デジタル・ドキュメント　6－1
（1997．3．28）

# 記事データからの分類知識獲得に関する実験シミュレーション

森本　由起子°　　間瀬　久雄　　辻　洋

(株) 日立製作所　システム開発研究所

本報告では、新聞記事自動分類システムに不可欠な分類知識の獲得方法に関する７種類の実験シミュレーションについて述べる。本システムでは、既に分類済みの教師データから自動抽出したキーワードおよびその重みをカテゴリ別に分類知識ベースに格納しておき、入力記事から抽出したキーワードとのマッチングにより各カテゴリとの類似度を計算して分類すべきカテゴリを決定する。この種の分類方式では、（1）キーワードの抽出方法だけでなく、（2）キーワードの抽出範囲、（3）教師データの選定方法が分類精度に与える影響が大きく、これらの要素を考慮して分類システムを構築することが不可欠である。これを実証するために、新聞記事データ１年分を用いて２２のカテゴリに自動分類する評価実験を行い、上記要素の分類精度に及ぼす効果･影響を測定した。その結果、上記３種類の要素により、分類精度が大きく変化することを確認した。

# Experimental Simulation for Classification Knowledge Discovery from Newspaper Articles

Yukiko Morimoto,　Hisao Mase,　Hiroshi Tsuji

Systems Development Laboratory,　Hitachi, Ltd.

This paper describes experimental simulation to test the feasibility of a keyword-based newspaper-article classification system in development. The system seeks to identify keywords which characterize a variety of categories from a series of classified training newspaper articles. Thus, the method of trainig articles selection is as important as how the keywords are derived.

We did seven kinds of experimental simulation on how to derive keywords and on how to select training articles. Our system generated knowledge bases from training articles for maximum one year which were then used to classify 37,000 articles into 22 categories. The results have shown that considering the keyword derivation and training articles selection is effective for improving classification collectness.

## 1. はじめに

大量の記事情報を扱う新聞社では、複数の情報源から電子データをリアルタイムで受信すると同時に、社外へ配信している。これらの記事情報を一つの窓口で受信し、内容別に自動分類して一括管理し、必要な情報を効率よく検索／抽出／配信したいという要求が高まっており、記事自動分類システム実現への期待が大きくなっている。

そこで、新聞記事を既存カテゴリに自動分類するシステムの構築を検討した。我々は既にキーワードに基づきテキストを既存カテゴリに自動分類するツールＦＬＵＴＥ（Filtering Lens software system for Unclassified TExts）を開発しており、今回、本ツールを新聞記事分類に適用した。

キーワードベースの自動分類システムには、分類のための知識が不可欠である。ＦＬＵＴＥでは、既に分類済みの教師データから自動抽出したキーワードおよびその重みをカテゴリ別に分類知識ベースに格納しておき、入力記事から抽出したキーワードとのマッチングにより各カテゴリとの類似度を計算して分類すべきカテゴリを決定する。従って、知識ベースを作成する際に教師データをどのようにして選定するかについて配慮することは、キーワードをどのように抽出するかについて配慮することと同じくらい重要であると考えている。

そこで本報告では、キーワードの抽出方法および教師データの選定方法に関する７種類の精度比較実験を行い、分類精度がどのくらい変化・改善するかを検証した。

キーワードの抽出方法に関する実験では、

共通語・不要語の選定・除去

（１） ひらがなのみからなる単語除去

（２） 特定文字数以下の単語除去

（３） キーワード抽出範囲の違い

（４） 重みの分布の正規化

の効果・影響に関して精度を比較した。

また、教師データの選定方法に関する実験では、

（１）教師データの作成時期

（２）教師データ量

の影響に関する検証を行なった。本実験では、新聞記事データ１年分を用意し、各記事を２２の分類カテゴリに自動分類したときの分類精度を比較することにした。

以下、２章では、テキスト分類ツールＦＬＵＴＥの概要について簡単に述べ、３章では上記７種類の実験方法および結果について詳細に述べる。

## 2. ＦＬＵＴＥの新聞記事自動分類への適用

ＦＬＵＴＥは、特許，記事，論文，WWWページ等のテキスト文書を既存のカテゴリに自動分類するためのツールである。ＦＬＵＴＥは、次の前提条件を満たすシステムに適用可能である。

