2006－MPS－61（9）
2006／9／15

アブストラクトを用いた論文分類システムの設計と実装

柏木　裕恵[†]　　髙田 雅美[†]　佐々木　明[††]　城　和貴[†]
[†]奈良女子大学　大学院人間文化研究科
{hiroe, takata, joe}@ics.nara-wu.ac.jp
[††]日本原子力研究開発機構
sasaki.akira@jaea.go.jp

概　要

本論文では，アブストラクトを用いた論文分類システムを提案する．オンラインで必要な論文を検索する際，アブストラクトの情報のみ閲覧可能である場合が多く，必要な論文を見つけ出すことは非常に困難である．従来の論文分類システムでは分類に論文が必須であるため，論文を入手していない論文検索段階では分類を行うことができない．しかし，我々の提案するシステムでは，アブストラクトさえあれば論文の分類が可能となるため，論文を検索する際に分類を行い，必要な論文だけを入手することができる．我々は，分類手法として LVQ を採用し，これまで困難とされてきた原子分子物理学分野の論文の分類実験を試みる．実験結果より，アブストラクトのみで論文分類を行うことができる提案手法の有効性を示す．

Design and implementation of a paper classification system using abstracts

Hiroe Kashiwagi[†]　Masami Takata[†]　Akira Sasaki[††]　Kazuki Joe[†]
[†] Graduate School of Humanities and Sciences, Nara Women's University
[††] Japan Atomic Energy Agency

Abstract

In the research field of text classification, classifying papers requires the body of the papers so far. When we look for some kind of papers on the web, we can obtain downloadable abstracts freely in general. But we cannot use the conventional text classification methods to make sure of the necessary paper, because we have just the abstracts without their bodies. Therefore we propose a paper classification system using just abstracts. We adopt LVQ, a pattern recognition technique, to implement a paper classification system. In this paper, we show our system can classify papers efficiently by abstract.

1. はじめに

近年，論文の電子化と検索エンジンの改善により，必要な論文をオンラインジャーナルから入手することが一般的となってきている．しかし，必要な論文を検索する際，多くの web ページにおいて，論文が有料であることやダウンロード可能な論文数が制限されているために，アブストラクトの情報しか閲覧できない場合が多い．よって，分類に論文が必須となる従来の論文分類システムでは，論文を入手していない論文検索の段階では分類を行うことができないという問題がある．この問題を解決するために，アブストラクトを用いた論文分類システムが求められている．アブストラクトのみで論文分類ができれば，論文を検索する際に分類を行い，必要な論文だけを入手することも可能となる．

アブストラクトを用いた論文の分類は，テキスト分類の一種である．現在までに様々な機械学習を用いたテキスト分類手法が提案されており，その有効性が示されている[1]．我々は分類手法として Learning Vector Quantization（LVQ）[2]を採用し，アブストラクトによる論文の分類を試みる．

以下 2 章で原子分子物理学分野の論文分類について述べ，3 章にて LVQ による論文分類システムを提案する．4 章で本システムの評価方法について述べる．5・6 章で評価・考察を行う．

2. 原子分子物理学分野の論文分類

本研究では，日本原子力研究開発機構，核融合科学研究所との共同研究[3]である特異表現に適応可能な論文分類システム開発の一環として，原子分子物理学分野の論文をアブストラクトのみで論文分類するための新しいシステムを開発する．原子分子物理学分野の論文には，原子分子と電子の衝突による電離や励起過程のデータ（以下，原子分子データと略記）を利用するための情報が記載されている．原子分子物理学分野において，毎年，原子分子データが発表されるジャーナルは Phys.Rev.A 誌をはじめ 20 種類程度である．その中の論文総数は 10^4 件/年のオーダーであるのに対して，収集の対象となる論文の数は 100 件/年程度である．人手による情報収集は，論文を読んで原子分子データが掲載されているか否かを判断しなければならないため，大変な労力を要する．この労力を軽減させるために，機械による認識が必要である．

