情 報 学 基 礎 39－4
（1995．9．14）

χ^2 法を用いた重要漢字の自動抽出と文書の自動分類

渡辺 靖彦　村田真樹　竹内雅人　長尾 真
京都大学工学部 電気工学第二教室
〒606-01 京都市左京区吉田本町
e-mail: watanabe@kuee.kyoto-u.ac.jp

あらまし
テキストが属する専門分野を決めるには、それぞれの専門分野を特徴づける用語（キーワード）を用いることが考えられる。しかし、日本語テキストの場合、単語を正しく切り出して用語（キーワード）を取り出すことはそれほど容易ではない。そこでわれわれは、特定の分野にかたよってあらわれる漢字に注目し、その組合せによってテキストを分類する方法を提案する。本稿では最初に、χ^2 法の考え方にもとづいて分野の識別に重要な漢字を自動的に抽出する方法について説明する。次に、その漢字の統計的情報を用いた単純なパターン分類の手法で日本語テキストを分類する方法について述べる。提案したテキストの自動分類法を評価するために行なった実験では、天声人語の47%、社説の74%、サイエンスの記事の85%の分類に成功した。
和文キーワード テキスト分類, 重要漢字, χ^2 法, 日本10進分類法, 百科事典

Document Classification Using Important Kanji Characters Extracted by χ^2 Method

Yasuhiko Watanabe　Masaki Murata　Masahito Takeuchi　Makoto Nagao
Department of Electrical Engineering II, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto 606-01, Japan
e-mail: watanabe@kuee.kyoto-u.ac.jp

Abstract
In this paper we describe a method of classifying Japanese text documents using important kanji characters. Text documents are generally classified on significant words (keywords) of the documents. However it is difficult to extract these significant words from Japanese text, because Japanese text is written without using blank space as delimiters and must be segmented into words. Therefore, instead of words, we used important kanji characters which appear more frequently in one category than the other. We extracted these important kanji characters by χ^2 method. Then, we tested our method. The correct recognition scores for editorial columns "TENSEI JINGO", editorial articles, and articles in "Scientific American (in Japanese)" were 47%, 74%, and 85%, respectively.
英文 key words document classification, important kanji character, χ^2 method, Nippon Decimal Classification, encyclopedia

1 はじめに

図書館や博物館で古い資料を整理していて、今まで知られていなかった重要な資料が発見されることがよくある。これは貴重な資料も分類・体系化されていなければ有効に利用することができないことを示している。同じことが計算機とネットワークの発達によって大量に生み出されている情報や文書についてもいえる。これらの情報や文書に分類コードなどの2次情報を付加して分類・体系化しなければ、効率的な検索はできない。

このため、科学技術論文等を対象にしてテキストを自動的に分類する研究が盛んに行なわれている。従来の自動分類の方法は

a. 表記の統計情報を用いた方法 [田村 88][鈴木 87]

b. 分類体系に依存した知識を用いる方法 [亀田 87]

の2つにわけることができる。統計情報を用いる方法は、分類済みの標本データからそれぞれの分野にかたよってあらわれる単語を調べ、その情報を手がかりにしてテキストを分類する。この方法はテキストの意味構造を扱わないので処理が比較的簡単になるが、精度が低くなるという問題がある。一方、分類体系に依存した知識を用いる方法では精度は高くなるが、知識が分類体系に依存しているので分野の拡張や変更が難しい。どちらの方法にも共通するのは、単語を意味の基本単位として分類を行なう場合が多いことである。このため、テキストを自動分類するには以下の問題がある。

1. 対象とすべき単語の種類が多い

   テキストが属する分野を識別するのに重要な手がかりになる単語はそれぞれの分野にかたよってあらわれる単語である。このような単語は主に専門用語である。専門用語は非常に数が多いため、テキストを分類する処理の対象にすべき単語の種類も非常に多くなる。このため、単語を意味の基本単位とする分類処理には大量の記憶容量と計算量が必要になる。さらに特定の分野にかたよってあらわれる単語の情報を獲得する方法も問題になる。