（１） カテゴリが予め定義されていること。

（２） カテゴリが互いに包含関係にないこと。

（３） 分類済み教師データが十分存在すること。

また、ＦＬＵＴＥの機能を次に示す（図2.1）。

（ア）キーワード自動抽出

テキストからその内容を特徴付けるキーワードを自動抽出し、出現頻度に基づく重みを各キーワードに割り当てる。

（イ）分類知識ベース自動生成

教師データから抽出されたキーワードから各カテゴリを特徴付けるキーワードおよびその重みを認定し、知識ベースに格納する。

（ウ）類似度計算によるカテゴリ決定

新規未分類文書からキーワードを自動抽出し、分類用知識ベースを参照して、適切な分類カテゴリを決定する。

その他、分類結果からの精度算出，分類結果を評価するための履歴出力などの補助機能を持つ。

## 3. 新聞記事自動分類実験

### 3.1.分類カテゴリ及び記事データの解析

本実験では、２２のカテゴリを分類体系として用いた。図3.1にそのカテゴリ一覧を示す。

本実験では、８８年１月～３月、９３年１０月～９４年３月までの最大９ヶ月分の分類済み記事データを知識ベースを作成するための教師データとし、９４年４月～６月までの３ヶ月分の記事を分類精度を測定するための評価データとした。

