アブストラクトを用いた原子分子物理学分野の論文分類支援システムの設計と実装

柏木裕恵[†,☆] 高田雅美[†]
佐々木明[††] 城和貴[†]

原子・分子と電子の衝突による電離・励起の断面積データは様々な分野で活用され，原子分子物理学分野の論文に記載されている．ゆえに，これらのデータを利用するためには原子分子物理学分野の数多くの論文から情報を収集し，目的のデータが掲載されている論文を探し出す必要がある．論文は一般にオンラインジャーナルより入手するが，論文の閲覧が有料である等の問題から原子分子物理学分野のすべての論文を入手することは実質不可能である．そこで本論文では無料で入手でき，誰でも閲覧可能であるアブストラクトを用いて，データが掲載されている論文を探し出す支援をするシステムを開発する．従来，論文分類手法として多く用いられている機械学習法を本システムにも適用し，アブストラクトだけを用いて論文を分類することが可能であることを検証する．

Design and Implementation of an Atomic and Molecular Physics Paper Classification Supporting System Using Abstracts

Hiroe Kashiwagi,[†,☆] Masami Takata,[†] Akira Sasaki[††]
and Kazuki Joe[†]

The cross section data of ionization and excitation by collision among atoms, mulecules and electrons is very useful in various research fields. We can obtain such data out of online jounals of atomic and molecular physics research as digital papers. To obtain the data, we need to classify online papers. However, the classification of the online papers is difficult because there are just abstracts available for free on web pages in general. In this paper, we design a paper classification supporting system to find papers including the data, using just abstracts. We apply a machine learning technique, which is a conventional text classification method, to the system in order to prove that our system using just abstracts is efficient.

1. はじめに

原子分子物理学分野の論文の中には，原子・分子と電子の衝突による電離・励起の断面積データ（以下，原子分子データと表記）が掲載されているものがある．原子分子データは様々な基礎研究や産業応用における重要な基礎データとして活用されている[1)]が，原子分子データを利用するためには，原子分子物理学分野の数多くの論文から情報を収集し，原子分子データが掲載されている論文を探し出す必要がある．論文はオンラインジャーナルより入手することが一般的となっている．原子分子物理学分野において，原子分子データが発表されるジャーナルは Phys. Rev. A 誌[2)]をはじめ 20 種類程度であり，その論文総数は 10^4 件/年のオーダであるのに対して，原子分子データが掲載されている論文の数は 100 件/年程度である．これは，毎年 100 件の論文を探索するために，10^4 件の論文を人間が読んで必要な論文であるか否かを判断しなければならないということを意味する．この作業は大変な労力や手間を必要とするため，機械による認識が必要である．また，オンラインジャーナルより論文を入手する場合，論文の閲覧が有料であったりダウンロード可能な論文数が制限されていたりすることから，原子分子物理学分野のすべての論文を入手することは実質不可能である．しかし，論文を分類するためには論文の内容に関する情報が必要である．そこで，論文の概要が書かれているアブストラクトに注目する．アブスト

† 奈良女子大学大学院人間文化研究科
Graduate School of Humanities and Sciences, Nara Women's University
†† 日本原子力研究開発機構量子ビーム応用研究部門
Japan Atomic Energy Agency, Quantum Beam Science Directorate
☆ 現在，三菱電機株式会社
Presently with Mitsubishi Electric Corporation

ラクトは無料で入手でき誰でも閲覧可能である．実際これまでに，論文のアブストラクトを用いて論文の分類を行う試み[3]があり，有効であることが示されている．そこで，本論文ではアブストラクトのみを用いて原子分子物理学分野の論文の中から原子分子データが掲載されている論文を探し出すための新しいシステムを開発する．我々はこれまでにも原子分子物理学分野の論文アブストラクトを少数用いた論文分類支援システムの開発を行い，試行錯誤を重ねてきている[4),5)]．

機械による論文の分類はテキスト分類[6)]の研究分野に属している．テキスト分類とは，論文や電子メール等のテキストドキュメント（以下，テキストと表記）をあらかじめ決められた2つ以上のカテゴリに分類する処理のことをいい，情報検索や自然言語の分野において非常に注目されてきた重要課題である．分類技術としては，1990年代より機械学習による手法が主流となっている．これは大量のテキストデータが利用可能になったことやコンピュータの性能が大幅に向上したことによるもので，これまでにテキスト分類に対する非常に多くの学習法が提案されている[7)]．代表的なものとして，Naive Bayes[8)]，決定木[9)]，ブースティング[10)]やサポートベクタマシン[11)]を適用した例があり，それらの有効性が示されている．また，近年ではカテゴリの重複を許してテキストを分類するためのモデル[12)]も提案されている．しかし，原子分子物理学分野の論文には化学式等の原子分子に関する特異な表現☆が含まれており，この特異表現（以下，化学式と略記）を一般的な単語と同じように機械に認識させることが容易でない．よって，原子分子物理学分野の論文分類において機械学習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に対して機械学習法を適用し，論文の分類を試みる．本研究は，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[13)]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である．

以下2章で論文分類支援システムについて説明し，3章においてシステムの評価方法について述べる．4章でシステムの評価を行い，5章でまとめる．

2. 論文分類支援システム

本研究において開発する論文の分類支援システムの目的は，人間がより効率的に原子分子データ掲載論文を探索できるようにすることである．たとえば，1,000件のアブストラクトを読んで原子分子データ掲載論文

