野生動物を紹介

◎読売新聞 92年11月27日 夕刊◇13面

テレビ放送;テレビ番組;野生動物

(4640) 92112700000

※ELNETの出力を著作権者に無断で複製・印刷することはできません。

[illegible]

👉 記事原文(ファクシミリで出力)

写真や図表、グラフ、データなども切り抜きと同じ「記事原文」だから一目瞭然。臨場感あふれる資料収集に威力を発揮します。

②ELNETの対象新聞・雑誌

新聞45紙、雑誌120誌以上を対象に1988年1月1日以降の350万件（うち新聞約340万件、雑誌約10万件）の記事をELNET採録基準に沿って採録している。新聞のうち、地方紙についてはその地方の主要記事、雑誌については記事原文提供可能な記事を中心に採録している。

新聞 45 紙

朝刊27紙から、必要な情報だけをFAXに自動送信。

ELMOR採録対象紙・27紙

朝日新聞
読売新聞
毎日新聞
日本経済新聞
産経新聞
電波新聞
日刊工業新聞

日本工業新聞
日経産業新聞
株式新聞
ニッキン
電気新聞
化学工業日報
薬事日報

日刊建設工業新聞
建設通信新聞
日刊自動車新聞
日経流通新聞
流通サービス新聞
織研新聞
日本繊維新聞
日本食糧新聞

日本農業新聞
日刊水産経済新聞
住宅新報
生産性新聞
東京新聞
93年12月より採録開始

朝日新聞(大阪本社)
読売新聞(大阪本社)
毎日新聞(大阪本社)
原子力産業新聞
食品新聞
山形新聞
福島民友
下野新聞
上毛新聞
神奈川新聞

山梨日日新聞
静岡新聞
信濃毎日新聞
北国新聞
富山新聞
京都新聞
南日本新聞
日刊スポーツ

雑誌 約 120 誌

AERA	週刊金融財政事情	興銀調査	日債銀レポート	労働経済判例速報	Food Life
サンデー毎日	税協	今月の問題点	ハートの経済情報	NTT R&D	食品工業
週刊朝日	CREDITAGE	埼玉県経済メモ(あさひ銀総研)	福島の進路	NTT技術ジャーナル	食品流通経済
週刊時事	Financial Adviser	たぐぎん調査	富士タイムズ	データ通信	総合食品
週刊読売	近代セールス	中小公庫月報	BEST management	ニューメディア	インダストリーレビュー
時事解説	金融ジャーナル	長銀レポート	北陸経済研究	JAAA REPORTS	EP REPORT
世界週報	金融ビジネス	調査(開銀)	オール投資	REPORT JARO	エネルギーフォーラム
THIS IS読売	月刊消費者信用	調査(三菱)	財界観測	アドバタイジング	太陽エネルギー
選択	IBJ経済・産業の動き	日本銀行月報	大和投資資料	広告月報	農林経済
地方行政	HOKUTO経済調査レポート(北都銀)	調査月報(あさひ銀)	投資経済	広告ジャーナル	科学朝日
エコノミスト	エコノレター	調査月報(住友信託)	投資月報	IATSS REVIEW	プレス技術
Japan Research Review	季刊かいぎん	調査月報(東海)	投資手帖	東北開発研究	暮らしと健康
経済政策情報	グラフによる経済レポート	調査月報(東洋信託)	日興投資時報	建設工事の動き	食の科学
週刊東洋経済	経済経営研究(開銀)	調査情報(住友信託)	野村週報	建築技術	厚生福祉
実業の日本	経済月報(三和)	調査情報(三菱信託)	ヒット銘柄100	月刊工事情報	内外教育
統計月報	経済月報(朝日生命)	調査時報(日債銀)	調査月報(日生)	不動産鑑定	月刊シンガポール
貿易之日本	経済情報(さくら)	調査通信(三井信託)	マネジメント21	機械技術	等
ESP	経済調査(大和銀)	調査特報(三菱)	工場管理	機械設計	
財界	経済トピックス(住友)	ちば経済季報	戦略コンピュータ	電機	
実業界	調査月報(中央信託)	東京銀行月報	税務経理	オートメーション	
金融財政	経済の動き(安田信託)	東銀週報	Venture Club	電子技術	

(1994年 2月現在)

③ELNETのサービス

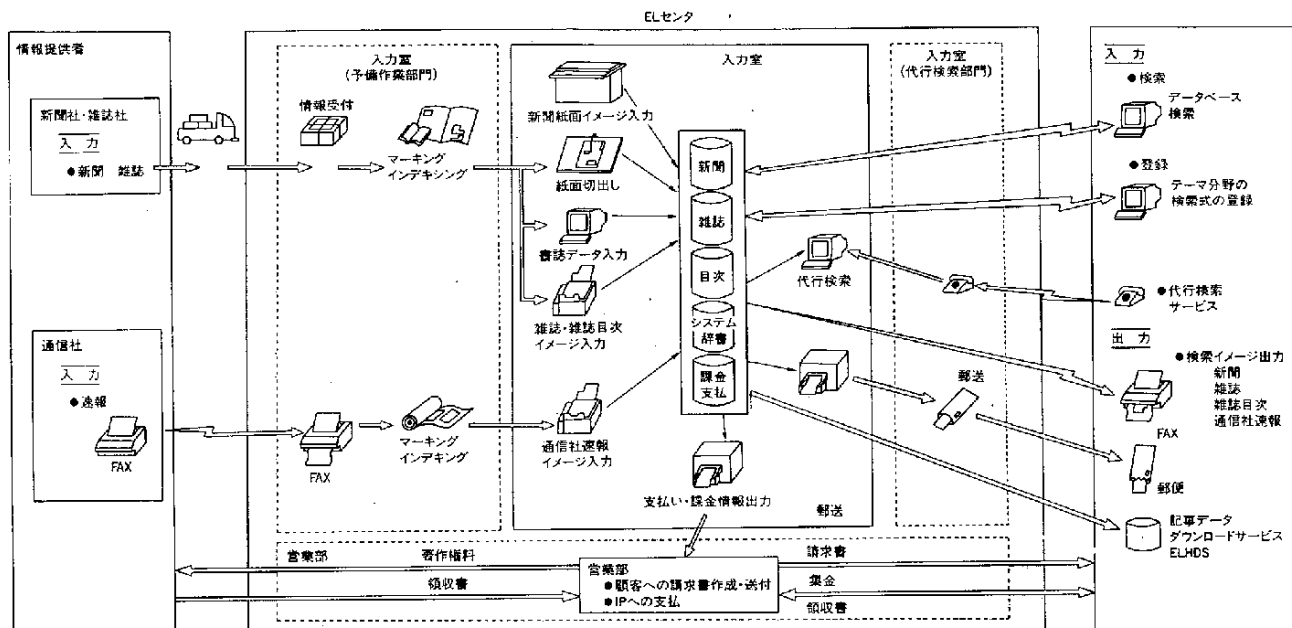
ELNETには以下の4つのサービスがあり、ELNET会員はこれらを自由に組み合わせて利用できる。

ELDB	350万件の記事（新聞45紙・雑誌約120誌）からパソコンで必要な記事を検索する。
ELMOR	当日の朝刊27紙（在京紙）から、毎朝必要な記事をFAXに自動送信する。
ELSDI	新聞45紙・雑誌約120誌から、翌朝、必要な記事をFAXに自動送信する。
代行サービス	専門サーチャーが利用者に代わって、必要な記事情報等を的確に検索する。

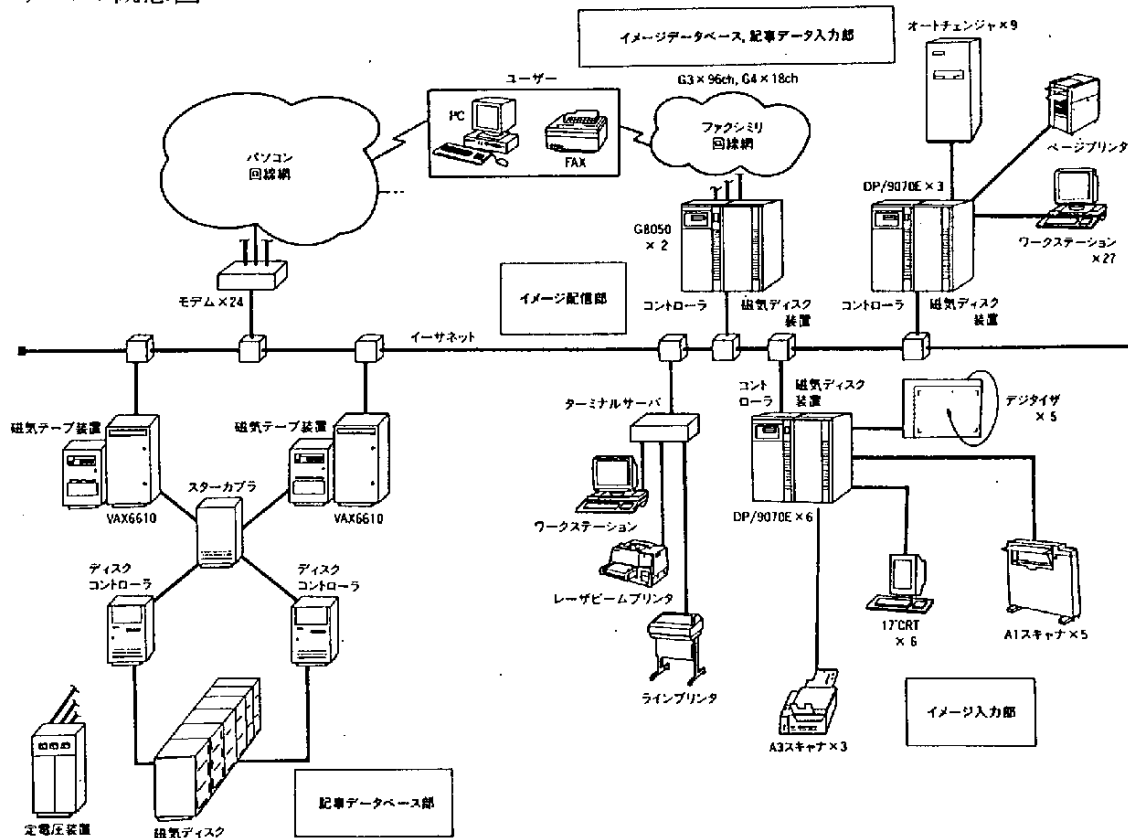
サービス一覧

	ELDB データベース	ELMOR イーエル モーニングレビュー	ELSDI イーエル 予約情報サービス	代行検索サービス	
サービス概要	ELNETで採録する新聞・雑誌 を対象に、ユーザーが検索 自分で行う検索	ELMOR対象の当日の朝刊を 対象に、登録されたキーワードに より必要な記事を自動出力 自動出力サービス	ELNETで採録する新聞・雑誌 を対象に、登録されたキーワード により必要な記事を自動出力	ELNETで採録する新聞・雑誌/ G-Searchの情報を対象に、 サーチャーが代行検索 サーチャーによる検索	
検索方法	必要ときにユーザーが パソコンから検索	あらかじめ登録された検索式により自動検索		必要ときに電話一本で 代行検索	
対象紙誌 ('93年12月現在)	新聞45紙 雑誌約120誌	新聞27紙(朝刊)	新聞45紙 雑誌約120誌	ELNET 新聞45紙 雑誌約120誌	G-Search G-Searchのメニュー によるファイル
データ内容	1988年1月1日以降前日までに 蓄積された全記事 (年間約60万件)	当日の朝刊 (日曜刊は原則として翌サービス 日のELMORで出力します)	前日に入力された全記事 (1週間分をまとめて郵送する サービスもあります)	1988年1月1日以降 前日までに蓄積された 全記事(年間約60万件)	G-Searchの設定 によるファイル (郵送サービスもあります)
サービス時間帯	平日 0:00~21:00 土曜 0:00~15:00	毎朝7時~10時頃に自動出力	毎日午前1時頃から自動出力	平日 10:00~18:00 土曜 10:00~14:00	
日曜・祝日はサービス休止					

