章立てに注目した論文の研究内容による自動分類

島 広幸　　建石 由佳

工学院大学情報学部

Email:j106061@ns.kogakuin.ac.jp,yucca@cc.kogakuin.ac.jp

１．はじめに

科学・技術論文を自動分類するとき，研究分野で分類を行っているものがほとんどである．例えば，[1]は研究分野のカテゴリの時間的変化を考慮に入れた論文のカテゴリ分類を行っている．また，論文の自動分類の研究ではなくても，自動分類の評価実験のために論文を用いるケースもある[2,3]．その際も論文は分野別に分類される．

しかし，論文は分野内でさらに研究内容により分類することができる．例えば，「文書分類」という研究分野には「属性選択」や「多重ラベリング」などの研究内容が挙げられる．

このように研究内容で分類することで，研究分野で分類している場合に比べて，より詳細に論文を整理することができる．したがって，論文が分野別に整理されているだけの場合より，検索を効率的に行うことができると考えられる．

そこで本研究では，論文を研究内容で自動分類する方法を提案する．一般的に論文は章立てされており，各章ごとに書かれる内容が異なる．この章立てを利用することで，論文から研究内容に関する記述部分を抽出し，その抽出した文章を一文書として扱うことで研究内容での自動分類を実現する．

本稿の構成は以下の通りである．まず 2 章で論文の構成とその構成要素の分類についての述べた後，提案手法の説明を行い，3 章で実験手順と結果を述べ，4 章で考察を行い，5 章で結論と今後の課題について述べる．

2. 提案手法

論文を自動分類する際，多くの研究が論文のアブストラクトあるいは本文を一文書として用いている．しかし，本研究では研究内容が記述されている部分を抽出し，その抽出された文章を一文書として分類を行う．

本章では論文の構成について述べ，その後，構成要素が「研究分野に関する記述」と「研究内容に関する記述」に分類できることを示した後，提案手法の説明を行う．

2.1 論文の構成要素

論文の本文は，序論部，本論部，結論部から構成される[4]．本研究では，本論部をさらに「提案手法」と「実験」に分割する．「文書分類」に関する論文を 101 編集めたが，そのうち 73 編が提案手法と実験方法を章立てて記述しているため，このように分割できると考えた．以下，本論文では 4 つの構成要素を「序論，提案手法，実験，結論」と記述する．

次に各構成要素の内容について説明する．「序論」に記述される内容は，研究の目的や研究を行う理由，研究分野の歴史や関連研究などである．「提案手法」には，研究の目的を実現するための方法，研究に用いる用語の定義などが記述される．次の「実験」には，評価実験の方法や実験データの内容，比較対象，実験結果などが記述される．最後に「結論」には，実験結果の分析や提案手法の問題点・改善方法，研究のまとめ，これからの課題などが記述される．

2.2 記述内容の分類

前述した構成要素 4 つをさらに「研究分野に関する記述」と「研究内容に関する記述」に分類する．

まず，「研究分野に関する記述」について説明する．研究の目的や研究の必要性は同じ分野内で内容が類似する．例えば，「情報検索」の分野では，多くの電子データからユーザが望む情報を効率良く見つけ出すという共通の研究目的がある．また，同一分野内の研究では提案手法を評価する際に共通の評価尺度が用いられることが多い．「文書分類」でテストデータがどれだけ正しく分類できたかを表す正解率などが例として挙げられる．このような同一分野内で内容が類似するものを「研究分野に関する記述」とする．つまり，前節で述べた 4 つの構成要素の「序論」と「実験」がこれにあたる．

次に「研究内容に関する記述」について「文書分類」の研究内容「属性選択」を例に説明する．「文書分類」では多くの場合，文書を単語集合(bag of words)に変換し，処理を行う．bag of words を使用した研究として[5]などがある．この単語集合の要素を選択する方法の研究が「属性選択」である．「属性選択」の論文には，従来の文書の表現法，つまり，文書から単語を抽出する方法などの問題点が「文書分類」の他の研究内容より多く記述される．つまり，論文には研究の内容ごとに記述される項目に偏りや，あるいは特定の研究内容にしか記述されない項目が生まれる．このような内容を「研究内容に関する記述」とする．また，前節で述べた 4 つの構成要素の中で「提案手法」がこれにあたる．

また，「結論」に書かれる内容は，前節で述べたとおり，実験結果の分析や提案手法の問題点・改善方法，研究のまとめ，これからの課題などである．この内容は「序論」や「実験」よりも「提案手法」に関係が強いと考えられるため，「結論」も「研究内容に関する記述」といえる．

