奈良女子大学大学院
修士論文

オンライン論文検索のための
アブストラクトを用いた
論文分類支援モデルの設計と実装

柏木　裕恵

人間文化研究科 博士前期課程
情報科学専攻　　I05-002

指導教官　：　城　和貴

2007年1月

概 要

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いが， Web 上にある大量のアブストラクトを人間がすべて読んで必要であるかを判断することは実質不可能である．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．そこで本修士論文では，アブストラクトを用いた論文分類支援モデルを提案する．分類手法としてこれまでに多くのテキスト分類の研究で用いられている機械学習法を採用する．

提案する論文の分類支援モデルの構成は以下のようになる．まず，探索対象であるか否かが既知であるトレーニングデータと未知であるテストデータを用意し，アブストラク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作成する．トレーニングデータの特徴ベクトルを使用して機械学習を行い，学習結果をテストデータに適用し，探索対象の論文のアブストラクトであるかを判断する．このシステムにより人間はシステムが探索対象であると判断した論文のみを読んで必要な論文を入手することが可能になるため，論文を読む労力を軽減することができる．

提案したモデルが有効であることを示すために，原子分子物理学分野の論文をアブストラクトのみで分類するシステムを開発する．ここで，提案モデルを原子分子物理学分野の論文へ適用する際に留意すべき点について述べる．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（O^{5+}, ${}^2S_{1/2}$ etc.）が含まれている．本修士論文において，この特異表現を化学式と略記する．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，化学式が書かれている部分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はない．そのため，論文を機械的に分類することは困難である．本研究では化学式に対して前処理を施すことにより，化学式に関する問題を解決する． システムの評価には，認識率，再現率，適合率を使用する．認識率とは，テストデータが正しく分類された率をいう．再現率は，探索対象であ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，適合率は，探索対象であると認識されたデータに対する本来の探索対象データの含有率を示す．本研究では，再現率がもっとも重要であるとする．特徴ベクトルはシステムの性能に大きく関わってくるため，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．

提案モデルを基に，原子分子物理学分野の論文を原子分子の特性データが掲載されている論文と掲載されていない論文に分類するシステムと，原子分子の特性データが掲載されている論文をさらに分類するシステムを開発する．これら 2 つのシステムの評価より，本モデルの有効性を立証する．

目 次

図 目 次

表 目 次

1　はじめに

近年，論文のオンライン化や検索エンジンの改善により，大量の論文の中からオンラインで論文を入手することが可能になっている．ところがそれに伴い，必要な論文だけを見出すことが困難になっている．これは，多くのオンラインジャーナルにおいて，論文検索時に我々が論文の内容として閲覧できるのは，題名やアブストラクトの部分のみであるため，アブストラクトの内容だけで必要な論文を判断することが難しいことに起因する．また，Web 上に存在する膨大な量のアブストラクトを，人間がすべて読んで必要であるかを判断することが不可能であることも要因のひとつである．これらの問題点を解決するために，機械による認識が必要になってくる．機械による論文の分類は，テキスト分類 [1] の研究分野に属する．現在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おり，その有効性が示されている [2]．しかし，機械学習法を用いた従来の論文分類システムを構築する際には論文そのものが必要であることから，アブストラクトによる論文検索時には役に立たない．そこで本修士論文において，アブストラクトのみを使用して機械学習を行う論文分類支援モデルを提案する．このモデルは，オンライン論文検索の際に人間が読まなければならないアブストラクトの量を減らし，人間にかかる負荷を軽減するためのモデルである．提案したモデルが有効であることを示すために，原子分子物理学分野の論文をアブストラクトのみで分類するシステムを開発する．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 [3] である．提案モデルを基に，原子分子物理学分野の論文を原子分子の特性データが掲載されている論文と掲載されていない論文に分類するシステムと，原子分子の特性データが掲載されている論文をさらに分類するシステムを開発し，分類対象に合う論文の特徴の表現方法を探し出すことにより，システムの有効性を検証する．これら 2 つのシステムの評価から，本モデルの有効性を示す．

以下 2 章にて，これまでのテキスト分類に関する研究や一般に用いられる評価尺度について説明し，3 章にて論文分類支援モデルを提案する．4 章，5 章においてモデルの適用方法とシステムの評価方法について述べ，6 章で論文分類支援モデルの評価を行う．

2　テキスト分類

テキスト分類 [1] 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う．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ある．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．機械学習法では，一般にトレーニングデータを用いて学習を行い，その学習結果を基に未分類のデータ（テストデータ）を分類する．代表的なものとして，Naive Bayes[4]，決定木 [5]，ブースティング [6] やサポートベクターマシン [7] を適用した例があり，それらの有効性が示されている．Naive Bayes は Bayes の規則を利用した分類器，決定木は可読性の高い分類器である．ブースティングは，分類精度の低い分類器を組み合わせることで高精度な分類器を得る方法のひとつであり，サポートベクターマシンは，トレーニングデータを正例と負例に分け，かつ，正負例間のマージンが最大になるように超平面を求める学習器である．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 [8] も提案されている．

テキスト分類器や情報検索モデルの評価には，一般に認識率や再現率，適合率と呼ばれる評価尺度を用いる．認識率とは，テストデータが正しく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，探索対象であると認識されたデータに対する，本来の探索対象データの含有率を示す．

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，

$$N\left(E\right) = N\left(E_1\right) + N\left(E_0\right)$$

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

が得られ，再現率，適合率は図 1 を用いて次のように定義できる．

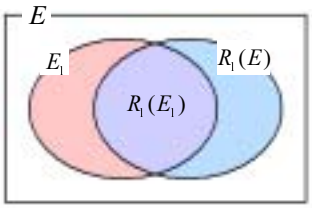

図 1: 再現率と適合率

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

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

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．本論文ではこのベクトルを特徴ベクトルとよぶ．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多い．本研究では単語や用語の出現頻度を用いる他，再現率を向上させるために，重要語との相関関係を用いる特徴ベクトルを作成する．

3　論文分類支援モデルの提案

本研究において提案する論文の分類支援モデルの目的は，人間がより効率的に論文を探索できるようにすることである．例えば，1000 件のアブストラクトを読んで必要な 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．分類手法として機械学習法を用いる．モデルは図 2 のような構成になっている．このモデルを基に開発するシステムにより，人間はシステムが探索対象であると判断した論文のみを読んで必要な論文を入手することが可能になるため，論文を読む労力を軽減することができる．以下にモデルの作成手順を示す．

Step1 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探索対象であるもの（D1）と探索対象でないもの（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．

Step2 D1，D2，D3 の各アブストラクトに前処理を施し，特徴ベクトルを作成する．

Step3 Step2 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．

Step4 Step3 で作成された分類器に D3 の特徴ベクトルを適用し，探索対象の論文のアブストラクトであるかを判断する．

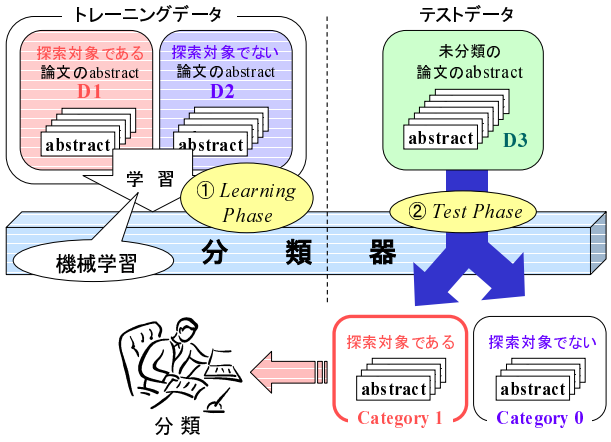

図 2: システム概要

Step2 の前処理として，論文に含まれている単語に対して，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，接続詞等を指す．あらゆる文章に含まれているため，論文の特徴を表す単語と

しての重要度は低い．したがって，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る．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[9] を使用した Perl モジュール [10] を用いてステミングを行う．

4　原子分子物理学分野の論文への適用

3 章で提案したモデルの有効性を検証するため，原子分子物理学分野の論文への適用を行う．

ここで，原子分子物理学分野の論文へ適用する際に留意すべき点について述べる．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[2]が含まれている．本論文において，この特異表現を化学式と略記する．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．よって化学式が書かれている部分を，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はないため，論文を機械的に分類することは困難である．ゆえに原子分子物理学分野の論文分類において，機械学習法を適用してきた例はない．本研究では化学式に対して前処理を施すことにより，化学式に関する問題を解決する．

以下，4.1 節で化学式に対する前処理について述べ，4.2 節で提案モデルを適用する 2 つのシステムについて説明する．4.3 節，4.4 節にて特徴ベクトル，今回用いる学習方法について解説する．

4.1　化学式に対する前処理

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のないように，原子分子物理学分野の論文に含まれる化学式に対して前処理を行う．我々は NICT 原子分子重要表現抽出システム [11] を用いて化学式に前処理を施す．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．我々は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．図 3 はその具体例である．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．

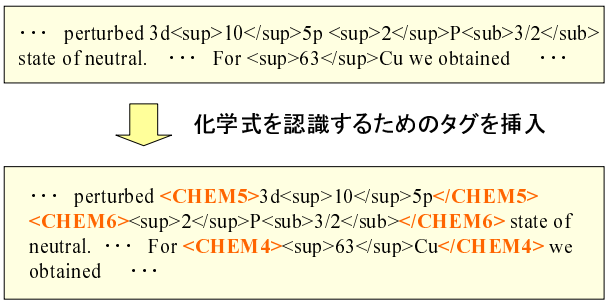

図 3: 化学式へのタグの挿入

[2]例えば O^{5+}，$1s^2 2s^2 2p^2$，$^2S_{1/2}$ etc.