2. 単語の正しい切り出しが難しい

   形態素解析の精度の問題の他に、複合名詞の問題がある。分類にはそれぞれの分野の専門用語が重要であるが、専門用語は複合名詞であることが多く、それらの用語を正しく切り出すことはむずかしい。なぜなら、形態素解析を行なうと、複合名詞はより基本的な単語に分割されてしまうおそれがあるからである。例えば、最近情報科学の分野で普及しはじめている「電子図書館」という用語も、形態素解析を行なうと「電子」と「図書館」の2つに分割されてしまう可能性がある。そして、このように分割されてしまった単語からもとの複合語の意味を正しく復元することはむずかしい。このため、分野の推定が困難になる。

以上のことから、単語を手がかりにテキストを分類するには高精度の形態素解析システムと膨大な語彙的知識が必要であり、このことが実用的な文書の自動分類システムの実現を困難にしていることがわかる。そこでわれわれは漢字が表意文字であることに注目し、単語のかわりに漢字を意味の基本単位とすることを考えた。すなわち、漢字の統計情報を用いて単純な処理によってテキストを分類する方法を提案する。漢字を意味の基本単位として科学技術文献を分類する方法は細野らによって報告されている [細野 84][細野 85]。しかし、細野らは漢字の頻度情報のみを用いて分野の識別に重要な漢字について考慮していなかったため、その分類精度は低かった。そこでわれわれは、χ^2 法の考え方にもとづいて分野の識別に重要な漢字を抽出することを考えた。

以下、2章では、漢字によるテキストの特徴表現および分類の妥当性について検討する。3章では、分野の識別に重要な漢字を抽出する方法と、抽出した漢字の統計情報によって各分野の特徴を表現する方法を説明する。4章ではさまざまなテキストを対象に実験を行ない、提案した自動分類法の有効性を確かめる。最後に5章では、テキストの分類情報の利用について考察する。

2 重要漢字を用いたテキストの分類

2.1 漢字によるテキストの特徴表現

テキストの内容を表現するためにキーワード、すなわち単語がよく用いられる。これは単語を意味の基本単位と見なし、それらを合成したものがテキスト全体の意味を表現するという考えにもとづいている。しかし単語は必ずしも意味の最小単位ではない。日本語の単語の多くは漢字を含み、それらの漢字の組み合わせによって単語の意味は表現される。英語では単語を意味の最小単位に分解したものをmorpheme(形態素) とよぶが、日本語では表意文字である漢字がこの morpheme に相当する意味の最小単位である。したがって、単語ではなく、漢字を意味の基本単位にしてもテキストの内容は表現できるのではないかとわれわれは考えた。

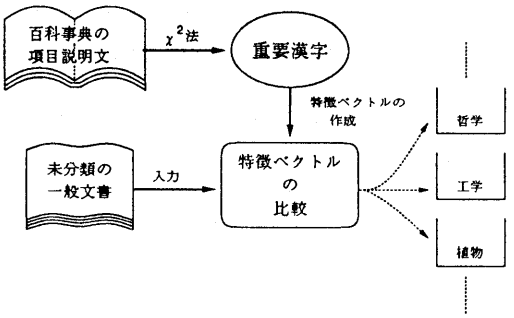

図 1: 重要漢字を用いたテキストの自動分類

漢字を意味の基本単位にすると、単語を手がかりに分類する方法のもつ2つの問題を回避することができる。すなわち、単語を切り出すのに比べ、漢字を抽出することは非常に容易である。また日本語で用いられる漢字はおよそ2,000種類程度なので、テキストを分類する処理の対象となる漢字の種類は2,000種類以下である。一方、単語を手がかりにして分類する方法に比べ、単語を構成する漢字を手がかりにする方法では分類の精度が低下すると予想される。以上のことから、漢字の統計情報を用いて、そして非常に簡単な処理によってどこまで有効にテキストを分類できるのかを明らかにすることに本研究の意味がある。

漢字によってテキストの内容を表現するためには、単語によってテキストの内容を表現する際のキーワードに相当する漢字を明らかにすることが重要である。キーワードに相当する漢字とは、それぞれの分野を特徴づける漢字である。例えば、『植物』という分野では「花」や「葉」などがキーワードに相当する漢字である。こうした漢字はテキストが属する分野の識別にも重要で、以後、重要漢字とよぶ。ある分野のテキストではよく用いられるが、他の分野のテキストではそれほど用いられない漢字は分野の識別に重要である。そこで、特定の分野にかたよってあらわれる漢字は重要漢字であるとみなし、χ^2法という統計的手法の考え方にもとづいてそれらの漢字を自動的に取り出す。