しかし，原子分子物理学分野の論文を機械的に分類することは，論文に含まれている化学式等の原子分子に関する特異な表現を機械に認識させるのが容易ではないため困難である．ゆえに原子分子物理学分野の論文分類において，機械学習手法を適用してきた例はない．そこで我々は，原子分子物理学分野の論文分類システムに LVQ を適用する．

LVQ は，入力データのパターン分類を目的とした教師ありの競合学習を行う手法である．解空間を分割するための参照ベクトルを用いて学習を行う．1990 年に Kohonen によって提案された．LVQ1，LVQ2.1，LVQ3 などの方法があり，これらの手法は，LVQ1 を基本として変形したものである．本研究ではLVQ1を用いて分類を行う．

3. LVQ による論文分類システム

テキストの分類に際しては, テキストの特徴を多次元のベクトルで表現することが多い. このベクトルを特徴ベクトルとよぶ. 本研究では, 各アブストラクトの単語に前処理を施し, 単語の頻出度情報を使用した特徴ベクトルを作成し，LVQ に適用する.

原子分子物理学分野の論文には, 化学式が含まれていることが多い. 化学式は, 空白の部分や特異な表現が含まれているため, 機械的にテキスト分類を行うと，1つの化学式が複数の化学式, あるいは単語として抽出されることがある. この問題を解決するために, 前処理として，NICT原子分子重要表現抽出システム[4]を用いて化学式を抽出する. このシステムでは, アブストラクトの化学式の部分を色付きで表示させるHTMLファイルが作成される. その出力結果を利用して, 化学式の部分に, 化学式であることを表すタグを挿入する. これらのタグによって機械的な化学式の認識を可能とする. この際, 化学式を次のように分類する.

CHEM1） 原子 *(e.g. Li，hydrogen)*
CHEM2） イオン *(e.g. Xe II，O^{5+})*
CHEM3） 分子 *(e.g. H_2O)*
CHEM4） 数字 + 原子 *(e.g. 3He，63Cu)*
CHEM5） CHEM1）の電子配置 *(e.g. $1s^2 2s^2 2p^2$)*
CHEM6） CHEM1）の微細構造 *(e.g. $^1S_0$，$^2S_{1/2}$)*
CHEM7） 数式 *(e.g. l=0，n=0)*
CHEM8） CHEM2）+ 数 +1 *(e.g. $2p^4 3snl$)*

2 つめの前処理として，化学式以外の単語に対して, ストップワードを除去しステミング処理を行う. 本研究では，最も広く利用されている有名な Porter stemming algorithm[5]を使用した Perl モジュール[6]を用いてステミングを行う.

以上の前処理を施した単語から, 特徴ベクトルを作成する．作成の際には，次の 6 種類の頻出度を用いる．なお, カテゴリ 1（原子分子データが掲載されているもの）のトレーニングデータを D1，カテゴリ 0（原子分子データが掲載されていないもの）のトレーニングデータを D2 と記す.

F1-a） D1 のアブストラクトに含まれる全単語の頻出度
F1-b） F1-a + 化学式の頻出度
F2-a） D1，D2 のアブストラクトに含まれる全単語の頻出度
F2-b） F2-a + 化学式の頻出度
F3-a） 理化学辞典[7]に掲載されている用語の頻出度
F3-b） F3-a + 化学式の頻出度

頻出度の算出には tf・idf 法[8]を用いる．特徴ベクトル F1-a, F1-b は，トレーニングデータセットによりベクトルの各要素が示す内容が変わる．また，要素数も一定にならない．これは特徴ベクトル F2-a，F2-b についてもいえることである. 理化学辞典には 23500 語が掲載されている．特徴ベクトル F3-a および F3-b では，それらをすべて使用するため，23500 次元，23508 次元となる.

