Library and Information Science No. 39 1998

図書をNDCカテゴリに分類する試み

An Experiment of Automatic Classification of Books Using Nippon Decimal Classification

石 田 栄 美
Emi Ishida

Résumé

In information retrieval, texts are usually retrieved by them with queries. In this study, an approach was suggested that texts are automatically classified into categories and retrieved by matching them with queries classified in the same way. For an efficient information retrieval using automatic classification, extracting methods of words from texts and matching methods are essential. Some extracting methods from Japanese texts have been suggested in natural languages processing. However, it is difficult to extract significant words from Japanese texts because Japanese texts are written without blank space separating words. As for matching methods, many weighting methods have been suggested as well as vector space models and probabilistic models.

This article reports the results of an experiment of classifying Japanese texts into Nippon Decimal Classification (NDC) categories based on the title information in Japanese MARC records. In this experiment, three extracting methods: —juman, MHSA, n-gram—are tested on a set of 1,000 books. Four weighting methods: —relative term frequency between categories, tf·idf and tf (max)·idf—are tested. The results indicate that the extracting method using juman achieved best and the best weighting method was the relative term frequency between categories, being able to select correct classification categories (upper three digits of NDC) for about 55.9% of 1,000 books.

石田栄美：慶應義塾大学大学院文学研究科図書館・情報学専攻，東京都港区三田 2-15-45
Emi Ishida: Graduate School of Library and Information Science, Keio University, 2-15-45, Mita, Minato-ku, Tokyo 108-8345, Japan
受付日：1999年8月9日 改訂稿受付日：1999年11月25日 受理日：1999年12月22日

I. はじめに

情報検索研究においては，従来，検索質問と検索対象集合中の個々の文献とを照合するアプローチが基本的にとられてきた。そこでは，検索質問や個々の文献中に出現する単語の出現頻度を統計的に分析することで，よりよい結果を導くための照合手法が多数提案されてきた。

本研究では，検索質問と個々の文献との照合を行う従来の検索手法のアプローチとは異なり，文献集合と検索質問の両方を分類することで利用者の要求に適合した文献群の提供を目指す。つまり，文献集合はあらかじめ主題ごとに分類しておき，検索質問も一種のテキストとみなし，新たに分類し，検索質問と同じ分類カテゴリに属する文献群を示せば，それが利用者の要求に適合した文献群であろうと考える。このアプローチは，分類という視点から情報検索を考えるというアプローチといえよう。

そのためには，まず，テキストを主題ごとに自動分類できることが必要である。分類を自動化する最大の利点は，大量の文献（テキスト）を一貫した基準で処理できることである。日本で出版される新刊図書ですら年間6万件を超えており，分類を図書以外のテキストにも広げようとするなら，すべてを人手にたよる手法は事実上不可能である。また，人がテキストを読んで主題を分類する際には，人によって判断基準が違うため，同じ主題を持ったテキストでも違う主題に分類されてしまうといった問題がある。たとえば，Japan MARC では，『インターネットイエローページ』（ナツメ社）と『みるみるわかるインターネット最新事情：イントラネット，Java，電子マネーからWebTVまで』（技術評論社）という図書はインターネットという同じ主題について書かれているにもかかわらず，付与されているNDCの分類記号はそれぞれ007.35と547.483であり，主題の異なる分類記号が付与されている。

さらに，自動分類は，「人が分類する」という行為に機械がどこまで近づけるかという方向だけでなく，大量のデータを用いることにより，テキスト群とそれが属するカテゴリとの関係から，人が分類する際には利用しない規則や知識を導き出すことができる可能性がある。これは，データマイニングとして研究されていることと同じ方向を目指すものである。

テキストの自動分類に関する研究は，電子化されたテキストが増加するのに伴い，1960年代ころからなされてきている。テキストを分類する方法には，大きく分けて，カテゴライゼーションとクラスタリングがある。カテゴライゼーションは既存のカテゴリにテキストを分類することであり，クラスタリングはテキスト集合の中から類似したテキストをグループ化していくことにより，テキストを分類することである。両者の分類方法とも，従来の研究において，様々な分類手法が提案されてきている。本研究では，カテゴライゼーションを対象にして，どのような分類手法が有効であるかを検討する。

また，テキストの自動分類の要素として，大きく関わってくるのは分類手法のほかに，分類対象となるテキストがある。従来の自動分類に関する

研究の言語の問題では，分類対象として，主に英語テキストが扱われてきた。日本語テキストは英語のように分かち書きされておらず，文を単語に切り分けることが難しいため，自動分類の研究が進まなかった。しかし，1980年代ころから日本語テキストの自然言語処理技術の研究が発展し，日本語の文を単語に切り分ける形態素解析の研究がさかんに行われるようになった。

日本語が分かち書きされていないことは自動分類の研究が遅れるというデメリットとして作用したが，他方，日本語の持つ特徴を生かした切り出し手法を考えることもできる。日本語は，漢字，かな，カタカナなど複数の文字種からなるが，漢字は表意文字であるので1文字でも主題を表わす場合がある。日本語の文から検索や分類の手がかりとなる文字列を切り出す場合，カタカナ文字列だけ，漢字文字列だけなど文字種によって切り出したり，漢字だけを用いたりでき，日本語の特徴は豊富な可能性を持っている。

また，検索や分類の場合，文から切り出した文字列は検索や分類の手がかりとして用いるだけであり，切り出された結果が必ずしも意味をもつ文字列である必要はない。検索では，後述するn-gramのように，文を数文字ずつの文字列に切り出し，それを索引語として用いている場合もある。このように，分類の手がかりとするための日本語の文からの文字列の切り出しには，いくつかの方法が考えられる。

本研究では，日本語テキストを用いてカテゴライゼーションを行い，その際に用いる重み付け手法の検討，および，日本語の文からの文字列の切り出し手法に関する検討を行う。

II. 先行研究

既存のカテゴリへテキストを分類するカテゴライゼーションは，1960年代ころから研究され始めた。これは，予め分類済みのテキスト集合からカテゴリごとの特徴を抽出し，テキストの特徴とカテゴリの特徴を照合することによって，最も類似した特徴を持つカテゴリにテキストを分類する手法である。ここで，テキストの特徴表現にどのような手法を用いるかが，分類の精度に大きく関わってくる。代表的なモデルには，情報検索研究で用いられている手法でもある確率モデルやベクトルモデルなどがある。