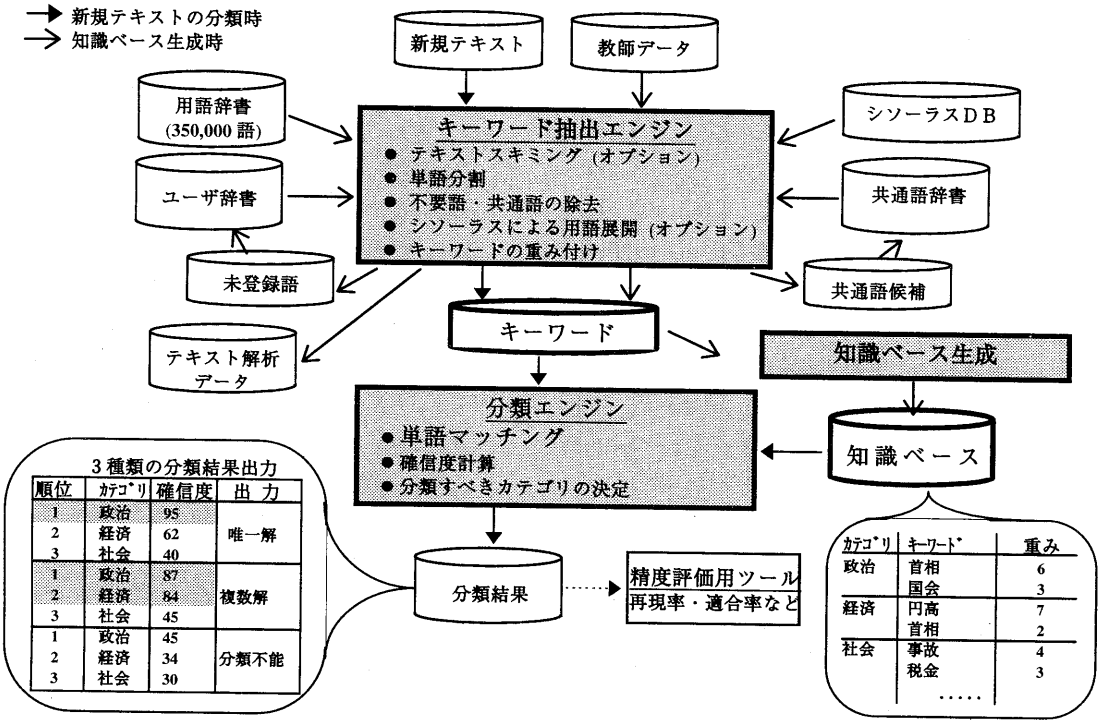


図 2.1 FLUTE の概要

| | | |
|---|---|---|
| 化学 | 情報・通信 | 資源・エネルギー |
| 建設 | 環境・公害 | 科学技術・文化 |
| 国際 | 社会・家庭 | 流通・サービス |
| 食品 | 金属・土石 | 繊維・木材・紙パルプ |
| 政治 | 経営・企業 | 機械・器具・設備 |
| 地域 | 経済・産業 | 農林・水産 |
| 地方 | 電子・電機 | 自然界 |
| 共通 | | |

図 3.1 カテゴリ一覧

表 3.1 記事データの記事件数

| | | |
|---|---|---|
| 教師データ | '88/1-3 | 32276 件 |
| | '93/10-12 | 36472 件 |
| | '94/1-3 | 34556 件 |
| 評価データ | '94/4-6 | 37462 件 |

表 3.2 キーワード種類数及び知識ベース規模

| | |
|---|---|
| 一記事あたりの平均カテゴリ数 | 3.5 カテゴリ |
| 知識ベースのキーワード種類数 | |
| 見出し＋第一段落 | 94028 種類 |
| 見出し＋本文全部 | 140923 種類 |
| 知識ベースの規模 | |
| 見出し＋第一段落 | 620719ﾚｺｰﾄﾞ |
| 見出し＋本文全部 | 1059721ﾚｺｰﾄﾞ |

| カテゴリ | キーワード | 品詞 | 重み |
|---|---|---|---|
| 政治 | 首相 | 名詞 | 15 |
| 経済 | 首相 | 名詞 | 8 |
| 社会・家庭 | 事故 | 名詞 | 20 |
| 企業・経営 | 決算 | 名詞 | 50 |
| 経済 | 決算 | 名詞 | 12 |

図 3.2 知識ベース (例)

表 3.1 に記事件数を示す。実験ではまず、ＦＬＵＴＥにより全教師用データからキーワード（名詞，動詞）を抽出した。テキスト解析には約３５万語の単語辞書を用いた。次に９４年１月～３月までの記事のキーワードデータから知識ベースを生成した。表 3.2 に上記記事データから抽出したキーワード種類数、及び生成された知識ベースの規模を示す。また、図 3.2 に知識ベースの例を示す。

### 3.2.評価尺度

分類結果の評価手法として、再現率、適合率が一般的である。再現率は付与すべきカテゴリのうち、どれだけ計算機が付与できたかを表し、適合率は計算機が付与したカテゴリのうち、どれだけ正解カテゴリが含まれているかを表す。再現率が高いほど分類の「もれ」が減少し、適合率が高いほど分類の「ノイズ」が減少する。一般に再現率と適合率の間にはトレードオフの関係がある。

本実験では、再現率／適合率を用いた二種類の評価尺度（以下、「正解率１」，「正解率２」と呼ぶ）を用いた。正解率１とは各記事に付与すべき正解カテゴリと同数のカテゴリを付与した場合の再現率（＝適合率）である。この評価尺度は、分類方式の有効性を比較評価するのに有効である。正解率２とは、あるしきい値よりも高い確信度（FLUTE が分類結果とともに出力する数値で、分類結果にどれだけ自信があるかを定量的に表した数値）をもつカテゴリを記事に付与した場合の再現率（＝適合率）である。実際には、各記事をいくつのカテゴリに分類すべきかは計算機には判らないので、確信度の高いカテゴリを記事に付与する。しきい値を高く設定すると、付与されるカテゴリの数が減少するため、再現率は下がるが適合率は上がる。逆に、低く設定すると、付与されるカテゴリの数が増加するため、再現率は上がるが適合率は下がる。本実験では、正解率２の算出のために用いるしきい値として、評価用データ３７４６２件に付与されるカテゴリの総数が、付与すべき正解カテゴリの総数と同じになるときの値を用いた。この評価尺度は、記事分類の全自動化を想定した場合の精度比較に有効である。