☆ たとえば O^{5+}，$1s^2 2s^2 2p^2$，$^2S_{1/2}$ 等．

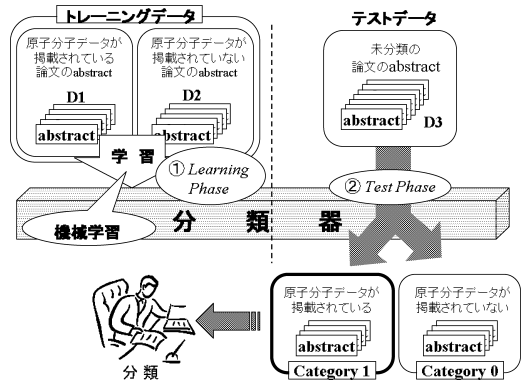


図1 システム概要
Fig. 1 System overview.

10件を探し出さなければならない場合に，100件のアブストラクトを読んで探し出せるようにする．テキストの特徴を多次元のベクトルで表現した特徴ベクトルを用い，機械学習法を使用して分類を行う．システムの構成としては，**図1**のようになっている．以下にシステムの作成手順を示す．

(1) トレーニングデータ（論文本体とアブストラクト），テストデータ（アブストラクトのみ，D3）を入手する．トレーニングデータは論文の内容から原子分子データが掲載されているもの（D1）と掲載されていないもの（D2）に分類する．D1をカテゴリ1，D2をカテゴリ0に割り当てる．

(2) D1，D2，D3の各アブストラクトに前処理を施し，単語や用語の出現頻度情報を使用した特徴ベクトルを作成する．

(3) (2)で作成されたD1，D2の特徴ベクトルを使用し機械学習を行う．

(4) (3)で作成された分類器にD3の特徴ベクトルを適用し，原子分子データが掲載されている論文のアブストラクトであるかを判断する．

以下2.1節，2.2節で(2)の前処理および特徴ベクトル作成について説明する．2.3節において本論文で用いる学習方法について述べる．

2.1 前 処 理

2.1.1 化 学 式

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．化学式には空白が含まれている場合や著者によって書き方が異なる場合があるため，論文から単語を抽出する際に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．このような問題を避けるため，NICT原子分子重要表現抽出システム[14)]を用いて化学式に前処理を施す．NICT

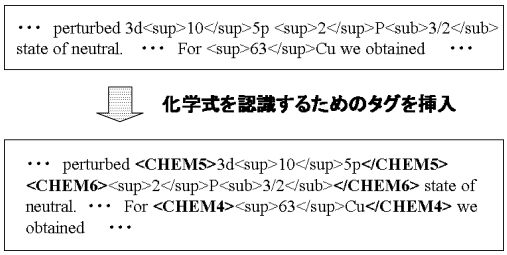


図 2 化学式へのタグの挿入
Fig. 2 Inserting tags to chemical symbols.

原子分子重要表現抽出システムは，アブストラクトに掲載されている化学式の部分を色付きで表示させる HTML ファイルを作成するためのシステムである．我々は NICT 原子分子重要表現抽出システムの出力結果を利用して**図 2** のように化学式の部分に化学式であることを表すタグを挿入し，これらのタグによって化学式の機械的な認識を可能とする．

また，化学式を次のように分類する．

CHEM1 原子 （e.g., *Li*, *hydrogen*）
CHEM2 イオン種 （e.g., *Xe II*, O^{5+}）
CHEM3 分子 （e.g., H_2O）
CHEM4 原子核 （e.g., ^3He, ^{63}Cu）
CHEM5 電子配置（e.g., $1s^22s^22p^2$）
CHEM6 スペクトル項（e.g., $^1S_0$, $^2S_{1/2}$）
CHEM7 数式 （e.g., *l=0*, *n=0*）
CHEM8 CHEM5 + 数 + l （e.g., $2p^43snl$）

2.1.2 ストップワード除去・ステミング処理

論文に含まれている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．

ストップワードとは冠詞，前置詞，接続詞等を指し，あらゆる文章に含まれている．したがって，論文の特徴を表す単語としての重要度は低いと考えられるため，文章から除去する．

ステミングとは単語の語幹を解釈する手法であり，この処理を行うことにより語幹の様々な変化形とマッチングさせることが可能となる．本研究では最も広く利用されている有名な Porter stemming algorithm [15] を使用した Perl モジュール[16] を用いてステミングを行う．

2.2 特徴ベクトル

テキストの分類に機械学習法を適用する際には通常，特徴ベクトルを用いるが，特徴ベクトルによりシステムの性能は大きく左右される．ゆえに，特徴ベクトルを作成する作業は機械学習法を使用するにあたって最も重要な作業である．特徴ベクトルの各要素はテキストに含まれている各単語やキーワードが出現するか否かという 2 値，あるいは，出現頻度を用いて重み付けを行い実数値で表す場合がある．本研究では単語や用語の出現頻度を用いて特徴ベクトルを作成する．さらに，原子分子物理学分野の論文には一般の論文には含まれていない化学式が含まれているため，化学式も論文の特徴を表す重要な表現であると考え，CHEM1〜CHEM8 の 8 種類に分類した化学式の出現頻度を特徴ベクトルの要素として加えることにする．したがって，本研究では以下の 6 種類の出現頻度を用いて特徴ベクトルを作成する．