④ E L サービスの業務概念図



⑤ E L システムの概念図



4.2 「MOR」登録検索式サンプル

4.2 「MOR」登録検索式サンプル

例1：税金関連の記事収集を目的としたもの

□検索式	S 1	30件	税金+?税+(税?#税引き利益)+国税庁
	S 2	12件	・税制調査会+?税制+減税?+増税?
	S 3	30件	S 1+S 2
□出力形式	FAX/ONLINE /B/		
□出力総数	記事データ	30件	記事原文37枚

例2：環境問題の記事収集を目的としたもの

□検索式	S 1	31件	地球環境+?汚染+環境保全+環境問題+環境対策
	S 2	47件	550/CL+ごみ?+?ごみ
	S 3	64件	S 1+S 2
□出力形式	FAX/ONLINE S3/B/ALL		
□出力総数	記事データ	64件	記事原文37枚

例3：外食産業関連記事の収集を目的としたもの

□検索式	S 1	3件	外食産業+日本料理+和食?+レストラン+食生活+すし+食
	S 2	4件	居酒屋+お好み焼き+グルメ+外食?+給食サービス+食文化
	S 3	0件	カフェテリア+おにぎり+仕出し弁当+惣菜+弁当+食材
	S 4	0件	ハンバーガーショップ+ドライブイン+ケータリング?
	S 5	0件	ファーストフード?+牛丼+もつ鍋+飲食店
	S 6	0件	ロイヤル+ロッテリア+松屋フーズ+日本マクドナルド
	S 7	2件	日本ケンタッキーフライドチキン+モスフードサービス
	S 8	0件	すかーらーく+西洋フードシステム+吉野屋?+麒麟シティ
	S 9	1件	ファーストキッチン+西友フーズ+デニーズジャパン
	S 10	4件	グルメ杵屋+東天紅+京樽+サッポロライオン+あさくま
	S 11	0件	ニュー・トーキョー+サト+藍屋+本曾路+リンガーハット
	S 12	1件	ドトールコーヒー+日本ペプシコフードサービス
	S 13	11件	S 1+S 2+S 3+S 4+S 5+S 6+S 7+S 8+S 9+S 10+S 11+S 12
□出力形式	FAX/ONLINE S 13//ALL		
□出力総数	記事データ	11件	記事原文17枚

4.3 本テストサンプル記事の見出し一覧

番号	字数	見 出 し
1	730	天声人語＝共和汚職事件
2	1, 365	社説＝税制改革論議をすぐ始めよ
3	88	日本新党、23人目の公認
4	404	日本IBMと東芝次世代の記憶カード開発小型化に成功
5	123	コーセースポーツ用化粧品発売
6	66	日本ヒューム管新社長に小林氏
7	1, 206	皇室報道(20)＝「やり得」に自主規制の網、だが・・・
8	644	中国のTV局「歴史的」番組生放送を初体験
9	371	服部君射殺事件父親傍聴し初公判「死を無駄にしないで」
10	1, 010	郵政・通信研が開発スタート首都圏直下型地震3次元予知システム地殻変動天体・衛星で観測
11	509	編集手帳＝高額納税者と脱税番付
12	1, 198	社説＝自己点検欠かせない大学改革
13	583	決算メモ モスフードサービス日本舗道ソニー・ミュージックエンタテインメント
14	179	シャープワープロ発売専用ペンで入力、編集
15	1, 184	ルーツ元気な経営者豊田佐吉トヨタグループ創業者(1)＝「発明熱」で立志伝実現
16	117	ノエビア植物の抽出液を配合ボディシャンプー発売
17	353	首都圏賃貸住宅の成約状況「低価格」の割合増加
18	426	きょうから千代田で介護者も同乗OK福祉タクシー運行開始
19	1, 470	93都議選一票の視線50年目の都政に(1)＝臨海開発関心薄く「票にならない」
20	447	予知連御前崎付近、沈降鈍る東海地震との関連注視
21	629	余録＝日本洋画壇の長老だった猪熊弦一郎さん
22	1, 333	社説＝消費者PL制導入は年内に決着を
23	224	大成建設、日立製作所、新明和工業大深度地下対応の駐車場発売
24	192	日本舗道3月期決算 官公需サマサマ
25	211	モスフードサービス3月期決算 モスチキン効果
26	619	ソニー・ミュージックエンタテインメント93年3月期「尾崎豊」「ひらり」で販売増
27	1, 493	930万人の選択それぞれの選挙区(1)＝新党、新人・・・混戦に拍車
28	720	日本ユニシス株価操作証券取引監視委 検察と最終協議新谷代表ら告発へ
29	942	自治省検討事業税対象広げ税率下げ給与や利子を加算
30	1, 984	景気回復への助走(2)＝数量先行の消費好循環生む低価格
31	659	春秋＝尾瀬ヶ原
32	2, 318	社説＝政治の混迷から出口を模索する西欧
33	513	通産省方針ASEANへの支援今秋までに具体策
34	298	93年3月期理事会長と理事長賞与木津信組、全額カット
35	763	伊、早期利下げ検討新政権、ERM復帰先送り
36	346	工作機械主要8社輸出には明るさ4月受注額25.6%減

番号	字数	見出し
37	1,422	国際ビジネス最前線＝インドの空、自由化で混戦民間の参入相次ぐ
38	780	カゴメ、会計基準を変更今期からリベートを売り上げから除外
39	519	ミズノリゾート事業で攻勢ゴルフ場など建設
40	662	A Vメーカー2社今上期、円高で収益大幅悪化
41	459	上新電割安感をドライに評価外国人持ち株比率上昇
42	1,023	サラリーマンミドルの情景＝第388話女性として(2)
43	138	オープン化粧品会長山下静夫氏死去
44	1,209	潮流＝デノミ論再燃1ドル100円時代に備え「1新円」
45	672	産経抄＝看護婦
46	983	主張＝臓器移植法は公開論議で
47	340	カーネギー平和財団A・ピエール氏警告米露関係の悪化要因に
48	218	ダイフク配当政策は自ら
49	117	雪印乳業豪州産のアイスクリーム発売
50	1,554	産業景気まあだだよ(5)＝干天の慈雨輸出の好調が下支え
51	1,843	どうしてですか育児休業3分の1社が支給連合は6割支給を提案
52	588	東大活性化に“新しい血”民間企業→法学部教授三菱商事法務部長代行・柏木さん
53	323	チャイム＝ごみをあさるカラス
54	600	産業春秋＝男女格差
55	1,081	社説＝動き出す「昆虫産業プロ」に期待
56	433	企業庁が設置関係団体と研究会中小企業のPL対策
57	131	元奥村組専務角田榮氏死去
58	1,514	シリーズ・経営本田技研工業(5)＝ホンダウエーの曲がり角国際化の壁米業界から三くだり半
59	1,436	ハイテク光と陰研究プロを検証(60)ゴミ発電(1)＝2010年に700万KW供給も
60	762	東芝が年内めど16メガDRAM月産70万個へ多ビットなど品種拡充100万個も射程に
61	579	東芝が発売応用ソフト内蔵ノートパソコン処理速度4倍に向上プリンター内蔵型も
62	470	日立造日鉄溶接と提携溶接ロボシステム拡販
63	656	日鉱金属など3社八戸の新亜鉛精錬所建設来月初に合併設立
64	632	濤和化学樹脂ペレットの新着色法加熱溶融が不要
65	311	小西酒造が予約販売大吟醸酒と輪島塗貝桶一対をセット
66	273	ホリー開閉式足場板付き脚立発売脚の長さ自由自在
67	278	いっぷく＝大阪ガス常務・ガス事業副本部長の宮腰久司さん値上げ問題に複雑な心境？
68	260	多摩中央信金が中小の景況調査今年度1四半期の売上高は改善の見通し
69	1,262	電子工業会が見通し97年時点の年間需要カラー868万台予測BS内蔵化率38%
70	467	「アジアテレコム93」シンガポールで開幕世界各国から355社出展最先端通信技術で競演
71	1,045	迫られる料金見直し岐路に立つNTT(5)＝「MA」拡大方向で見直し
72	292	古野電気がプレジャーボート用の無線機20日から販売

番号	字数	見出し
73	404	日立など3社大深度地下機械式駐車場を開発
74	698	村田機械ビジネス用FAX発売
75	539	NECフォトカプラー増産グループ会社九州電子月産5600万個体制に
76	1,154	広がる戸建てシステムバス市場96年には3千億円に各社、商品開発を急ぐ
77	525	ポリドールがCD集5点クラシックでストレス解消
78	1,061	クラリオンが21機種高精度ナビゲーションシステムなど発売
79	927	通産輸入債務保証が始動第1号蘭の半導体装置円高が追風輸入拡大に弾み
80	1,423	新技術立国ニッポン21世紀への選択(47)第2部急げテクノインフラ整備=官民プロ
81	614	三菱重工産機部品を海外調達輸出拡大へコスト対応
82	509	他山石語=公園と緑
83	595	4月の企業倒産不況型が36%増3ヵ月連続で1000件超す
84	212	三井海上ユニセフ募金とセットの積立保険来月1日発売
85	288	ニチメンが中国で刺繍、縫製加工で合弁会社
86	407	日本ビクター25インチのうす型カラーTVを発売
87	499	コンパクトマイクロソフト製OS採用DOS/Vと3.1パソコン24種発売
88	1,715	マルチメディア台所事情夢と現実(10)=NTT映像通信端末取込めば広がり
89	281	日化協、3人体制新副会長に大西實・富士写社長
90	317	東京・池袋のトヨタオートサロンアムラックス東京累計来場者が700万人に
91	779	工技院物質研、三笠製薬が開発有機物質の分離・濃縮技術感熱性高分子を利用
92	637	この人=新井組社長に就任した花房正次郎氏
93	845	東証分析個人投資家は低位株狙い
94	534	三菱化成透明樹脂使った駆動装置開発ガラス張り広告などフィルム状、超薄型
95	1,892	企業の興亡(25)第1部落日の大企業変わるモノ作り1=都市離れ、生き返った工場
96	1,038	思い秘めて続トップ交代93(3)=西武百、役員降格は改革への決意
97	951	NTT、X線で小型化CMOS0.2ミクロン回路を開発
98	610	日本IBM発声練習支援ソフトなど発売 障害者にも使いやすく
99	375	カシオ初心者用小型ワープロ発売正しい指使い教えます
100	449	アルプス電気DCCモジュールでフィリップスと提携
101	649	自工会米大教授の論文発表米経済活性化に寄与日系企業の工場移転
102	1,025	核心インタビュー=三菱重工業菅原信一取締役環境装置、今後の取り組みは産業廃棄物の処理軸に
103	596	主要8社の4月前年比工作機械受注25.6%減海外には明るさも
104	1,683	生活を変える身近なハイテク=ナムコテクスチャーマッピングコスト抑えゲームに
105	747	全日空エコノミーでゆったり3クラス制見直し効率化
106	1,449	住宅・住設機器メーカー企業出版曲がり角「まとめ買い減る」事業縮小なら「道楽」の批判
107	446	アルダスマック対応DTPソフト 発売
108	327	伊勢丹副社長に城森氏迎える財務強化へ三菱銀専務