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（$e.g.\ 1s^2 2s^2 2p^2$）

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 （$e.g.\ 2p^4 3snl$）

4.2 2 つの論文分類支援システムの設計

本論文では 3 章で述べた提案モデルを基に，原子分子物理学分野の論文のためのシステムを 2 つ開発する．

原子や分子は，原子番号や質量数，あるいは化学的組成に固有の特性を有している．本論文ではこれらの特性データを原子分子データと表記する．原子分子データは様々な基礎研究や産業への応用における重要な基礎データとして活用されている [12]．大量のデータを扱うために，現在，核融合科学研究所において，原子分子データを検索可能なデータベースを構築する計画がすすめられている [3]．

原子分子データの多くは原子分子物理学分野の論文に記載されている．データベースへ登録する原子分子データは，ジャーナルに掲載されている論文から収集する．これは，ジャーナルに掲載されている論文はもっとも信頼性があると考えられるからである．原子分子物理学分野の論文には，原子分子データを記載していない論文も含まれているため，原子分子データを記載している論文と記載していない論文に分類する必要がある．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 [13] をはじめ 20 種類程度である．その論文総数は 10^4 件/年のオーダーであるのに対して，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，人間が 10^4 件のアブストラクトを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．よって，この作業の機械化が必要である．

核融合科学研究所で構築がすすめられている原子分子データベースには，原子分子データが含まれている論文に記載されている表やグラフからデータ

を抽出し，そのデータを登録している．原子分子データが掲載されている論文の中でも，表やグラフの形式から，データベースへの登録に向いている論文と向いていない論文がある．よって，データベース登録に向いている論文であるか否かを判断することも原子分子データベース構築には必要な作業である．人間がアブストラクトだけを読んで，この分類を行うことは非常に困難であるため，こちらも機械による分類が求められる．

したがって 1 つめのシステムとして，原子分子物理学分野の論文を原子分子データを含む論文と含まない論文に分類するシステムを開発し，これをシステム A とする．2 つめのシステムは，原子分子データを含む論文を原子分子データベースへの登録に向いている論文と向いていない論文の分類システムであるとし，システム B と表記する．システム A は，汎用性のあるシステムである．一方，システム B は原子分子データベースにデータを登録するために特化されたシステムである．ゆえに本論文では，システム A の評価より他分野の論文分類への提案モデルの応用可能性を示し，システム B により分野を限定した場合の有効性を立証する．

4.3 特徴ベクトル

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．この特徴ベクトルはシステムの性能に大きく関わってくるため，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．以下，本研究において使用する特徴ベクトルの作成方法に関して詳述する．4.3.1 項で単語・用語の出現頻度を用いる場合，4.3.2 項でアブストラクトから専門用語辞書を作成して用いる場合について述べる．4.3.3 項にて単語と化学式の組み合わせを使う場合を説明し，4.3.4 項では化学式と文章数の関係を用いる場合を解説する．

4.3.1 単語・用語の出現頻度を用いる場合

一般に各単語や用語の出現頻度として考えられるのは，以下の 3 種類である．

$F_{(D1)}$ D1 のアブストラクトに含まれる全単語の出現頻度

$F_{(D1+D2)}$ D1，D2 のアブストラクトに含まれる全単語の出現頻度

$F_{(Dic)}$ 専門用語辞典に掲載されている見出し語の出現頻度

本論文で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っているため，$F_{(Dic)}$ 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 [14] を用いる．また，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．よって，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．ゆえに，本論文において特徴ベクトルの要素として用いる出現頻度には，先ほど述べた 3 種類の特徴ベクトルと以下の 3 種類を用いる．

$F_{(D1)+Chem}$	$F_{(D1)}$ + 化学式の出現頻度
$F_{(D1+D2)+Chem}$	$F_{(D1+D2)}$ + 化学式の出現頻度
$F_{(Dic)+Chem}$	$F_{(Dic)}$ + 化学式の出現頻度

$F_{(D1)+Chem}$，$F_{(D1+D2)+Chem}$，$F_{(Dic)+Chem}$ 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したベクトルであるので，$\left(F_{(D1)+Chem}\text{の要素数}\right) = \left(F_{(D1)}\text{の要素数}\right) + 8$ となる．

特徴ベクトル $F_{(D1)}$，$F_{(D1)+Chem}$ 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 $F_{(D1+D2)}$，$F_{(D1+D2)+Chem}$ 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 23500 語が掲載されており，提案する分類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，23500 次元，23508 次元のベクトルとなる．

4.3.1.1 TF・IDF 法を使用する場合

各特徴ベクトル $F_{(D1)}$，$F_{(D1)+Chem}$，$F_{(D1+D2)}$，$F_{(D1+D2)+Chem}$，$F_{(Dic)}$，$F_{(Dic)+Chem}$ の要素を，情報検索分野においてよく用いられている TF・IDF 法 [15] を使用して作成する．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毎の頻度で表すものである．本論文では，各単語の出現頻度を文書中の全単語の出現数で割った，式 (2) で表される相対頻度を重みとして採用する．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している．

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

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

IDF(Inverse DocumenTFrequency) 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．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している．

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

TF･IDF 法を用いて作成した特徴ベクトルを，それぞれ $F^{tf}_{(D1)}$，$F^{tf}_{(D1)+Chem}$，$F^{tf}_{(D1+D2)}$，$F^{tf}_{(D1+D2)+Chem}$，$F^{tf}_{(Dic)}$，$F^{tf}_{(Dic)+Chem}$ と表す．

4.3.1.2　文章頻度を使用する場合

TF・IDF 法は文書頻度を使用し，文書集合の中での重要な語というものに留意している．ここで，アブストラクト内の文章構造について考えてみる．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の文章には，類似した語が含まれていると推定される．しかし TF・IDF 法では，アブストラクト内のどの文章にどの語が含まれているかという点やアブストラクトに記載されている文章数に関しては考慮されていない．そこで，ひとつの文章に対する語句の相対頻度とその語句が含まれている文章数を掛け合わせることにより得られる値を使って特徴ベクトルを作成するという新たな方法を試みる．この方法を以下の式で定義する．

$$w^d_t = \sum_{u \in d} \left(\frac{tf\,(t,u)}{\sum\limits_{i \in u} tf\,(i,u)} \right) \times \left(\log \frac{N_s(d)}{sf(t)} + 1 \right) \tag{4}$$

式 (4) において，w^d_t は文書 d における索引語 t の重み，i は文章に出現する索引語を表す．$N_s(d)$ は文書 d に含まれている全文章数，$sf(t)$ は索引語 t が出現する文章数を指している．この方法は他の文書との関係を一切考慮していないため，認識率，再現率，適合率に影響が出ると思われる．しかし，ひとつのアブストラクトのみで特徴ベクトルを作成できるため，データセットに依存しない特徴ベクトルを作成できる．この方法を，4.3.1 項で述べた 6 種の出現頻度に対して適用し，これらの特徴ベクトルを $F^{sf}_{(D1)}$，$F^{sf}_{(D1)+Chem}$，$F^{sf}_{(D1+D2)}$，$F^{sf}_{(D1+D2)+Chem}$，$F^{sf}_{(Dic)}$，$F^{sf}_{(Dic)+Chem}$ と表す．

4.3.2　アブストラクトから専門用語辞書を作成し用いる場合

アブストラクトに掲載されている単語･用語から専門用語辞書を作成し，その辞書を使って特徴ベクトルを作成する．専門用語辞書について，本研究で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”[16] 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 [17] を全面的に組みなおしたものである．英文の形態素解析ソフトとしては，”Brill’s Tagger”[18] を用いる．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 4 のような形で

出力される．また TermExtract には，学習機能がある．この機能は，いままでに処理対象としたテキストからの情報をファイルとして蓄積し，スコアを計算する際に用いるものである．特徴ベクトルの各要素の重みは，カテゴリ 1 のトレーニングデータに含まれている用語のスコアの自然対数をとる．この特徴ベクトルを $F^{log}_{(Special)}$ とする．

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

図 4: TermExtract の出力ファイル

4.3.3 単語と化学式の組み合わせを使用する場合

単語と化学式の組み合わせを使用した特徴ベクトル作成方法を述べる．この特徴ベクトルを $F^{correlation}_{(Word+5Chem)}$ と表記する．4.3.4 項で述べる化学式と文章数の関係を用いる特徴ベクトルと合成して，システム B を評価する際に使用する．システム B で対象とするアブストラクトの数は，システム A で対象とするアブストラクトよりも少ない．よって 4.1 節で述べたように化学式を 8 種まで細かく分けると，化学式各々の数が少なくなってしまう．8 種の化学式のうち，CHEM1∼CHEM4 の化学式は成分的に類似しているため，特徴ベクトル $F^{correlation}_{(Word+5Chem)}$ を作成する際には，CHEM1∼CHEM4 の化学式を 1 種類の化学式としてまとめ，CHEM1-4 とおくことにする．したがって，特徴ベクトル $F^{correlation}_{(Word+5Chem)}$ では，CHEM1-4，CHEM5，CHEM6，CHEM7，CHEM8 の 5 種類の化学式を用いる．

特徴ベクトル $F^{correlation}_{(Word+5Chem)}$ は，各アブストラクトにおいて，

(1) 化学式の含まれている文章中にある単語を抽出

(2) (1) で抽出された各単語と CHEM1-4 の化学式のセットがひとつの文章中にあった場合，その次元の特徴ベクトル要素を 0.2 とする．但し，重複しているものは数えない．この操作を，CHEM5∼CHEM8 の化学式においても同様に行う．これにより，5 種類の同次元数の特徴ベクトルが作成される．この特徴ベクトルの各要素の最大値は 1 となる．

(3) CHEM1-4∼CHEM8 の各化学式においてできた特徴ベクトルを，足し合わせて合成する．

という手順で作成する（図 5 参照）.

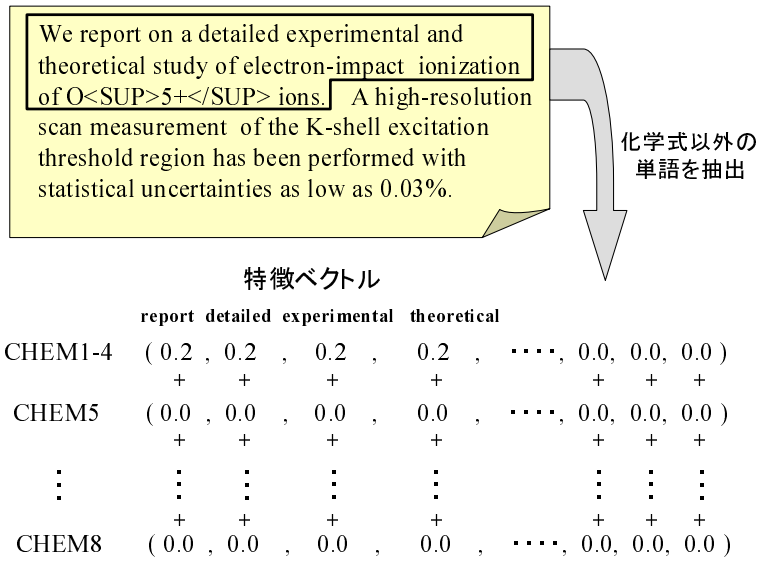

図 5: 特徴ベクトル $F^{correlation}_{(Word+5Chem)}$ の作成方法

4.3.4 化学式と文章数の関係を用いる場合

化学式と文章数の関係を用いて特徴ベクトルを作成する．具体的には，各アブストラクトの全文章数に対する化学式の含まれていない文章数の割合を使用する．この特徴ベクトルを $F^{num}_{(Sentence)}$ とする．特徴ベクトル $F^{num}_{(Sentence)}$ は 1 次元のベクトルであるため，4.3.3 項で説明した特徴ベクトル $F^{correlation}_{(Word+5Chem)}$ に含めて用いる．特徴ベクトル $F^{correlation}_{(Word+5Chem)}$ と特徴ベクトル $F^{num}_{(Sentence)}$ を合わせた特徴ベクトルを $F^{correlation}_{(Word+5Chem)} + F^{num}_{(Sentence)}$ と記述する．

4.4 学習方法

我々は今回，システムの学習方法として，他の機械学習法よりも比較的単純な構造をもつ Learning Vector Quantization(LVQ)[19] を採用する．これは，単純な機械学習法を用いてもアブストラクトのみで論文を分類することが可能であることを示すためである．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．LVQ1，LVQ2.1，LVQ3 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．本システムでは LVQ1 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参照ベクトルによっては長くな