LVQで使用する参照ベクトルは, ランダムに生成しても学習が可能であるが，学習に要する時間が長くなる. 本研究ではすべての参照ベクトルをトレーニングデータより作成する．カテゴリ毎にランダムに5つの特徴ベクトルを選択し，その平均ベクトルを求め，これを参照ベクトルとする. LVQによる学習は，学習させた参照ベクトルによって, 97%以上のトレーニングデータを正しく分類できるようになるまで行う．20回学習させて，97%以上のトレーニングデータを正しいカテゴリに分類できないようであれば20回で学習を打ち切る．LVQ の学習係数の初期値は0.8とし，学習回数に応じて0.04ずつ減少させる.

4. 評価方法

本論文の評価実験では, 市川が収集した原子分子データが掲載されている論文 379 件[9]のうち，ジャーナル Phys.Rev.A 誌に掲載されている 126 件をカテゴリ 1 として用いる．また，Phys.Rev.A 誌（vol.41～62）に掲載されている全てのアブストラクトのうち, カテゴリ 1 以外のものをカテゴリ 0 とする. このカテゴリ 0 の全アブストラクトは，市川が全てチェックしたもので, データ数は 15944 件となる.

本システムの性能評価を行うために, 全アブストラクトの中からランダムに選び出されたトレーニングデータセットとテストデータセットを用いて実験を行い, その際の認識率，再現率（Recall rate），適合率（Precision rate）を求める．本研究では，再現率を重視する．これは, 我々の提案するシステムでは, 原子分子物理学分野の論文のアブストラクトを正しくカテゴリに分類することよりも, 本来カテゴリ1 に属している論文のアブストラクトを可能な限り大量に収集することが重要であるためである.

5. 実験

5.1 特徴ベクトル作成方法の違いによる比較実験

600 件のアブストラクトを使用した際の認識率，再現率，適合率を調べる．600 件のデータは，カテゴリ 1 が 126 件，カテゴリ 0 が 474 件で構成される．トレーニングデータとテストデータを共に各 300 件用いる（カテゴリ 1 : カテゴリ 0 = 63 : 237 とする）. LVQ で用いる参照ベクトルの数は 200 とし[10]，そのうちの半分がカテゴリ 1 に，残りの半分がカテゴリ 0 に属するものとする. この実験で使用する特徴ベクトル F3-a は，全アブストラクトに掲載されていない理化学辞典の用語に属する要素を省いたため，1178 次元である.

5.1.1 項において，特徴ベクトルの作成方法の違いによる認識率，再現率，適合率の比較，さらに 5.1.2 項にて, 特徴ベクトル作成に化学式を使用した場合と使用しなかった場合の比較を行う.

5.1.1 特徴ベクトル F1，F2，F3 の違いによる比較

図 1 に特徴ベクトル F1-a の要素数の変化を示す. 図 1 から, 特徴ベクトル F1-a の要素数は 1006～1205 の間で変動していることがわかる．これは，システムの性能がトレーニングデータセットに大きく依存することを意味する．それに対し，理化学辞典の用語を基とする特徴ベクトル F3 は，理化学辞典に掲載されている用語やその数が決まっているため, 要素が示す内容や要素数は変化しない.

図 2～図 4 は，トレーニングデータセットを 100 個作成し, それぞれ学習させた結果の平均値をグラフで表したものである．特徴ベクトル F3 を使用した場合の実験