2.3 提案手法

論文を研究内容で分類するために「研究内容に関する記述」を用いて論文の分類を行う．しかし，本研究では「研究内容に関する記述」のうち，「提案手法」のみを用いる．「考察」を除く理由は，「考察」の内容が各研究固有の問題点などであるため，研究内容ごとに分類する際には不要と考えるためである．

例として本稿の本文の章立てを利用し，論文から「研究内容に関する記述」を抽出する方法を説明する．本稿の章立ては表 1 のような構成になっている．この章立てを 4 つの構成要素に対応付けると表 2 のようになる．表 2 内の番号は表 1 で示した章立ての章番号である．したがって，分類時に用いる「研究内容に関する記述」は「2.提案手法」となる．

表 1:本稿の章立て

1.はじめに
2.提案手法
3.評価実験
4.考察
5.まとめ

表 2:章と構成要素の対応

1:序論
2:提案手法
3:実験
4,5:結論

また，本研究では文書の特徴として名詞を用いる．名詞の抽出は形態素解析器 MeCab[1]を用いて行った．また，分類器には文書分類でよく用いられる単純ベイズ分類器を用いた[10,11].

3. 評価実験

3.1 実験データ

本研究は論文の研究内容による分類のため，処理を適用する文書集合は分野ごとに分類されている必要がある．本実験では，JDream Ⅱ[2]および CiNii[3]より「文書分類」に関する論文を集めた．集めた論文数は 101 編である．集めた論文は PDF 形式であるため，OCR ソフトを用いてテキストファイルへの変換を行った。また、「研究内容に関する記述」部分は人手で抽出した．学習および評価を行うため，集めた論文を人手で 6 つのカテゴリに分類した．6 つの分類は「属性選択，重み

[1] http://mecab.sourceforge.net/
[2] http://pr.jst.go.jp/Jdream2/
[3] http://ci.nii.ac.jp/

表 3:実験データ

カテゴリ名	文書数(編)
属性選択	25
重み付け	5
特定ジャンル	17
多重ラベリング	7
アルゴリズム	29
その他	17

表 4:実験結果(正解率:%)

手法(テストデータ数)	クローズドテスト	オープンテスト
提案手法(5)	94.7	43.2
比較手法(5)	96.8	39.2
提案手法(10)	95.1	46.0
比較手法(10)	96.9	45.8

付け，特定ジャンル，多重ラベリング，アルゴリズム，その他」である．各カテゴリの文書数は表 3 に示す．

「属性選択」とは，文書を単語集合に変換する際に単語集合の要素の選択や追加を行い文書分類の精度の向上を図る研究のことである．例えば，[6]ではシソーラスを用いて単語の上位概念を取得し，単語集合に加えることで同義語を考慮した文書分類を行っている．次に「重み付け」とは，分類を行うとき，単語の出現頻度を用いるのではなく，単語や文書の重要度などを何らかの処理により単語に対応する数値を計算し，その数値を用いて分類を行おうとする研究のことである．文書の重要度を用いる研究の例として[7]が挙げられる．[7]は時系列的な文書に対し，新規性を考慮したクラスタリングを行うため，時間の経過にしたがって文書の価値を逓減させる手法を提案している．「特定ジャンル」とは，本研究のように特定のジャンルの文書を対象とした研究のことである．「多重ラベリング」は，1 つのデータと複数のカテゴリに対応させる研究のことである．「アルゴリズム」とは，今まで行われていないクラスタリングアルゴリズムを提案している研究である．例として[8,9]などが挙げられる．「その他」とは，上記のどのクラスにも属さない研究内容の論文とした．

3.2 実験方法

まず，テストデータを 5 個あるいは 10 個ランダムに選択する．残りを学習データとし，ナイーブベイズ分類器で学習を行った．その後，テストデータの分類を行い，その正解率を算出する．正解率の以下の式によって算出する．

$$\text{正解率} = \frac{\text{正しく分類されたデータ数}}{\text{テストデータ数}}$$

以上の作業を 50 回行った．また，比較実験として論文の本文全てを用いた場合の実験も同様に行った．

3.3 実験結果

提案手法と比較実験の平均正解率を表 4 に示す．

4. 考察

表 3 よりわかるように，提案手法と論文の本文を用いた場合の比較実験の結果，正解率にほとんど差は見られなかった．この結果より，提案手法が比較実験より精度が良い分類が行えるとは言えない．