番号	字数	見 出 し
109	814	新ビジネス訓一滴の水（１）＝明光商会社長高木礼二氏より良く生きる修業の場が会社
110	780	電波障害による製品事故通産が原因究明へ通検内に施設を整備
111	335	クラレ乾燥収縮率が小さい綿Ｔシャツ開発
112	339	日本特殊塗料３月期決算減収減益を計上
113	499	帝人マクラーレンにタイヤウオーマー供給発熱繊維使い性能アップ
114	1,038	日本スピンドルが一貫システムバグフィルターの飛灰を無人処理再飛散なく作業環境改善
115	978	厚生省医薬品などのＦＤ申請９５年春本格実施へパイロットを開始
116	700	野村事務所抗菌・防カビ剤品揃えＭＲＳＡなど専任部署設けて対応
117	1,972	原子力と放射線作業（下）＝被ばく低減元を絶つ発想で効果
118	632	焦点＝円高差益
119	760	日米環境セミナー一層の対話促進で合意ＣＯ２固定化で協力検討
120	196	リコー高性能レーザープリンター発売
121	542	ＩＨＩ、きょう命名式ＳＰＢ方式ＬＮＧ船第１船は「ポーラ・イーグル」
122	659	三菱系（ギャラン店カープラザ）の業績好調９３年３月期各社前年並みＲＶ快走で年販新
123	549	オペル生産、販売減少も今年は黒字を維持
124	397	中部４月の輸入車登録２０．６％の大幅な伸び日本ブランド以外は微増
125	322	首都圏新空港研需要動向など把握秋にはシンポ開催も
126	690	ＪＨ圏央道鶴ヶ島ＪＣＴ６月にＰＣ橋を発注
127	122	名古屋市国際展示場３号館３日、１０ＪＶで入札
128	794	ＪＨ大阪建設局今年度にルート地形図作成
129	1,173	中央公契連特別会員制度を創設各省官房長クラスで入札・契約の連携強化
130	846	道建協工事用機械が３５％災害事故発生状況調査
131	142	弘電社新社長に成吉信氏
132	533	千葉県推進会議幕張熱供給、かずさセンター施設など三角構想５年度方針決定
133	950	広島防衛５年度計画江田島隊舎など建設自衛隊関係施設は２７件
134	923	社説＝高齢者住宅の進化のために
135	920	適正取引推進機構「期限付き借家」で活用指針適用要件など示す
136	338	首都圏機構成約・前年比戸建ても同水準単価は下落
137	1,054	紡協パキスタン綿糸ダンピング提訴か通産プロジェクトチーム設置へ生活産業局審議官ヘッドに迅速対応
138	602	東洋紡産業廃棄物処理費半減へ
139	700	三面鏡＝新素材開発
140	510	CHOYAのドレスシャツ日清紡「ＳＳＰ」加工導入高い安定性、安全性
141	878	ズームショット＝大手婦人服専門店チェーン出店状況出店抑え、既存店強化退店は収束の傾向
142	1,629	医薬品等のフロッピー申請開始へ
143	1,461	創薬科学研究現状と評価（下）＝「分子生物学」研究が課題鶴藤丞氏（サイトシグナル研究所長）に聞く
144	760	暖流寒流＝サッポロビールの複合商業施設「生活工房・サッポロファクトリー」が開業

番号	字数	見 出 し
145	995	中小企業庁などが調査中小卸売業の商慣行百貨店への協賛金売上高の1・6%に
146	568	コーセースポーツ化粧品改良紫外線カット耐水性強化
147	407	たかの友梨クリニック家庭やオフィスにエステ無料出前専門家派遣し実演“営業”は一切なし
148	588	東急百、都内全店にシステムギフト受注電算管理伝票作成など省力
149	1,982	秀和保有株処理でどうなる伊勢丹(下)＝三菱側、「熟柿」狙うラストチャンス、価格で攻防
150	798	電通ブロックス自然や街の風景リアルに映し出す疑似窓を開発商業施設などに的
151	693	HＣのみつたが運営日本海側で最大級の“スーパーセンター”脚光5000平方メートルに生活品満載
152	316	想 ニュートーカーの福田恒雄社長
153	873	有田VOCが導入テーマパークの商品管理にPOSに代わる低価格システム
154	1,461	問屋サバイバル地方流通担う(37)＝丸三米穀外食・小売店を強化
155	294	ふぁいる＝大卒初任給の伸び最低の2.1%
156	1,162	開発トビックス＝富士製薬の液体手袋「ノングローブ」手に塗り乾燥、汚れ防ぐ細かい手作業もOK
157	635	グラフ解説消費者行動＝男性「家電」、女性は「身の回り品」
158	993	オリンパスが新情報ネット「ユーバス」開發生産から小売りまで結ぶ
159	363	都内百貨店4月売上高衣料10.6%減紳士いぜん不振、18.1%減
160	502	東レ・和装用新合繊「シルック・ライテシモ」開発軽量感、着心地アップ秋冬から初年度1万5千反
161	402	ワコール貸衣装店などに1店3年契約で時価1000万円のティアラをレンタル
162	677	阪神百貨店「バイヤーズアイ」再構築工場直ルート完全買取りへ
163	489	私の職場＝京都丸紅土岐訓代さん商品企画資料作成から販売まで
164	848	3月期決算発表本格化予想上回る収益回復公共投資関連が強い
165	1,366	JR東日本東急と比較東急は将来のお手本
166	949	JAの組織整備実行方策づくり“加速”来月まで41県完了直接利用肥料・農業に集中
167	1,203	論説＝強硬さを増す米国の対日戦略
168	856	今、農村の女性たちは関東農政局管内アンケートから(中)＝給料制励みと生きがいに
169	342	一村逸品 パート2＝岐阜県可児農協味噌漬物ざわりよいキクイモ漬物
170	1,680	日本IWCへの対応課題脱退論含め方策検討も

4.4 原文イメージ、OCR読み取り原文、 採用キーワードのサンプル

(1) 記事原文の切り出しイメージ

- ・朝日新聞（93年5月18日付）社説

(2) OCRが読んだ社説原文

- ・太い下線部分は採用されたキーワード

(3) 採用キーワードの一覧表

- ・OCRで読みだした記事を「INDEXER」にかけてEJ辞書登録語と合致した108語をキーワード候補語として抽出
- ・キーワード候補語にEJ独自の重み付けをして絞り込み、高点順に19語をキーワードとして選定した

社説

税制改革論議をすぐに始めよ

政府税制調査会(首相の諮問機関)の加藤寛会長は、さきの総合経済対策で実施が見送られた所得税減税について、「消費課税や資産課税とのバランスを考へるなど、税制全体の基本的なありかたを議論する中で検討したい」という趣旨の見解を明らかにした。先週末に開かれた税調総会後の記者会見での発言である。

いまでもなく政府税調は、毎年度の税制改正答申とは別に、長い目で見た税制のあるべき姿を描く中期答申を、委員の任期(三年)の最後の年にまとめることになっている。今年はその年にあたっており、加藤会長の発言は、この答申づくりを指していると思われる。

しかし、政局の先行き不透明から、いつ税調審議を始めるのか見通しが立っていない。総選挙の実施時期によっては、税制全体の本格的な見直しそのものが見送られる可能性すら、否定できない状況だ。

いまの税金の仕組みは、さまざまな不公平やゆがみを抱えている。それを正していくのは、政府税調の当然の役割だ。総選挙を控えた各政党の思惑を超え、ただちに税

制改革論議に取りかかるべきである。

私たちはかねて所得税減税の早期実施を主張してきた。税調にも昨年来、この問題を真っ正面から取り上げるよう求めてきたが、当の税調は、財政当局の意を受け、深入りを避けてきた。

だが、所得税減税が依然として重要な課題であることを強調しておきたい。

日本経済は、不況の中にも一部に明るさが見え始めているが、まだ景観は許されない。本格的な景気回復が実現するかどうかは、設備投資と並んで民間最終需要の二本柱である個人消費の立ち直りが、そのカギをにぎっている。

しかも、累進税率が適用される所得税の場合、一九九〇年以降は減税が実施されないまま、所得の伸びに応じた実質増税が続いていることを忘れてはならない。

八七、八九両年の大幅減税は、消費税導入の見返りであると同時に、それまでの実質増税を取り戻す意味があった。その後、減税見送りにともない、とくにサラリーマンの給与所得の税負担率は、三十数年ぶりの高水準に達している。

もともと所得課税では、サラリーマンの負担が重すぎる「クロヨン」問題が、不公平の象徴として批判を浴びてきた。国税庁発表の高額納税者には無縁の話かも知れないが、そこに実質増税が加わって、税制への不信感は強まっている。こうした状態を一日でも早く改善することは、個人消費を好転させるのにも役立つはずだ。

加藤会長もその必要性を認めているように、税調は減税を出発点に論議を進めるべきだ。その際、税率そのものはもちろん、五段階の税率区分や課税最低限など、所得税全体の見直しが欠かせない。

問題は、財源をどう確保するかである。

減税規模によって話が変わってくるが、加藤会長は消費税の増税を選択肢のひとつに挙げていた。しかし、金持ち優遇の批判を浴びている利子や株式売却益の分離課税を総合課税に改めて、増収増をはかることが優先されねばなるまい。

かりに消費税率を手直しするにしても、帳簿方式や簡易課税制度など、同税のさまざまな欠陥を正していくことが先決だ。

振り返ると、税調審議は、時間をかけても結論を先送りするケースが少なくなかった。今回、「中期答申」の名を借りて、時間かせぎは許されない。減税を軸とした税制改革論議を急いでもらいたい。

社説＝税制改革論議をすぐ始めよ

政府税制調査会（首相の諮問機関）の加藤寛会長は、さきの総合経済対策で実施が見送られた所得税減税について、「消費課税や資産課税とのバランスを考えるなど、税制全体の基本的なありかたを議論する中で検討したい」という趣旨の見解を明らかにした。先週末に開かれた税調総会後の記者会見での発言である。いうまでもなく政府税調は、毎年度の税制改正答申とは別に、長い目で見た税制のあるべき姿を描く中期答申を、委員の任期（三年）の最後の年にまとめることになっている。今年はその年にあたっており、加藤会長の発一召は、この答申づくりを指していると思われる。しかし、政局の先行き不透明から、いつ税調審議を始めるのか見通しが立っていない。総選挙の実施時期によっては、税制全休の本格的な見直しそのものが見送られる可能性すら、否定できない状況だ。いまの税金の仕組みは、さまざまな不公平やゆがみを抱えている。それを正していくのは、政府税調の当然の役割だ。総選挙を控えた各政党の思惑を超え、たたちに税制改革論議に取りかかるべきである。私たちはかねて所得税減税の早期実施を主張してきた。税調にも昨年来、この問題を買っ圧面から取り上げよう求めてきたが、当の税調は、甫政当局の意を受け、深入りを避けてきた。たが、所得税減税が依然として重要な課題であることを強調しておきたい。日本経済は、不況の中にも一部に明るさが見え始めているが、また楽観は許されない。本格的な景気回復が実現するかどうかは、設備投資と並んで民間最終需要の二本柱である個人消費の立ち直りが、そのカギをにぎっている。しかも、累進税率が適用される所得税の場合、一九九〇年以降は減税が実施されないまま、所得の伸びに応じた実質増税が続いていることを忘れてはならない。八七、八九兩年の大幅減税は、消費税導入の見返りであると同時に、それまでの実質増税分を取り戻す意味があった。その後の減税見送りにともない、とくにサラリーマンの給与所得の税負担率は、三十数年ぶりの高水準に達している。もともと所得課税では、サラリーマンの負担が咀すきる「クロヨン」問題が、不公平の象徴として批判を浴びてきた。国税庁発表の高額納税者には無縁の話かも知れないが、そこに実質増税が加わって、税制への不信感が強まっている。こうした状態を一日でも早く改善することは、個人消費を好転させるのにも役立つはずだ。加藤会長もその必要性を認めているように、税調は減税を出発点に論議を進めるべきだ。その際、税率そのものはもちろん、五段階の税率区分や課税最低限など、所得税全体の見直しが欠かせない。問題は、財源をどう確保するかである。減税規模によって話が変わってくるが、加藤会長は消費税の増税を選択肢のひとつに挙げている。しかし、金持ち優遇の批判を浴びている利フや株式売却益の分離課税を総合課税に改めて、税収増をはかることが優先されねはなるまい。かりに消費税率を手直しするにしても、帳簿方式や簡易課税制度など、同税のさまざまな欠陥を正していくことが先決だ。振り返ると、税調審議は、時間をかけても結論を先送りするケースが少なくなかった。今回、「中期答申」の名を借りて、時間かせぎは許されない。減税を軸とした税制改革論議を急いでもらいたい。