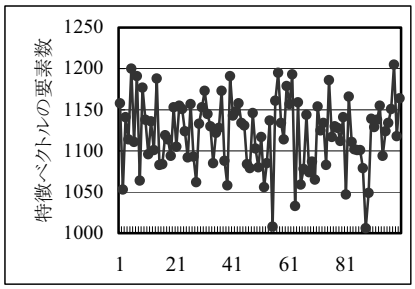

図 1：特徴ベクトルの要素数の変化（実験 5.1）

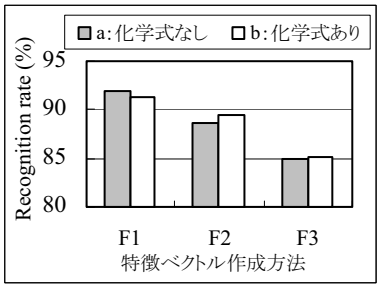

図 2：認識率（実験 5.1）

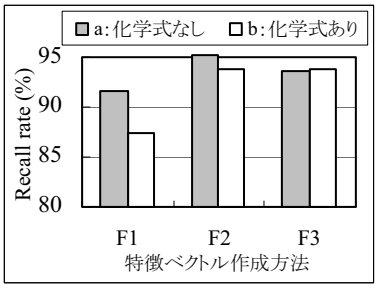

図 3：再現率（実験 5.1）

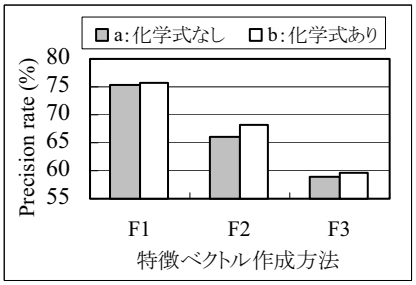

図 4：適合率（実験 5.1）

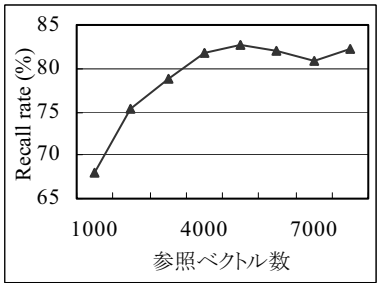

図 5：再現率（実験 5.2）

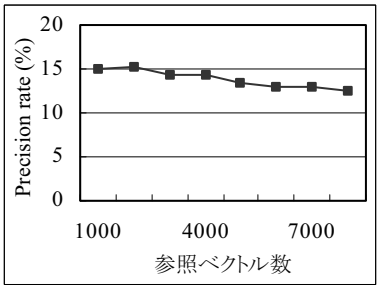

図 6：適合率（実験 5.2）

結果より，適合率は低いが，我々が重視している再現率は 90%を超えている．よって，特徴ベクトル F3 が最適である．

5.1.2 化学式使用に違いによる比較

原子分子物理学分野において化学式は非常に重要な表現であり，専門性が高いと考えられる．しかし，特徴ベクトル F1 や F2 に関する実験結果を比較してみると，化学式が重要な役割を果たしているとは言い切れない．これは，特徴ベクトル F1，F2 ではアブストラクトに掲載されている単語を使用しているため，化学式よりも専門性の高い単語が含まれているためである．

一方，特徴ベクトル F3 に注目すると，図 2～図 4 より特徴ベクトル F3-a と F3-b では，認識率，再現率，適合率のすべてにおいて，化学式を使用した場合に若干よい結果が得られている．これは，特徴ベクトル F3 が理化学辞典の用語を基に作成されていることに起因すると思われる．理化学辞典には，原子分子物理学分野で使用される用語だけでなく，生物学分野や化学分野等を含む広い分野で使用される用語が多く含まれている．使用している化学式はイオンや電子配置などの特有の表現を含んでいる．ゆえに特徴ベクトル F3 を使った実験の際には，理化学辞典の用語ではなく化学式に専門性が認められ，認識率，再現率，適合率に影響を及ぼすものと考えられる．

5.1.1 項，5.1.2 項より，理化学辞典の用語と化学式の出現頻度を基に作成された，特徴ベクトル F3-b を最適な特徴ベクトルであるとする．

5.2 参照ベクトル数の違いによる比較実験

4 章で述べた 16070 件のデータセットをすべて使用して実験を行う．トレーニングデータとテストデータとして 8035 件ずつ用い，そのうち，カテゴリ 1 のアブストラクトを 63 件とする．最適な参照ベクトル数を調べるために，参照ベクトル数を 1000～8000 とし，1000 ずつ増やして実験を行う．特徴ベクトルは F3-b を用いる．実験 5.2 においては，特徴ベクトル F3-b は 3519 次元のベクトルである．