$F_{(D1)}$	D1 のアブストラクトに含まれる全単語の出現頻度
$F_{(D1)+Chem}$	$F_{(D1)}$ + 化学式の出現頻度
$F_{(D1+D2)}$	D1，D2 のアブストラクトに含まれる全単語の出現頻度
$F_{(D1+D2)+Chem}$	$F_{(D1+D2)}$ + 化学式の出現頻度
$F_{(Dic)}$	専門用語辞典に掲載されている見出し語の出現頻度
$F_{(Dic)+Chem}$	$F_{(Dic)}$ + 化学式の出現頻度

我々は論文分類の対象として原子分子物理学分野の論文を扱うため，$F_{(Dic)}$ の専門用語辞典は物理・化学分野中心の用語が含まれている理化学辞典[17] を用いる．特徴ベクトル $F_{(D1)+Chem}$，$F_{(D1+D2)+Chem}$，$F_{(Dic)+Chem}$ は特徴ベクトル $F_{(D1)}$，$F_{(D1+D2)}$，$F_{(Dic)}$ にそれぞれ 8 種類の化学式を加えて作成するベクトルであるので，($F_{(D1)+Chem}$ の要素数) = ($F_{(D1)}$ の要素数) + 8 となる．

特徴ベクトル $F_{(D1)}$ 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り，要素数も一定にならない．これは特徴ベクトル $F_{(D1)+Chem}$，$F_{(D1+D2)}$，$F_{(D1+D2)+Chem}$ についてもいえることである．特徴ベクトル $F_{(Dic)}$，$F_{(Dic)+Chem}$ で使用する理化学辞典には見出し語として 23,500 語が掲載されており，本システムではそれらをすべて使用する．よって，特徴ベクトル $F_{(Dic)}$，$F_{(Dic)+Chem}$ はそれぞれ 23,500 次元，23,508 次元のベクトルとなる．

我々は 6 種類の単語・用語の出現頻度と TF・IDF 法[18] を利用して特徴ベクトルを作成し，本システムの評価を行う．

TF・IDF 法は単語の出現頻度から文書内の単語の重要性を測る方法であり，各単語の重み付けに一般的に用いられる技法である．TF（Term Frequency）法は単語の頻度をもとに重み付けする技法であり，式 (1) のようにある文書中に含まれる単語ごとの頻度で表すものである．本論文では式 (2) で表される各単語の出現頻度を文書中の全単語の出現数で割った相対頻

度を重みとして採用する．IDF（Inverse Document Frequency）法は単語の特定性をもとに重み付けする技法であり，ある単語が全文書中のどれくらいの文書に出現するかを表すもので，式 (3) によって定義される．これら 2 つを掛け合わせる技法が TF・IDF 法である．

$$w_t^d = tf\,(t,d) \tag{1}$$

$$w_t^d = \frac{tf\,(t,d)}{\sum_{i\in d} tf\,(i,d)} \tag{2}$$

$$idf\,(t) = \log\frac{N}{df\,(t)} + 1 \tag{3}$$

式 (1)，式 (2) の w_t^d は文書 d における索引語 t の重み，式 (2) にある i は文書に出現する索引語を表す．式 (3) において，N は対象となる文書集合に含まれている全文書数，$df\,(t)$ は索引語 t が出現する文書数を示す．

2.3 学習方法

我々は今回，システムの学習方法として Learning Vector Quantization（LVQ）[19] を採用する．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，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．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ため，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セットより作成する．カテゴリごとにランダムに 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学習を行い，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．

3. 評価方法

3.1 データセット

システムの評価に使用するデータセットは，市川リスト☆に記載されている原子分子データ掲載論文 379 件[20] のうち，ジャーナル Phys. Rev. A 誌[2] に掲載されている 126 件をカテゴリ 1 として用いる．ジャーナル Phys. Rev. 誌は物理学者の業界では最もメジャーな論文誌で，A〜E の 5 つの分野に分かれており，原子分子物理学分野は A に属している☆☆．Phys. Rev. A 誌（Vol.41〜62）に掲載されているすべ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すべてチェックしたもので，データ数は 15,944 件である．

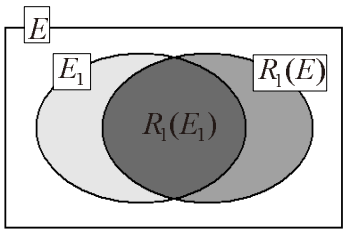


図 3 再現率と適合率
Fig. 3 Recall and precision rates.

3.2 評価尺度

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，適合率を求める．認識率とはテストデータが正しいカテゴリに分類された率をいう．再現率（Recall rate）は探索対象である（カテゴリ 1 に属している）と認識されるべきデータが正しく分類された率であり，適合率（Precision rate）はカテゴリ 1 に属していると認識されたデータに対する本来のカテゴリ 1 のデータの含有率を示す．なお，本論文では原子分子データ掲載論文をカテゴリ 1 としている．

E ：全テストデータセット
E_i ：カテゴリ i のテストデータセット
$N(X)$ ：データセット X の要素数
$R_1(X)$ ：データセット X においてカテゴリ 1 であると認識されたデータセット
$R_1^T(X)$ ：データセット X においてカテゴリ 1 であると認識されたカテゴリ 1 のデータセット
$R_1^F(X)$ ：データセット X においてカテゴリ 1 であると認識されたカテゴリ 0 のデータセット