採用キーワード一覧

キーワード	重みづけ	出現位置	頻度	評価点
会長	4322	1 21	4	1
所得税減税	4209	1 43	3	1
税制改革	9016	0 -997	3	1
税制	4316	1 76	4	1
減税	3912	8 705	5	1
「社説」	9011	0 -1000	1	1
政府税制調査会	6500	1 1	1	1
加藤寛	6000	1 18	1	1
首相	4502	1 9	1	2
サラリーマン	3620	9 820	2	2
総会	4004	1 128	1	2
国税庁	6009	10 909	1	2
消費課税	4012	1 54	1	2
論議	9000	0 -993	4	2
政府税調	6100	2 152	2	2
個人消費	3614	7 652	2	2
税調	7506	1 126	6	2
総選挙	4105	3 307	2	2
資産課税	4010	1 59	1	2

(注) ・ 朝日新聞社説(93/5/18)からの自動索引したKW

・ 重み付け点数

千、百の位が名詞分類、出現位置から評価した点数
 高点順に制限KW数まで採用

下2桁はMOR検索式に出現する頻度。百点台まで
 同点の場合はMOR頻度の高点順に採用

・ 出現位置

左が本文中の段落、右が文頭からの文字数を示す
 見出しは0段落、マイナスは見出しの文字位置で
 1000から逆算する

・ 頻度

記事中のKWの出現回数。

・ 評価点

自動索引後に7人で3段階評価した点数

1は必須、2は関連、3は不要KW

4.5.1 OCR誤読の統計データ

- ①正しい読み：記事原文の文字
- ②誤読：テキストデータとしてOCRより出力された文字
- ③頻度：誤読した文字列と正しい読みの組み合わせが全記事に出現した回数
- ④全出現数：誤読した文字列と同じ文字列が出現した回数
(誤読ではなく②と同じ文字列が出現した回数と③を足したもの)

<例>

	①正しい読み	②誤読	③頻度	④全出現数
6	体	休	8	55

上の例では6番目のデータで「体」と読むべき所を「休」と読んだ回数が8回あり、「休」という出力が170記事で55回あった事を示す。

	正しい読み	誤読	頻度	全出現数
1	▼	v	10	23
2	だ	た	60	1,919
3	題	遁	1	1
4	賢	賛	1	12
5	言	一召	1	1
6	体	休	8	55
7	惑	感	1	58
8	真	貢	1	1
9	正	圧	1	25
10	上	ヒ	5	22
11	財	甫	3	3
12	重	咀	1	1
13	ぎ	き	5	575
14	び	び	5	6
15	子	f	3	9
16	ば	は	7	2,263
17	間	問	7	107
18	日	=	9	212
19	本	木	2	29
20	十	+	82	88
21	東	束	19	26
22	五	. h	6	6
23	顔	舶	1	27
24	補	袖	3	4
25	前	[	1	57
26	前	[朋	1	1
27		'	13	129
28	立	h J	3	3
29	予	狛	1	1
30		,	4	3,851
31	三	・一	1	2
32	三	. 一	2	2
33	目	目	14	164
34	I	一	14	202
35	用	川	3	38
36	力	力	17	217
37	一	一	38	202
38	=		1	1,063
39	=	一	1	202
40	五	互	1	12
41	五	. H	1	1
42	三	、.]	1	1
43	三	一. 一	1	2
44	億	惚	1	1
45	量	=	1	212
46	三	一一、	1	1
47	五	、h H	1	1
48	び	ひ	3	31
49	が	カ S	1	1
50	か	カ>	8	8
51	化	イヒ	2	2
52	ず	す	12	995
53	止	[F.	1	1

	正しい読み	誤読	頻度	全出現数
54	乳	乎し	1	1
55	卒	T r	1	1
56	年	I	1	66
57	日	ト	1	21
58	6	O	3	47
59	ほほ	ほほ	2	11
60	ター	ター	2	11
61	ビ	ヴ	1	10
62	ご	こ	10	823
63	味	昧	19	22
64	『	「	1	618
65	』	一	1	930
66		一	11	930
67	日	口	7	43
68	号	巨	1	18
69		。	1	23
70	宮	ウー	1	1
71	間	間	1	202
72	螻	捲	1	2
73	『	〔	1	57
74	』	〕	1	40
75	題	盟	1	3
76	手	丁	1	12
77	中	申	9	59
78	へ	ペ	1	6
79	艾	文	1	39
80	市	巾	3	4
81	(三	a 一一	1	1
82	早	里	1	2
83	ブ	ヴ	4	10
84	五	H	3	33
85	予	r	2	19
86	言	一 [. - . -	1	1
87	組	紺	1	1
88	六	人	1	276
89	省	雀	4	4
90	“	ツ	23	363
91	”	’	38	129
92	ピ	ビ	14	168
93	工	エ	4	134
94	ー	l	9	47
95	ブ	プ	8	221
96	っ	づ	1	38
97	◆	・	2	641
98	た	だ	1	520
99	表	衷	1	1
100	1	一	3	1,278
101	史	吏	1	1
102	裏	裏	1	1
103	す	ず	3	79
104	合	台	5	111
105	責	貴	5	9
106	○	O	70	166

	正しい読み	誤読	頻度	全出現数
107	。	。。	3	8
108	○	○	1	196
109	万	刀	4	104
110	だ。	。」○f	1	1
111	○	○	1	166
112	増	璫	1	1
113	触	角虫	1	1
114	が	f) ズ	1	1
115	W	。一	1	1
116	」	一	1	202
117	1	」	1	40
118	が	力ズ	5	5
119	」	.	1	59
120	年	」	1	40
121	れ	ね	1	25
122	一	1	17	47
123	漠	漠	1	1
124	聞	聞	1	2
125	エ	エ	11	134
126	ッ	ツ	14	61
127	。	○	2	196
128	ヤ	ヤ	8	34
129	ト	ト	6	21
130	=	一・一	1	1
131	ば	いよ	1	2
132	バ	ノズ	1	1
133	ブ	ブ	12	66
134	居	届	1	7
135	は	ほ	1	132
136	ぜ	セ	2	94
137	め	の	1	4,352
138	境	現	1	160
139	る	う	1	707
140	民	氏	1	84
141	錦	鋪	1	2
142	題	門	1	1
143	壮	仕	1	47
144	題	預	1	2
145	ば	ば	2	173
146	いた。	ゝツ」○〜	1	1
147	運	迎	1	15
148	お	わ	1	228
149	今	令	4	9
150	パ	パ	3	111
151	▲	△	5	5
152	昨	昨	3	3
153	賞	貫	1	4
154	…	…	6	31
155	提	捉	2	4
156	責	貞	1	3
157	損	担	1	28
158	一	」	3	40
159	造	薦	1	4

	正しい読み	誤読	頻度	全出現数
160	法	伝	1	23
161	情	伯	2	14
162	和	相	1	85
163	1	l	5	47
164	1	i	3	42
165	0	O	26	166
166	1	—	8	202
167	5	6	1	37
168	1	q	1	20
169	%	勿	1	1
170	1 1	II	1	3
171	%	物	2	128
172			5	1,063
173		・	3	641
174	会	<ム	1	1
175	ソ	ン	5	752
176	エ	工	11	244
177	増	憎	2	3
178	十	ト	1	21
179	会	< [	1	1
180	会	< [ム	1	1
181	倍	伯	2	14
182	連	辻	1	1
183	百	自	4	200
184	善	清	1	21
185	豊	曲. Ω	1	1
186	影	彫	1	1
187	響	響	1	2
188	証	] 訛	1	1
189	二	一一	13	73
190	前	[・	1	3
191	期	朋	3	6
192	三	・・、	1	2
193	億	億	1	9
194	上	土	3	28
195	最	巖	1	2
196	三	一。一	2	2
197	三	一一一	7	20
198	△	V	5	92
199	得	行	3	194
200		ゝ	1	14
201		r	1	19
202		"	1	11
203		\	1	2
204		.	5	59
205	原	眼	1	2
206	原	胞	1	2
207	会	ムム	1	1
208	原	奥	1	8
209	ブ	フ	3	211
210	ル	ソレ	1	2
211	“	”	1	16
212	”	w	1	16

	正しい読み	誤読	頻度	全出現数
213	っ	つ	10	352
214	い	レゝ	1	1
215	情	’ 問	1	1
216	原	県	1	115
217	場	堀	1	2
218	債	偵	1	1
219	治	泊	1	7
220	省	胃	1	1
221	末	宋	5	5
222	討	肘	5	6
223	省	酋	5	6
224	げ	け	26	441
225	三	’ 二	1	1
226	着	善	7	30
227	一	i	14	42
228	最	巖	1	1
229	力	カ	2	202
230	財	尉	6	6
231	そ	ぞ	7	32
232	日	H	3	33
233	順	願	1	2
234	一	!	2	9
235	道	遭	1	7
236	一	一	6	1, 278
237	言	----- [	3	2
238	右	布	1	9
239	言	-----]	2	3
240		}	1	7
241	連	更	1	13
242	合	脇	1	1
243	【	[	11	57
244	】	]	9	40
245	下	ド	2	156
246	入	人	11	276
247		・	1	6
248	財	・	2	6
249	げ	U	2	20
250	財	宙	1	6
251	営	常	2	59
252	○	O	1	47
253	ユ	ユ	16	48
254	ワ	リ	4	291
255	三	. [-	1	1
256	一	1	2	64
257	二	一 [	3	10
258	ぼ	ほ	13	132
259	チ	ヂ	2	3
260	ボ	ボ	10	61
261	3 0	叩	1	2
262	三	] 二	1	1
263		一’	1	3
264	か	カ	1	202
265	うだ。	Yづ こ。	1	1

	正しい読み	誤読	頻度	全出現数
266	一	鴟	1	1
267		\、	1	2
268	ジ	シ	2	304
269	う	っつ	1	1
270	▼	V	20	92
271	『	円	2	280
272	』	」	1	624
273	方	力	25	217
274	賃	ロ	2	174
275	こ	二	12	513
276	ご	一」	1	3
277	弁	井	4	32
278	器	掃	2	5
279	大	人	3	276
280	」	一	1	930
281	普	菖	1	1
282	ら	f	1	9
283	首	酋	1	6
284	価	イ曲	1	1
285	ル	ノレ	1	1
286	う	」っ	5	5
287	う	Yっ	2	3
288	年	牢	2	2
289	百	巨	2	18
290	生	牛	3	24
291	見	児	4	22
292	万	h i	1	1
293	一	.	1	59
294	百	丙	1	2
295	雨	巾	1	4
296	百	目	1	164
297	だ。	‘こOf’	1	1
298	言	———亜	1	1
299	音	立日	1	1
300	口	目	3	164
301	垣	拍.	1	1
302	隆	降	1	28
303	車	中	1	231
304	得	柑	1	1
305	う	っ	1	352
306	た。	こoナ	3	3
307	二	..	1	6
308	う	vっ	1	1
309	年	什	5	5
310	賃	貨	2	37
311	た。	」っこot	1	1
312	請	謂	3	3
313	こ	二	1	105
314	九)	&	1	5
315	じ	P	1	77
316	負	価	1	122
317	障	瞳	1	2
318	や	°	1	2