本手法の問題点の 1 つとして，単語頻度が考えられる．表 5 に実験中に取得した各カテゴリの頻出単語上位 5 個を示す．表 5 の通り，全てのカテゴリに「文書」，5 つのカテゴリに「分類」，他にも多くの文書分類に関する単語が入っていて，カテゴリ特有の単語(例：多重ラベリングの「ラベル」)がほとんど出現していない．そのため，各カテゴリを特徴づけるような単語が出現していても，その単語頻度が小さいためにカテゴリ推定に影響を与えず，誤分類の原因となると考えられる．その解決方法として，全てのカテゴリに共通する単語の影響を小さくするため，各カテゴリから特徴となる単語を選択し，その単語に重みを付けることが考えられる．

また，学習データが多い方の精度が悪くなると

表 5:各カテゴリの頻出単語

頻度順位	属性選択	重み付け	特定ジャンル	多重ラベリング	アルゴリズム	その他
1	文書	文書	分類	分類	分類	分類
2	分類	クラスタ	文書	文書	クラス	文書
3	単語	類似	電報	データ	ラベル	単語
4	分野	分野	抽出	クラスタ	文書	確率
5	ベクトル	クラスタリング	重要	単語	学習	記事

いう実験結果であった.

この原因として，テストデータが少ないため，1 つの誤りの影響が大きく点が挙げられる．したがって，より多くのデータで実験を行う必要がある.

実験の結果，分類精度にはほとんど差がないと言えるが，提案手法と比較実験では用いる文章量が異なるので，提案手法の方が比較実験に比べ，より少ない文章で同等の精度と示すことができたと言える.

5. まとめ

本研究では，章立てに注目した論文の研究内容による自動分類を提案した．論文の構成を利用し，研究内容に関する記述を抽出し，その部分を一文書として分類を行うことで実現を試みた．評価実験の結果は本文全体を用いた場合と同等の精度であった．しかし，提案手法は比較実験に比べ少ない文章量で同程度の精度を示すことができた.

今後は，より精度を向上させる方法と学習データと分類精度の関係の解明を行っていきたい.

6. 参考文献

[1] 榊剛史，石塚満，松尾豊：制約付きクラスタリングを用いた論文分類，人工知能学会全国大会論文集，Vol.20th，pp.1A1-1，2006.

[2] 小熊淳一，内海彰：語の共起情報を用いた文書クラスタリング，人工知能学会全国大会論文集，Vol.19th，pp.2E1-01，2005.

[3] 藤野昭典，上田修功，斉藤和巳：最大エントロピー原理に基づく付加情報の効果的な利用によるテキスト分類，情報処理学会論文誌，Vol.47，No.10，pp.2929-2937，2006.

[4] 藤野昭典，上田修功，斉藤和巳：文書の構成要素モデルのアンサンブル学習に基づくテキスト分類，電子情報通信学会技術研究報告，Vol.104，No.349，pp.69-74，2004.

[5] 藤井遼，櫻井彰人：文学作品推薦のための文書分類，人工知能学会全国大会論文集，Vol.22nd，pp.2G2-03，2008.

[6] 上嶋宏，三浦孝夫，塩谷勇：同義語,多義語の考慮による文書分類の精度向上，電子情報通信学会論文誌，Vol.J87-D-1，No.2，pp.137-144，2004.

[7] 石川佳治，北川博之：忘却の概念に基づくインクリメンタルな文書クラスタリング手法, 情報処理学会研究報告, Vol.2001, No.70, pp.313-320，2001.

[8] 平博順, 向内隆文 : Support Vector Machine によるテキスト分類，情報処理学会研究報告，Vol.98，No.99，pp.173-180，1998.

[9] 高木昇，時永祥三：遺伝的プログラミングによる分類関数近似を用いた文書分類とその応用，電子情報通信学会技術研究報告，Vol.105，No.503，pp.13-17，2006.

[10] Tedy Segaran(原著)，當山 仁健(翻訳)，鴨澤 眞夫(翻訳)：集合知プログラミング，オライリー・ジャパン，pp.134-138，2008.

[11] 元田浩，津本周作，山口高平，沼田正行：データマイニングの基礎，オーム社，pp.35-39，2006.

[12] 古郡 廷治：論文・レポートの文章作成技法―論理の文章術，日本エディタースクール出版部，pp.176-177，2006.