以上の記号を用いると，

$$N\,(E) = N\,(E_1) + N\,(E_0)$$

$$R_1\,(E) = R_1^T\,(E) + R_1^F\,(E) = R_1\,(E_0) + R_1\,(E_1)$$

が得られ，再現率，適合率は，**図 3** を用いて次のように定義できる．

$$Recall = \frac{N\left(R_1^T\,(E)\right)}{N\,(E_1)} = \frac{N\,(R_1\,(E_1))}{N\,(E_1)}$$

☆ 原子分子物理の専門家が原子分子データを収集し，データベース化することを目的として，1995〜1999 年の間に原子分子物理の主要な学術雑誌中のデータが記載されている論文のリストを作成したもの．

☆☆ Phys. Rev. A には他に物理光学分野の論文が含まれている．

$$Precision = \frac{N\left(R_1^T(E)\right)}{N(R_1(E))} = \frac{N(R_1(E_1))}{N(R_1(E))}$$

3.3 再現率と適合率

再現率と適合率はトレードオフの関係にある．再現率の向上を図るとカテゴリ 1 であると認識されるデータが増えるため，一般に適合率は低下する．一方，適合率の向上を目指すとカテゴリ 1 から省かれるデータが多くなるために，再現率が低下する．したがって，どちらを重要とするかは評価を行うシステムの目的を考慮して決定する必要がある．我々のシステムでは，原子分子データ掲載論文のアブストラクトを正しいカテゴリに分類することよりももれなく収集することが重要であるため，再現率を重視する．

4. 性 能 評 価

4.1 特徴ベクトル作成方法の違いによる比較実験

3.1 節で述べたデータセットのうち，600 件のアブストラクトを使用して実験を行う．600 件のデータにはカテゴリ 1 に属する 126 件を含み，カテゴリ 0 に属するものに関してはランダムに選ばれた 474 件からなっているとする．カテゴリ 1：カテゴリ 0 = 63：237 とするトレーニングデータセットとテストデータセットを各 100 セット用意し，それらを用いる際の認識率，再現率，適合率の平均を比較する．参照ベクトル数は 200 とし，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．

4.1.1 文書集合の違いによる比較

2.2 節で述べた IDF 法を表す式 (3) には対象となる文書集合に含まれる全文書数 N が含まれている．この文書集合を P とする．一般に情報検索分野で用いられる場合には P は検索対象となる文書全体の集合をさしているが，本論文ではトレーニングデータセットとテストデータセットという 2 つの文書集合を使用しているために，様々な文書集合の組合せが考えられる．特徴ベクトル要素の値は計算に用いる文書集合に依存するため，どの文書集合を選択するかということは本システムの評価に関わる重要な課題である．文書集合の組合せは，2 つのカテゴリが混在している集合のみでなっている集合と混在していない集合が含まれる集合に分けて考えることができる．TF・IDF 法によって他の文書との関連性を示す部分である文書頻度を計算する場合，カテゴリが混在している集合においては，カテゴリ 1 に属しているアブストラクトの数が少ないことからカテゴリ 1 のアブストラクト内の語が重要であると認識される率が高いが，混在していない集合では，カテゴリ 1 のアブストラクトに掲載されている重要な語が一般的な語として認識されてしまうおそれがある．ゆえに，カテゴリが混在している集合のみを使用して作成された特徴ベクトルの方が優れていると仮定できる．紙面の都合上，実験結果は掲載しないが，過去に我々が行った実験の結果よりこの仮定は正しいことが証明されている．

また，本システムを実用化する場合，未分類のデータに適用する前にトレーニングデータによって学習を行っておく必要がある．したがって，トレーニングデータの特徴ベクトルを作成する際に適用する未分類のデータを含む文書集合を用いるのは実用的ではない．よって，本論文における実験ではシステムを実用化する場合を考慮し，トレーニングデータの特徴ベクトルを作成する場合にはトレーニングデータセットに含まれる全文書数を使用し，テストデータの特徴ベクトルを作成する際にはテストデータセットに含まれる全文書数を用いる．

4.1.2 単語・用語の出現頻度の違いによる比較

特徴ベクトルを作成する際に使用する単語・用語の出現頻度の違いにより認識率，再現率，適合率を比較する．**図 4**，**図 5**，**図 6** は 4.1 節に記述したデータセットを用いた場合に得られた認識率，再現率，適合率の平均値を並べたグラフである．

図 4，図 5，図 6 より，どの特徴ベクトルを用いた場合にも再現率が 90%以上の値になっていることが確認でき，認識率，適合率に関しては特徴ベクトル $F_{(D1)}$，$F_{(D1+D2)}$ を使用した場合に高い値が出ていることが認められる．その一方で，特徴ベクトル $F_{(D1)}$，$F_{(D1+D2)}$ には 2.2 節で述べたとおり要素数が一定でないという問題がある．要素数が一定でないこと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，理化学辞典の見出し語をもとに作成された特徴ベクトルは，理化学辞典に掲載されている見出し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．本実験で使用した特徴ベクトル $F_{(Dic)}$，$F_{(Dic)+Chem}$ の要素数は 1,182，1,190 である．これは，全アブストラクトに掲載されていない理化学辞典の見出し語に属する要素は本実験で分類を行う際には何の意味もなさないことから，特徴ベクトルより省いたためである☆．特徴ベクトル $F_{(Dic)}$ を用いた際の実験結果を確認すると，適合率が若干低い値ではあるが，本研究で重視している再現率は特徴ベクトル $F_{(D1)}$，$F_{(D1+D2)}$ と比