	正しい読み	誤読	頻度	全出現数
319	済	溶	1	26
320	を	牛	1	12
321	度	庶	1	5
322	一	I	4	66
323	〇	O	3	47
324	右	石	1	28
325	貧	負	2	24
326	貧	黄	1	2
327	毘	毘	3	3
328	夕	夕	10	15
329	波	彼	3	5
330	“	。	8	2,452
331	榮	柴	1	4
332	1 7	Π	1	3
333	1 1	H	1	33
334	一	口	1	43
335	十	土	1	28
336	二	二	14	513
337	i	一	1	1,278
338	十六	土ハ	1	2
339	デ	一ア	1	4
340	ン	ノ	3	3
341	濤	鴻	1	127
342	器	羽	2	9
343	貝	與	1	1
344	L	亡	1	10
345	1	I	4	66
346	百	白	11	41
347	二	ニ	9	105
348	程	稗	2	2
349	シ	ン	4	4
350	テ	ア	1	1
351	三	一	1	930
352	三	一二	3	12
353	り	O	9	166
354	千	モ	1	131
355	百	〔	4	57
356	百	〔丘	1	1
357	五	丘	1	1
358	千	丁	1	12
359	十	ト	2	481
360	千	F	1	38
361	二	一. 一	1	2
362	七	L	2	42
363	十	廿	1	2
364	千	千	16	27
365	ラ	う	5	707
366	富	富	3	3
367	I	i	3	42
368	ア	ア	6	276
369	い。	ノ)	1	4
370	=	目	1	358
371	士	上	1	289

	正しい読み	誤読	頻度	全出現数
372	下	丁	1	12
373	一	5	1	39
374	ラ	一ワ	6	6
375	オ	オ	2	112
376	ニ	一	1	930
377	ニー	一下	3	4
378	が	か	32	1,063
379	ラ	) リ	1	1
380	一	、	1	3,851
381	立	q	1	20
382	泉	限	1	42
383	リ	り	1	753
384	フ	ソ	1	68
385	ら	ろ	1	126
386	ラ	) ワ	1	1
387	フ	リ	1	291
388	ラ	] ワ	1	1
389	散	敬	1	4
390	'	'	1	129
391	カー	カー	2	14
392	可	珈	2	2
393	ク	タ	10	264
394	0 0	〇〇	3	20
395	I	1	1	64
396	テ	一ア	3	4
397	キ	牛	4	24
398	ダ	グ	9	108
399	エ	干	1	27
400	エン	干ノ	1	1
401	入	へ	6	114
402	り	U	1	20
403	入	→へ	1	1
404	置	買	2	59
405	事	串	1	4
406	号	u 5	1	1
407	エー	エー	1	3
408	う	[つ	1	1
409	工	. L	1	1
410	役	投	1	76
411	立	上 J	1	1
412	石	。 L	1	1
413	口	口	7	43
414	緯	綺	1	1
415	だ。	こ〇J'	2	2
416	正	王	1	11
417	た。	こ〇チ'	1	1
418	>	'	1	129
419	I	一	2	1,278
420	“	ア	1	28
421	ない。	l、o φ'	1	1
422	晶	品	1	279
423	“	*	1	8
424	ビ	ピ	1	65

	正しい読み	誤読	頻度	全出現数
425	札	杜	1	17
426		[	1	57
427	三・	一一丁	2	4
428	累	素	1	32
429	服	叩	1	2
430	服	朋	2	6
431	パ	バ	9	113
432	S	Q	1	2
433	千	エ	1	134
434	ジ	フ	1	211
435	だ。	こ of	1	1
436	た。	つこ。	1	1
437	ご	う1」	1	1
438		]	1	40
439	六	土ハ	1	2
440	る	ろ	1	126
441	面	而	1	1
442	高	古	2	19
443	高	古向	2	2
444	三	一一 [	2	4
445	井	井	2	2
446	演	潰	1	1
447	顧	廠	1	3
448	。	.	10	59
449	責	寅	1	1
450	着	俯	1	1
451	略	賂	1	1
452	三	. .]	1	1
453	ゲ	ケ	1	53
454	コ	ー	1	1, 278
455	三	], ー	1	1
456	プ	。フ	3	8
457	可	河	2	6
458	二	、一	2	59
459	二	. 一	1	6
460	〇〇	〇〇	1	20
461	ニ一	二一	1	11
462	l	i	1	42
463	p	P	1	77
464	言	一一]]]	1	1
465	計	訃	2	2
466	目	青	1	11
467	き	ぎ	1	34
468	言	雷	1	5
469	顧	顯	2	3
470	!	"	2	11
471	!」	乙	1	1
472	ち	ぢ	1	2
473	運	連	2	89
474	十	一	3	202
475	た	こ	1	823
476	末	未	3	25
477	業	業	1	1

	正しい読み	誤読	頻度	全出現数
478	二	〔	1	2
479	書	轡	1	2
480	I	工	1	244
481	島雅	篇	1	3
482	、	・	1	641
483	4	一	1	202
484	エ	壬	1	1
485	エデ	壬ア	1	1
486	財	町	1	23
487	言	宮	2	19
488	ボ	ホ	2	42
489	走	疋	1	1
490	飛	戒	1	3
491	イ	「	1	618
492	が	M	1	90
493	増え	乍凡	1	1
494	ら	り	1	753
495	ど	こ	1	823
496	合	百	1	241
497	で	て	5	1,780
498	ど	と	2	2,025
499	ク	タ	2	15
500	ゴ	コ	1	181
501	三	一。	1	1
502	ダ	タ	5	264
503	づ	つ	2	352
504	バ	グ	1	108
505	ご	7」	2	3
506	三	一三	1	11
507	ぞ	そ	1	313
508	イ	イ	1	357
509	ペ	ベ	1	109
510	グ	ク	7	328
511	ペ	ベ	1	65
512	可	司	2	8
513	ば	ぱ	1	4
514	現	俱	3	4
515	だ。	こ〇t	1	1
516	着	糖	1	1
517	冷	冶	1	1
518	着	糲	1	1
519	題	鹿	1	2
520	。	一	1	1,278
521	急	憩	1	2
522	差	艶	1	1
523	。	。	1	2
524	“	イ	2	144
525	長	畏	2	2
526	ま	未	1	25
527	蔵	藏	1	1
528	“	、	1	11
529	言	肯	1	2
530	。	。	1	23

	正しい読み	誤読	頻度	全出現数
531	環	珊	3	3
532	2	ヨ	1	77
533	エ	平	1	85
534	ネ	不	1	119
535	意	息	1	6
536	た。	つこ○	1	1
537	温	湿	1	8
538	《	<	1	72
539	》	V	1	92
540	議	議	1	110
541	ほ	は	3	2,263
542	ド	ト	1	481
543	ペ	。へ	1	1
544	二・	一丁	1	6
545	二二	一三	1	11
546	り	リ	1	291
547	活	后	1	1
548	三	一一丁	1	4
549	ラ	一フ	2	2
550	場	湯	1	21
551	～	一	4	202
552	開	〔帽	1	1
553	連	練	1	7
554	社	杜	9	17
555	社	祉	2	8
556	二	ミ	1	1
557	六	〕ハ	1	1
558	い	〔い	1	2
559	大	太	11	24
560	検	儉	1	8
561	二	一二	1	12
562	庫	車	2	91
563	▽	ワ	12	62
564	順	碩	1	1
565	】	一	3	930
566	屋	窒	1	17
567	▽	V	6	92
568	【	〕	1	40
569	三	一千	1	3
570	、	’	1	129
571	▽	v	1	23
572	▽	7	2	40
573	階	烏	1	1
574	階	昔	2	3
575	二	一{	1	1
576	書	鵬	1	1
577	省	歪	1	1
578	)」	こ	1	20
579	)。	。.)	1	1
580	書」	！	2	9
581	ご	C	1	135
582	省	吉	1	30
583	省	首	3	35

	正しい読み	誤読	頻度	全出現数
584	』	6	1	37
585	最	鼠	1	1
586	な	カ	1	202
587	さ	ざ	1	30
588	期	荊	2	2
589	期	勘	1	1
590	省	肯	1	2
591	十	ヤ	1	57
592	三	一一	1	1
593	三	一一 [	1	1
594	期	則	1	11
595	含	命	1	12
596	い	一一	1	4
597	——	一	4	202
598	品	晶	4	9
599	加	如	1	2
600	提	呵	1	1
601	一	一一	1	13
602	ト	一	1	1, 278
603	査	徨	1	1
604	果	里	1	2
605	に	一	1	1, 278
606	具	臭	1	1
607	一	i	2	42
608	一	l	3	47
609	酒	洒	1	2
610	上	士	1	2
611	一	一	5	202
612	三・三	一一丁三	1	1
613	ニエ	一王	2	2
614	三	[一一	1	2
615	顧	顯	1	3
616	言	一——]	1	1
617	由	山	1	31
618	千	午	1	12
619	三	一二	1	1
620	当	徂一	1	1
621	二一	] 下	1	1
622	置	一	1	202
623	三二	目 [	1	1
624	組	糾	1	9
625	上	L	1	42
626	年	庄	3	25
627	立	h v	1	1
628	白	白	1	2
629	三	]、.	1	1
630	月	川	3	38
631	計	廿	1	2
632	五	、 h	1	2
633	単	弔	3	5
634	デ	テ	1	252
635	高	] 口回	1	1
636	シ	・シ	2	2

	正しい読み	誤読	頻度	全出現数
637	参	琴	1	1
638	る	一 o	1	1
639	在	往	2	4
640	市	…	1	3
641	向	回	1	165
642	合	へ口	1	1
643	生	主	1	97
644	に	'	1	129
645	の	(U	1	1
646	っ	リ	1	291
647	産	「賄	1	1
648	べ	、へ	1	2
649	自	臼	1	2
650	主	=	1	212
651	者	晋	1	1
652	頭	「	1	57
653	合	ゝ =	1	1
654	最	口取	1	1
655	位	泣	1	161
656	を	斗」 =	1	1
657	を	斗」	1	1
658	目	=	3	212
659	商	(間	1	1
660	力	刀	1	104
661	在	作	1	129
662	軽	軒	1	2
663	戸	. ハ	1	1
664	立	L v	1	1
665	場	揚	1	3
666	貨	貸	2	24
667	得	担	1	28
668	ブ	. フ	1	2
669	百	r =	1	1
670	額	順	1	21
671	主	+	1	88
672	年	ハ	1	55
673	三	一、一	1	1
674	与	写	1	36
675	三	. 「「	1	1
676	単	キ	1	64
677	撥	撓	1	1
678	」	' 一	1	2
679	て	で	1	1,782
680	三	一「一	1	1
681	綸	輪	1	7
682	作	“乍	1	1
683	担	+」」	1	1
684	袖	有h	1	1
685	目	「	1	1
686	強	弓螢	2	2
687	ば	ゝよ	1	2
688	ブ	ヴ	1	1
689	ば	C J	1	1