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セットより作成する．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学習を行う．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．

5　評価方法

5.1　分類対象の論文

システムの評価に使用する論文について述べる．原子分子物理学分野の論文のうち，市川が収集した原子分子データが掲載されている論文 419 件 [20] のアブストラクトと Phys.Rev.A 誌（vol.41∼62）[13] に掲載されている論文のアブストラクトを使用する．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ある．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[3]．本論文で開発する 2 つのシステムは，それぞれ分類の対象とする論文が異なる．システム A の評価の際に使用するトレーニングデータセットを DA_L，テストデータセットを DA_T とし，システム B の評価で用いるトレーニングデータセットを DB_L，テストデータセットを DB_T とする．データセットの内容を 5.1.1 項，5.1.2 項で説明する．

5.1.1　システム A に用いるデータセット

先ほど述べた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 419 件のうち，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．Phys.Rev.A 誌（vol.41∼62）に掲載されているカテゴリ 1 以外の論文のアブストラクトをカテゴリ 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 件である．

これらのカテゴリ 1 のアブストラクト 126 件とカテゴリ 0 のアブストラクト 15944 件から，トレーニングデータセットとテストデータセットを各 10 セット作成する．これらのデータセットは，8035 件ずつランダムに選んだもので，そのうちカテゴリ 1 のアブストラクトを 63 件ずつ含んでいるとする．i 番目のトレーニングデータセット，テストデータセットをそれぞれ $DA_L(i)$, $DA_T(i)$ と表す．

データセット $DA_L(1) \sim DA_L(10)$，$DA_T(1) \sim DA_T(10)$ とは別に，600 件のアブストラクトを使用するデータセットを作成する．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 件から成っているとする．トレーニングデータとテストデータを 300 件ずつとして，データセットを各 100 セット作成する．600 件のアブストラクトより作成したデータセットは, i 番目のトレーニングデータセットを $DA'_L(i)$，テストデータセットを $DA'_T(i)$ と表記する．データセット $DA'_L(1) \sim DA'_L(100)$，$DA'_T(1) \sim DA'_T(100)$ は，カテゴリ 1 のアブストラクト 63 件を，カテゴリ 0 のアブストラクトを 237 件含んでいる．

[3] Phys.Rev.A には他に，物理光学分野の論文が含まれている

5.1.2 システム B に用いるデータセット

5.1 節で述べた市川が収集した原子分子データが掲載されている論文 419 件を用いる．システム B では，システム A において原子分子データが掲載されていると判断されたデータを使用することになるため，カテゴリ 1 でない論文も多少含まれていると考えられる．したがって，システム B のための実験では，原子分子データが掲載されていない論文 15 件を交えた 434 件のアブストラクトを使用する．434 件のアブストラクトのうち，原子分子データベースへの登録に向いている論文 183 件をカテゴリ 1 とし，それ以外の論文 251 件をカテゴリ 0 と設定する．トレーニングデータとテストデータを各 217 件とし，それぞれランダムに選択したデータセットを 100 セット作成し使用する．トレーニングデータセットは 217 件のうちカテゴリ 1 を 92 件含み，テストデータセットは 217 件中カテゴリ 1 を 91 件含む．100 セットのトレーニングデータセットを DB_L とし，テストデータセットを DB_T と表す．

5.2 評価尺度

本モデルの性能評価を行うために，5.1 節で述べた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．

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．提案するシステムでは，探索対象である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，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，再現率を重視する．

6　本モデルの性能評価

本章では，4.2 節で述べたシステム A，システム B を提案モデルを基に開発し，システム A，B の評価から提案モデルの性能評価を行う．特徴ベクトルは，4.3 節で挙げたベクトルを使用する．

6.1　システム A についての評価

システム A では，TF・IDF 法を使用する特徴ベクトル，文章頻度を使用する特徴ベクトル，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から専門用語辞書を作成することによって得られる特徴ベクトルを使用して評価を行う．また，2 つの特徴ベクトルを組み合わせた場合や 2 つの LVQ を用いる場合についての評価も行う．さらに，人間がアブストラクトのみで論文を分類した場合の模倣実験を行い，機械学習による論文分類との比較を行う．最後に特徴ベクトルの次元数について考察する．

6.1.1　TF・IDF 法を使用する場合

6.1.1.1　使用する文書集合，出現頻度の違いによる比較実験

5.1.1 項で述べたトレーニングデータセット $DA'_L(1) \sim DA'_L(100)$ とテストデータセット $DA'_T(1) \sim DA'_T(100)$ を使用する場合の認識率，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．

文書集合の違いによる比較

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．しかし本論文では，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，以下のような集合が考えられる．

P_L　トレーニングデータの文書集合

P_T　テストデータの文書集合

P^{Cx}　カテゴリ x に属している文書集合

P_L^{Cx}　カテゴリ x に属しているトレーニングデータの文書集合

P_T^{Cx}　カテゴリ x に属しているテストデータの文書集合

P_{All}　トレーニングデータセットとテストデータセットを合わせた文書集合

TF・IDF 法による重みを計算する際にどの文書集合を用いるかによって，文書頻度の値が異なってくる．すなわち特徴ベクトルの要素の値が変わってくる．したがって，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ってくるため重要な課題である．そこでこれらの文書集合を組み合わせて，各特徴ベクトルについての実験を行う．

文書集合の組み合わせは，(P_{All})，(P_L,P_T)，(P^{C1},P^{C0})，$(P_L^{C1},P_L^{C0},P_T^{C1},P_T^{C0})$，$(P_L^{C1},P_L^{C0},P_T)$ の 5 つを用いる．特徴ベクトル $F^{tf}_{(D1)}$，$F^{tf}_{(D1)+Chem}$，$F^{tf}_{(D1+D2)}$，$F^{tf}_{(D1+D2)+Chem}$，$F^{tf}_{(Dic)}$，$F^{tf}_{(Dic)+Chem}$ を使用した際の認識率，再現率，適合率の平均値をそれぞれ図 6〜図 8，図 9〜図 11，図 12〜図 14 に示す．