☆ 実用化する際には，全アブストラクトに掲載されていない理化学辞典の見出し語の数は把握できないため，全 23,500 語に属する要素を用いる．

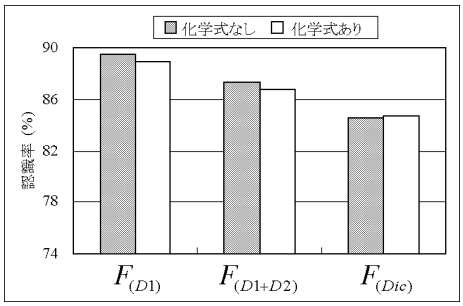


図 4 実験 4.1 における認識率
Fig. 4 The recognition rates in experiment 4.1.

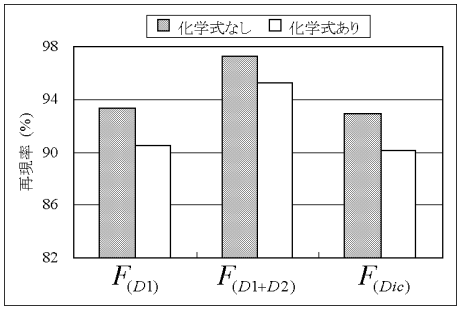


図 5 実験 4.1 における再現率
Fig. 5 The recall rates in experiment 4.1.

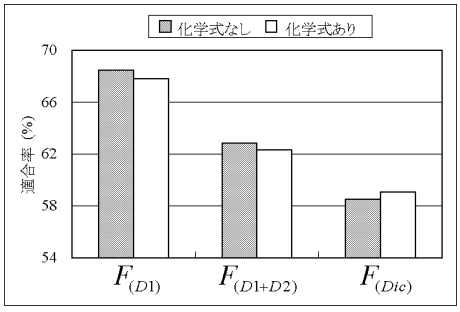


図 6 実験 4.1 における適合率
Fig. 6 The precision rates in experiment 4.1.

較して遜色のない結果になっている．よって，特徴ベクトル $F_{(Dic)}$ が最適であると考えられる．

4.1.3 化学式使用の違いによる比較

特徴ベクトル作成時に化学式を使用する場合と使用しない場合での認識率・再現率・適合率の比較を行う．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が，図 4，図 5，図 6 の化学式を使用する場合としない場合での実験結果を比較してみると，再現率に対して化学式が重要な役割を果たしているとはいえない．これはアブストラクトに掲載されている単語や理化学辞典に掲載されている見出し語に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と考えられる．

4.2 参照ベクトル数の違いによる比較実験

3.1 節で述べた 16,070 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を 8,035 件ずつ用い，どちらのデータセットにも 63 件のカテゴリ 1 のアブストラクトを含めるようにする．トレーニングデータ数とテストデータ数を各 300 とした実験 4.1 においては参照ベクトル数を 200 としたが，今回はデータ数がそれぞれ約 8,000 件と多いため，最適な参照ベクトル数を調べる必要がある．そこで，参照ベクトル数を 1,000〜8,000 とし 1,000 ずつ増やして実験を行う．本実験で用いる特徴ベクトル $F_{(Dic)}$，$F_{(Dic)+Chem}$ は全アブストラクトに掲載されていない理化学辞典の見出し語に属する成分を省いたため，それぞれ 3,557 次元，3,565 次元のベクトルになる．

図 7，**図 8**，**図 9** はトレーニングデータセットを 10 セット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．認識率はすべての場合において 95%以上の値になっており，参照ベクトルの数に関係なくほぼ一定の値になっている．再現率については参照ベクトル数が増加するに従ってグラフが一定の状態に達している．適合率は徐々に低くなっているが，再現率が一定状態になっている段階，つまり，参照ベクトル数 3,000〜8,000 のときには大きな差はみられない．特徴ベクトル作成時に化学式を使用する場合としない場合を比較すると，実験 4.1 の結果と同様に化学式を使用しない場合により高い再現率が得られている．この実験結果により，原子分子データが記載されている論文を探索する際の化学式の重要度は高くないと判断できる．本実験ではデータとして Phys. Rev. A 誌に収録されている論文を扱っているが，Phys. Rev. A 誌には原子分子物理学分野だけでなく物理光学分野も含まれている．したがって，化学式は原子分子物理学分野の論文と物理光学分野の論文を分類する際には役に立つのではないかと思われる．

4.3 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

実験 4.2 では，過去に我々が行ってきた研究結果[5]により，LVQ に用いる参照ベクトルのカテゴリの割合をカテゴリ 1：カテゴリ 0 = 1：1 としている．文献 5) 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を得ている．しかし，今回は約 8,000 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

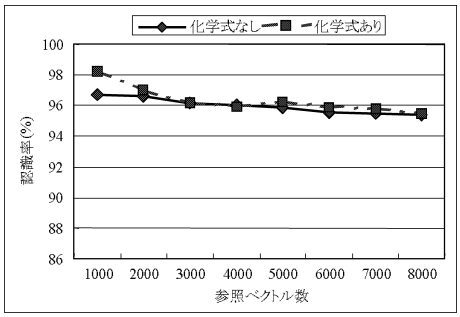


図 7 実験 4.2 における認識率
Fig. 7 The recognition rates in experiment 4.2.