OCR誤読文字

93.5.28 170記事

	正しい読み	誤読	頻度	全出現数
690	都	者 l	1	1
691	戸	1 イ	1	1
692	”	N	1	77
693	好	女子	1	4
694	が	力 S	1	2
695	男	唱	1	3
696	く	ぐ	1	19
697	行	イ子	1	1
698	画	[向	1	1
699	作	イ作	1	1
700	打	キ丁	1	1
701	忙	↑亡	1	1
702	カ	力	1	217
703	在	任	1	48
704	合	ムロ	1	1
705	鉄	銑	7	7
706	益	査	1	80
707	」	L	1	42
708		、	1	16
709	市	申	1	59
710	漬	流	1	1
711	岸	串	3	4
712	氷	水	3	86
713	ン	、	1	129
714	]	一	1	930
715	員	]	1	40
716	員	」 員	1	1
	合 計		2,071	90,926

4.5.2 O C R 誤読補正の結果

誤読補正結果

記事No	元の語	補正した語
4	東芝	東芝
7	モンタ	モンク
7	持集	特集
8	申国	中国
8	日本	日本
9	量刑一	量刑
9	r 定	予定
9	東バトンルージュ郡裁判所	東バトンルージュ郡裁判所
10	せき	せぎ
10	フィリピン海	フィリピン海
10	プレート	プレート
10	レーザー光線	レーザー光線
10	通信総台研究所	通信総合研究所
10	せき	せぎ
10	フィリピン海	フィリピン海
12	一大学	大学
12	総台政策学部	総合政策学部
13	味	味
13	台弁事業	合弁事業
13	味	味
16	ホップ	ホップ
18	車一いす	車いす
18	ベッド	ベッド
19	セネコン	ゼネコン
22	製造物貞任	製造物責任
23	新明相工業	新明和工業
28	購入	購入
30	H経平均株価	日経平均株価
30	労働力不足	労働力不足
30	ブドウ	ブドウ
30	労働力不足	労働力不足
30	ホリイトカラー	ホワイトカラー
30	ブドウ	ブドウ
32	癒善	癒着
32	そく	ぞく
32	癒善	癒着
35	利下げ	利下げ
36	東芝機械	東芝機械
37	日本	日本
37	一都市	都市
37	スチユリーデス	スチユワーデス
39	スポーツ用品	スポーツ用品
44	一カ	一カ
44	一カ	カ
44	デノミ研究議員懇一談会	デノミ研究議員懇談会
44	コンピュータ	コンピュータ
44	一カ	一カ
46	日井連	日弁連
48	特別配当	特別配当
50	プラスチック	プラスチック

記事No	元の語	補正した語
50	大口需要家	大口需要家
50	輸入	輸入
50	プラスチック	プラスチック
51	口金	口金
52	一日付	日付
53	トップ	トップ
53	有刺鉄線	有刺鉄線
59	ピン	ピン
59	ピン	ピン
59	地力公共団体	地方公共団体
59	ピン	ピン
59	NED O	NEDO
60	スタ	スタ
64	力法	方法
69	グラフィックス	グラフィックス
69	LD I ROM	LD-ROM
69	グラフィックス	グラフィックス
69	LD I G	LD-G
69	カラーテレビ	カラーテレビ
70	移動体通信	移動体通信
72	一マリンVHF	マリンVHF
77	ベストセラー	ベストセラー
78	カセットデツ牛	カセットデツキ
78	クウリオン	クラリオン
78	オールインリン	オールインワン
78	CD I ROM	CD-ROM
78	クウリオン	クラリオン
79	輸入拡大	輸入拡大
79	日本半導体製造装置買協会	日本半導体製造装置協会
79	輸入拡大	輸入拡大
80	プロジェクト	プロジェクト
80	持許情報	特許情報
80	窓目	窓口
80	プロジェクト	プロジェクト
82	ベンチ	ベンチ
82	弁当	弁当
82	ビジネスマン	ビジネスマン
82	ベンチ	ベンチ
84	日本ユニセフ協会	日本ユニセフ協会
85	ニチメン	ニチメン
85	合併会社	合併会社
85	ニチメン	ニチメン
86	コンパクト化	コンパクト化
88	N I I SDN	N-I SDN
92	責任	責任
95	東京製造所	東京製造所
95	東芝	東芝
95	東京工場	東京工場
95	興味	興味
95	東京製造所	東京製造所
95	東芝	東芝
95	東京工場	東京工場

記事No	元の語	補正した語
95	興味	興味
96	記者クラブ	記者クラブ
97	マイクロプロセッサー	マイクロプロセッサー
97	エックス線	エックス線
97	日本電信電話	日本電信電話
97	マイクロプロセッサー	マイクロプロセッサー
97	エックス線	エックス線
100	ミニコンボ	ミニコンボ
101	日本メーカー	日本メーカー
102	クル	クル
104	川途	用途
106	プレーキ	ブレーキ
106	興味	興味
106	プレーキ	ブレーキ
106	興味	興味
107	東京	東京
111	ストツタ	ストック
113	マタラーレン	マクラーレン
113	コーネッタス	コーネックス
113	マタラーレン	マクラーレン
113	コーネッタス	コーネックス
114	焼却プラント	焼却プラント
114	日本スピンドル製造	日本スピンドル製造
114	焼却プラント	焼却プラント
114	日本スピンドル製造	日本スピンドル製造
114	コンベヤー	コンベヤー
114	セノコ	セッコ
115	DOS I V	DOS-V
115	パイロットスタディ	パイロットスタディ
115	ヘアカラー	ヘアカラー
116	プラスチッタ	プラスチック
116	投人	役人
116	輸入販売	輸入販売
116	プラスチッタ	プラスチック
123	ヘル	ヘル
125	シンボジウム	シンポジウム
126	年度末	年度末
132	オーター	オーカー
132	第三セクタ	第三セクタ
132	オーター	オーカー
132	第三セクタ	第三セクタ
137	ダンピング提訴	ダンピング提訴
137	日本ニット工業組合連合会	日本ニット工業組合連合会
137	Cろ	C3
138	東洋	東洋
140	ナショナルブランド	ナショナルブランド
141	のほり	のほり
142	パソコン通信	パソコン通信
142	フロッピーディスク	フロッピーディスク
142	パソコン	パソコン
142	フロッピーディスク	フロッピーディスク
144	ガラス工房	ガラス工房

記事No	元の語	補正した語
147	不二ビューティ	不二ビューティ
147	連	運
154	運搬中	運搬車
156	ブルー	ブルー
158	r 定	予定
158	東京	東京
158	ヘース	ベース
158	東京	東京
159	f 供服	子供服
159	百貨店	百貨店
159	ニケ	ニケ
160	善尺	着尺
160	申空糸	中空糸
160	善尺	着尺
163	ブランド	ブランド
165	ポイント	ポイント

4.6 自動抽出キーワードの「漏れ」 「ノイズ」分析

記事番号：（ ）が170記事の通し番号

種別

1 桁目～1:統制キーワード、2:フリーキーワード

2 桁目～1:一般、2:地域、3:企業名、4:人名

漏れ原因

点 : KW数制限のための点数が低い

日本語解析: 日本語解析の切り出し、補正、置き換え、追加、削除

創出KW : 本文に無い

辞書 : 辞書の登録・分類

中黒 : 中黒がある（今回の実験では日本語解析処理対象から除外）

省略語 : 本文中では省略語

機械の必須KWの漏れ

記事番号	種別 キーワード	漏れ原因	備 考
N0005(1)	110 判決	点	裁判があるので問題なし
N0006(2)	110 所得税	点	所得税減税があるので問題なし
N0020(3)	110 衆院選	誤読が原因で切りミス	同党の衆院選 h J 候補決定者
N0040(4)			
N0048(5)	110 発売	点	新製品があるので問題なし
N0050(6)	110 社長	日本語解析	
	110 社長人事	創出KW	
	110 人事	創出KW	
N0069(7)	110 メディア	点	報道、テレビがあるので問題なし
N0070(8)			
N0080(9)	120 米国	点	
	240 服部剛丈	辞書	(姓と名の判別強化)
N0090(10)	110 地震	点	直下型地震があるので問題なし
	210 レーザー反射衛星	新出KW	
N0094(11)	130 リクルート	中黒	
N0103(12)	230 K D D 研究所	辞書	KDD研究所(F1)
N0120(13)	110 会社決算	創出KW	
N0140(14)	110 発売	点	新製品があるので問題なし
N0143(15)			
N0147(16)	110 発売	点	新製品があるので問題なし
N0150(17)	110 調査	創出KW	
	130 アットホーム	日本語解析	辞書にC3あり。形容動詞形と判断
	110 調査結果	創出KW	
N0160(18)	110 身障者	点	
N0163(19)	110 臨海副都心	点	
	110 都市開発	創出KW	
N0180(20)	210 東海	点	東海地震があるので問題なし
N0182(21)			
N0191(22)	140 船田元	中黒	
	240 船田経企庁長官	辞書	
	110 社会党	日本語解析	本文中は"自民、社会、公明、民社"
N0200(23)			
N0207(24)	110 会社決算	創出KW	
N0208(25)			
N0210(26)	110 決算	創出KW	
	110 会社決算	創出KW	
N0220(27)	110 選挙	創出KW	
	120 港区	点	
	110 都議選	点	

記事番号	種別 キーワード	漏れ原因	備 考
	120 東京	点	
	120 中央区	点	
	120 文京区	点	
	120 新宿区	点	
N0230(28)	210 刑事告発	日本語解析	
N0238(29)	110 抜本改革	点	
N0240(30)	130 富士総合研究所	中黒	
	130 長銀総合研究所	中黒	
N0246(31)	220 尾瀬	点	尾瀬ヶ原があるので問題なし
N0247(32)	110 民主主義	点	
	110 政党政治	点	
N0260(33)	140 森喜朗	辞書	
	240 森通産相	辞書	
	110 経済支援	創出KW	
N0270(34)			
N0280(35)	210 E R M	辞書	
N0290(36)			
N0300(37)	230 ジャグソンエアラ	中黒	
	110 航空路線	日本語解析	
	230 モデルフト	中黒	モディ・ルフト
	130 ボーイング	日本語解析	G@ボーイング7371200型
			ユーザ辞書にボーイング737が登録されていたため
	230 イーストウエスト	中黒	イースト・ウエスト・エアラインズ社
	230 ジェットエアウエーズ	中黒	ジェット・エアウエーズ
N0310(38)			
N0320(39)	110 リゾート開発	日本語解析	リゾート事業があるので問題なし ／リゾート／開発事業
N0340(40)	110 企業業績	創出KW	
N0350(41)	110 株	日本語解析	持ち株、株価があるので問題なし
	110 投資家	点	外国人投資家があるので問題なし
	110 持株比率	辞書	本文中は「持ち株比率」 「持株比率」は旧辞書になし 最新辞書ではF1
N0370(42)	110 芸術	点	芸術活動があるので問題なし
	110 ボランティア	点	ボランティア団体、ボランティア活動を
り出しているがいずれも点数で落			
N0380(43)	230 オッペン化粧品	日本語解析	オッペン化粧品会長
N0389(44)	110 円	点	