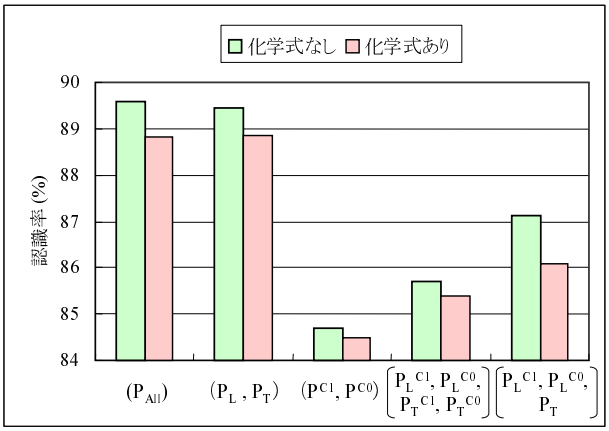

図 6: 特徴ベクトル $F^{tf}_{(D1)}$，$F^{tf}_{(D1)+Chem}$ の認識率

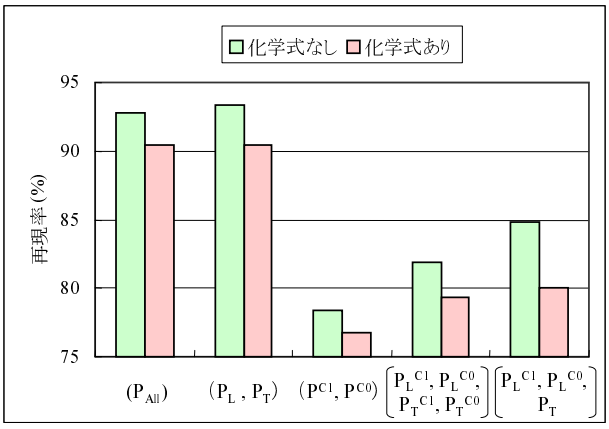

図 7: 特徴ベクトル $F^{tf}_{(D1)}$，$F^{tf}_{(D1)+Chem}$ の再現率

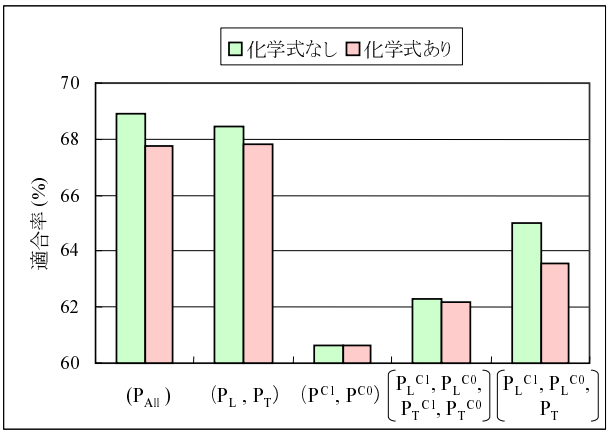

図 8: 特徴ベクトル $F^{tf}_{(D1)}$， $F^{tf}_{(D1)+Chem}$ の適合率

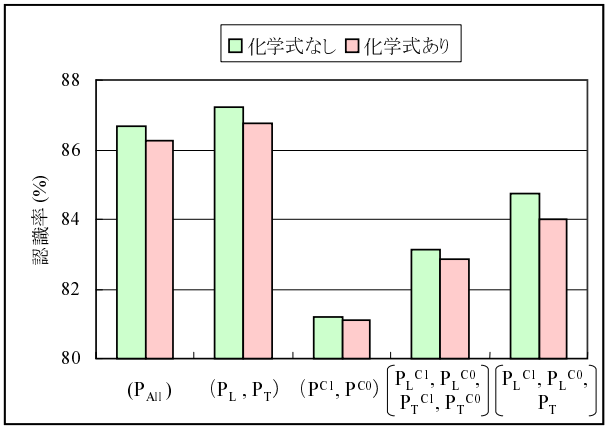

図 9: 特徴ベクトル $F^{tf}_{(D1+D2)}$， $F^{tf}_{(D1+D2)+Chem}$ の認識率

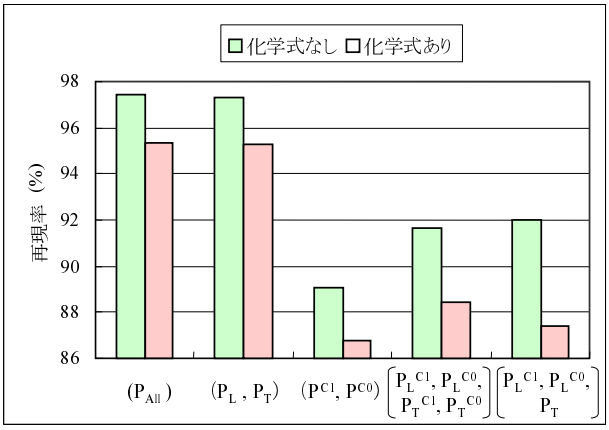

図 10: 特徴ベクトル $F^{tf}_{(D1+D2)}$, $F^{tf}_{(D1+D2)+Chem}$ の再現率

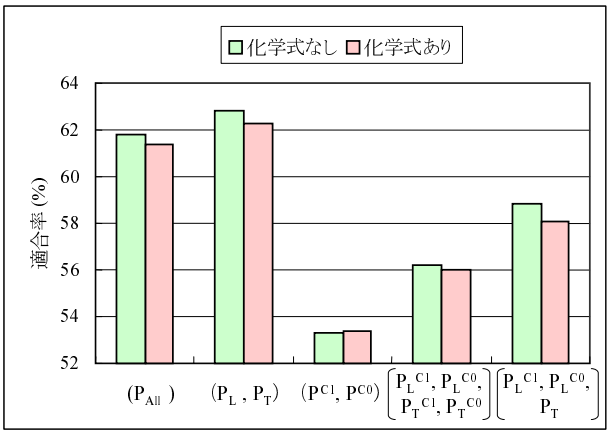

図 11: 特徴ベクトル $F^{tf}_{(D1+D2)}$, $F^{tf}_{(D1+D2)+Chem}$ の適合率

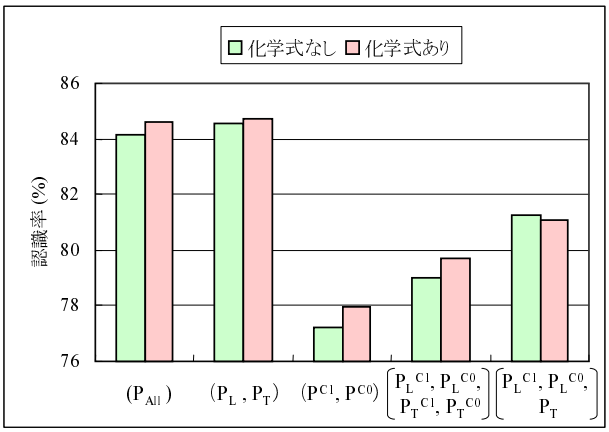

図 12: 特徴ベクトル $F^{tf}_{(Dic)}$， $F^{tf}_{(Dic)+Chem}$ の認識率

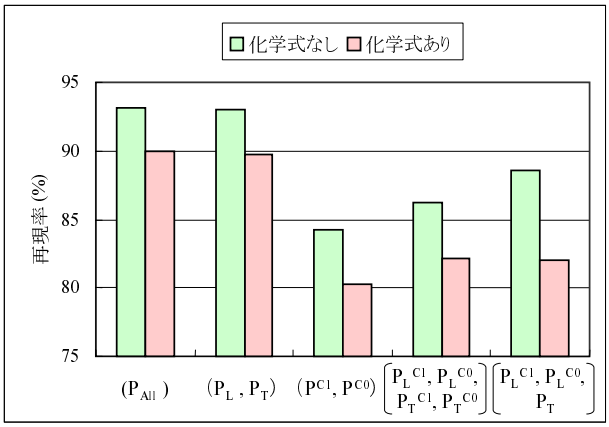

図 13: 特徴ベクトル $F^{tf}_{(Dic)}$， $F^{tf}_{(Dic)+Chem}$ の 再現率

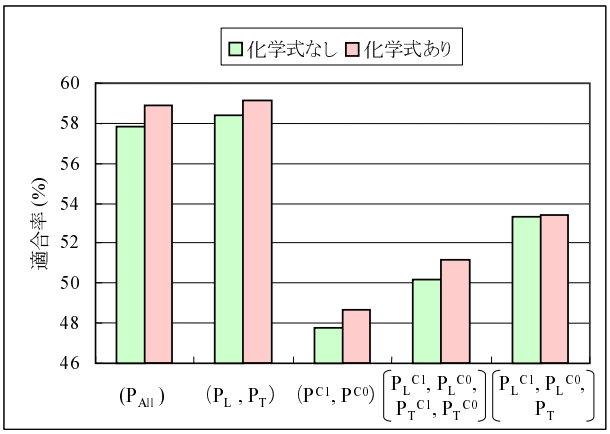

図 14: 特徴ベクトル $F^{tf}_{(Dic)}$，$F^{tf}_{(Dic)+Chem}$ の適合率

図 6~ 図 14 より，認識率，再現率，適合率について，集合 (P_{All}) と集合 (P_L, P_T) を用いた場合によい結果が得られている．一方，集合 (P^{C1}, P^{C0})，$(P_L^{C1}, P_L^{C0}, P_T^{C1}, P_T^{C0})$，$(P_L^{C1}, P_L^{C0}, P_T)$ を使用した際には低い値となっている．集合 (P_{All})，(P_L,P_T) は 2 つのカテゴリが混在している集合のみから成っており，集合 (P^{C1},P^{C0})，$(P_L^{C1},P_L^{C0},P_T^{C1},P_T^{C0})$，$(P_L^{C1},P_L^{C0},P_T)$ は混在していない集合が含まれている．これは，TF・IDF 法におい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際に，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ているが，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ことが原因であると思われる．また，集合に含まれるデータ数が多いほど他の文書との関係を考えた重みを計算できるため，優れた特徴ベクトルを作成できるといえる．これらの理由により，我々はカテゴリが混在している集合 (P_{All})，(P_L, P_T) を用いる方が優れた特徴ベクトルであると考える．

ここで，特徴ベクトルの実用性について考える．本システムを実用化する場合，未分類のデータに対して適用することになる．このときデータのカテゴリは未知である．よって，集合 P^{C1},P_T^{C1},P_T^{C0} は，適用するデータのカテゴリがわかっていることを前提に作られた集合であるため使えない．すなわち集合 (P^{C1},P^{C0}) と集合 $(P_L^{C1},P_L^{C0},P_T^{C1},P_T^{C0})$ を用いて作成した特徴ベクトルは使用できない．さらに本システムでは，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，適用するデータを含んだ集合を用いて特徴ベクトルを作成する必要がある集合 (P_{All}) は，適用するデータがないと学習を行えないため，実用的ではない．ゆえに実用可能であるのは，集合 (P_L,P_T) と集合 (P_L^{C1},P_L^{C0},P_T) を使用して

作成された特徴ベクトルである．

以上の理由により，集合 (P_L,P_T) を使用して作成した特徴ベクトルが最適であると考えられる．以降，集合 (P_L,P_T) を用いて作成した特徴ベクトルについての比較を行う．

出現頻度の違いによる比較

特徴ベクトル $F^{tf}_{(D1)}$，$F^{tf}_{(D1)+Chem}$，$F^{tf}_{(D1+D2)}$，$F^{tf}_{(D1+D2)+Chem}$，$F^{tf}_{(Dic)}$，$F^{tf}_{(Dic)+Chem}$ の違いによる認識率，再現率，適合率を比較する．図 15，図 16 は特徴ベクトル $F^{tf}_{(D1)}$，$F^{tf}_{(D1+D2)}$ の要素数の変化を示す．

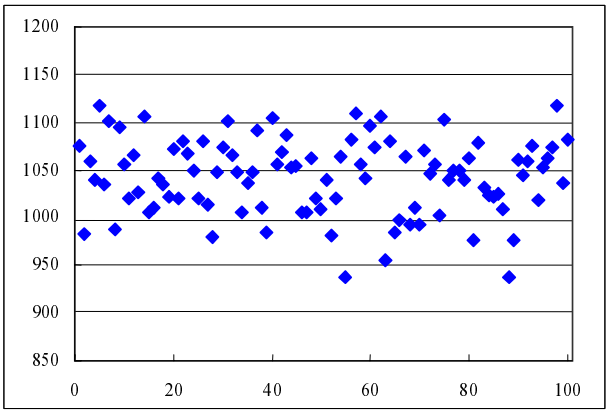

図 15: 実験 6.1.1.1 における特徴ベクトル $F^{tf}_{(D1)}$ の要素数の変化

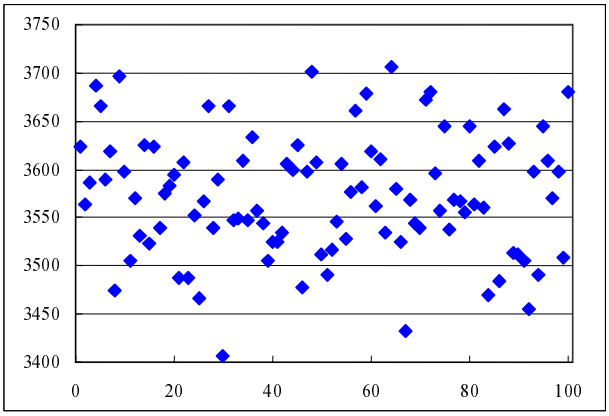

図 16: 実験 6.1.1.1 における特徴ベクトル $F^{tf}_{(D1+D2)}$ の要素数の変化

図 15，図 16 より，特徴ベクトル $F^{tf}_{(D1)}$，$F^{tf}_{(D1+D2)}$ の要素数はそれぞれ，937~1117 要素，3406~3706 要素に分布しており，4.3.1 項で述べたとおり，

要素数が一定でないことが確認できる．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，理化学辞典の見出し語を基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 $F^{tf}_{(Dic)}$，$F^{tf}_{(Dic)+Chem}$ の要素数は 1182，1190 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が，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[4]．

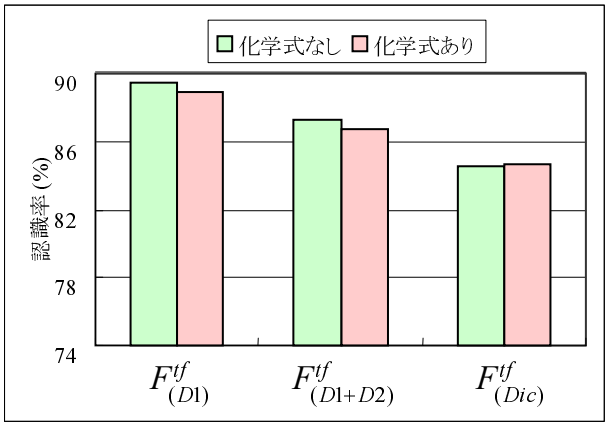

図 17: 実験 6.1.1.1 における認識率

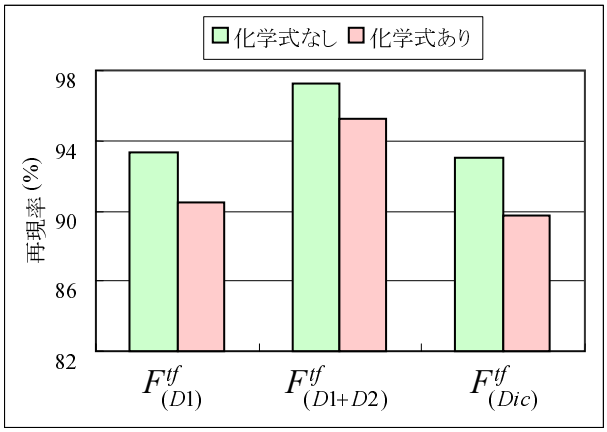

図 18: 実験 6.1.1.1 における再現率

[4]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は把握できないため，全 23500 語の要素を用いる