2.2 重要漢字を用いた分類方法

テキストを漢字の統計情報を用いて分類する方法は次の通りである(図1)。

1. 分類済みのテキストを学習サンプルとして用意する。本研究では日本10進分類法(NDC)の分類体系にしたがって分類した百科辞典の項目説明文を学習サンプルとして用いた。
2. 分野によってかたよってあらわれる漢字を重要漢字としてχ^2法で抽出する。
3. 抽出した重要漢字の種類と頻度情報から、各分野の特徴を表す特徴ベクトルを獲得する。
4. 未分類のテキストに含まれる漢字の種類と頻度を調べ、各分野の特徴ベクトルと比較してテキストの分類先を決定する。

雪割草　　ゆきわりそう
〈雪割草〉とは寒いときに雪を割って咲く花の意のため、早春に咲く植物、たとえばイチリンソウ、ショウジョウバカマなどをさすこともあるが、サクラソウ科の多年草1種の和名ともなっている。またキンポウゲ科の山草のミスミソウの別称でもある。⇒サクラソウ
堀田　満

図 2: 百科事典の項目説明文の例

3 重要漢字の自動抽出

本章では最初に、学習サンプルとして用いる大量の分類済みのテキストを作成する方法を説明する。次に、その学習サンプルから重要漢字を抽出する方法と、各分野の特徴ベクトルを獲得する方法を述べる。

3.1 著者の専門分野による百科事典の項目説明文の分類

重要漢字と各分野の特徴ベクトルを獲得する学習サンプルとして百科辞典の項目説明文を用いた。実験に利用した百科事典は、平凡社の『世界大百科事典CD-ROM版』である。『世界大百科事典CD-ROM版』は6,722人の著者によって執筆されたおよそ8万個の項目から構成されている。この百科事典全体の文字数は約6,520万文字、そのうち漢字は約2,520万文字である。

図2に項目の説明文の例を示す。項目の説明文にはその項目がどの分野に属するかは記述されていないが、その著者の氏名が説明文の末尾に示されている。著者の専門分野は著者一覧にまとめて示されているので、項目がどの専門分野に属しているかがわかる。例えば、図2の著者の専門分野は植物であるので、【雪割草】は植物の分野に属する項目であることがわかる。

『世界大百科事典CD-ROM版』の著者の専門分野は418個あり、項目説明文はそのいずれかに分類できた。し

5(00)	工学・技術	類	
54(0)	電気工学	綱	
548	情報工学	要	
548.2	計算機工学	分目	細目
548.23	記憶装置	厘目	

NDC は日本で広く用いられている図書館分類法で、階層的な分類が行なわれている。その分類は最上位から類(10区分)、綱(100区分)、要(1000区分)および細目から構成されていて、約8,500ある分類項目には区分けに応じた10進数の番号が与えられている。

図 3: 日本10進分類法(NDC)の階層構造の例

かし、その418個の専門分野からは直接重要漢字を取り出さなかった。なぜなら、分野の細かさがさまざまだったからと、それぞれの分野に属している項目の数に大きな差があったからである。例えば、英米文学という専門分野がある一方で、SF(Science Fiction)という専門分野がある。また、数百の項目が属する専門分野がある一方で、わずか数個の項目しか属していない専門分野もあった。そこでわれわれは著者の専門分野の名称が日本10進分類法(NDC)の分野の名称に似ていることに注目し、著者の専門分野をNDCの分野に対応づけ、NDCの階層構造を利用して著者の専門分野をまとめることを考えた。図3にNDCの階層構造の例を示す。このNDCの階層構造を利用して、418個の専門分野を以下の手順で42個にまとめた。

1. 418個ある著者の専門分野のうち206分野はNDCの分野の名称と完全に一致するので、そのまま対応づける。NDCの分野の名称と一致しない残りの212個は人手によってNDCの分野と対応づける。
2. NDCの綱の区分を基準に分野の統合を行ない、418個の専門分野を59個の分野にまとめた。
3. 細かすぎると判断した分野を人手によってまとめて42個の分野を作成した。