切

記事番号	種別 キーワード	漏れ原因	備 考
N0393(45)	110 円相場	点	
	110 N H K	誤読	
	110 調査結果	創出KW	
N0394(46)	110 臓器移植	点	
	110 脳死	点	
	110 脳死臨調	点	
N0410(47)	110 輸出		武器輸出があるので問題なし
	120 米国	誤読、中黒	
	120 ロシア	点	3 回出現
N0420(48)			
N0425(49)	110 発売	点	新製品があるので問題なし
N0441(50)	110 内需	点	
	130 昭和電工	日本語解析	
	110 セメント業界	点	
	130 三菱化成	中黒	
N0450(51)			
N0460(52)	110 人事	日本語解析	人事交流があるので問題なし
	240 柏木昇	新出KW	
	110 大学人事	創出KW	
N0480(53)			
N0488(54)	110 国連平和維持活動	点	
N0489(55)			
N0490(56)	110 通産省	中黒	通産省・中小企業庁
N0500(57)			
N0504(58)	130 日産自動車	省略語	本文中では「日産」
	130 トヨタ自動車	省略語	本文中では「トヨタ」
	130 三菱自動車工業	省略語	本文中では「三菱」
	210 米自動車工業会	辞書	本文中は「米国自動車工業会」
N0520(59)	210 ゴミ発電	点	出現 4 回
	110 エネルギー	日本語解析	
	110 ごみ処理	創出KW	
N0530(60)	110 増産	創出KW	
	110 体制強化	創出KW	本文中では「生産体制を強化」
	110 1 6 M D R A M	辞書	16メガDRAMを同義語処理に
N0540(61)			
N0570(62)			
N0600(63)	110 新会社	点	
	230 エムアイエムホールディングス	誤読	エム・アイ・エムホールーアイノグス
N0610(64)	230 濤和化学	新出KW	
	210 M C 法	新出KW	

記事番号	種別 キーワード	漏れ原因	備 考
	110 技術開発	創出KW	
N0620(65)	110 発売	創出KW	本文中では「販売を開始した」
	110 新製品	創出KW	
N0630(66)	110 発売	点	新製品があるので問題なし
N0640(67)	210 常務	中黒	
N0660(68)	110 調査	点	
	110 中小企業	点	
	110 調査結果	創出KW	
N0666(69)	110 需要見通し	点	
	110 映像機器	点	
N0670(70)	110 通信機器	点	
	130 日立製作所	省略語	本文中では「日立」
	110 展示会	創出KW	
N0675(71)	210 単位料金区域	点	
	110 電話料金	創出KW	
N0680(72)	110 発売	点	新製品があるので問題なし
N0690(73)	110 駐車場		機械式があるので問題なし
	110 共同開発	創出KW	
N0710(74)	110 発売	点	新製品があるので問題なし
N0720(75)			
N0730(76)			
N0740(77)	110 クラシック音楽	創出KW	本文中では「クラシック」
N0750(78)			
	110 カーオーディオ	誤読	高級カーオーディオ
N0756(79)	130 日本エーエスエム	誤読	
N0758(80)	110 共同開発	点	
	110 工業技術院	中黒	
	130 ロールスロイス	誤読	英ロールスロイス
	130 石川島播磨重工業	誤読	
	130 I B M	誤読	
N0760(81)			
N0762(82)			
N0770(83)			
N0780(84)	110 発売	点	新製品があるので問題なし
N0790(85)	230 江蘇省服装進出口	新出KW	
	230 南京蘇綿服装有限	新出KW	
	230 南京市江寧県第二	新出KW	
	230 蘇州合田刺繍有限	新出KW	
N0810(86)			
N0820(87)	110 発売	点	

記事番号	種別 キーワード	漏れ原因	備 考
N0827(88)	110 I S D N	誤読	
N0840(89)	210 副会長	日本語解析	新副会長
	130 富士写真フイルム	中黒	
	110 人事	創出KW	
N0850(90)	120 東京	中黒	東京・東池袋
	220 東池袋	中黒	
	110 ショールーム	創出KW	
N0860(91)	210 感熱性高分子	新出KW	
	210 ポリオキシエチレ	新出KW	ポリエキシエチレンノニルフェニルエーテル
	210 ポリビニルメチル	点	ポリビニルメチルエーテル
	110 共同開発	創出KW	
N0870(92)			
N0900(93)	110 調査結果	創出KW	
	110 株式売買	創出KW	
N0910(94)	110 装置開発	日本語解析	駆動装置と開発があるので問題なし
	210 透明樹脂	日本語解析	未知語として切り出し
N0911(95)	110 みなとみらい 2 1	点	9 段目
	110 再開発	点	
	110 経済協力	日本語解析	
	130 三井不動産	誤読	
N0914(96)			
N0930(97)	110 次世代	点	
N0940(98)			
N0950(99)	110 発売	日本語解析	接辞(P)と分類。新製品があるので問題なし
N0960(100)	110 業務提携	日本語解析	"業務提携する"と切り出し
N0970(101)	210 トランスプラント	点	
N0980(102)			
N0990(103)	110 受注動向	創出KW	
N1010(104)			
N1030(105)	210 エコノミークラス	点	
N1040(106)	110 住宅設備機器	点	
N1060(107)	130 アップルコンピュータ	日本語解析	米アップルコンピュータ
	110 新製品	点	
N1070(108)	110 役員人事	創出KW	
N1090(109)			
N1091(110)	110 電波障害	点	
	210 製品事故	新出KW	製品、事故があるので問題なし
	210 通商産業検査所	点	
N1100(111)			

記事番号	種別 キーワード	漏れ原因	備 考
N1110(112)	110 会社決算	創出KW	
N1120(113)	110 F1 (モータースポーツ)	日本語解析	F1チーム
N1140(114)	210 ごみ焼却	誤読	都市ごみ焼却
N1150(115)			
N1160(116)	110 拡販	点	
	110 抗菌剤	中黒	
	110 市場投入	創出KW	市場に投入。市場があるので問題なし。
	110 防かび剤	中黒	本文では抗菌・防カビ剤
N1178(117)	110 放射線	点	
	110 原子力発電所	点	
N1182(118)	110 円高	日本語解析	円高差益があるので問題なし
N1190(119)	110 省エネルギー	日本語解析	省エネルギー技術があるので問題なし
	110 技術開発	点	
	110 地球環境	点	
	110 環境問題	創出KW	
N1200(120)	110 発売	日本語解析	接辞(P)と分類。新製品があるので問題なし
	130 アップルコンピュータ	省略語	本文中では「アップル社」
N1210(121)			
N1230(122)	110 乗用車	点	
	130 三菱自動車工業	省略語	本文中では「三菱系、三菱自動車」
N1240(123)			
N1250(124)	130 トヨタ自動車	省略語	本文中では「トヨタの」
N1260(125)			
N1270(126)	210 PC橋	新出KW	
	210 首都圏中央連絡自動車道	中黒	
	210 鶴ヶ島ジャンクション	新出KW	
N1280(127)	110 新築工事	創出KW	
N1290(128)	110 道路建設	創出KW	
N1299(129)	110 入札	点	
	110 公共工事	点	
N1300(130)	110 調査	日本語解析	接辞(P)と分類。調査結果があるので問題なし
	110 日本道路建設業協	中黒	(団体・団体のときの処理)
N1310(131)	110 社長人事	創出KW	
N1320(132)	120 千葉	日本語解析	後処理で対処可 千葉県→千葉
	210 新産業三角構想推進会議	新出KW	
	110 都市計画	点	
	210 幕張新都心	点	
	210 かずさアカデミアパーク	日本語解析	
N1330(133)	110 防衛施設局	点	
	110 建設計画	創出KW	建設工事があるので問題なし

記事番号	種別 キーワード	漏れ原因	備 考
N1339(134)			
N1340(135)			
N1350(136)	110 首都圏	中黒	
N1373(137)	110 ダンピング	中黒	
N1380(138)	110 廃棄物処理 110 環境保護	日本語解析 創出KW	産業廃棄物処理があるので問題なし
N1390(139)	110 ポリエステル長繊維	日本語解析	
N1400(140)	110 発売 110 秋冬物	点 点	新製品があるので問題なし
N1410(141)			
N1420(142)			
N1430(143)	210 サイトシグナル研究所 240 鶴藤丞 110 高齢化社会 110 遺伝子 110 インタビュ	新出KW 日本語解析 点 点 創出KW	サイト、シグナル、研究所長と分割した 鶴藤丞氏と切った
N1439(144)	110 ビール 110 複合施設	日本語解析 創出KW	本文中では「複合商業施設」
N1440(145)	110 調査結果	創出KW	
N1450(146)	110 化粧品	日本語解析	スポーツ化粧品があるので問題なし
N1460(147)	110 エステティック	点	
N1470(148)	110 管理システム 110 データベース 110 システム開発	日本語解析 点 日本語解析	システムがあるので問題なし システム、システム導入があるので問題なし
N1472(149)	110 株	日本語解析	接辞(P)とした。保有株がある
N1480(150)	210 バーチャルウインド	誤読	
N1490(151)	230 みつた 110 売り場面積	新出KW 点	
N1500(152)			
N1510(153)	230 有田ヴィオーシ 110 ソフト開発	中黒 日本語解析	
N1520(154)	110 倉庫 120 兵庫 230 神明 110 販路開拓 110 品質管理	点 日本語解析 日本語解析 創出KW 創出KW	後処理 兵庫県→兵庫 一般として切った
N1530(155)	110 調査 110 調査結果	点 創出KW	本文中では「調査をした結果」
N1540(156)	110 手袋 110 製品開発	日本語解析 創出KW	液体手袋があるので問題なし

記事番号	種別 キーワード	漏れ原因	備 考
N1550(157)	110 調査	日本語解析	接辞(P)とした
	110 調査結果	創出KW	
N1560(158)	230 ロープ	誤読	ロープ
	110 新規事業	創出KW	
	210 情報ネットワーク	創出KW	
N1561(159)	110 日本百貨店協会	誤読	=本
	120 東京	誤読	東京
N1570(160)	110 素材開発	創出KW	
N1580(161)	110 宝飾品	創出KW	
N1586(162)	110 事業戦略	創出KW	
N1590(163)	240 土岐訓代	新出KW	
N1592(164)	110 決算見通し	創出KW	
	110 会社決算	創出KW	
N1593(165)	110 経営計画	点	
N1600(166)			
N1602(167)	120 米国	日本語解析	一般(G)で切った
	110 農業交渉	点	
	110 日米関係	創出KW	
N1608(168)	210 給料制	新出KW	
	120 千葉	日本語解析	後処理で対処可 千葉県→千葉
	120 東京	日本語解析	後処理で対処可 東京都→東京
	110 調査		
	120 栃木	日本語解析	後処理で対処可 栃木県→栃木
	120 山梨	日本語解析	後処理で対処可 山梨県→山梨
	120 埼玉	日本語解析	後処理で対処可 埼玉県→埼玉
	120 長野	日本語解析	後処理で対処可 長野県→長野
N1610(169)	120 岐阜	日本語解析	後処理で対処可 岐阜県→岐阜
	210 キクイモ	点	
N1611(170)	110 捕鯨	点	
	120 ノルウエ	点	

ゴミ

ゴミの種別

①誤読

②キーワード候補語抽出ミス

品詞分解および後処理における抽出ミス

③種別ミス

④本来の意味ではなく、より広汎な意味で用いるもの

①血液の「血」②「血」がさわぐ

①のケースだけ抽出したいにもかかわらず②のケースを抽出してしまったためのゴミ

⑤④の逆で限定的に使っている場合だけ検出したいにもかかわらず一般的な使い方の場合も抽出してしまったためのゴミ

例：環境 自然環境の場合だけをKWとして抽出したい。

⑥複数の意味をもつもの

⑦主題でない

(a)本来、何かとの論理積として使うべきもの

例：利用、効果、システム

これらは、むしろゴミではなくつけるべき。なぜなら、これがあったために主題でないものがヒットすることは考えられない。

(b)例示として上げているもの。

例：「結婚」と同様、パートナー選びが難しいよう。

通常総代会（「株式会社」の「株主総会」に相当）

(c)その他

1.誤読

003	木	006	目	007	モンク	015	漢	029	宋
036	芝(1)	041	宋	054	P K[F3]	059	NED[F3]	071	リア[F3]
072	写真(2)	074	白	075	宋	071	あ	077	目
078	オールインワン	088	レクトロニクス	094	ブレ	103	宋	106	鳥(1)
114	オキ	114	シン	114	セツコ	123	ヘル	134	カル
136	京都	140	ブランド	148	宋	158	一体	158	白
159	ニケ	160	タム(1)	160	空	166	力	170	あ
170	水(7)								