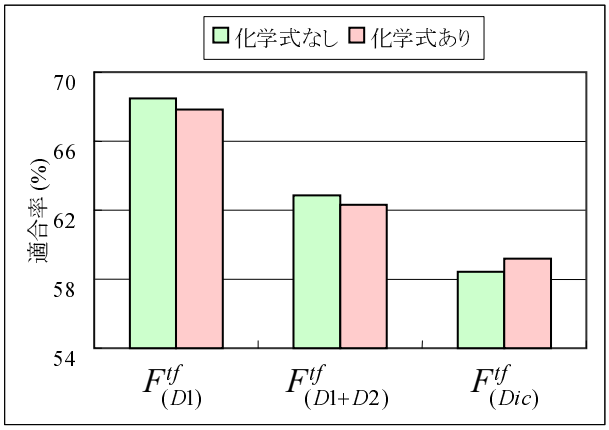

図 19: 実験 6.1.1.1 における適合率

図 17～図 19 は，文書集合 (P_L,P_T) に関して得られた認識率，再現率，適合率の平均値を並べたグラフである．図 18 より，どの特徴ベクトルを用いた場合にも再現率が 90%以上かそれに近い値になっている．認識率，適合率に関しては，特徴ベクトル $F^{tf}_{(D1)}$，$F^{tf}_{(D1+D2)}$ を使用した場合に高い値が出ている．しかし特徴ベクトル $F^{tf}_{(D1)}$，$F^{tf}_{(D1+D2)}$ には，先に述べたようにトレーニングデータセットに依存するという問題がある．特徴ベクトル $F^{tf}_{(Dic)}$ を用いた際の実験結果より，適合率が低い値ではあるが，本研究で重視している再現率は，特徴ベクトル $F^{tf}_{(D1)}$，$F^{tf}_{(D1+D2)}$ と比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 $F^{tf}_{(Dic)}$ が最適であるとみなす．

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．しかし，化学式を使用して作成した特徴ベクトルと使用せずに作成した特徴ベクトル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．

6.1.1.2　参照ベクトル数の違いによる比較実験

5.1.1 項で述べたデータセット $DA_L(1) \sim DA_L(10)$ と $DA_T(1) \sim DA_T(10)$ を使用する．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 6.1.1.1 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 8000 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 1000~8000 とし，1000 ずつ増やして実験を行う．本実験には特徴ベクトル $F^{tf}_{(Dic)}$，$F^{tf}_{(Dic)+Chem}$ を用いる[5]．

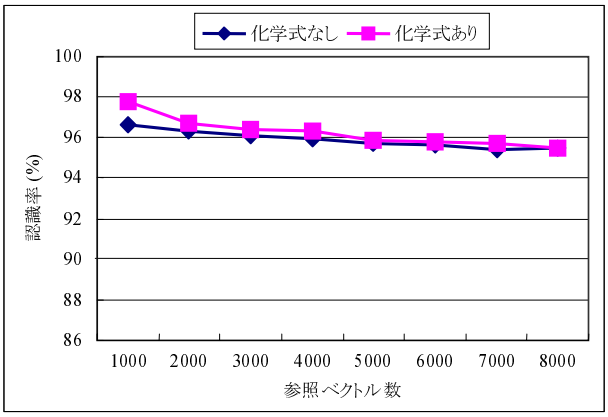

図 20: 実験 6.1.1.2 における認識率

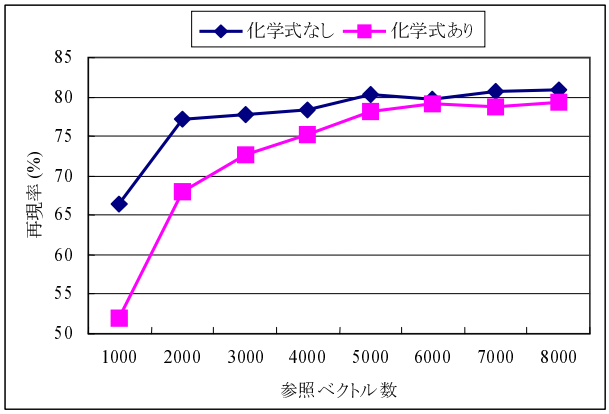

図 21: 実験 6.1.1.2 における再現率

[5]実験 6.1.1.2 において特徴ベクトル $F^{tf}_{(Dic)}$ は 3558 次元，$F^{tf}_{(Dic)+Chem}$ は 3564 次元のベクトルである

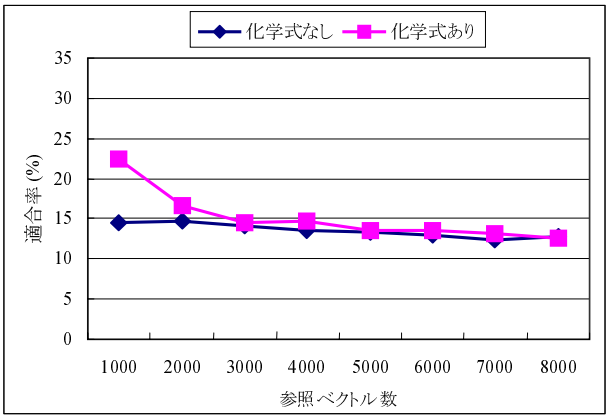

図 22: 実験 6.1.1.2 における適合率

図 20～図 22 は，10 セットのトレーニングデータセットを用いて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したが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 5000~8000 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 6.1.1.1 の結果と同様に再現率において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 に収録されている論文を扱っている．Phys.Rev.A 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．

6.1.1.3　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 6.1.1.2 では参照ベクトル数が 8000 の場合に，再現率は 80.95%になっているが適合率は 12.68%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に注目してみる．これまでは，過去に行ってきた研究結果 [21] 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 [21] 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 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 2000 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

ときの認識率，再現率，適合率の変化を調べる．トレーニングデータセット，テストデータセット共に実験 6.1.1.2 で使用したものを使用し，特徴ベクトル $F^{tf}_{(Dic)}$ を用いる．図 23〜図 25 は平均値をとったものである．

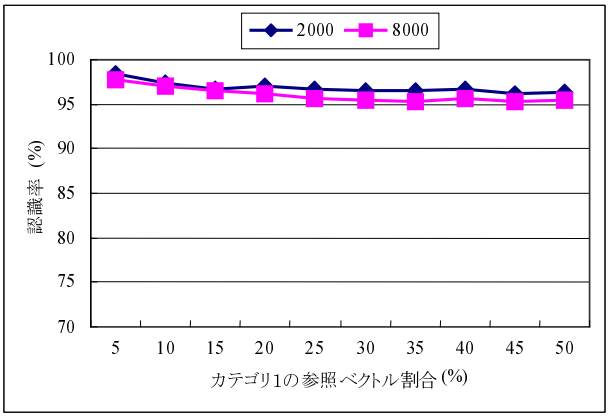

図 23: 実験 6.1.1.3 における認識率の変化

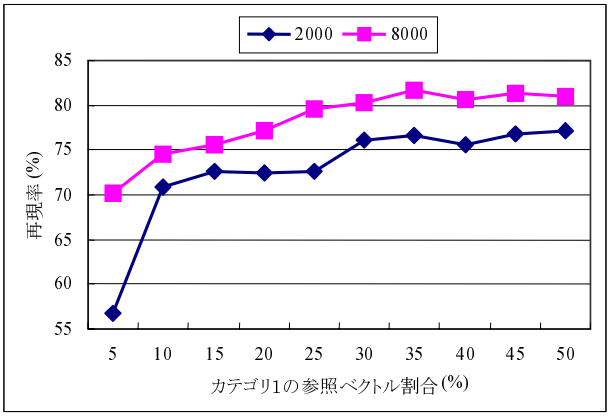

図 24: 実験 6.1.1.3 における再現率の変化

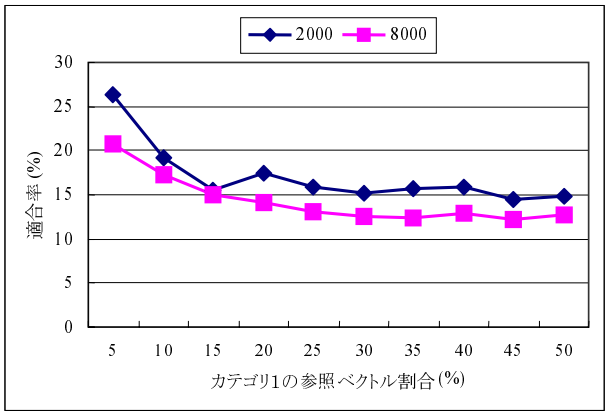

図 25: 実験 6.1.1.3 における適合率の変化

図 23∼図 25 より，参照ベクトルの数が 2000 と 8000 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 26.40%，20.77%になっているものの，再現率が 56.67%，70.16%でカテゴリ 1 の割合が 50%のときと比べると 10∼20%低下している．また参照ベクトル数が 8000 の場合，カテゴリ 1 の参照ベクトルの割合が 35%の際に最高の再現率になっているが，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る．参照ベクトル数 2000 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 8000 でカテゴリ 1 の参照ベクトルの割合が 35%であるときが最適であると判断できる．このとき，認識率は 95.28%，再現率は 81.75%，適合率は 12.33%である．しかし，カテゴリ 1 に属している 63 件のアブストラクトを 2800 の参照ベクトルによって探し出すことになる．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．

6.1.2 文章頻度を使用する場合

データセット $DA'_L(1) \sim DA'_L(100)$，$DA'_T(1) \sim DA'_T(100)$ の各 100 セットを用いる実験の結果を示す．認識率，再現率，適合率の平均値を比較する．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．図 26∼図 28 に特徴ベクトル $F^{sf}_{(D1)}$，$F^{sf}_{(D1)+Chem}$，$F^{sf}_{(D1+D2)}$，$F^{sf}_{(D1+D2)+Chem}$，$F^{sf}_{(Dic)}$，$F^{sf}_{(Dic)+Chem}$ を用いた場合の認識率，再現率，適合率の平均値の結果を示す．