図 5，図 6 は，トレーニングデータセットを 10 個作成し，それぞれ学習させた結果の平均値をグラフ化したものである．認識率に関しては，参照ベクトルの数に関係なく 95%以上でほぼ一定の値になっていたため省略する．再現率は，参照ベクトル数が 5000 と 8000 の場合に最もよい結果を示している．適合率は，参照ベクトル数が増加するにつれて，徐々に低くなっているが，それほど差がみられないため，最適な参照ベクトル数は 5000 と 8000 の 2 つであると考えられる．

5.3 参照ベクトルの属するカテゴリの割合の違いによる比較実験

参照ベクトル数 5000 と 8000 の場合において，再現率は共に約 82%になっているが，適合率はどちらも 13%程度とやや低い．そこで，参照ベクトルが属しているカテゴリの割合を考慮して比較実験を行う．

これまでは，過去に我々が行ってきた研究結果により，カテゴリの割合はカテゴリ 1：カテゴリ 0＝1：1 としている．文献[11]においては扱っているアブストラクトデータの総数が 364 で，そのうちの 127 がカテゴリ 1 に属しているアブストラクトであるため，この割合が最適であるという実験結果が示されている．しかし，今回は約 8000 のデータの中から約 60 のカテゴリ 1 のデータを探し出す作業であることから，1：1 という割合は適していない可能性がある．そこで，参照ベクトルの数が 5000 と 8000 の場合に，参照ベクトルが属しているカテゴリの割合を変化させたときの再現率・適合率の変化を図 7，図 8 に示す．認識率については，変化がほぼ認められなかったため省略する．

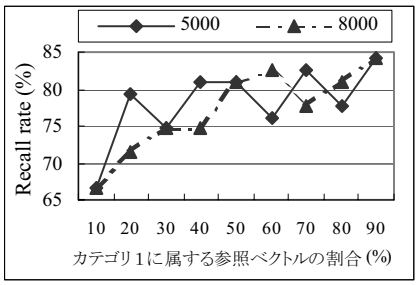

図 7：再現率の変化（実験 5.3）

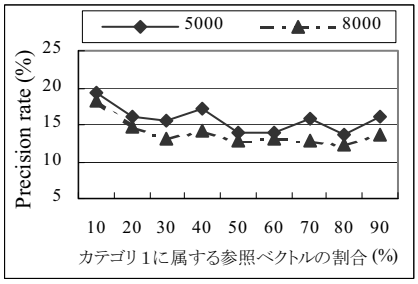

図 8：適合率の変化（実験 5.3）

図 7，図 8 より，参照ベクトル数がどちらの場合においても，再現率，適合率が振動していることが確認できる．これは，参照ベクトルによって解空間を明確に分割できていないことを示している．原因としては，使用したデータの総数に対し，カテゴリ 1 のデータ数が極端に少ない点があげられる．この問題については，本システムを利用して，より多くのカテゴリ 1 の論文を探し出し，カテゴリ 1 のデータ数を増やしていくことによって解決できると考えている．

6. 考察

我々は各種実験を行ってきたが，ここで，特徴ベクトル F3-a のベクトルデータについての考察を行う．

実験 5.2, 5.3 で使用した各アブストラクトの特徴ベクトル F3-a は 3519 次元である．それらのベクトルのうち，1 件のベクトルにしか値のない成分が 893 あった．これは分類には何の意味もなさない成分である．よって明らかに 2626 次元には縮約されるといえる．また，10 件以下の数件にしか値のない成分も数多くあった．このような成分は他の成分との相関が極端に小さくなり，分類指標としての意味がない．分類の目的にもよるが，ある程度の数のアブストラクトに値のないような成分は除かれるべきである．次元数が増大すると，一般に統計モデルの安定性は非常に悪くなり，意味のないモデルになるといわれている．よって，次元数を縮尺することは効果があると考えられる．