確率モデルは，Maron[1]らが提案したモデルをもとにしており，予め分類済みのテキストやテキスト集合に出現する単語の出現頻度を用いて，テキストが各カテゴリに分類される確率を求める。確率モデルでは，この確率の求め方が分類精度に大きくかかわる。確率モデルによる分類は，Hamillら[2]が，条件付き確率を用いて *Chemical Abstracts* の文献を自動分類した実験をはじめとして，様々な研究[3),4)]が行われている。

以下では，まず，カテゴライゼーションで用いられている最も代表的な手法であるベクトルモデルについて述べる。次に，日本語テキストからの文字列の切り出し手法について述べる。

A. ベクトルモデル

ベクトルモデルは，カテゴリの特徴を表現したカテゴリ特徴ベクトルと分類対象となるテキストを表現したテキスト特徴ベクトルの類似度を計算することによって，類似度が最も大きいカテゴリにテキストを分類する。カテゴリ特徴ベクトルは，カテゴリに予め分類済みのテキスト集合中で出現する単語の統計的情報をもとに求められる単語とカテゴリとの重みで表現される。重みは，テキストやテキスト集合中で単語の重要性を表現している。そのため，重みが大きくなればその単語はテキストやテキスト集合中で重要であり，小さくなればそれほど重要でないということがいえる。

カテゴリ C_i における特徴ベクトルは $C_i=(w_{i1}, w_{i2}, \ldots, w_{ij})$ $(i=1, 2, 3, \ldots, \mathrm{N})$ のように表わされる。このとき，w_{ij} は単語 T_j のカテゴリ C_i における関連度を表わす。テキスト特徴ベクトルは $q_i=(w_{qi1}, w_{qi2}, \ldots, w_{qij})$ $(j=1, 2, 3, \ldots, \mathrm{M})$ のように表わされ，w_{qij} は単語 T_j のテキスト q_i における重要度である。

ベクトルモデルにおいて，分類精度に大きくかかわるのは，重み付け手法である。重み付け手法

は，カテゴリごとに予め分類済みのテキストを用いて，そのテキスト中における単語の出現頻度情報をもとに単語にカテゴリごとの重みを付けるものである。

重み付け手法には，tf·idf[5]やカテゴリ間での単語の相対出現率を重みに利用する方法や，ニューラルネットワークなどを用いて，分類してある文書の特徴を分析することにより重みを学習する方法[6),7)]などが提案されている。以下では，tf·idfとカテゴリ間での相対出現率を重みにする手法について述べる。

1. tf·idfによる重み

tf·idfは，各テキスト内での単語の出現頻度tfと逆文献頻度idfの積である。この重みは，各テキストにおいて出現頻度が高く，テキスト集合全体での出現が低い単語の重みが大きくなる。tf·idfは，検索・分類の分野で様々な研究が行われている。

塩見ら[8)]は，重要語とシソーラスを対応付けることにより文書データを階層的に分類し，自動的に適切なメニューを作成するメニュー検索システムを提案している。この実験では，重要語の選択に，tf·idfを用いている。この実験で作成したメニュー検索システムで，評価用テストコレクションであるBMIR-J1を検索したところ，73%のデータを検索できた。また，徳永ら[9)]は，新しいテキストの索引語の重み付け手法として，idfを改良したWIDFという重み付け手法を提案している。idfが単語が出現するテキスト数をもとに重みを求めるのに対し，WIDFは単語のテキストでの出現回数をもとに重みを求める方法である。『現代用語の基礎知識』のテキストの自動分類をした結果，tf·idfを用いるよりも7.4%精度が改善された。

tf·idfは情報検索の分野で幅広く用いられており，その考え方を用いた様々な式が数多く提案されている。しかしながら，これは検索質問とテキストとの照合において重要語を決定し，テキストの順位付けなどをするための手法であり，テキストの自動分類という立場で有効な手法であるかどうかの検討はあまりされていない。

2. カテゴリ間での単語の相対出現率による重み

カテゴリ間での単語の相対出現率による重みとは，単語の出現頻度をカテゴリ間で比較したときに，出現頻度の偏りが大きかったカテゴリに対して大きな重みを付けるという考え方である。

渡辺ら[10)]は，特定の分野に偏って出現する漢字に注目し，新聞記事を分類している。この実験では，分野の重要漢字をχ^2法の考え方にもとづいて自動的に抽出し，抽出した重要漢字に分野内での出現頻度をそのまま重みとして，分類を行った。学習用サンプルとして百科事典の項目説明文を用い，新聞記事を42の分野に分類したところ，朝日新聞の天声人語2,000件で47%，社説3,000件で74%，日経サイエンスの論文記事162件で85%の分類に成功している。また，河合[11)]は，単語に意味属性を持たせ，各分類分野ごとに偏って出現する意味属性を分類システムが学習することによって，新聞記事の分類を行っている。各分野における出現頻度から重みへの変換にχ^2法を用いている。その他には，藤井ら[12)]の研究もある。

カテゴリにおける重要単語を抽出する際に，χ^2法の考え方を用いた手法はいくつか提案されている。しかしながら，テキストの自動分類において，カテゴリ間での単語の相対出現率による重み付けは適用例が少なく，適当な手法が確立されていないのが現状である。

3. 重み付け手法を比較した研究

複数の重み付け手法を比較したものとして，Larson[13)]が行ったMARCレコードの書名と件名標目から図書にアメリカ議会図書館分類法(LCC)の分類記号を付与する研究がある。自動分類では，語幹処理方法と重み付け手法，分類対象の表現方法の3つの手法を組み合わせて用いている。この研究では，3種類の語幹処理方法と4種類の重み付け手法と5種類の検索対象の表現方法のすべてを組み合わせた60通りで実験を行った。語幹処理方法は，キーワードの接尾辞を