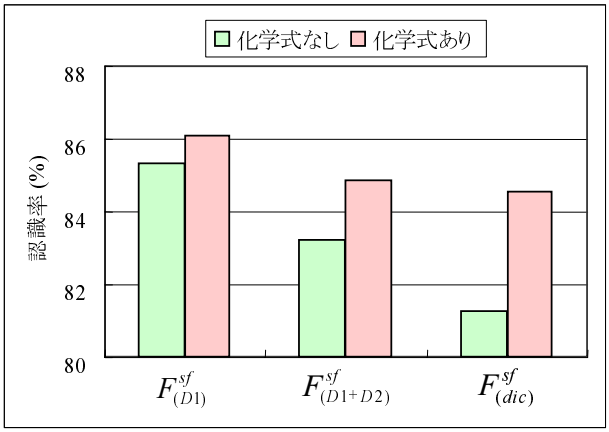

図 26: 実験 6.1.2 において 600 件のデータを使用した際の認識率

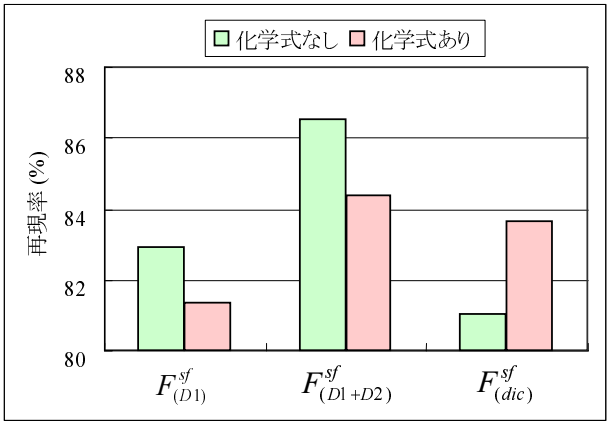

図 27: 実験 6.1.2 において 600 件のデータを使用した際の再現率

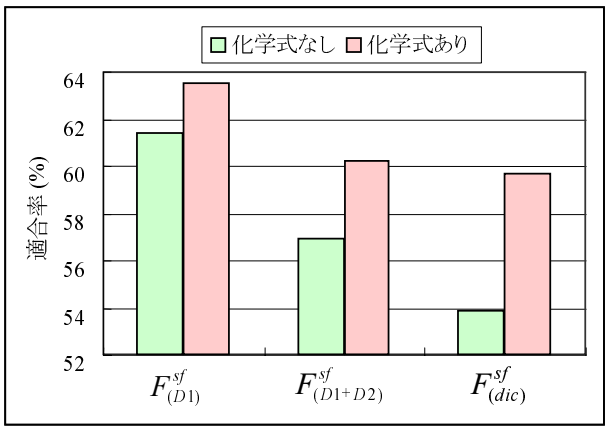

図 28: 実験 6.1.2 において 600 件のデータを使用した際の適合率

図 26～図 28 より，特徴ベクトル $F^{sf}_{(Dic)}$，$F^{sf}_{(Dic)+Chem}$ を使用する場合の認識率，再現率，適合率において，化学式を含んだ特徴ベクトル の方が秀でた結果を示している．この実験結果より，理化学辞典を使用した特徴ベクトルについて，TF・IDF 法を使用し文書集合全体に対する重要性をみている際には生かされていない化学式の特徴が，ひとつのアブストラクト内では重要な役割を果たしていると仮定できる．

この仮定が正しいかどうかを確認するために，データセット $DA_L(1) \sim DA_L(10)$ と $DA_T(1) \sim DA_T(10)$ を用いた場合の実験結果を示す．TF・IDF 法を用いる場合と同様に, 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 $F^{sf}_{(Dic)}$，$F^{sf}_{(Dic)+Chem}$ を用いる．図 29～図 31 は，学習させた結果の平均値を示したものである．

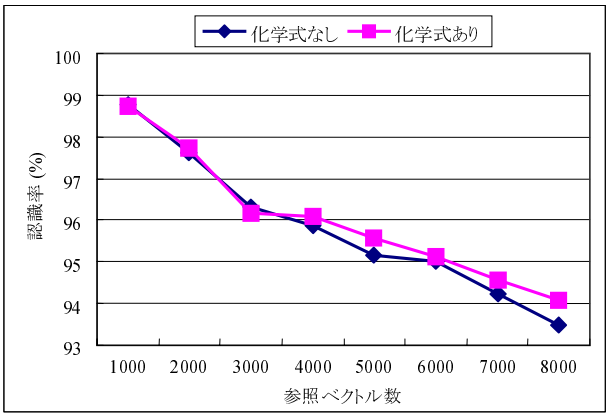

図 29: 実験 6.1.2 において 16070 件のデータを使用した際の認識率の変化

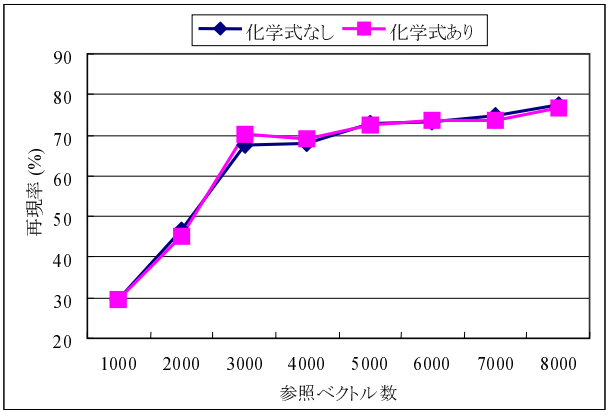

図 30: 実験 6.1.2 において 16070 件のデータを使用した際の再現率の変化

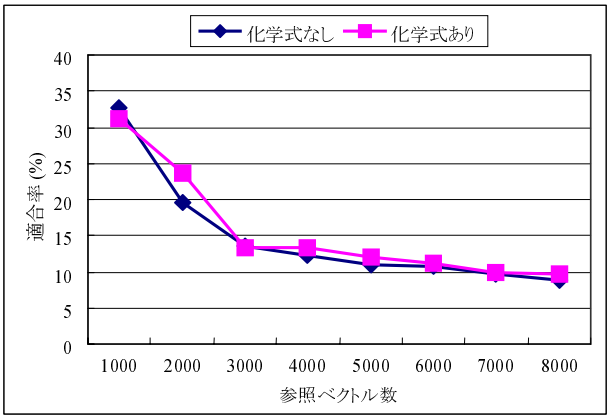

図 31: 実験 6.1.2 において 16070 件のデータを使用した際の適合率の変化

化学式を含んだ特徴ベクトル $F^{sf}_{(Dic)+Chem}$ を使用する場合と含まない特徴ベクトル $F^{sf}_{(Dic)}$ を使用する場合を比較すると，認識率，再現率，適合率のすべてにおいて，それぞれ僅差であることが確認できる．よって，化学式が重要な役割を果たしているとは言い切れない．これは先ほどの「ひとつのアブストラクトの中では化学式に重要性が認められる」という仮定が誤っていることを意味する．ゆえに化学式は，論文に原子分子データが掲載されているか否かを判断する際には有用ではない．

再現率に注目すると，特徴ベクトル $F^{sf}_{(Dic)}$ で 77.30%，特徴ベクトル $F^{sf}_{(Dic)+Chem}$ で 76.83%という結果が参照ベクトル数 8000 の場合に得られている．よって，他の文書との関連性を考慮していない特徴ベクトルを用いても，80%近い再現率を出すことが可能であることを実証できたといえる．

6.1.3　アブストラクトから専門用語辞書を作成し用いる場合

特徴ベクトル $F^{log}_{(Special)}$ を用いて実験を行う．機械学習を行う際に，特徴ベクトルのすべての要素が 0 になるアブストラクトはカテゴリ 0 であるとする．データセットは，実験 6.1.1.2 で使用したデータセットのうち，トレーニングデータセットは $DA_L(1) \sim DA_L(3)$ の 3 セットを，テストデータセットは $DA_T(1) \sim DA_T(3)$ の 3 セットを使用する．参照ベクトル数は，1000~5000 とし，1000 ずつ増やして実験を行う．図 32 は認識率，再現率，適合率の平均値のグラフである．

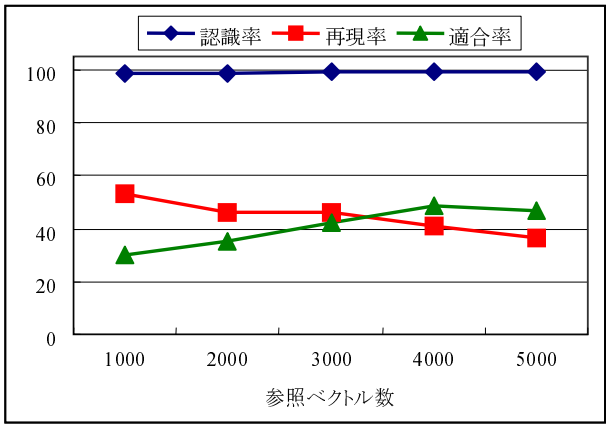

図 32: 特徴ベクトル $F^{log}_{(Special)}$ を使用した場合の認識率，再現率，適合率

図 32 より，認識率は参照ベクトル数に関わらず，ほぼ 100%を示している．適合率は 30%以上の値になっているが，再現率は 60%を下回る．再現率が低い理由として，専門用語自動抽出システムの仕様上，アブストラクトに記載されている単語のステミング処理を行えないために，同じ単語でも別の単語として認識されている可能性が高いことが考えられる．また，参照ベクトル数が増えるにつれて再現率が低下している．これは特徴ベクトルの要素数が約 1000 であること，カテゴリ 0 のトレーニングデータに含まれている用語を考慮していないことが原因としてあげられる．要素数が少ないことから，参照ベクトルが増加すると，異なるカテゴリに属している参照ベクトル同士であっても類似した参照ベクトルが徐々に作成されてしまい，それらが互いに衝突して再現率を下げているものと思われる．

6.1.4　その他の特徴ベクトル作成方法

理化学辞典の用語と専門用語自動抽出システムで抽出された用語を合わせて使用する．各要素の重みを TF･IDF 法で計算する特徴ベクトルを $F^{tf}_{(Dic)+(Special)}$

とし，理化学辞典の用語はTF･IDF法で計算し，専門用語自動抽出システムで抽出された用語は，抽出された際に共に出力されたスコアの自然対数を重みとして，それぞれ正規化したベクトルを合成する特徴ベクトルを $F^{log}_{(Dic)+(Special)}$ と表す．この2つの方法で作成した特徴ベクトルを使用した際の実験結果は図33，図34のようになる．

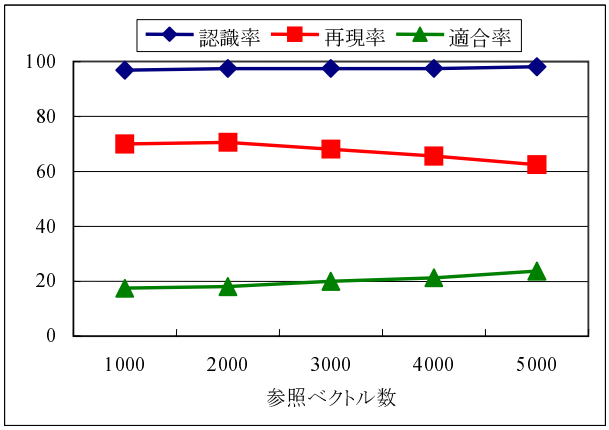

図 33: 特徴ベクトル $F^{tf}_{(Dic)+(Special)}$ を使用した場合の認識率，再現率，適合率

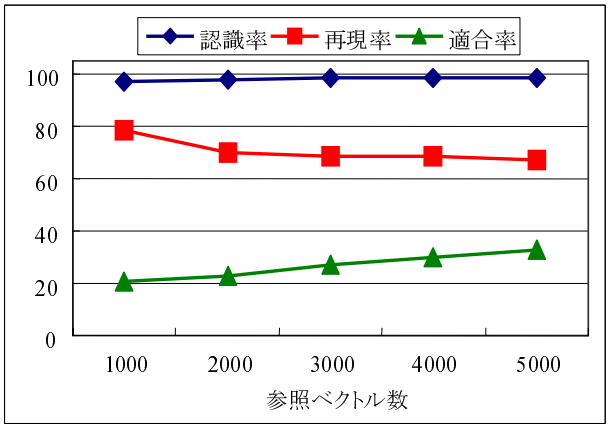

図 34: 特徴ベクトル $F^{log}_{(Dic)+(Special)}$ を使用した場合の認識率，再現率，適合率

再現率を向上させるために特徴ベクトル $F^{tf}_{(Dic)}$ に新たな専門用語の要素を加えたが，再現率はむしろ低下しているということが図33，図34より明らかである．これは加えられた専門用語が，数少ないカテゴリ1のアブストラクト