手順2でえた59分野を手順3で人手によって42分野にまとめたのは、59分野の分類の細かさに違いがあったからである。例えば、59個の専門分野の中には「物理学」「電気工学」「ドイツ文学」があったが、「物理学」に比べて「電気工学」と「ドイツ文学」は分類がやや細かいと感じられる。そこで、人手によって専門分野をまとめ、分類の細かさができるだけ一定になるようにした。

3.2 χ^2 法による重要漢字の選択

出現頻度の大きな漢字を重要漢字として取り出そうとすると、どの分野にも高頻度であらわれる漢字、すなわち分野の識別力が低い一般的な漢字を抽出してしまうおそれがある。そこで、出現頻度の偏りが大きな漢字を重要漢字として抽出する。

χ^2 検定の χ^2 値は漢字の出現頻度の分野による偏りを示す指標として用いることができる。すなわち、漢字一つ一つについて各分野ごとの頻度を標本値とし、その漢字の出現確率は全分野を通じて等しいと仮定して χ^2 値を求める。もしその χ^2 値が十分大きな値になれば特定の分野に集中してあらわれる漢字ということになり、分野の識別に有効な漢字とみなせる。具体的には漢字 i の χ^2 値は、以下の式で求める。

$$\chi_i^2 = \sum_{j=1}^{l} \chi_{ij}^2 \tag{1}$$

$$\chi_{ij}^2 = \frac{(x_{ij} - m_{ij})^2}{m_{ij}} \tag{2}$$

$$m_{ij} = \frac{\sum_{j=1}^{n} x_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}} \times \sum_{i=1}^{m} x_{ij} \tag{3}$$

ただし

- m: 異なり漢字数
- n: 分野数
- x_{ij}: 漢字iの分野jにおける頻度
- m_{ij}: 漢字iの分野jにおける理論度数

理論度数とは、全分野に等確率でその漢字が出現した場合の出現頻度である。漢字ではないが、χ^2 法を用いて重要なキーワードを抽出する研究が行なわれており[長尾 76]、χ^2 法がキーワードの抽出に有効であることが確かめられている。

しかし、(1)式の χ_i^2 値からは分野全体に対して出現頻度に偏りのある漢字はわかっても、どの分野を特徴づける重要漢字であるのかはわからない。したがって、(1)式の χ_i^2 値を用いて重要漢字を取り出すと、各分野の重要漢字を平均して取り出せたか確かめられず、重要漢字を抽出できない分野が発生するおそれがある。そこで、それぞれの分野における出現頻度の理論度数からのずれ、すなわち、(2)式の χ_{ij}^2 を用いて重要漢字を抽出した。χ_{ij}^2 の値が大きい漢字はその分野にかたよって出現している漢字である。そこ