今回使用したデータセットは，カテゴリ 1 のデータ数とカテゴリ 0 のデータ数にかなりの差があるため，次元数の縮約には十分な注意が必要である．よって次元数の縮約は行わなかった．次元数の縮約に関しては，今後検討していく予定である．

7. まとめ

本論文では，アブストラクトを用いた論文分類システムを提案し，実装した．その結果，論文本体ではなくアブストラクトを用いても論文を分類できることを実証した．また，原子分子物理学分野の論文を分類する手法として LVQ を適用することにより，原子分子物理学分野の論文に対しても機械学習が有効であることを示した．理化学辞典の用語と 8 種類の化学式の出現頻度を基に特徴ベクトルを作成した場合に，認識率 95%，再現率 80%，適合率 15%という良好な結果を得ることができた．原子分子物理学分野では，今まで 8000 件の論文の中から 60 件しかないカテゴリ 1 の論文を人手により探し出している．我々の提案したシステムを用いる場合，検索すべき論文を 8000 件から 320 件に減らすことができ，その中から人手で 60 件を探し出せばよい．これにより，大きな労力を使わずに効率的に必要な論文を収集できる．

今回，Phys.Rev.A 誌に掲載されている論文の分類を行い，成果をあげることができた．したがって他のジャーナルにおいても，本研究の手法が有効であることが十分に考えられる．しかし，再現率が 80%であるため，60 件のうち 12 件の論文は探し出せないことになる．ゆえに再現率をさらに向上させる必要がある．我々は，今回の実験では用いなかった，理化学辞典に掲載されていない原子分子物理学分野特有の用語を調査し，特徴ベクトルの要素として加えることにより，この問題を解決できると考えている．

謝　辞

特徴ベクトルデータの分析に御協力いただきましたお茶の水女子大学吉田裕亮教授に心より感謝いたします．

参 考 文 献

[1] Sheng Gao, Wen Wu, Chin-Hui Lee, Tat-Seng Chua : Maximal Figure-of-Merit Learning Approach to Text Categorization, ACM SIGIR, P174-181 (2003).

[2] HUT –CIS – Research – SOM_PAK, LVQ_PAK, http://www.cis.hut.fi/research/som-research/nnrc-programs.shtml

[3] Atomic and Molecular Data Research Center, NIFS, http://dpc.nifs.ac.jp/amdrc/index-j.html

[4] 佐々木明，村田真樹 他：論文アブストラクトから原子分子の状態の情報を検出, 抽出する方法の研究, Journal of Plasma and Fusion Research, Vol.81, No.9 (2005).

[5] M. F. Porter : An algorithm for suffix stripping, Program, vol.14, No.3, pp.130-137 (1980).

[6] SWISH::Stemmer http://search.cpan.org/dist/SWISH-Stemmer/

[7] 長倉三郎 他（編）: 岩波理化学辞典第５版 CD-ROM 版，岩波書店（1999）

[8] Salton, G. and McGill, M.J. : Introduction to Modern Information Retrieval, McGraw-Hill Book Company, (1983).

[9] Y. Itikawa, ADNDT 80, 117 (2002)

[10] Hiroe Kashiwagi, Chiemi Watanabe, Akira Sasaki and Kazuki Joe : Text Classification for Constructing an Atomic and Molecular Journal Database by LVQ, International Conference on Parallel and Distributed Processing Techniques and Applications, Vol. II, pp.481-487 (2005).

[11] 柏木裕恵，渡辺知恵美，佐々木明，城和貴：Learning Vector Quantization (LVQ) によるテキスト分類の試み, IPSJ Symposium Series Vol.2004, No.12 pp.103-106 (2004).