### 3.3.実験内容一覧

本報告では、キーワードの抽出方法および教師データの選定方法に関する７種類の精度比較実験を行い、分類精度の変化を検証した。

キーワードの抽出方法に関する実験では、

（１） 共通語・不要語の選定・除去
（２） ひらがなのみからなる単語除去
（３） 特定文字数以下の単語除去
（４） キーワード抽出範囲の違い
（５） 重みの分布の正規化

の効果・影響に関して精度を比較した。

また、教師データの選定方法に関する実験では、

（１）教師データの作成時期
（２） 教師データ量

の影響に関する検証を行なった。以下、個々の実験方法および実験結果について詳細に説明する。

### 3.4.共通語・不要語の選定・除去

#### 3.4.1. 実験の目的

キーワードとして不適当な単語を予め選定し除去することにより、キーワードを洗練化する。ここで「共通語」とは、多くのカテゴリに共通して出現する単語をさす。従って、共通語は教師用データの分野・内容・量などにより変動する。また、「不要語」とは、文章によらずキーワードとなり得ない単語をさす。不要語は共通語と違い、教師データに依存しない。

本実験では、２２カテゴリすべてに共通して出現する単語を共通語と定義した。また、単語辞書の中から手作業で抽出した一般的と思われる単語３２８７語を不要語とした。

図 3.3 に、共通語、不要語データの一例を示す。また、表 3.3 に知識ベースの規模の変化を示す。

#### 3.4.2.実験結果及び考察

表 3.4 に実験結果を示したように、共通語、不要語を除去した場合、分類精度が１．０％から７．６％向上した。また、不要語よりも共通語を除去した方が効果が大きかった。共通語は記事文章から選定したものであるため、記事の分類に対しては効果的に働いたと考えられる。したがって、これらの共通語を、他の分類体系あるいは他の文書（論文など）の分類で適用した場合、効果は比較的少ないと考える。

また、共通語の選定において、あるカテゴリに極端に多く出現する単語は共通語としないなど、カテゴリ別の重みの分布を考慮する必要がある。

### 3.5. ひらがなのみからなる単語の除去

#### 3.5.1. 実験の目的

ひらがなのみからなる単語が分類のためのキーワードとして適当であるか否かを検証し、不適当な場合、予め除去することにより分類精度を向上させ

る。除去したひらがなのみからなる単語の一例を図3.4に示す。また、知識ベースの規模を表3.5に示す。

◎共通語
11月　PR　TEL　アイデア　イメージ
会議　街　人々　調達　冬
波紋　発生　表現　訪問　名前
有名　用途　．．．
◎不要語
月曜日　最短　西北　要点　翌月　来期　来年
裏　利用　連夜　六月　．．．

図3.3　共通語、不要語データの一例

表3.3　不要語／共通語除去前後の知識ベース規模

| キーワード抽出範囲：見出し＋本文全体 | | |
|---|---|---|
| | 単語数 | 知識ベースの規模 |
| 除去前 | − | 1059721 ﾚｺｰﾄﾞ |
| 共通語 | 3627 語 | 929927 ﾚｺｰﾄﾞ |
| 不要語 | 3287 語 | 1034980 ﾚｺｰﾄﾞ |
| 共通＋不要語 | 6404 語 | 966406 ﾚｺｰﾄﾞ |