表 1: 百科事典から抽出した重要漢字

分野名	重要漢字 (左の漢字ほど χ^2 値が大きい)																			
図書館学	書	版	館	冊	庫	本	紙	丁	図	糊	刊	刷	印	巻	帖	誌	文	蔵	折	獄
哲学	哲	論	学	思	而	想	朱	理	儒	熹	教	的	神	倫	義	派	念	識	孟	彼
心理学	心	我	理	精	習	療	学	眠	能	的	識	児	欲	知	析	神	己	意	験	象
宗教	寺	教	宗	仏	神	僧	禅	聖	仰	祭	蓮	派	願	信	会	皇	祈	檀	浄	陀
社会	族	社	会	婚	人	民	孫	番	儀	住	耕	牧	狩	礼	的	婦	団	呪	姻	男
政治	政	党	治	国	議	権	民	閣	会	主	制	挙	員	選	義	票	軍	争	戦	委
経済	資	労	税	業	済	険	働	金	産	貨	企	価	債	銀	券	経	株	額	財	需
法	法	訴	権	訟	裁	判	条	刑	債	犯	罪	為	憲	務	審	事	罰	項	責	請
軍事	艦	戦	軍	隊	砲	撃	兵	弾	攻	爆	銃	核	航	略	敵	艇	空	闘	搭	潜
教育	育	校	教	学	科	児	習	童	制	等	課	師	盲	年	授	稚	塾	設	員	私
商業・流通	売	商	品	販	卸	店	業	族	舗	買	購	費	泊	消	宿	取	者	顧	産	企
風俗・民俗	餅	漬	煮	楽	茶	踊	香	飯	粥	神	煎	醤	俗	菓	豆	料	忌	食	祭	菜
科学史	学	医	究	研	科	授	術	剖	論	技	博	病	痘	賞	年	蘭	理	物	験	然
数学	数	式	幾	関	線	値	算	点	角	何	積	分	微	定	図	次	列	直	限	程
情報科学	械	憶	処	理	索	計	題	問	算	言	御	識	機	採	報	適	記	解	最	論
天文	星	陽	惑	太	天	測	暦	恒	球	月	銀	軌	光	河	鏡	遠	矮	波	宙	午
物理	電	磁	子	波	粒	光	振	力	量	核	度	体	速	熱	荷	射	質	動	気	温
化学	酸	溶	塩	素	化	水	液	硫	晶	応	反	合	炭	子	沸	硝	電	錯	結	媒
地学	岩	気	震	鉱	海	雲	堆	氷	層	噴	候	地	火	石	褶	晶	風	測	雨	温
考古	墓	墳	器	石	跡	斧	土	棺	葬	遺	銅	鏃	掘	塚	槨	甕	址	陶	陵	竪
生物	酵	胞	遺	酸	細	猿	糖	質	物	生	類	淘	鎖	謝	伝	素	脂	食	雄	体
植物	花	葉	茎	植	咲	枝	萼	草	種	栽	苞	緑	裂	色	培	藻	状	胞	被	芽
動物	虫	卵	翅	類	雌	巣	腹	昆	色	褐	魚	肢	吻	幼	殖	雄	鳥	骨	尾	哺
医学	症	血	病	腫	患	臓	傷	療	炎	癌	疾	筋	脳	肺	痛	髄	腸	膜	診	骨
工学	電	船	坑	車	圧	水	機	翼	炉	送	用	速	力	削	路	航	燃	回	舵	波
農林水産	壌	林	培	栽	農	飼	肥	繭	漁	種	畜	苗	森	穫	樹	産	菌	刈	圃	耕
管理技術	郵	送	聞	企	営	刊	告	業	鉄	便	誌	新	社	益	資	駅	車	輪	報	簿
化学工業	鋼	溶	料	耐	繊	炉	酸	油	紡	剤	熱	紙	材	維	鋳	脂	金	染	燃	用
精密機械他	服	衣	袖	着	襟	帽	巾	綿	衿	濯	裾	縫	用	装	袋	丈	毛	襲	繍	飾
建築	建	築	堂	棟	殿	屋	壁	塔	葺	柱	造	廊	室	住	瓦	居	窓	宅	梁	寺
美術	絵	画	彫	美	像	術	描	漆	蒔	刻	仏	陶	彩	釉	屏	窯	飾	曼	茶	様
環境・都市	宅	市	都	住	災	街	火	公	計	防	地	画	消	整	設	区	園	施	策	土
写真・印刷	写	刷	真	印	版	像	撮	字	紙	光	凹	謄	凸	感	色	用	鋳	肖	稿	聾
音楽・舞踊	楽	曲	奏	音	歌	舞	弦	拍	演	箏	声	謡	旋	唄	鼓	踊	作	囃	笛	譜
娯楽・芸能	劇	映	演	撲	舞	伎	督	郎	画	駈	競	俳	棋	賭	座	作	優	踊	瑠	戯
言語	語	詞	音	言	字	文	韻	彙	声	漢	舌	母	辞	仮	晴	英	唇	詩	話	書
西洋文学	詩	劇	人	文	説	作	彼	恋	評	悲	家	寓	女	抒	批	魔	叙	編	愛	世
東洋文学	歌	諸	俳	句	蕉	詩	話	撰	芭	譚	文	詠	説	亭	巻	娼	集	狂	郎	江
地理	川	山	地	県	島	町	南	北	岸	郡	西	港	東	市	峠	泉	湖	丘	麓	口
古代史	王	帝	前	僧	征	隷	訌	奴	民	市	軍	政	族	世	彼	朝	盟	神	支	都
西洋史	党	政	民	領	国	革	年	王	議	会	邦	権	帝	世	命	農	軍	争	戦	蜂
東洋史	氏	藩	荘	幕	朝	郡	皇	府	領	城	町	田	授	武	臣	姓	国	政	官	惣