に対して使われたのではなく，カテゴリ 0 のアブストラクトを正しく分類するために使われたのではないかと考えられる．特徴ベクトル $F^{tf}_{(Dic)+(Special)}$，$F^{log}_{(Dic)+(Special)}$ のどちらの場合についても，再現率が徐々に低くなり適合率が高くなっている．このような結果は，実験 6.1.3 の結果にもみられる．したがって，専門用語自動抽出システムで抽出した用語の特徴ベクトル成分が影響しているものと思われる．

6.1.5　2 つの LVQ を用いる場合

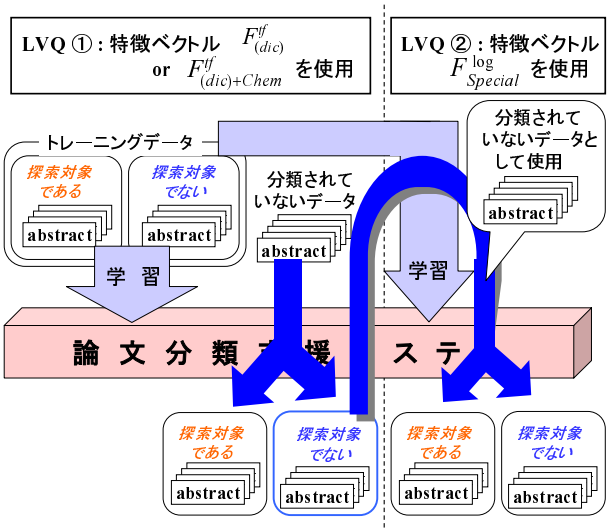

図 35: 2 つの LVQ の適用イメージ

実験 6.1.1.2 と実験 6.1.3 の実験結果により，特徴ベクトル $F^{tf}_{(Dic)}$，$F^{tf}_{(Dic)+Chem}$ を使用する場合の再現率と特徴ベクトル $F^{log}_{(Special)}$ を使って機械学習を行う際の再現率では，特徴ベクトル $F^{tf}_{(Dic)}$，$F^{tf}_{(Dic)+Chem}$ を用いる方が高い再現率を得られると考えられる．ここで特徴ベクトル $F^{tf}_{(Dic)}$，$F^{tf}_{(Dic)+Chem}$ でシステムがカテゴリ 1 であると認識したアブストラクトと，特徴ベクトル $F^{log}_{(Special)}$ による学習でカテゴリ 1 であると判断したアブストラクトの比較を行う．すると特徴ベクトル $F^{tf}_{(Dic)}$，$F^{tf}_{(Dic)+Chem}$ で正しく分類できているアブストラクトが特徴ベクトル $F^{log}_{(Special)}$ で分類できていないのではなく，それぞれ異なるアブストラクトを正しく認識していることが明らかになったため，図 35 のように，はじめに特徴ベクトル $F^{tf}_{(Dic)}$ あるいは特徴ベクトル $F^{tf}_{(Dic)+Chem}$ を使って分類を行い，カテゴリ 1 であると判断されたアブストラクトはカテゴリ 1 に分類し，カテゴリ 0 と判断されたアブストラクトについては特徴ベクトル $F^{log}_{(Special)}$ を用いて再度分類を行い，カテゴリ 1 であると認識された

アブストラクトはカテゴリ 1 に，カテゴリ 0 であるとされたアブストラクトはカテゴリ 0 に分類する方法を試みる．参照ベクトル数を各々1000~5000 とし，すべての組み合わせを考慮して実験を行う．この方法を使って得られる認識率，再現率，適合率の平均値を表 1~ 表 6 に示す．表データセットについては，$DA_L(1) \sim DA_L(3)$ と $DA_T(1) \sim DA_T(3)$ を用いる．

$F^{tf}_{(Dic)} \backslash F^{log}_{(Special)}$	5000	4000	3000	2000	1000
1000	95.97	96.22	96.35	96.47	96.49
2000	95.61	95.84	95.96	96.08	96.10
3000	95.52	95.75	95.86	95.97	95.98
4000	95.37	95.60	95.71	95.82	95.83
5000	95.26	95.48	95.58	95.69	95.70

表 1: 特徴ベクトル $F^{tf}_{(Dic)}$ と特徴ベクトル $F^{log}_{(Special)}$ を用いた場合の認識率 (%)

$F^{tf}_{(Dic)} \backslash F^{log}_{(Special)}$	1000	2000	3000	4000	5000
1000	80.95	77.78	76.19	74.60	73.54
2000	86.24	84.66	83.07	82.54	80.95
3000	87.30	85.19	83.60	83.07	81.48
4000	84.66	83.07	80.95	81.48	80.42
5000	85.19	83.60	80.95	81.48	80.42

表 2: 特徴ベクトル $F^{tf}_{(Dic)}$ と特徴ベクトル $F^{log}_{(Special)}$ を用いた場合の再現率 (%)

$F^{tf}_{(Dic)} \backslash F^{log}_{(Special)}$	1000	2000	3000	4000	5000
1000	14.07	14.52	14.74	14.98	14.92
2000	13.81	14.21	14.47	14.78	14.61
3000	13.56	13.98	14.09	14.38	14.23
4000	12.93	13.33	13.42	13.77	13.68
5000	12.67	13.09	12.99	13.38	13.29

表 3: 特徴ベクトル $F^{tf}_{(Dic)}$ と特徴ベクトル $F^{log}_{(Special)}$ を用いた場合の適合率 (%)

$F^{tf}_{(Dic)+Chem} \backslash F^{log}_{(Special)}$	1000	2000	3000	4000	5000
1000	96.96	97.22	97.35	97.47	97.47
2000	95.73	95.95	96.09	96.19	96.20
3000	95.81	96.03	96.16	96.25	96.26
4000	95.75	95.96	96.09	96.17	96.18
5000	95.05	95.27	95.37	95.45	95.47

表 4: 特徴ベクトル $F^{tf}_{(Dic)+Chem}$ と特徴ベクトル $F^{log}_{(Special)}$ を用いた場合の認識率 (%)

$F^{tf}_{(Dic)+Chem} \backslash F^{log}_{(Special)}$	1000	2000	3000	4000	5000
1000	75.66	72.49	71.43	68.25	65.61
2000	82.01	79.89	78.31	78.31	77.78
3000	82.54	78.84	77.78	76.72	75.13
4000	80.95	77.78	78.31	77.25	76.19
5000	85.19	82.54	82.01	82.01	80.95

表 5: 特徴ベクトル $F^{tf}_{(Dic)+Chem}$ と特徴ベクトル $F^{log}_{(Special)}$ を用いた場合の再現率 (%)

$F^{tf}_{(Dic)+Chem} \backslash F^{log}_{(Special)}$	1000	2000	3000	4000	5000
1000	18.40	19.68	20.63	21.27	20.52
2000	13.53	13.92	14.19	14.51	14.49
3000	13.82	13.99	14.26	14.42	14.25
4000	13.57	13.83	14.32	14.44	14.37
5000	12.15	12.35	12.56	12.74	12.67

表 6: 特徴ベクトル $F^{tf}_{(Dic)+Chem}$ と特徴ベクトル $F^{log}_{(Special)}$ を用いた場合の適合率 (%)

表 1～表 6 より，全体的に特徴ベクトル $F^{log}_{(Special)}$ を参照ベクトル数 1000 で使用した際の再現率が高いことが確認できる．これは特徴ベクトル $F^{log}_{(Special)}$ のみを用いた実験 6.1.3 においての実験結果でも参照ベクトル数 1000 のときに最もよい再現率が出ていたことからも納得できる．この実験によって，特徴ベクトル $F^{tf}_{(Dic)}$ や $F^{tf}_{(Dic)+Chem}$ だけでは検索しきれなかったカテゴリ 1 のアブストラクトを特徴ベクトル $F^{log}_{(Special)}$ で補うことにより，さらによい再現率を導くことが可能であることを示すことができる．また，再現率が向上し

たことにより，特徴ベクトル $F^{tf}_{(Dic)}$，$F^{tf}_{(Dic)+Chem}$ と特徴ベクトル $F^{log}_{(Special)}$ では分類基準が異なるということも実証できる．特徴ベクトル $F^{tf}_{(Dic)}$（参照ベクトル数 3000）と特徴ベクトル $F^{log}_{(Special)}$（参照ベクトル数 1000）を使用することで，認識率 95.52%，再現率 87.30%，適合率 13.56%という結果を得ることができる．これは実験 6.1.1〜実験 6.1.4 の中では最高の再現率である．

6.1.6 人間による論文分類の模倣実験

人間がアブストラクトを読んで論文を分類する場合には，原子分子データが掲載され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論文を判断していると思われる．そこで，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，それらを合計して各アブストラクトのスコアを算出し，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，論文分類時に人間が脳で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ついて，4.3.2 項で使用した Perl モジュール”TermExtract”を使用する．機械学習法を用いる場合との比較を行うために，トレーニングデータセット $DA_L(1) \sim DA_L(10)$ とテストデータセット $DA_T(1) \sim DA_T(10)$ を用いる．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．キーワードのスコアは以下の 3 つの方法により計算する．

方法 I ト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，その情報を用いてトレーニングデータとテストデータのスコアを計算

方法 II トレーニングデータセットを使って TermExtract の学習機能により蓄積させた情報はトレーニングデータに対してのみ適用し，テストデータについて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スコアを計算

方法 III TermExtract の学習機能を使用せずにスコアを計算

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，Category1 のトレーニングデータ 63 件のうち，63 件を正しく Category1 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する．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 36〜図 38 のグラフに示す．