取り除く語幹処理方法，複数形を単数系に直すという一部だけの語幹処理方法，件名標目に統制する方法が使われている。用いた重み付け手法は，単純な単語の出現，tf･idf，確率モデルの考え方を用いた確率型，カテゴリ間での相対出現率をもとにした相対出現率型である。分類対象の表現方法は，書名，すべての件名標目，最初の件名標目の3種類を組み合わせた5種類である。383件の評価用データに対し分類を行ったところ，重み付け手法にカテゴリ間での相対出現率をもとにした相対出現率型を用い，分類対象の表現方法として最初の件名標目を用いた場合の結果が最もよく，正しい分類記号が1位にランクされた割合は46.6% であり，10位までにランクされた割合は74.4% であった。

B. 文字列の切り出し手法

分類の手がかりとするために日本語の文から文字列を切り出すための手法は自然言語処理の分野で研究がされており，代表的な切り出し手法には，形態素解析とn-gramがある。

形態素解析とは，文を適切な形態素に分割する処理のことである。形態素とは，意味を持つ最小の要素のことであり，文を構成している。形態素解析には，意味を持つ単語を切り出すことができるという特徴があるが，辞書を持たなければならず，辞書にない単語は切り出せない，適切に切り出せない場合があるなどの問題がある。文を形態素に切り分け，品詞情報などを付与する代表的な形態素解析システムには，京都大学長尾研究室のjuman[14]と奈良先端科学技術大学院大学松本研究室のchasen[15]がある。

その他，形態素解析システムの変形として，長谷部らが開発したMHSA[16]がある。これは，文を既に辞書にある語に分割する。他のシステムとの相違点は，文中の文字列が辞書にある単語に複数マッチすれば，すべてのパターンで切り出すことである。この辞書には，漢字，かな，カタカナ，英数字の1文字の語も含まれているので，未知語と呼ばれる辞書にはない文字列が出現した場合は，1文字ずつ切り出される。このシステムは，文から複数のパターンで文字列を切り出すので，切り出した文字列の種類数が多くなるという問題があるが，結果が不適切な切り出し方だけでないという特徴がある。

表1 それぞれの切り出し手法で切り出された文字列

切り出し手法	切り出された文字列
juman	CD/で/学ぶ/楽譜/の/楽しみ方/:/アレンジ/しながら/楽譜/を/マスター
MHSA	CD/で/学/学ぶ/ぶ/楽/楽譜/譜/の/楽/楽し/楽しむ/し/む/方/:アレンジ/す/する/る/な/なが/ながら/が/がら/ら/楽/楽譜/譜/を/マスター
n-gram (n=2)	CD/D で/で学/学ぶ/ぶ楽/楽譜/譜の/の楽/楽し/しみ/み方/方:/:ア/アレ/レン/ンジ/ジし/しな/なが/がら/ら楽/楽譜/譜を/をマ/マス/スタ/ター

n-gramは，文を1文字ずつずらしながらn文字ずつ切り出す手法である。n-gramはn個の連続する文字列であり，一般的に，n=2であるbigramやn=3であるtrigramが用いられている。これは，機械的に文から文字列を切り出す手法なので，形態素解析などのように大規模な辞書を持つ必要がないことが利点となる。しかし，切り出された文字列の数が多くなったり，意味を表わさない場合があるという問題がある。

表1は，書名『CDで学ぶ楽譜の楽しみ方: アレンジしながら楽譜をマスター』(ナツメ社)をそれぞれの手法を用いて切り出した結果を示したものである。jumanにより切り出された文字列のほとんどは意味を持つ文字列となっている。MHSAで切り出された文字列は意味を持つ文字列とひらがな1文字など意味を持たない文字列の両方が切り出されている。n-gram (n=2) で切り出された文字列は，ほとんどが意味を持たない文字列である。

III. 図書を NDC カテゴリに自動分類する実験

A. 分類実験の概要

本研究では，テキストの自動分類であるカテゴライゼーションのベクトルモデルを研究した。一事例として，書名から図書を分類する実験を行った。分類では，前述したように，分類の手がかりとなる書名からの文字列の切り出し方法と，重み付け手法が分類精度に大きく関わってくる。以下の実験では，これらの手法の中で，どの手法が有効であるかを検討する。分類の手がかりとして書名中の文字列を用い，分類先のカテゴリとしては既存のカテゴリとして，日本十進分類法 (NDC) の分類記号を用いた（以下では，これを NDC カテゴリと呼ぶ)。

分類の手がかりとして，書名を用いるのは，図書の書名，目次，序章などの中で，書名が最も簡潔に図書の主題を表わすと言われているためである。

また，分類先の主題カテゴリとして NDC を用いるのは，日本の図書館の多くが用いている分類法であって普遍性があり，数千のカテゴリが存在するためである。自動分類の場合，分類先のカテゴリ数が多いほど分類の精度を一定以上の水準に保つことが困難になる。しかし，実際に分類を情報検索の一手法として実現するためには，多数のカテゴリであっても，一定水準の分類精度が必要とされるため，NDC を分類先のカテゴリとして選択した。

実際の分類実験においては，最初に，分類対象書名が与えられたときに，それがどの NDC カテゴリに分類されるかを決定するための重み付けデータベースを作成する。重み付けデータベースは，既に NDC が付与されている図書を使って求めた，書名中の文字列の NDC カテゴリとの重みの集合である。分類とは，重み付けデータベースを用いて，分類対象書名中の文字列と NDC カテゴリとの重みの総和を算出し，NDC カテゴリをその総和が高い順にランク付けすることである。

本実験では，まず，文字列の切り出し手法の中から，juman, MHSA, n-gram (n＝2) を用いて，比較した。次に，重み付け手法として，カテゴリ間での文字列ごとの相対出現頻度を求め，それをそのまま重みとする 2 種類の重み付け手法（相対出現率型）と tf･idf[5] と tf･idf の一種である tf (max)･idf[5] により重みを求める 2 種類の重み付け手法（tf･idf 型）を用いて分類実験を行った。

B. 分類実験

自動分類の手順を図 1 に示す。これは，重み付けデータベースを作成する学習フェーズと作成した重み付けデータベースをもとに分類対象書名の分類先を決定する分類フェーズとからなる。学習フェーズで作成する重み付けデータベースとは，単語と NDC カテゴリとの重みがセットになったものである。これは，すでに NDC が付与されている学習用集合から求められる。