で、各分野で χ^2_{ij} の値が大きい漢字から必要な数だけその分野の重要漢字として取り出す。(2) 式の χ^2_{ij} を用いて 42 個の分野それぞれから重要漢字を 20 個ずつ抽出した結果を表 1に示す。

3.3 重要漢字による特徴ベクトルの作成

42 個の分野から取り出した重要漢字の異なりを特徴軸にとり、各特徴軸の値には重要漢字の頻度をとる特徴空間を考える。この特徴空間を用いて、未分類のテキストの内容と 42 個の分野それぞれの特徴を特徴ベクトルで表現し、ベクトル間の角度を測定することでテキストの自動分類を行なう。

未分類のテキストの内容は、この特徴空間内のベクトル x で以下のように表現する。

$$x = (f_1, f_2, \ldots, f_i, \ldots, f_n) \tag{4}$$

f_i は重要漢字 i の出現頻度で、n は χ^2 法によって取り出した重要漢字の総数である。

42 個の専門分野の特徴も、この特徴空間におけるベクトルによって表現する。例えば、分野 j の特徴ベクトル v_j は以下のように表現する。

$$v_j = (f_{1j}, f_{2j}, \ldots, f_{ij}, \ldots, f_{nj}) \tag{5}$$

f_{ij} は分野 j における重要漢字 i の出現頻度である。

テキストの分類は、未分類のテキストの内容を表すベクトルに最も近い特徴ベクトルをもつ分野にそのテキストを分類することで実現する。すなわち、分野 i の特徴ベクトル v_i と未分類のテキストの内容を表すベクトル x との間の角度 $\theta(v_i, x)$ は

$$\theta(v_i, x) = \cos^{-1}\left(\frac{v_i \cdot x}{|v_i||x|}\right) \tag{6}$$

と表現されるが、このとき

$$\min_i \theta(v_i, x)$$

となる特徴ベクトル v_i をもつ分野に未分類のテキストを分類する。

以上の方法で獲得した各分野の特徴ベクトルを評価するために、学習サンプルとして用いた百科辞典の項目説明文を分類する実験を行なった。実験は、各分野から取り出す重要漢字の数を 5 〜 100 個とさまざまに変化させて行なった。その結果を図 4に示す。分類精度は、重要漢字を各分野から 40 個以上取り出した場合で、およそ 60% とかなり低い。これは、百科辞典の項目説明文には、図 2の例のように 100 〜 200 文字程度の比較的短いテキストが多いためで、分野を推定するのに有効な量の漢字を取り出せなかったのが原因であると考えている。現段階では分類するテキストの長さと分類精度の関係は明らかではない。しかし、予備的な実験では漢字ののべ出現数が 200 文字以下のテキストでは分類精度が低下するという結果を得ている。今後、さまざまな長さのテキストを分類し、適切な分類が行なえるテキストの長さを明らかにする予定である。

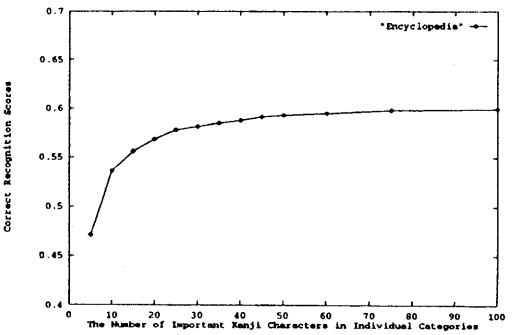

図 4: 百科事典の項目説明文の自動分類の結果

4 一般文書の自動分類

本章では、3章で作成した特徴ベクトルを用いて一般文書を分類する実験を行ない、提案したテキストの自動分類法が有効であることを示す。

4.1 実験と結果

一般文書の自動分類の実験対象として、次の 3種類のテキストを用いた。

1. 朝日新聞の天声人語 (約 2,000 件)
2. 朝日新聞の社説 (約 3,000 件)
3. 日経サイエンスの論文記事 (162 件)