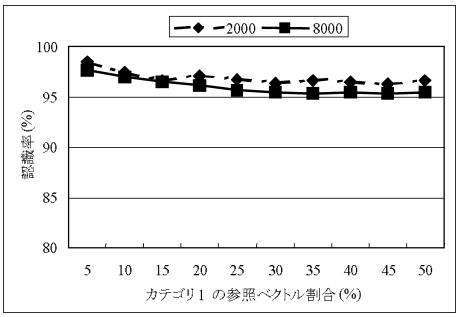


図 10 実験 4.3 における認識率の変化
Fig. 10 The recognition rates in experiment 4.3.

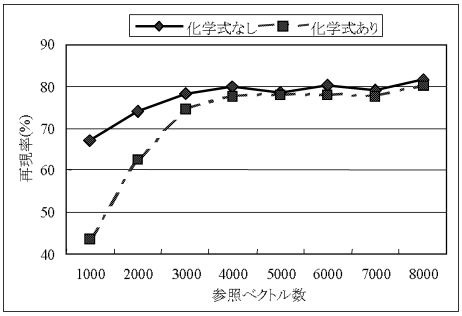


図 8 実験 4.2 における再現率
Fig. 8 The recall rates in experiment 4.2.

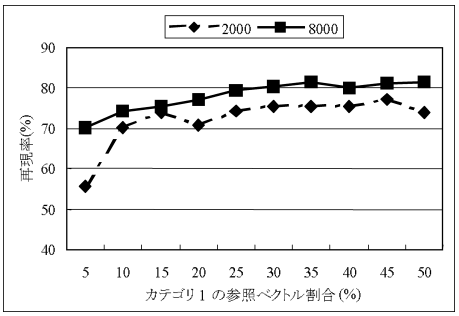


図 11 実験 4.3 における再現率の変化
Fig. 11 The recall rates in experiment 4.3.

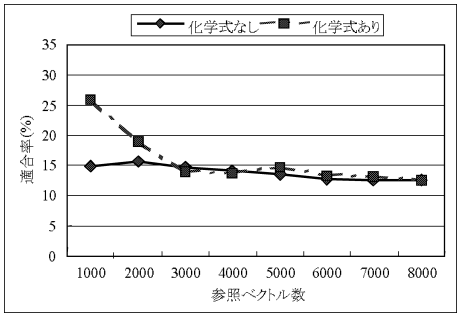


図 9 実験 4.2 における適合率
Fig. 9 The precision rates in experiment 4.2.

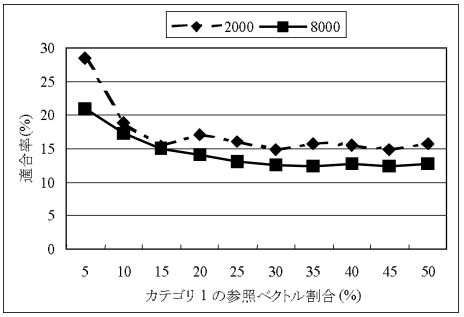


図 12 実験 4.3 における適合率の変化
Fig. 12 The precision rates in experiment 4.3.

い可能性がある．そこで，トレーニングデータセット，テストデータセットともに実験 4.2 で使用したものを用いて，参照ベクトルの数が 2,000 と 8,000 の場合にカテゴリ 1 に属している参照ベクトルの割合を 5〜50%に変化させたときの認識率，再現率，適合率の変化を調べる．**図 10**，**図 11**，**図 12** は結果の平均値をとったものである．

図 10，図 11，図 12 より，参照ベクトルの数が 2,000 と 8,000 のどちらの場合にも，カテゴリ 1 の割合が 5%であるときに適合率が 20%以上であるものの再現率が 55.56%，70.16%であり，実験 4.2 と比べると 11〜19%低下している．また，参照ベクトル数が 8,000 の場合において再現率のグラフ全体からは 50%に近づくにつれて一定になっていく様子が確認でき，参照ベクトル数 2,000 の際の再現率についてもグラフが一定状態になっているといえる．結果として，参照ベクトル数が 8,000 でカテゴリ 1 の参照ベクトルの割合が 50%であるときに最高の再現率になっており，最適な割合であると判断できる．このとき，認識率は 95.42%，再現率は 81.59%，適合率は 12.66%になっているが，参

function	950539.00
transfer	193344.00
transfer rate	135309.44
two parameter	19019.42
phenomenological model	16006.46
proton	3750.00

図 13 TermExtract の出力ファイル
Fig. 13 Output file from TermExtract.

照ベクトルの数が 8,000 でカテゴリ 1 の参照ベクトルの割合が 50%であるということは，カテゴリ 1 に属している 63 件のアブストラクトを 4,000 の参照ベクトルによって探し出すということである．これは非効率的な方法ではあるが，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ことが原因であると思われる．したがって，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問題であると考えている．

4.4 人間による論文分類の模倣実験

人間が論文を分類する場合には，必要としている論文に含まれていると推測されるいくつかのキーワードが論文に含まれているか否かで必要な論文を判断していると思われる．同様に，アブストラクトによる分類の際にも，アブストラクトに含まれるいくつかのキーワードによって必要な論文を判断すると考えられる．そこで我々は，アブストラクトのみを用いる場合の，機械学習による論文分類と人間による論文分類を比較するために，いくつかのキーワードにスコアをつけ，それらを合計して算出された各アブストラクトのスコアによって論文を分類する方法を試みる．この方法により，アブストラクトによる論文分類時に人間が行っている判断を模倣できると考えられる．