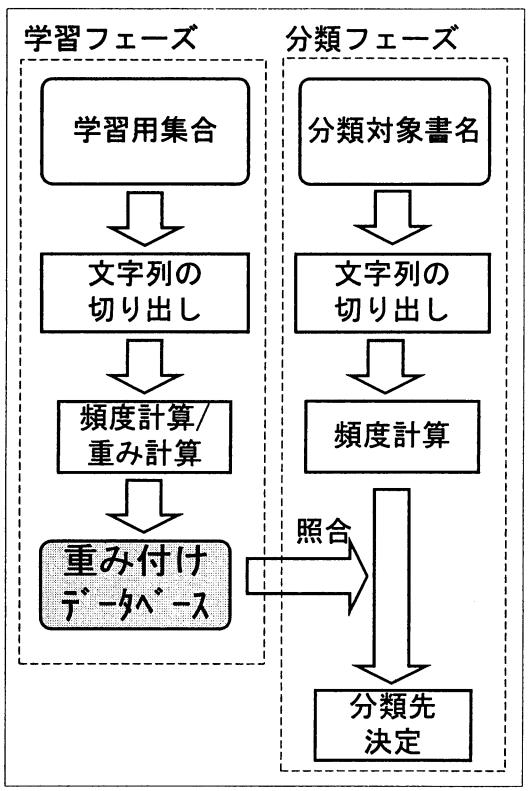

図 1　自動分類の手順

1. 学習フェーズ

学習フェーズにおける重み付けデータベースの作成は以下の手順で行う。

(1) 学習用集合から，書名とNDCのセットを取り出す。
(2) 書名から文字列を切り出す。
(3) 文字列の各NDCカテゴリごとの重みを求める。
(4) 文字列と各NDCカテゴリとの重みのペアを作成する。これを重み付けデータベースと呼ぶ。

重みを求めるには何通りかの手法があるが，本研究では，相対出現率型とtf･idf型のそれぞれについて2種類ずつ，合わせて4種類の計算方法を用いた。それぞれの計算方法は，

(1) カテゴリ間での文字列の相対出現率をそのまま重みとするもの（相対出現率型1）
(2) 相対出現率型1をカテゴリに属する書名数で正規化したもの（相対出現率型2）
(3) tf(max)･idf[5]を用いたもの（tf･idf型1）
(4) tf･idf[5]を用いたもの（tf･idf型2）

である。

(1) 相対出現率型1

カテゴリ C_i $(i=1, 2, 3, \ldots, \mathrm{N})$ における文字列 t_j $(j=1, 2, 3, \ldots, \mathrm{M})$ の出現率による重み w_{ij} は，以下の式で求める。

$$w_{ij}=\frac{F_{ij}}{\sum_{i=1}^{N} F_{ij}} \tag{1}$$

ここで，F_{ij} は文字列 t_j のカテゴリ C_i における出現回数である。この重みは，カテゴリ間での文字列の相対出現率であり，出現回数の大きさに比例して大きくなる。

(2) 相対出現率型2

カテゴリ C_i $(i=1, 2, 3, \ldots, \mathrm{N})$ における文字列 t_j $(j=1, 2, 3, \ldots, \mathrm{M})$ の相対出現率を書名数で正規化する重み w_{ij} は，以下の式で求める。

$$w_{ij}=\frac{F_{ij}}{\sum_{i=1}^{N} F_{ij}} \cdot \frac{\sum_{i=1}^{N} D_i}{D_i} \tag{2}$$

ここで，D_i はカテゴリ C_i に分類されている書名数である。この重みは，文字列の相対出現率が大きいほど，カテゴリに分類されている書名数が少ないほど大きくなる。

(3) tf･idf型1

カテゴリ C_i $(i=1, 2, 3, \ldots, \mathrm{N})$ における文字列 t_j $(j=1, 2, 3, \ldots, \mathrm{M})$ の tf(max)･idf による重み w_{ij} は，以下の式で求める。

$$w_{ij}=tf(\max)\cdot idf = \frac{F_{ij}}{\max_{i=1, 2, \ldots, M} F_{ij}} \cdot \left(\log \frac{X}{\chi_j}+1\right) \tag{3}$$

ここで，X は総カテゴリ数，x_j は文字列 t_j が出現するカテゴリ数である。出現頻度 tf はカテゴリと文字列の関連性であり，逆文献頻度 idf はカテゴリ中での文字列の特定性を表したものである。この重みは，カテゴリ内での文字列の出現回数が多いほど，その文字列が出現するカテゴリが少ないほど大きくなる。

(4) tf･idf型2

カテゴリ C_i $(i=1, 2, 3, \ldots, \mathrm{N})$ における文字列 t_j $(j=1, 2, 3, \ldots, \mathrm{M})$ の tf･idf の重み w_{ij} は，以下の式で求める。

$$w_{ij}=tf\cdot idf=\frac{F_{ij}}{\sum_{j=1}^{M} F_{ij}} \cdot \left(\log \frac{X}{\chi_j}+1\right) \tag{4}$$

tf･idf型1との相違点は，文字列の出現回数を割るときの値である。tf･idf型1ではカテゴリ C_i に出現する文字列の最大の出現回数で割るのに対し，tf･idf型2ではすべての文字列の総出現回数で割っている。そのため，tf･idf型1と比べると，tf･idf型2は，tfの値がより小さくなる。

2. 分類フェーズ

図書の分類先は，以下のようにして決定する。

(1) 分類対象の書名を文字列に切り出す。
(2) 書名中の文字列の出現回数を求める。
(3) 重み付けデータベースと文字列を照合し，文字列とNDCカテゴリごとの重みの総和を求める。
(4) 重みの総和が高い順に，NDCカテゴリを

ランク付けする。

書名とそれぞれのNDCに対する重みの総和は以下のように求める。まず，各カテゴリ C_i ($i=1, 2, 3, \ldots, N$) を $C_i=(w_{i1}, w_{i2}, \ldots, w_{ij})$ ($j=1, 2, 3, \ldots, M$) と表わす。このとき，w_{ij} は上のそれぞれの重み付け手法で求めた文字列 t_j の重みとなる。書名を $q_i=(w_{qi1}, w_{qi2}, \ldots, w_{qij})$ ($j=1, 2, 3, \ldots, M$) と表わす。このとき，w_{qij} は書名中における，文字列 t_j の出現回数である。書名の各カテゴリにおける重みの総和は以下の式で求めることができる。

$$Sim(C_i, q_i)=\sum_{j=1}^{M} w_{ij}w_{qij} \qquad (5)$$

この総和が高いNDCカテゴリほど図書との関連性が高いと言える。本研究では，総和が高い順にNDCカテゴリを1位から10位までランクづけし，Japan MARCに付与されているNDC分類記号と比べ，一致すれば正解とした。