表3.4　不要語／共通語除去の効果

| | 正解率1 | 正解率2 |
|---|---|---|
| 除去前 | 65.32% | 61.34% |
| 共通語 | 72.72% | 66.62% |
| 不要語 | 67.39% | 62.55% |
| 共通＋不要語 | 72.93% | 66.83% |

あいさつ　あいまい　あおむけ　あかすり
あかつき　あかり　あくび　あぐら
あけぼの　あこがれ　あさぎり　あさひ
．．．．

図3.4　ひらがなのみからなる単語の一例

表3.5　ひらがな語除去前後の知識ベース規模

| キーワード抽出範囲：見出し＋本文全体 | | |
|---|---|---|
| | ひらがな語 | 知識ベースの規模 |
| ひらがな語除去前 | − | 1059721 ﾚｺｰﾄﾞ |
| ひらがな語除去後 | 4546 語 | 1025971 ﾚｺｰﾄﾞ |

表3.6　ひらがな語除去の効果

| | 正解率1 | 正解率2 |
|---|---|---|
| ひらがな語除去前 | 65.32% | 61.34% |
| ひらがな語除去後 | 66.08% | 61.71% |

### 3.5.2.　実験結果及び考察

表3.6に実験結果を示したように、ひらがなのみからなる単語を除去した場合、1％以下の分類精度の向上しか見られなかった。これは、新聞記事には、ひらがなからのみなる単語があまり出現しないことが原因であると考えられる。

## 3.6.特定文字数以下の単語の除去

### 3.6.1.　実験の目的

この実験の目的は、キーワードを構成する文字数と分類精度との間の関係を検証し、特定の文字数以下の単語を除去することにより、分類精度を向上させることにある。図3.5に特定文字数の単語の一例を示す。また、表3.7に知識ベースの規模を示す。

### 3.6.2.　実験結果及び考察

表3.8に実験結果を示したように、特定の文字数以下の単語を除去した場合、1．3％～7．4％の分類精度の向上が見られた。特に、2文字以下、3文字以下の単語を除去した場合、向上が顕著であった。これは、2文字以下、3文字以下からなる単語には、記事内容を特徴付けない一般的な単語が多く含まれているためである。しかし一方で、重要なキーワードも多く含まれているので、単語の文字数だけでキーワードか否かを決定するではなく、他のパラメータと組み合わせて決定することにより、分類精度をより向上させることができると考える。

## 3.7.キーワード抽出範囲の違いの影響

### 3.7.1.　実験の目的

新聞記事には、重要な内容は冒頭に記述するという特徴がある。そこで、キーワードの抽出範囲を見出し＋第一段落のみとし、分類精度がどのように変化するかを検証する。また、知識ベースを作成する場合には、大量のデータが必要であり、データ量が多ければ多いほど、そのデータ解析に、時間／コストがかかる。そこで、解析量が少なくても、分類精度を保持することが出来れば、知識ベースの作成、メンテナンスに、時間／コストをかけずに済むというメリットがある。

表3.9に知識ベースの規模を示す。

| 1文字 | 2文字 | 3文字 | 4文字 |
|---|---|---|---|
| 雨 | 安産 | ＣＣＤ | ソケット |
| 胃 | 衣装 | ランチ | ドル相場 |
| 円 | 永遠 | 営業店 | 移動電話 |
| 燕 | 華僑 | 久居市 | 焼き付け |
| 塩 | 怪人 | 芸術品 | 名所旧跡 |
| 音 | 学歴 | 皇太子 | 米ＣＩＡ |

図 3.5　特定文字数の単語の一例

表 3.7　特定文字数の語除去前後の知識ベース規模

| キーワード抽出範囲：見出し＋本文全体 | | |
|---|---|---|
| | 単語数 | 知識ベースの規模 |
| 除去前 | － | 1059721 ﾚｺｰﾄﾞ |
| 1文字以下の語除去後 | 3023 | 1017073 ﾚｺｰﾄﾞ |
| 2文字以下の語除去後 | 27255 | 767458 ﾚｺｰﾄﾞ |
| 3文字以下の語除去後 | 50262 | 586030 ﾚｺｰﾄﾞ |
| 4文字以下の語除去後 | 90109 | 303700 ﾚｺｰﾄﾞ |