キーワードの抽出において，我々はテキストデータから専門用語を取り出すための Perl モジュール “TermExtract” [21] を使用する．TermExtract は東京大学中川研究室・横浜国立大学森研究室で開発された用語抽出システム[22] を全面的に組みなおしたもので，TermExtract にテキストデータを適用すると，専門用語とその重要度が**図 13** のような形で出力される．TermExtract には学習機能があり，この機能はそれまでに処理対象としたテキストからの情報を蓄積し，スコアを計算する際に用いるものである．我々は英文の形態素解析ソフトとして “Brill’s Tagger” [23] を使用し，機械学習法を用いる場合との比較を行うために 4.2 節で使用した各 10 セットの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う．ト

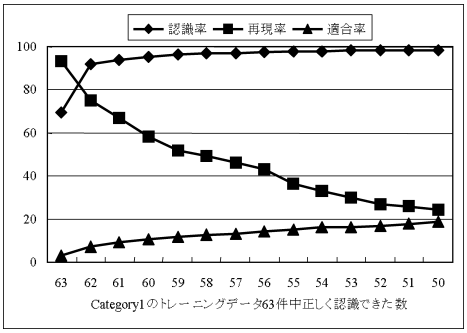


図 14 “TermExtract” を用いる場合のテストデータの認識率・再現率・適合率
Fig. 14 Recognition, Recall and Precision rates with “TermExtract”.

レーニングデータセットを使って TermExtract の学習機能により情報を蓄積させ，その情報を用いて，再度 TermExtract によりトレーニングデータとテストデータに含まれるキーワードのスコアを計算する．各アブストラクトのスコアはカテゴリ 1 のトレーニングデータに含まれているキーワードのスコアの自然対数を合計して算出する．

カテゴリ 1 とカテゴリ 0 の境界値はトレーニングデータの再現率が 80〜100%になるように設定する．具体的には，カテゴリ 1 のトレーニングデータ 63 件のうち，63 件を正しくカテゴリ 1 であると認識できる境界値，62 件を正しく認識できる境界値，というように境界値を設定していくものとし，トレーニングデータの再現率が 80〜100%になる際のテストデータの認識率，再現率，適合率を調べる．トレーニングデータ，テストデータを各 10 セット使用して得られた実験結果の平均値を**図 14** のグラフに示す．スコアの算出方法に関して，テストデータのスコアはテストデータセットを使って蓄積させた情報を用いて算出する場合や TermExtract の学習機能を使用せずに算出する場合の実験も行ったが，図 14 とほぼ変わらない結果が出たため省略する．

図 14 より，再現率が 90%を超える値であるときには認識率，適合率ともに低い値になっており，対照的に，認識率が 90%を超える値である場合や適合率が 20%近くになる際の再現率は 80%をきる結果となっている．これは 1 つの境界値のみを用いてカテゴリを分けたことが原因であると思われる．この実験結果により，カテゴリ 1 に属しているトレーニングデータのアブストラクトに掲載されている語が多く含まれていることとカテゴリ 1 に属していることとが，必ずしも必要条件を満たすわけではないということが確認できる．カテゴリを正しく認識させるためにはより複雑な

分類構造が必要であると考えられるため，本実験によりアブストラクトのみで論文を分類する際には，人間が分類するよりも機械学習が有効であることを示すことができる．

4.5 特徴ベクトルの次元数

我々は各種実験を行ってきたが，ここで，特徴ベクトル $F_{(Dic)}$ のベクトルデータについての考察を行う．16,070 件すべてのデータを使用した際の実験に用いた各アブストラクトの特徴ベクトルは 3,557 次元であるが，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 902 ある．これは分類には何の意味もなさない成分であるので，特徴ベクトルは明らかに 2,655 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り，このような成分は，他の成分との相関が極端に小さくなるため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ことから，次元数を縮約することは効果があると考えられる．

しかし，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，本論文では次元数の縮約を行っていない．次元数の縮約に関しては今後検討していく予定である．

5. ま と め

本論文では，アブストラクトだけを用いて原子分子物理学分野の論文を分類することが可能であることを検証した．論文分類のための方法として LVQ を採用した結果，以下のことが分かった．専門用語辞典に掲載されている見出し語の出現頻度をもとに作成した特徴ベクトルを使用した際に，認識率 95.42%，再現率 81.59%，適合率 12.66%という良好な結果を得ることができた．この結果は，10,000 件の論文のうち必要な論文が 78 件しかない文書集合から必要である論文を探索する場合に，今までは人間が 10,000 件の論文を読んで探し出していた作業を，LVQ によるシステムを用いる場合には 503 件の論文を読んで 64 件の論文を探し出す作業に置き換えることが可能であることを意味する．本システムにより人間は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ことを立証できた．

本論文では原子分子物理学分野の論文の分類に理化学辞典を使い，優秀な分類結果を得ることができた．よって，他の分野の論文を分類したい場合には，その分野に合った専門用語辞典を使用すればシステムの有効性が得られると思われ，様々な分野の論文を用いてシステムの評価を行っていくことが今後の課題である．また，本論文ではシステムに用いる機械学習法として LVQ を採用し優れた性能を確保できたが，今後は他の機械学習法を用いてシステムを評価していくことも課題の 1 つである．