また，分類先のNDCカテゴリとして，NDCの学習用集合に出てくるすべての分類記号（全桁）と上位3桁の2通りを設定対象とした。

C. 分類結果

1. 学習用集合と評価用集合

重み付けデータベースを作成する学習用集合として，Japan MARC中の1997年刊行分の中でNDCが与えられている49,433件から，9類を除いた38,011件を用いた。9類は文学であり，書名中の単語からカテゴリを判断することは困難であると思われるので除いた。手法の評価をするための分類対象は，Japan MARCの1998年刊行の1,000件である。学習用集合，および，評価用集合のNDC付与状況を表2と表3に示す。

表2から，学習用集合では3類が全体の件数の28.8%，8類が2.4%の割合を占めている。カテゴリ数は，上位3桁では674種類，全桁では6,214種類であり，かなり多いことがわかる。

表3の評価用集合では，7類が全体の件数の24.3%，8類が2.6%の割合を占めており，データの分布は均等ではない。また，学習用集合との分布とも異なっている。1998年刊行のデータを用いたのは，学習用集合と同様に，実際のデータを用いるためであり，また，これは実験時に得ることができた最も新しいデータであった。

表2　学習用集合のNDCの区分と付与状況（1997年刊行図書）

NDC	件数	割合	3桁		全桁	
			種類	平均件数	種類	平均件数
0 総記	2,255	5.9%	40	56.4	230	9.8
1 哲学	2,631	6.9%	79	33.3	447	5.9
2 歴史	4,862	12.8%	62	78.4	540	9.0
3 社会	10,931	28.8%	84	130.1	1,713	6.4
4 自然	4,416	11.6%	89	49.6	868	5.1
5 技術	4,896	12.9%	98	50.0	903	5.4
6 産業	2,796	7.4%	75	37.3	731	3.8
7 芸術	4,308	11.3%	86	50.1	578	7.5
8 言語	916	2.4%	61	15.0	204	4.5
計	38,011	100.0%	674	56.4	6,214	6.1

表3　評価用集合のNDCの区分と付与状況（1998年の刊行図書1000件）

NDC	件数	割合	3桁		全桁	
			種類	平均件数	種類	平均件数
0 総記	105	10.5%	15	7.0	34	3.1
1 哲学	49	4.9%	17	2.9	28	1.8
2 歴史	78	7.8%	20	3.9	46	1.7
3 社会	192	19.2%	54	3.6	153	1.3
4 自然	46	4.6%	19	2.4	35	1.3
5 技術	187	18.7%	29	6.4	67	2.8
6 産業	74	7.4%	23	3.2	40	1.9
7 芸術	243	24.3%	42	5.8	90	2.7
8 言語	26	2.6%	15	1.7	18	1.4
計	1,000	100.0%	234	4.3	511	2.0

2. 実験結果

(a) 文字列の切り出し手法別の分類結果

文字列の切り出し手法が分類精度にどの程度影響するかを確かめるために，juman，MHSA，

表4 切り出し手法別にみたNDCカテゴリへの分類結果（NDCの上位3桁）

切り出し手法	juman	MHSA	n-gram (n=2)
第1位での正解率(%)	55.9	45.6	49.4
第10位までの正解率(%)	80.3	73.0	75.9

n-gram (n=2) を用いた。ここでは，切り出し手法の違いをみることを目的としており，重み付け手法はすべて相対出現率型1を用いた。

文字列の切り出し手法を変えて，NDCの上位3桁のカテゴリに評価用データを分類した結果を表4に示す。第1位での正解率は，分類実験の結果，1位にランクづけされたNDCカテゴリのうちJapan MRACレコードで付与されているNDCと一致した割合である。第10位までの正解率は分類実験の結果の第10位までのいずれかにおいて一致した割合である。

表4から，第1位での正解率は，jumanは55.9%，MHSAは45.6%，n-gramは49.4%である。同様に，第10位までの正解率は，80.3%，73.0%，75.9%である。これらの結果から，jumanが最も分類精度が高く，ついで，n-gram，MHSAの順になっていることがわかる。

(b) 重み付け手法別の分類結果

次に，重み付け手法を変えて結果を比較した。書名からの文字列の切り出しには，2.(a)の結果で最も分類精度が高かったjumanを用いた。

NDCの上位3桁のカテゴリに評価用データを分類した結果を表5に示す。相対出現率1における第1位での正解率は55.9%と最も高く，相対出現率型2，tf･idf型1，tf･idf型2の順になっている。第10位までの正解率をみると，相対出現率型1では，80.3%とかなり良い成績となった。

次に，NDCの全桁のカテゴリを対象に分類した結果を表6示す。NDCカテゴリは6,214種類ある（表2参照）にもかかわらず，相対出現率型1の第1位での正解率は45.9%であり，第10位までの正解率は69.8%であった。一方，tf･idf型1の第1位での正解率は6.6%，tf･idf型2では，6.5%となっており，相対出現率型に比べて非常に低い。

表5 重み付け手法別にみたNDCカテゴリへの分類結果（NDCの上位3桁）

重み付け手法	相対出現率型1	相対出現率型2	tf･idf型1	tf･idf型2
第1位での正解率(%)	55.9	23.5	16.8	7.6
第10位までの正解率(%)	80.3	67.1	49.1	35.5

表6 重み付け手法別にみたNDCカテゴリへの分類結果（NDCの全桁）

重み付け手法	相対出現率型1	相対出現率型2	tf･idf型1	tf･idf型2
第1位での正解率(%)	45.9	16.3	6.6	6.5
第10位までの正解率(%)	69.8	46.1	40.6	24.4

表5と表6から相対出現率型1は，上位3桁で6割，全桁でも半分近く，Japan MARCで付与されているNDCと同じカテゴリに図書を分類しており，相対出現率型1は相対出現率型2に比べると分類精度が高くなっている。それに比べて，tf･idf型は両方とも，かなり低い。全桁では，6%しか分類できていない。