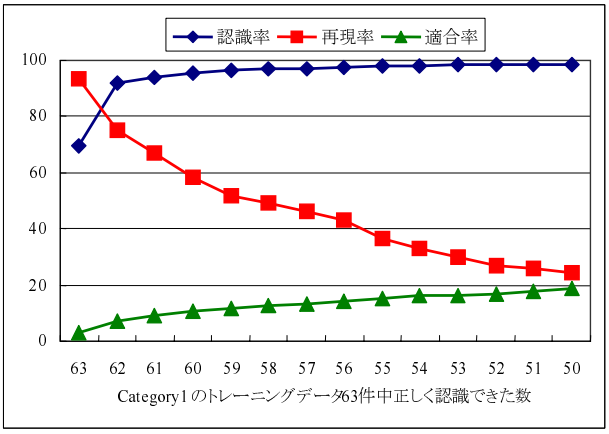

図 36: 方法 I によるテストデータの認識率・再現率・適合率

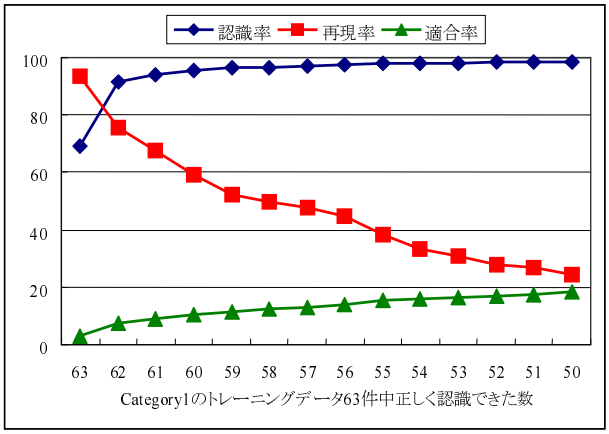

図 37: 方法 II によるテストデータの認識率・再現率・適合率

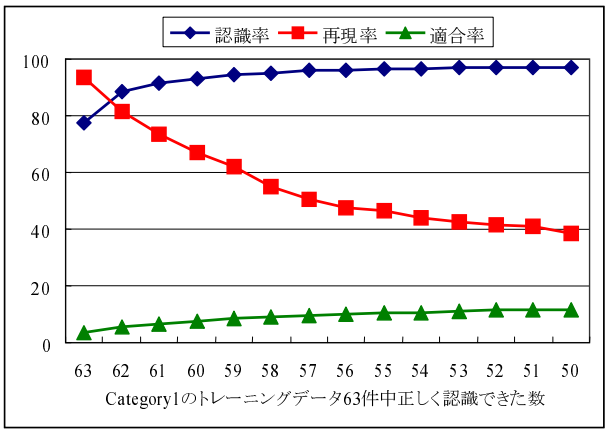

図 38: 方法 III によるテストデータの認識率・再現率・適合率

図 36~ 図 38 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いる．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，ひと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．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，単純にカテゴリ 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．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．

6.1.7 特徴ベクトルの次元数

これまで各種実験を行ってきたが，ここで，特徴ベクトル $F^{tf}_{(Dic)}$ のベクトルデータについての考察を行う．16070 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 3558 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 902 ある．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 2656 次元に縮約可能であるといえる．また，10 件以下の数件にしか値のない成分も数多くある．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．よって次元数の縮約は行わない．

6.2 システム B についての評価

単語と化学式の組み合わせを使用する特徴ベクトル，化学式と文章数の関係を使用する特徴ベクトルを合わせたベクトル $F^{correlation}_{(Word+5Chem)} + F^{num}_{(Sentence)}$ と TF・IDF 法を用いる特徴ベクトル $F^{tf}_{(D1+D2)}$ を用いる．理化学辞典の見出し語 23500 語のうち，434 件のアブストラクトに掲載されている語は 582 語しかなく，特徴ベクトルの次元としては低すぎるため，今回の実験では特徴ベクトル $F^{tf}_{(Dic)}$ は使用しない．データセットは，DB_L と DB_T を使用して，参照ベクトル数を 50，100，200 と変化させた際の認識率，再現率，適合率の平均値を求め，評価を行う．

まず，特徴ベクトル $F^{tf}_{(D1+D2)}$ を用いた際の認識率，再現率，適合率を調べる．実験結果を表 7 に記載する．

参照ベクトル数	認識率（%）	再現率（%）	適合率（%）
50	72.82	51.08	76.63
100	73.23	54.56	75.12
200	73.96	59.02	73.97

表 7: 特徴ベクトル $F^{tf}_{(D1+D2)}$ を使用した場合の認識率，再現率，適合率の平均値

表 7 より，どの参照ベクトル数であっても再現率が 50~60%になっていることが確認できる．これは再現率を重視している場合，適した方法ではなかったことを意味する．認識率に関しても同じことがいえる．この実験結果より，データベースへの登録に向いているか否かを分類するには，単純な特徴ベクトル作成方法では対応できないということが実証されたといえる．そこで，特徴ベクトル $F^{correlation}_{(Word+5Chem)} + F^{num}_{(Sentence)}$ と特徴ベクトル $F^{tf}_{(D1+D2)}$ を使い，システム A における実験 6.1.5 のときと同様に 2 つの LVQ を用いる実験を行う．はじめに特徴ベクトル $F^{correlation}_{(Word+5Chem)} + F^{num}_{(Sentence)}$ を使用して学習させた LVQ をテストデータに適用し，この LVQ がカテゴリ 1 であると認識したアブストラクトをカテゴリ 1 に分類する．カテゴリ 0 であると認識されたアブストラクトについては，特徴ベクトル $F^{tf}_{(D1+D2)}$ を使用して学習させた別の LVQ に適用した結果を採用する（図 35 参照）．分類結果の平均値を表 8 に示す．

参照ベクトル数	認識率（%）	再現率（%）	適合率（%）
50	65.94	75.05	57.37
100	63.75	79.16	54.78
200	62.35	82.01	53.42

表 8: 2 つの LVQ を使用した場合の認識率，再現率，適合率の平均値

表 8 より，認識率はどの参照ベクトルの数においても 60%台である．再現率については参照ベクトル数が増加するにつれて，高くなっていることが確認できる．特に参照ベクトル数が 200 のときに再現率は 80%を超えている．特徴ベクトル $F^{tf}_{(D1+D2)}$ のみを使用した場合と比較すると，認識率低下の度合いに比べて再現率向上の程度の方が大きい．よって本システムに関しては 2 つの LVQ を使用してカテゴリを認識させる方法が適していると考えられる．

7　まとめ

本論文では，アブストラクトを用いた論文分類支援モデルを提案し，原子分子物理学分野の論文を分類する際に適用した．その結果，論文そのものではなく，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，本モデルが有効であることを示した．

原子分子物理学分野の論文を原子分子データを含む論文を含まない論文に分類するシステムにおいては，専門用語辞典に掲載されている見出し語の出現頻度を基に作成した特徴ベクトルを使用した際に，認識率 95.28%，再現率 81.75%，適合率 12.33%という良好な結果を得ることができた．この結果は，10000 件の論文のうち入手したい論文が 78 件しかない文書集合から入手したい論文を探索する場合に，今までは人間が 10000 件の論文を読んで探し出していた作業を，我々の提案したシステムを用いる場合には，517 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．しかし，再現率が 81.75%であるため，78 件のうち 14 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．再現率を向上させるため，様々な方法での特徴ベクトル作成を試みた．そのひとつとして，専門用語自動抽出システムを使用して，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から専門用語とスコアを取り出し，スコアを基に作成した特徴ベクトルを作成した．その特徴ベクトルと専門用語辞典の見出し語を用いて作成した特徴ベクトルの 2 種類のベクトルを 2 つの LVQ にそれぞれ適用し学習させることで，認識率 95.52%，再現率 87.30%，適合率 13.56%という結果が得られ，再現率の向上に成功した．

さらに本論文において，原子分子データを含む論文の中で，原子分子データベースへの登録に向いている論文と向いていない論文に分類するシステムの開発を行った．このシステムでは，トレーニングデータの論文のアブストラクトに掲載されている化学式と各単語の相関関係を基にした特徴ベクトルに各アブストラクトの全文章数に対する化学式の含まれていない文章数の割合から作成した特徴ベクトルを加えたベクトルと，トレーニングデータの論文のアブストラクトに掲載されている全単語の出現頻度を用いて作成された特徴ベクトルを 2 つの LVQ に各々適用することにより，認識率 62.35%，再現率 82.01%，適合率 53.42%という結果を得た．

提案した論文分類支援モデルの評価において，本論文では，原子分子物理学分野の論文の一般的な分類の際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本モデルの有効性が得られると思われる．さらに，分類したい分野の論文のアブストラクトから専門用語を抽出・利用することが，再現率向上に有用であることを確認できた．すべての場合の評価において，機械学習法として LVQ を採用し，優れた性能を確保することが

できた．ゆえに他の機械学習法を適用しても効果があると考えられる．

謝辞

本研究を遂行するにあたり，指導教官である城和貴教授には，多大なるご指導とご協力を頂き，さらに本研究だけでなく，研究室生活においても大変お世話になりました．厚くお礼申し上げます．

また，論文のアブストラクトを提供いただきました，日本原子力研究開発機構の佐々木明氏，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 他，ご協力いただいた皆様に心より感謝いたします．

本学 高田雅美 助手には，本研究をすすめるにあたって，様々な角度からアドバイスをいただきました．ありがとうございました．

最後になりましたが，城研究室の皆様とは寝食を共にする等，とても楽しい研究生活を送ることができました．本当にありがとうございました．

参考文献

[1] 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).

[2] 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).

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

[4] Lewis, D. D.: Naive (Bayes) at Forty : independence Assumption in Information Retrieval, *Proceedings of the 10th European Conference on Machine Learning (ECML-98)*, pp. 4–15 (1998).

[5] Lewis, D. D. and Ringuette, M.: A comparison of two learning algorithms for text categorization, *In The Third Annual Symposium on Document Analysis and Information Retrieval*, pp. 81–93 (1994).

[6] 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).

[7] 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).

[8] 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).

[9] Porter, M.: An algorithm for suffix stripping, *Program*, Vol. 14, No. 3, pp. 130–137 (1980).

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

[11] 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, pp. 717–722 (2005).

[12] 加藤 隆子他: 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).

[13] *APS physics Physical Review A*. http:// pra.aps.org/.

[14] 長倉 三郎他 (編): 岩波 理化学辞典 CD-ROM 版, 岩波書店, 5 edition (1999).

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

[16] *TermExtract*. http://gensen.dl.itc.u-tokyo. ac.jp/termextract.html.

[17] 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).

[18] *Brill's Tagger*. http://research.microsoft.com/~brill/.

[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] 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).

研究業績

国際会議

- "Text Classification for Constructing an Atomic and Molecular Journal Database by LVQ",
  Hiroe Kashiwagi, Chiemi Watanabe, Akira Sasaki and Kazuki Joe,
  The 2005 International Conference on Parallel and Distributed Processing Techniques and Applications, Vol.I, pp.481–487(2005.6)

- "Design and implementation of an evolutional data collecting system for the atomic and molecular databases",
  Akira Sasaki, Kazuki Joe, Hiroe Kashiwagi, Chiemi Watanabe, Manabu Suzuki, LukasPichl, Masatoshi Ohishi, Daiji Kato, Masatoshi Kato, *4th International Conference on Atomic and Molecular Data and Their Applications*, pp.348–351(2005)

国内発表会（査読あり）

- "Learning Vector Quantization (LVQ) によるテキスト分類の試み",
  柏木裕恵，渡辺知恵美，佐々木明，城和貴，第 *11* 回数理モデル化と問題解決シンポジウム，Vol.2004，No.12 pp.103-106 (2004.10)

国内発表会（査読なし）

- "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，高田雅美，佐々木明，城和貴，
  情報処理学会第 *61* 回数理モデル化と問題解決研究会，2006-MPS-61(9)，pp.33–36(2006.9)