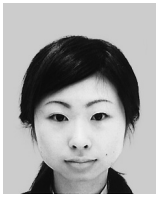
謝辞 特徴ベクトルデータの分析にご協力いただきましたお茶の水女子大学吉田裕亮教授に心より感謝いたします．

参 考 文 献

1) 加藤隆子ほか：プラズマ原子・分子過程の展望，プラズマ・核融合学会誌，Vol.75, No.10, p.1124 (1999).
2) *APS physics Physical Review A*. http://pra.aps.org/
3) Aizawa, A.: The Feature Quantity: An Information Theoretic Perspective of Tfidf-like Measures, *Proc. ACM SIGIR 2000*, pp.104–111 (2000).
4) Kashiwagi, H., Watanabe, C., Sasaki, A. and Joe, K.: 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).
5) 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).
6) 永田昌明，平 博順：テキスト分類—学習理論の「見本市」，情報処理学会誌，Vol.42, No.1, pp.33–37 (2000).
7) Sheng, G., Wen, W., Chin-Hui, L. and Tat-Seng, C.: Maximal Figure-of-Merit Learning Approach to Text Categorization, *ACM SIGIR*, pp.174–181 (2003).
8) Lewis, D.D.: Naive (Bayes) at Forty: Independence Assumption in Information Retrieval, *Proc. 10th European Conference on Machine Learning* (*ECML-98*), pp.4–15 (1998).
9) Lewis, D.D. and Ringuette, M.: A comparison of two learning algorithms for text categorization, *The 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp.81–93 (1994).
10) 平 博順，春野雅彦：トランスダクティブ・ブースティング法によるテキスト分類，情報処理学会論文誌，Vol.43, No.6, pp.1843–1851 (2002).
11) 平 博順，春野雅彦：Support Vector Machine

によるテキスト分類における属性選択，情報処理学会論文誌，Vol.41, No.4, pp.1113–1123 (2000).
12) 上田修功，斉藤和巳：類似テキスト検索のための多重トピックテキストモデル，情報処理学会論文誌，Vol.44, No.SIG14, pp.1–8 (2003).
13) *Atomic and Molecular Data Research Center, NIFS*. http://dpc.nifs.ac.jp/amdrc/index-j.html
14) 佐々木明，村田真樹ほか：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, pp.717–722 (2005).
15) Porter, M.: An algorithm for suffix stripping, *Program*, Vol.14, No.3, pp.130–137 (1980).
16) *SWISH::Stemmer*. http://search.cpan.org/dist/SWISH-Stemmer/
17) 長倉三郎ほか（編）：岩波 理化学辞典 CD-ROM 版，5th edition，岩波書店 (1999).
18) Salton, G. and McGill, M.: *Introduction to Modern Information Retrieval*, McGraw-Hill Book Company (1983).
19) *HUT – CIS – Research – SOM_PAK, LVQ_PAK*. http://www.cis.hut.fi/research/som-research/nnrc-programs.shtml
20) Itikawa, Y.: ANNOTES BIBLIOGRAPHY ON COLLISIONS WITH ATOMIC POSITIVE IONS: EXCITATION AND IONIZATION, 1995–1999, *Atomic Data and Nuclear Data Tables*, Vol.80, No.1, pp.117–146 (2002).
21) *TermExtract*. http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html
22) 中川裕志，森　辰則，湯本紘彰：出現頻度と連接頻度に基づく専門用語抽出，自然言語処理学会論文誌，Vol.10, No.1, pp.27–45 (2003).
23) *Brill's Tagger*. http://research.microsoft.com/~brill/

(平成 19 年 2 月 2 日受付)
(平成 19 年 3 月 23 日再受付)
(平成 19 年 4 月 4 日採録)



柏木　裕恵（正会員）

奈良女子大学理学部情報科学科卒業，2007 年同大学大学院人間文化研究科情報科学専攻博士前期課程修了，現在，三菱電機株式会社に勤務.

高田　雅美（正会員）

1977 年生．2004 年奈良女子大学大学院人間文化研究科複合領域科学専攻修了．博士（理学）を同大学より取得．2004 年独立行政法人科学技術振興機構戦略的創造研究推進事業において，京都大学大学院情報学研究科にて委嘱研究員．2006 年奈良女子大学大学院人間文化研究科助手．2007 年奈良女子大学大学院人間文化研究科助教．数値計算ライブラリの開発，分散メモリ環境を対象とする並列プログラムの開発に関する研究に従事.

佐々木　明

電気通信大学電気通信学部卒業．1989 年電気通信大学新形レーザー研究センター助手，1991 年工学博士（東京工業大学），1996 年日本原子力研究所関西研究所光量子科学研究センター研究員，2000 年同副主任研究員，2005 年日本原子力研究開発機構量子ビーム応用研究部門研究副主幹，現在に至る．専門はプラズマ物理学，原子分子データベース.

城　　和貴（正会員）

大阪大学理学部数学科卒業．日本 DEC，ATR 視聴覚研究所（日本 DEC より出向），（株）クボタ・コンピュータ事業推進室で勤務の後，1993 年奈良先端科学技術大学院大学情報科学研究科博士前期課程入学，1996 年同研究科後期課程修了，同年同研究科助手．1997 年和歌山大学システム工学部講師，1998 年同助教授．1999 年奈良女子大学理学部情報科学科教授，現在に至る，博士（工学博士）．情報処理学会論文誌「数理モデル化と応用」編集委員長.