D. 結果の考察

1. 異なって分類された図書の書名の特徴

重み付け手法に相対出現率型1を用いた実験で，Japan MARCの分類記号と異なって分類された図書302件の書名の特徴を調べた。特徴をまとめた結果を表7に示す。カタカナ文字列，英数字，固有名詞を含む書名，人手でも分類が困難の割合は，学習用集合に含まれない分類記号を持つ図書数を除いた場合の割合である。全評価用データでは，カタカナ文字列を含む割合が1,000件中496件 (49.6%)，英数字を含む割合が189件(18.9%)，固有名詞を含む割合が172件(17.2%)であり，異なって分類された書名の割合とほとんど変わらないことがいえる。ただし，固有名詞を

表7　Japan MARCの分類記号と異なるカテゴリに分類された図書の書名の特徴*

特徴	件数	割合(%)	例	分類記号
カタカナ文字列を含む	134	51.2	**『ヤングのライスクック』**	**596.5**
英数字を含む	56	21.6	**『Kanga Saying: カンガの教え』**	**753.3**
固有名詞を含む	44	17.0	**『棟方志功の世界』**	**733.087**
人手でも分類が困難	7	2.7	**『あったかさん』**	**369.28**
学習用集合に含まれない分類記号を持つ	43	14.2	**『デジタル商品・用語辞典』**	**549.033**

* 特徴は重なりがある。

含む写真集などの書名は,「写真」という単語が含まれているため，正しく分類されていた。一方，書名の大部分が固有名詞の場合は，異なって分類されていた。

表7に例としてあげられた書名が，実際にはどのように分類されたのかを示す。

『ヤングのライスクック』は,「ヤング/の/ライスクック」に切り出される。それぞれの文字列と重み付けデータベースとを照合すると,「ヤング」はカテゴリ368.61と498.3に対して，それぞれ0.33と0.67の重みがあり,「の」はほとんどのカテゴリに出現しているため，そのカテゴリごとに非常に小さい重みがある。「ライスクック」は，重み付けデータベース中に存在しない単語であった。この結果，それぞれのカテゴリごとの重みを総和すると,「ヤング」の重みの影響を最も大きく受け，関連性が一番高いカテゴリは，498.3になる。

『棟方志功の世界』は,「棟方/志/功/の/世界」に切り出される。「棟方」はカテゴリ733にのみ出現しており，重みは1である。「志」は23のカテゴリに出現しており，最大の重みを持つのはカテゴリ222.043であった。「功」は13のカテゴリに出現しており，最大の重みを持つのはカテゴリ498.3であった。「の」は，ほとんどのカテゴリに対して非常に小さい重みがあった。「世界」は，多くのカテゴリに出現しており，それぞれが小さい重みであった。この結果，それぞれのカテゴリごとの重みを総和すると,「棟方」の重みの影響を最も大きく受け，関連性が一番高いカテゴリは733になる。

この302件の分野別の割合をみた。総記においては，異なって分類された図書の割合が4.9%と評価用データに占める割合(10.5%)に比べて低かったが，その他の分野に関してはほとんど差が見られなかった。特定の主題が特に分類されにくいということではなかった。

2.　文字列の切り出し手法

表4の結果から，文字列の切り出し手法の違いによる分類は，jumanが最も高い分類精度であり，次いで，n-gram, MHSAの順になっている。ここでは，この理由について考察する。

まず，それぞれの切り出し手法の特徴を述べる。切り出された文字列の種類数は，jumanは32,163種類，MHSAは35,661種類，n-gramは108,991種類であった。このように，jumanとn-gramでは文字列の種類数が3倍以上も差がある。また，jumanによって切り出された文字列はほとんどが意味を持つ集合であるのに対し，MHSAは意味を持つものと持たないものとの混合，n-gramはほとんどが意味を持たない集合である。

このような違いから，jumanの分類精度が高かった理由としては，意味を持った文字列を切り出すためと考えられる。書名は，抄録や新聞記事などに比べて文字列として情報量は少ない一方，通常は特徴的な主題を表わす単語が含まれている場合が多い。そのため，書名から意味を持った主題を表す特徴的な単語を切り出すことにより，NDCカテゴリの特徴，分類対象書名の特徴も表わしやすくなる。

次に，n-gram の分類精度が高かった理由としては，n-gram は意味を持たない文字列の集合ではあるが，手がかりとなる単語の数が多いことが考えられる。また，漢字は表意文字であるので，1 文字だけでも主題を表わすことができる場合があるため，n-gram のように文字列を細かく切り出しても，その中に漢字が含まれているので，ある程度，漢字だけで図書の主題を表わすことができると考えられる。

MHSA は，juman で切り出された文字列の種類数とほとんど同じであるが，意味を持たない文字列がノイズになってしまった可能性がある。

3. 重み付け手法

表 5, 表 6 から，重み付け手法として，相対出現率型 1, 2 を用いる方が，tf･idf 型 1, 2 を用いるよりも分類精度が高いことを示した。この理由を，それぞれの重みの求め方の違いから述べる。ここで，相対出現率型 1, 2 は，重みを求める考え方は同じなので，相対出現率型 1 を用いて説明をする。同様に，tf･idf 型 1, 2 では，tf･idf 型 1 を用いて説明を行う。

本研究で用いた相対出現率型の重みは，カテゴリ間での任意の文字列の相対出現率であるので，カテゴリ間での出現回数の差が重みの大きさに比例する。たとえば，図 2 に示すように，文献集合全体でカテゴリ A, B, C があり，文字列 1 の出現頻度がそれぞれ 15 回，0 回，5 回だったとすれば，カテゴリ A における文字列 1 の重みは w_{A1}＝15/(15＋0＋5)＝0.75 となる。同様に，カテゴリ B における文字列 1 の重みは 0.0，カテゴリ C における文字列 1 の重みは 0.25 となる。このように，本研究で用いた相対出現率型の重みの考え方は，カテゴリ間での比較したときに，文字列がどのカテゴリとの関連性が強いかということである。