表 3.8　特定文字数の語除去の効果

| | 正解率1 | 正解率2 |
|---|---|---|
| 除去前 | 65.32% | 61.34% |
| 1文字以下の語除去後 | 67.57% | 62.63% |
| 2文字以下の語除去後 | 72.15% | 65.81% |
| 3文字以下の語除去後 | 72.79% | 66.74% |
| 4文字以下の語除去後 | 68.81% | 65.12% |

表 3.9　ｷｰﾜｰﾄﾞ抽出範囲限定前後の知識ベース規模

| | |
|---|---|
| 見出し＋第一段落 | 620719 レコード |
| 見出し＋本文全部 | 1059721 レコード |

表 3.10　キーワード抽出範囲限定の効果

| | 正解率1 | 正解率2 |
|---|---|---|
| 見出し＋第一段落 | 65.32% | 61.34% |
| 見出し＋本文全部 | 67.16% | 62.60% |

### 3.7.2.実験結果及び考察

表 3.10 に実験結果を示したように、キーワードの抽出範囲を見出し＋第一段落のみとした場合、分類精度が１．３％から１．８％向上することがわかった。このことから、キーワードの抽出範囲を、見出し＋題一段落に絞っても、分類精度を保持できることがわかった。

## 3.8.　重みの分布の正規化

### 3.8.1.　実験の目的

キーワードに付与された重み（出現頻度）の価値がカテゴリによらず同一になるように重みを正規化（補正）することにより、分類精度が向上するか否かを検証する。

### 3.8.2. 実験方法

従来のキーワード重み付け方法の問題点は、カテゴリ別の記事数に格差があると、カテゴリによって重みの分布幅に差ができてしまうことであった（図 3.6）。その結果、あるカテゴリで最大の重みを持つキーワードでも、記事数の多いカテゴリのキーワードと比較すると、それほど大きな重みでないという状況が起こる。

このように、重みの価値がカテゴリ間で異なってしまっている状況を解決し、重みの価値をカテゴリによらず統一するために、カテゴリ毎に重みの分布をある一定幅に正規化（補正）することを考えた。図 3.6 では、多くの記事データを教師用データとして持つカテゴリＸに含まれるキーワードの重みは、１～１０００まで広範囲に分布しており、記事データの少ないカテゴリＹについては、１～５０までの狭い範囲に分布している。これらの分布を正規化することは、１～Ｎという同一の幅にマッピングすることを意味する。

重みの分布を１～Ｎ（Ｎ＞１）の間に正規化する方法として、以下の二種類が考えられる。

（１）各カテゴリについて、キーワードの重みの最大値と最小値に着目し、最大値がＮに、最小値が１にそれぞれマッピングされるように、各キーワードの重みの値を相対的に補正する。

（２）各カテゴリについて、重みの分布から各キーワードの重みの偏差値を求める。偏差値がある値Ｍとなるときの重みを最大値とし、最大値がＮに、最小値が１にそれぞれマッピングされるように、各キーワードの重みの値を相対的に補正する。

上記（１）の方法では、重みの最大値が２番目に大きい重みの値と比べて極端に大きい場合、に示すように、正規化後の重みの分布が偏ってしまう。一方、上記（２）の方法では、正規化後の重みの分布が偏らない。そこで、本実験では（２）の方法を採用した。