実験結果を図 5、図 6 、図 7に示す。実験は、各分野から取り出す重要漢字の数を 5 〜 100 個とさまざまに変化させて行なった。

実験に使用した天声人語と社説のテキストは、NDC とは異なる 3 レベルの階層的な分類体系にしたがって分類されている。そこで、天声人語と社説の分野と 3章で設定した 42 個の分野とを人手によって対応づけ、実験結果と比較した。また、天声人語も社説も複数のテーマにまたがる内容を扱うことがあるので、1 つの記事が複数の分野に分類されていることがある。例えば『あふれる片仮名の言葉』という記事は「政治－地方行政－首長」「経済－経済総類－企

業」「文化－文化総類－言語」という3つの分野に分類されている。このため、実験結果がそれらのいずれかと一致すれば正解とする。

サイエンスの論文記事は分類されていなかったので、人手によって分類し、実験結果と比較した。

4.2 検討

図6、図7で示すように、社説のテキストの分類精度は最高で74%、サイエンスの記事では最高で85%であった。統計情報のみを用いた方法で、大量かつ幅広い範囲のテキストを分類した結果としてはかなりよい結果である。一方、天声人語のテキストは最高で47%しか分類に成功せず、他の2つのテキストに比べて分類精度が悪い。この原因として次の2つが考えられる。

1. テキストの種類

   天声人語は随筆や小説などのテキストに近く、平易な単語が用いられる。一方、社説やサイエンスは学術論文に近く、専門用語が多く用いられている。専門用語を構成している漢字はその分野の特徴を表す重要漢字であることが多いので、社説やサイエンスの記事の正解率が高くなると考えられる。

   しかし、学術文献であっても重要な概念を表す単語がカタカナなどの漢字以外の文字で構成されている場合がある。このようなテキストに対しては本手法は有効ではないと考えられる。そのようなテキストには単語を意味の基本単位とし、構文情報や文脈情報あるいは単語間の共起関係などを用いて分類しなければならないだろう。

2. 複数のテーマにまたがる内容

   天声人語では内容が複数のテーマにまたがる場合が多い。特に主要テーマを説明する前に、内容的な関連の薄い導入部が存在することが多い。例えば『出処進退の美学』という記事では、衆議院議長の辞職という主要テーマを述べるために、スポーツ選手の引退の話を導入として用いている。内容が複数のテーマにまたがると、それぞれのテーマの内容を表す漢字がまざりあい、本手法では分類が正しく行なえない。

   そこで、章や節などの意味的な区切りを利用してテキストを分割し、分割したテキストをそれぞれ分類することを考えた。実験の対象に「人工知能と人間」という本を選び、本全体と各章の自動分類結果を表2に示す。この本はNDCでは「情報科学」に分類されるので、本全体の分類結果は正しい。また、3章と5章では「情報科学」以