一方，tf･idf 型における tf･idf の重みの求め方は，カテゴリ内でのその文字列の出現率が高く，かつ，カテゴリ全体でその文字列が出現するカテゴリ数が少ない場合に重みが高くなる。たとえば，文字列 1 のカテゴリ A での重みを求める場合は，図 3 に示すように，tf はカテゴリ A 中での文字列 1 が他の文字列に比べてどの程度多く出現しているかをもとに求められ，idf は文字列 1 が出現するカテゴリ数をもとに求められる。この場合，カテゴリ A における文字列 1 の重みは w_{A1}＝tf(max)･idf＝0.20 となる。同様にカテゴリ B における文字列 1 の重みは 0.0，カテゴリ C にお

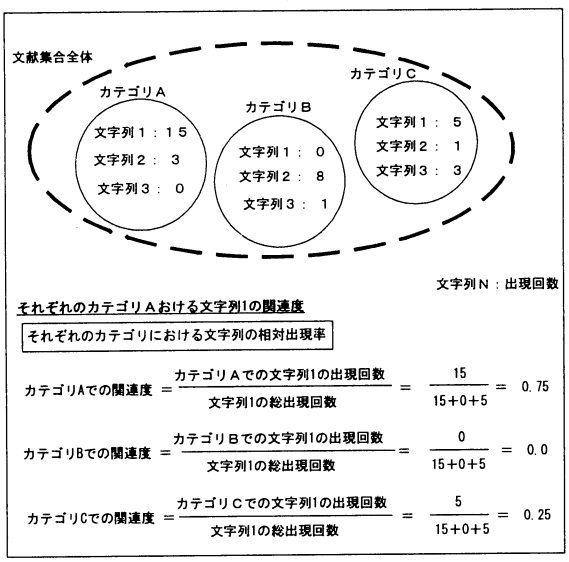

図 2　相対出現率型の重みの求め方

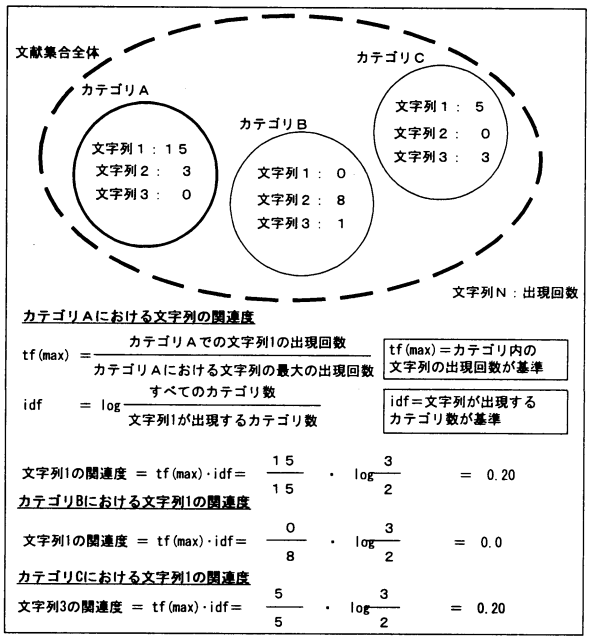

図3　tf･idf型の重みの求め方

ける文字列1の重みは0.20となる。つまり，tf･idfの重みの考え方は，カテゴリ内で他の文字列と比べて，どの程度，関連性が高いかということである。

図書の分類の場合は，カテゴリ間で図書との関連性が高いカテゴリを比べることによって，分類先を決定している。その手がかりとして，書名中の文字列の各カテゴリとの重みを用いている。そのため，重みを求める際には，カテゴリ内で他の文字列と比較するよりも，カテゴリ間で文字列の出現回数を比較した方が，分類には適していると考えられる。このため，tf･idf型で重みを求めるよりも，相対出現率型を用いた方が，結果が良かったと考えられる。

なお，相対出現率型2よりも相対出現率型1の方が結果が良かった。表2に示したように，評価用集合には大きな偏りがある。NDC上位3桁を分類先のカテゴリとしたとき，評価用集合において，千以上の書名数があるカテゴリもあれば，ひとつの書名しかないものもあった。この場合，正規化を行うと，同じ出現回数でも，重みに数千倍以上の差がついてしまうことになる。以上のような理由から，相対出現率型2よりも，相対出現率型1のほうが結果が良かったと考えられる。

4. 分類精度の向上

III.D.1で示したように，書名からの文字列を切り出す際に固有名詞を切り出せないこと，分類対象書名から切り出された文字列が重み付けデータベースに存在しないことなどが，図書を正しいカテゴリに分類できない原因として考えられる。分類精度を高めるためには，文字列の切り出しシステムや重み付けデータベースについて，以下のようなことが考えられる。

文字列の切り出しシステムでは，切り出しの際に，参照している辞書に固有名詞やカテゴリに特徴的に用いられる単語などを追加することが考えられる。

重み付けデータベースを作成する際には，学習用データ量を変化させることなどが考えられる。学習用データを増やせば，分類対象書名から切り出された文字列が重み付けデータベースに存在しないことを防げる。しかし，データ量を増やしすぎることによって，ノイズが発生してしまう可能

性も考えられる。また，本研究では，書名の文字数が少ないために，学習用データ，評価用データともに，書名から切り出された全ての文字列を用いた。その結果，ある程度の分類ができることがわかったので，名詞だけを用いるなど品詞による文字列の選択を行うことが考えられる。これは，品詞によって主題の特徴をより表わす場合があると考えられるからである。精度を高めるための重み付けデータベースの作成には，これらの方法が考えられるが，更なる実験と考察が必要である。

5. Larson の研究との比較

II.A.3 で述べた Larson[13] の研究において最も精度が高かった分類手法の組み合わせと比較した結果を表 8 に示す。この表から，本実験で用いた手法と同じ重み付け手法での精度が高かったことがわかる。また，本実験は，Larson のものよりカテゴリ数が千以上多いこと，Larson が対象としたZ（書誌及び図書館学）よりも分類カテゴリの範囲が広いこと，Larson は分類の手がかりとして件名標目を用いているのに対し，本実験では書名だけを用いていることなどの相違点がある。

一般的に，カテゴリの主題の幅が広いこと，カテゴリ数が多いことなどは分類を困難にするものであり，件名標目は書名よりも図書の主題を簡潔かつ的確に表しているものなので，分類の手がかりとしては有効ではないかと考えられる。よって，分類結果を一概に比較することはできないが，百分率で 1 ポイント以内の差しかないのであるから，Larson の実験よりも，本実験の分類精度の方が上回っていると推定できる。