偏差値は一般的に知られている算出方法で計算する。重みの最大値を決定する偏差値Ｍを変化

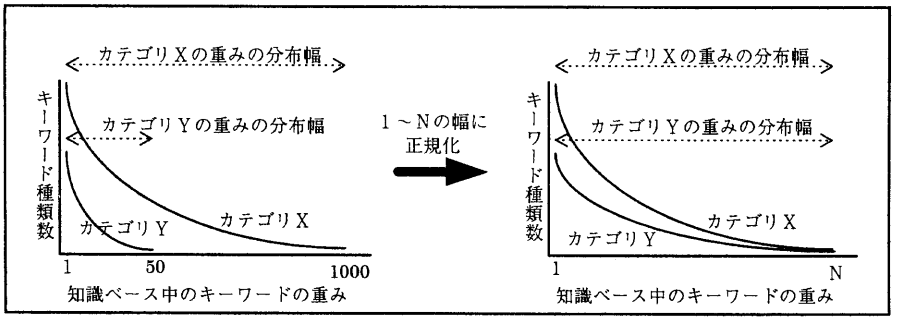


図 3.6　キーワードの重みの関係

表 3.11　実験方法

| 実験項番 | 実験内容 |
|---|---|
| 0 | 正規化をしない |
| 1 | 偏差値M＝２００として正規化 |
| 2 | 偏差値M＝１５０として正規化 |
| 3 | 偏差値M＝１００として正規化 |

表 3.12　キーワードの重み正規化の効果

| | 正解率１ | 正解率２ |
|---|---|---|
| 正規化なし | 65.32% | 61.34% |
| 偏差値２００ | 73.82% | 67.00% |
| 偏差値１５０ | 74.21% | 67.48% |
| 偏差値１００ | 74.77% | 68.13% |

させて分類実験を行ない、結果を比較した。

実験方法を表 3.11 に示す。なお、正規化の幅を示すNの値を１００とした。すなわち、正規化後の重みは、すべて１以上１００以下である。

### 3.8.3.　実験結果及び考察

表 3.12 に実験結果を示したように、重みの正規化をした場合、4.3%～14.2%の分類精度向上が見られた。正規化によりカテゴリ間の記事数の格差を是正することができたと考える。また、本手法により、記事数の比較的少ないカテゴリについて分類精度を飛躍的に向上させることができた。重みの正規化をしない場合、記事数の多いカテゴリについては精度が良いが、そうでないカテゴリについては精度が極端に悪かった。これは、記事数の多いカテゴリの方が、知識ベースにおけるキーワードの種類数が多く、また、比較的高い重みが付与されるためである。本方式は、全体の分類精度を向上させるだけでなく、カテゴリ別の分類精度を均質化するのにも有効な手段であると言える。

## 3.9.教師用記事データの作成時期

### 3.9.1.　実験の目的

時が経つとともに記事の内容は変化するので、知識ベースやカテゴリの更新保守は必要不可欠な作業となる。しかし、知識ベースの更新により分類精度を維持することと、その作業コストの間にはトレードオフの関係がある。従ってメンテナンス作業は必要最少限に止めるのが望ましい。

本実験の目的は、知識ベースの継続的保守の必要性を検証することである。すなわち、実運用時において、どの程度の期間毎に知識ベース等の更新、改良を行なえば良いかを検証することである。

表 3.13 に本実験で使用した記事データの件数及び知識ベースの規模を示す。

### 3.9.2.実験結果及び考察

表 3.14 に実験結果を示したように、９４年１月から３月までの記事を使った場合の方が８８年１月から３月までの記事を使った場合よりも、精度が約３％良かった。すなわち、評価データに近い記事を使って知識ベースを作成した方が分類精度が良かった。これらのことから、知識ベースは定期的に更新する必要があると考える。

## 3.10.教師用記事データ量の影響

### 3.10.1.　実験の目的

記事自動分類システムにおいて、大量の記事データを用いて知識ベースを充実させることにより、分類精度を向上させることが可能である。その反面、大量の記事データの保守負担や、処理時間の増大等の問題が生じる。

本実験では、分類精度を安定させるためには、どの程度の教師用データがあれば必要十分であるかを検証する。表3.15に本実験で使用した記事データの件数及び知識ベースの規模を示す。