[]はKW種別

()はヒット数

(12)

2. キーワード候補語抽出ミス

009	ひろ[F3]	012	電気通信[F3] (9)	015	いいの[F3]	016	ラル[F3]	030	ながら[F3]
036	森	044	える[F3]	045	はざま[F3]	046	える[F3]	050	ながら[F3]
050	ともや[F3]	053	すぎた	054	産業(産業春秋)	063	太平洋	065	王朝文
068	多摩	075	アック[F3]	076	戸(1)	082	ながら[F3]	86	うす
104	シム	114	スピンドル	121	港	130	道	136	戸
137	糸	140	マス	143	サイト[C1]	143	シグナル(F1)	147	エステ[F3]
148	東急[F3]	149	あさひ[F3]	153	まるなか[F3]	153	日本システム [F3]	154	あ[F3]
155	ゾーン[F3]	158	バス[C1]	167	硬さ[C1]	169	味	170	いいの[F3]

[]はKW種別 ()はヒット数 (10)

3. 種別ミス

014	WD[F3]	018	千代田[F3]	027	藤井[C3]	027	トップ[C3]	050	ラッシュ [F3]
052	トランスナシ ョナル[F3]	059	KW[F3]	062	ロボ[F3]	063	MIM	068	中央[F3]
081	産機[F3](1)	101	米[C1](4)	101	E S I [F3]	101	アダムス[F3]	104	内藤[C3]
114	バン[F3]	118	流通[F3](1)	122	リペロ[F3]	124	ブランド[F3](2)	125	シンボ[F3]
167	エスピ[F3]	170	R M S [F3]						

[]はKW種別 ()はヒット数 (8)

4. 広汎な意味(例: 血) <元々の意味に限定される場合だけKWを付ける>

007	網	007	音	019	視線	035	特効薬(9)	044	時代
052	血(3)	058	壁	059	光(3)	074	感覚	88	夢
104	感覚	109	水(8)	115	パイロット(2)	126	I C (3)		

<他に考えられるもの>: 匂い、花、卵、衝突 ()はヒット数 (28)

5. 狭義(本来の意味とは別の特別な場合で使われるときのみKWとする)

007	環境(15)	083	環境(13)	092	環境(7)				
-----	--------	-----	--------	-----	-------	--	--	--	--

<他に考えられるもの>: 採用 ()はヒット数 (35)

6. 複数の意味<どちらの意味でもとられるもの>

126	ランプ(2)	163	N C						
-----	--------	-----	-----	--	--	--	--	--	--

()はヒット数 (2)

7-a.主題でない～何かと論理積で使うべきもので主題となり得ないもの

005	効果	008	効果	009	人	011	人	011	処分
011	基準	012	システム	018	人	029	検討	032	現象(2)
033	協力	033	姿勢	038	実績	039	攻勢	045	人
046	認識	051	利用	053	人	056	認識(1)	062	組織(1)
064	変化	067	問題	067	心境	068	改善	069	内臓
071	人	073	利用	077	人	077	記録(1)	091	利用
092	社員	093	結果	093	売買代金	098	人(1)	099	円
102	軸	103	数字	106	企業(2)	107	データ	107	管理
107	位置(1)	110	通	111	データ	112	計上	117	元
117	発想	117	作業	118	時間(1)	119	検討	122	結果
125	経済	126	分別	135	推進機構	135	適用要件	137	組織
138	実績	139	事業	141	傾向	141	大手	142	時間(1)
149	管理	149	組織	150	設置	155	結果	158	提供
160	アップ	161	時価	164	本格化	164	結果	165	像
166	加速	166	利用	166	機能	168	人		

()はヒット数 (11)

7-b.主題でない～比喻

013	結婚(1)	034	株式会社	034	株主総会(2)	048	家賃(3)		
-----	-------	-----	------	-----	---------	-----	-------	--	--

()はヒット数 (6)

7-c.主題でない～その他

001	金	003	区	006	ヒューム管	008	政府	008	ミネラルウォーター(11)
009	死	011	手帳	012	情報(1)	012	国際	018	ベッド
019	地域	019	国政選挙	021	ロボット(5)	021	仕事(1)	021	動物
021	人	021	戦後	022	ゆとり(5)	022	行事	023	日本
026	日本	027	名前	029	道路(9)	031	色(1)	032	第二次大戦
032	作法(3)	032	社会党	037	空	037	墜落事故(2)	037	経済(1)
038	米国(1)	038	日本	040	景気回復(5)	040	黒字	043	社長
044	和光(2)	047	今井博	049	ピーターズ	051	病気	055	プロ
057	寺(1)	058	P R(2)	059	プロ	060	国内	063	日鉱
063	日本	063	社長(1)	065	価格	069	カラー	070	世界
070	競演	073	新明和	074	電話機(2)	076	商品	077	税込み
077	円	077	仕事(1)	078	円	080	プロ	082	食堂(4)
082	ウィークデー	082	レストラン(8)	086	カラー(1)	088	台所(5)	092	大学(3)
105	ビジネス	106	ブレーキ(3)	107	日本法人	108	社長(2)	109	たばこ(6)
109	食事(3)	112	東大	112	経常利益	112	売上高	113	アップ
114	日本	114	採用(6)	116	生活空間	116	武器	118	円安
121	航路	122	減益	123	VWD	133	弾薬(5)	140	S S P
141	ズーム	142	机	143	研究所長	144	暖流	144	芝生(1)
144	木	144	北欧	144	弁当(3)	144	トンネル	144	樹木(1)
146	カット	146	価格	146	顧客	147	クリニック	152	想
156	水(7)	158	ネット	161	ヘッド(1)	163	意匠(5)		

()はヒット数

(118)

4.7 インデクサー処理後の追加処理

インデクサー処理後の追加処理

1. 置き換え、追加などの処理

(1). 見出しよりのKW切り出し

- a. 見出しをスペース、=、括弧、句読点をデリミタとして分解し、EL辞書にあれば採用する。
- b. 見出しの、=より前の文字列を「」で囲み、カットキーワードとする。

(2). 置き換え

- a. 本文の全文を対象とし、辞書に指定した文字列があれば、指定した語に置き換える。

米 → 米国
英 → 英国
仏 → フランス

(3). 追加

- a. 本文の全文を対象とし、辞書に指定した文字列があれば、指定した語を追加する。

自民+ (自) → 自民党
自由民主党 → 自民党
公明+ (公) → 公明党
共産+ (共) → 共産党
民社+ (民) → 民社党
自・社+自、社 → 自民党+社会党
社公民+社・公・民+社、公、民 → 社会党,公明党,民社党
(社)+(社会) → 社会党
工場を建設 → 工場建設
跡地を利用 → 跡地利用
建設を計画 → 建設計画
周年を記念 → 周年記念
技術を開発 → 技術開発
システムを開発 → システム開発
システムを導入 → システム導入
情報を提供 → 情報提供
顧客を開拓 → 顧客開拓
工場の設備 → 工場設備
市場に投入 → 市場投入
ゴミ → ごみ
リストラ → リストラクチャリング
ソフト → ソフトウェア
オリジナル品 → オリジナル商品
CO₂ → 二酸化炭素
NO₂ → 二酸化窒素
NO_x → 窒素酸化物
DOS/V → DOS-V
千代田区+中央区+新宿区+港区 → 東京
品川区+目黒区+世田谷区+大田区 → 東京
中野区+練馬区+板橋区+文京区 → 東京
豊島区+北区+荒川区+葛飾区 → 東京
江戸川区+江東区+墨田区+台東区 → 東京
渋谷区+杉並区+足立区 → 東京
名古屋市 → 愛知
宮沢首相+宮沢総理 → 宮沢喜一
武藤外相 → 武藤嘉文
後藤田法相 → 後藤田正晴

林蔵相 → 林義郎
 森山文相 → 森山真弓
 丹羽厚相 → 丹羽雄哉
 田名部農相+田名部農水相 → 田名部匡省
 森通産相 → 森喜朗
 越智運輸相 → 越智伊平
 小泉郵政相 → 小泉純一郎
 村上労働相 → 村上正邦
 中村建設相 → 中村喜四郎
 村田自治相 → 村田敬次郎
 河野官房長官 → 河野洋平
 鹿野総務庁長官 → 鹿野道彦
 北北海道沖縄開発庁長官+北北海道開発庁長官+北沖縄開発庁長官 →
 北修二
 中山防衛庁長官 → 中山利生
 船田経済企画庁長官+船田経企庁長官 → 船田元
 中島科学技術庁長官+中島科技庁長官 → 中島衛
 林環境庁長官 → 林大幹
 井上国土庁長官 → 井上孝
 平岩経団連会長 → 平岩外四,経団連,会長
 永野日経連会長 → 永野健,日経連,会長
 速水経済同友会代表幹事 → 速水優,経済同友会,代表幹事

- b. 本文の第一段落を対象とし、辞書に指定した文字列があれば、指定した語を追加する。

発売 → 新製品
 開発*販売 → 新製品
 開発*発表 → 製品開発,新製品

- c. 本文の見出しを対象とし、辞書に指定した文字列があれば、指定した語を追加する。追加した語のEL辞書との突合はおこなわない。

「一村逸品」 → 一村一品運動

(4). 削除

- a. 本文の全文を対象とし、指定した文字列があれば、KW候補語より削除する。
=写真

4.8 調査研究チームメンバーと スケジュール

●調査研究チームメンバー●

チーム長 斉藤 裕 (副社長 情報・技術担当)

河合 正晃 (常務 技術担当)
 村田 弘 (取締役 情報担当)
 新山 迪雄 (取締役 営業担当)
 今別府芳暢 (技術部 担当部長)
 渡野辺 勇 (技術部 課長職)
 小島 博志 (情報部 機械化担当)
 菅谷 茂 (情報部 課長職)
 高梨 明 (情報部 シニアインテグラー)
 松延 健 (情報部 機械化担当)
 島田 誠一 (営業部 担当部長)
 大田 康博 (営業部 担当部長)
 岩元 正史 (営業部)
 小島志圭央 (営業部)

●実施スケジュール●

作業工程 \ '93 月別	7	8	9	10	11	12	1	2	3
①調査機器類の手配	↔								
②調査対象新聞の選定	↔								
③イメージをテキスト化		↔							
④誤読文字の調査			↔						
⑤誤読文字を正文字に変換				↔					
⑥テキストデータから キーワード抽出					↔				
⑦キーワードの品質評価						↔			
⑧システム化の検討							↔		
⑨レポート作成・提出								↔	

禁無断転載

平成6年3月発行

発行 財団法人 データベース振興センター

東京都港区浜松町二丁目4番1号

世界貿易センタービル7階

TEL 03-3459-8581

委託先 (株)エレクトロニック・ライブラリー

東京都品川区西五反田八丁目11番13号

TEL 03-3779-1218

印刷所 (株)城南コピーセンター

東京都品川区西五反田二丁目22番12号