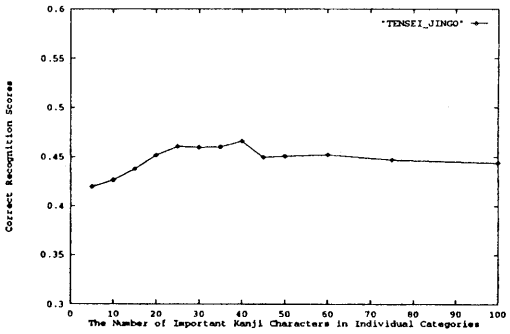

図 5: 天声人語を自動分類した結果

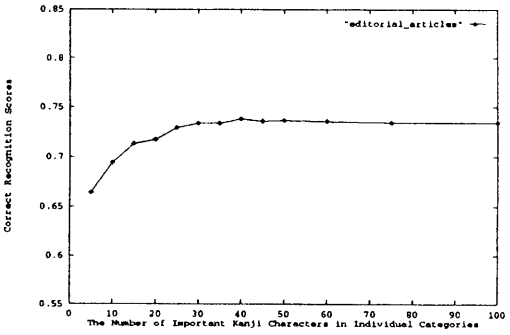

図 6: 社説を自動分類した結果

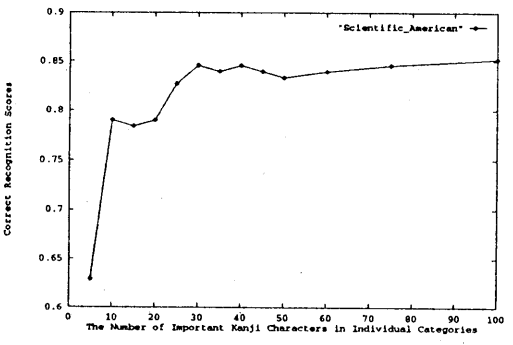

図 7: サイエンスの論文記事を自動分類した結果

表 2:「人工知能と人間」の分類結果

章	タイトル	分類結果
第1章	コンピュータができること	情報科学
第2章	認識への挑戦	情報科学
第3章	言葉への挑戦	言語
第4章	理解することへの挑戦	情報科学
第5章	人工知能と哲学	心理学
「人工知能と人間」全体		情報科学

外の結果が得られたが、それらの章では人間の言葉と心理的側面が述べられているのでその分類結果は妥当である。このように、意味的な区切りを見つけてテキストを分割すれば、複数のテーマにまたがるテキストでも正しく分類することができる。

5 おわりに

本研究では漢字の統計情報によってかなりの精度でテキストを分類できることを確かめた。また、日本語テキストに含まれている漢字によって、そのテキストの内容を表現することができるという見通しがえられた。

テキストの分類情報はさまざまな目的に利用できる。その1つに、新聞記事のおおまかな分類への利用が期待されている。新聞記事の自動分類は、今後発展が期待されている電子新聞の検索や配布のために重要な技術になると考えられる。そのほかに、目次情報を用いた図書検索への利用がある。4.2節ではテキストの分類精度をあげるために章や節などの目次情報を利用したが、こうした目次情報は図書の検索にも有効であることが確かめられている [長尾 92]。目次情報による図書検索では、図書や雑誌の章や節のタイトルに含まれる単語をその図書の内容を示すキーワードとして扱っている。テキストの分類情報は、章や節のタイトルの中で用いられる単語の意味を精密に解釈するのに役立つ。例えば、章や節のタイトルの中に「言語」という単語が出てきた時、その意味は自然言語を含めて非常に広い範囲におよぶ。その中から「計算機言語」あるいは「プログラミング言語」の意味で検索したいときは、章・節のタイトルの中に「言語」という単語を含むものの中からその章・節のテキストが「工学」あるいは「情報科学」の分野に分類されている文献を検索すればよい。

新聞記事や図書の検索への利用のほかに、かな漢字変換における変換候補の選択にもテキストの分類情報は利用できる。すなわち、編集しているテキストの分類情報にもとづいて、漢字の変換候補の順位を変更するのである。例えば、編集しているテキストが「植物」の分野に属しているならば、「はな」というかな入力に対して、「鼻」や「華」ではなく「花」を優先して出力するのである。分類情報は、編集しているテキストに含まれる漢字の統計情報から容易に獲得できる。

謝辞 本研究のために『世界大百科事典 CD-ROM 版』の使用を快諾してくださいました平凡社の各位に感謝いたします。また図書館分類法について貴重な助言をしていただいた光華女子大学の谷口敏夫助教授、京都大学工学部の中尾富貴子氏、呑海さおり氏、由本慶子氏、および京都大学理学部の影山貴子氏、丹下晴美氏に感謝いたします。

参考文献

[細野 84] 細野 他: 漢字出現頻度に基づいた日本語文献の定量的分析, 第 28 回情報処理学会全国大会論文集 3M-1 (1984).

[細野 85] 細野 他: 漢字の出現頻度情報を用いた日本語文献の自動分類, 情報処理学会研究報告 85-NL-47 (1985).

[亀田 87] 亀田, 藤崎: テーマ・キー概念・キーワード間の階層構造を利用する新聞記事情報の分類・検索システム, 情報処理学会論文誌 Vol.28 No.11 (1987).

[長尾 76] 長尾, 水谷, 池田: 日本語文献における重要語の自動抽出, 情報処理 Vol.17 No.2 (1976).

[長尾 92] 長尾: 目次情報などを利用した図書・文献検索方式, 情報の科学と技術 Vol.42 No.8 (1992).

[鈴木 87] 鈴木, 小橋, 深谷: ビジネス通知文書の自動分類, 信学技報 OS87-42 (1987).

[田村 88] 田村 他: 統計的手法による文書自動分類, 第 36 回情報処理学会全国大会論文集 6U-5 (1988).