表3.13 教師ﾃﾞｰﾀ作成時期考慮前後の知識ﾍﾞｰｽ規模

| キーワード抽出範囲：見出し＋本文全体 | | |
|---|---|---|
| 教師用データ | キーワード種類数 | レコード数 |
| '94/1-4 | 126379語 | 923959 ﾚｺｰﾄﾞ |
| '88/1-4 | 136244語 | 966406 ﾚｺｰﾄﾞ |

表3.14 教師ﾃﾞｰﾀ作成時期考慮の効果

| 教師用データ | 正解率1 | 正解率2 |
|---|---|---|
| '94/1-4 | 75.66% | 68.95% |
| '88/1-4 | 72.04% | 65.78% |

### 3.10.2.実験結果及び考察

表3.16に実験結果を示したように、2週間（6461件）から1ヶ月分（12578件）の教師用データ量で分類精度が安定することがわかった。1件の記事あたりのカテゴリ数が平均3．5カテゴリであることを考慮すると、知識ベース作成に必要な1カテゴリあたりの記事数は、平均約1000件である（2週間分教師用データの場合）。また、教師用データの記事数が6ヶ月分の場合、分類精度が頭打ちとなり、分類精度がかえって悪くなるという結果となった。教師用データが増加すると、キーワードが充実する反面、ノイズとなるキーワードも増える。このトレードオフが影響していると考えられる。

## 4. おわりに

新聞記事を既存のカテゴリに自動分類するシステムの構築を検討した。キーワードベースの自動分類システムでは、教師データの選定方法はキーワード抽出方法とともに分類精度に大きく影響を及ぼすことを実験により確認した。

表3.15 教師データ量考慮前後の知識ベース規模

| キーワード抽出範囲：見出し＋本文全体 | | | |
|---|---|---|---|
| 教師用データ | 記事数 | キーワード種類数 | レコード数 |
| 1週間分 ('94/3/25-31) | 3018件 | 41481語 | 243563 |
| 2週間分 ('94/3/16-31) | 6461件 | 61749語 | 387924 |
| 1ヶ月分 ('94/3) | 12578件 | 85787語 | 560867 |
| 3ヶ月分 ('94/1-3) | 34556件 | 136244語 | 966406 |
| 6ヶ月分 ('93/10-'94/3) | 71028件 | 182334語 | 1361958 |

表3.16 教師データ量考慮の効果

| 教師用データ | 正解率1 | 正解率2 |
|---|---|---|
| 1週間分 | 67.54% | 63.60% |
| 2週間分 | 73.65% | 67.91% |
| 1ヶ月分 | 74.88% | 68.72% |
| 3ヶ月分 | 75.66% | 68.95% |
| 6ヶ月分 | 75.63% | 68.50% |

**謝辞**

本研究の機会と、本研究への貴重な御意見、御助言を頂いた（株）日本経済新聞社システム局、データバンク局の関係者の方々に感謝致します。また、本実験にご協力頂いた（株）日立製作所情報システム事業部の関係者の方々に感謝します。

## 5. 参考文献

[1] 辻 洋他：テキスト自動分類エキスパートシステムの一構成法：情報処理学会第49回全国大会講演論文集（第3分冊,3-93,1994.9）

[2] 間瀬 久雄他：テキスト分類支援ツール FLUTEの開発（1）－機能と構成－：情報処理学会第52回全国大会講演論文集（第3分冊,3-303,1996.3）

[3] 森本 由起子他:テキスト分類支援ツールFLUTEの開発（2）－障害事例分類への適用－：情報処理学会第52回全国大会講演論文集（第3分冊,3-305,1996.3）

[4] 森本 由起子他：新聞記事自動分類システム構築の検討と評価：情報処理学会第53回全国大会講演論文集（第3分冊,3-205,1996.9）