IV. おわりに

A. 応用可能性

本研究では，日本語テキストの自動分類の一事例として，図書を NDC カテゴリに分類する実験を行った。III.D.4 で述べたような方法で分類の精度が高まれば，本研究の応用可能性として，以下のことが考えられる。

1. 分類作業の支援

本実験では，実験的なデータではなく，実際の Japan MARC のデータを使った結果であり，すぐに分類作業に利用できる。Japan MARC に収録されている 1997 年刊行図書 64,583 件のうち NDC が付与されているのは 49,433 件 (76.5%) である。国立国会図書館が全ての図書に NDC を付与できないのは，分類作業に労力を要することも一因となっていると考えられる。

今回の実験では，書名を入力することにより，図書を NDC カテゴリに分類できる可能性を示した。よって，入力した書名に対し，候補となる NDC を表示する分類支援システムを考えることができる。

また，前述のように，Japan MARC 中で与えられている NDC には，同一主題であっても異なる NDC が与えられている場合があり，人手による分類では常に一貫性が問題になる。自動分類を用いることにより，一貫性のある分類作業が期待できる。

2. 情報検索への応用

本実験で行った分類は，書名のような短い単位

表 8 Larson の研究と本研究の比較

研究名	分類体系	対象範囲	カテゴリ数	用いた手法
Larson	LCC	書誌及び図書館学 (Z)	5,765	カテゴリ間での相対出現率＋定数
本研究	NDC	1 類から 8 類	6,214	カテゴリ間での相対出現率
研究名	評価用データ数	分類の手がかり	分類結果（1 位）	分類結果（10 位）
Larson	383	件名標目	46.6%	74.4%
本研究	1,000	書名	45.9%	69.8%

のテキストで分類記号を与えることが可能であることを示した。このことは，データベースに収録されている文献ばかりでなく，検索質問にも分類記号を与えることができるので，検索に利用できる。

全文検索などの完全照合手法では，同義語など意味が類似しているが表記や表現が異なる語があることは最大の問題である。この方法では，単語そのものを扱わないので，この問題を回避できるという大きな利点がある。

B. 今後の課題

自動分類システムとしての実用化のためには，(1) 出版社名など書名以外のデータの利用，(2) 文学作品の判定などの課題がある。出版社名の利用に関しては，出版社により出版している図書の分野が限定される傾向があるので，それを分類の手がかりとして盛り込むことが考えられる。文学作品の判定には，出版社名を利用すること，同様に，著者名を利用することなどが考えられる。また，文学作品の書名を分析し，その特徴などを利用することも考えられる。

情報検索システムとして用いるには，さらに分類精度を高め，分類手法を改良するなどの工夫が必要である。そのために，相関索引の利用などが考えられる。その前段階として，論文や新聞記事など，日本語の他のテキストでも，この分類手法が適切であるか，分類実験を積み重ねる必要がある。

分類先のカテゴリとしてNDCでは不十分であるが，全分野を含み，大量の学習用集合を得られるのは現在のところJapan MARCだけであるというのが現状である。しかし，NDCの一部を精しくするなどの改善が必要である。

謝　辞

様々なご示唆を頂いた慶應義塾大学文学部の上田修一先生，倉田敬子先生，亜細亜大学の安形輝先生，鉄道総合研究所の野末道子先生，作新学院大学女子短大部の久野高志先生に感謝致します。

引用文献

1) Maron, M. E., Kuhns, J. L. "On Relevance Probabilistic Indexing and Information Retrieval". Journal of the American Society for Information Science. Vol. 7, p. 216–244 (1960)
2) Hamill, K., Zamora, A. "The Use of Titles for Automatic Document Classification". Journal of The American Society for Information Science. Vol. 31, No. 6, p. 252–277 (1994)
3) 岩山　真，徳永健信．"自動文書分類のための新しい確率モデル"．情報処理学会情報学基礎 33–9. p. 47–52 (1994)
4) 野本　忠司．"確率モデルによる主題の自動抽出"．情報処理学会自然言語処理 108–1. p. 1–6 (1995)
5) 岸田和明．情報検索の理論と技術．勁草書房．1998. 314 p.
6) 山崎健文．イドダガン．"誤り駆動型学習とシソーラスを用いた文書自動分類"．情報処理学会自然言語処理 120–14, p. 89–96 (1997)
7) 福本文代，鈴木良弥．"語の重み付け学習を用いた文書の自動分類"．情報処理学会論文誌．Vol. 40, No. 4, p. 1782–1791 (1999)
8) 塩見隆一ほか．"シソーラスを用いた文書データの自動分類法"．情報処理学会自然言語処理 117–14. p. 99–104 (1997)
9) 徳永健伸，岩間　真．"重み付きIDFを用いた文書の自動分類について"．情報処理学会自然言語処理 100–5. p. 33–40 (1994)
10) 渡辺靖彦ほか．"χ^2 法を用いた重要漢字の自動抽出と文書の自動分類"．情報処理学会情報学基礎 39–4. p. 25–32 (1995)
11) 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, p. 1114–1122 (1992)
12) 藤井洋一ほか．"共起情報を利用した文書の自動分類"．情報処理学会自然言語処理 118–16. p. 97–10 (1997)
13) Larson, R. Ray. "Experiments in Automatic Library of Congress Classification". Journal of The American Society for Information Science. Vol. 43, No. 2, p. 130–148 (1992)
14) 京都大学情報学研究科長尾研究室．"日本語形態素解析システムJUMAN"．[1999.05.06]〈http://www-nagao.kuee.kyoto-u.ac.jp/nlresource/juman.html〉
15) 奈良先端科学技術大学院大学自然言語処理学講座松本研究室．"茶筅ホームページ"．[1999. 05. 06]〈http://cl.aist-nara.ac.jp/lab/nlt/chasen.html〉
16) Hasebe, K., Nakamoto, K., and Yamamoto, T. "An Information Retrieval System on Internet for Languages without Obvious Word Delim-

iters", Proceedings of International Symposium on Digital Libraries 1995. p. 181–185 (1995)