要素技術とその効果を用いた学術論文の自動分類

福　田　悟　志[*1]

難　波　英　嗣[*2]

竹　澤　寿　幸[*3]

[*1] ふくだ　さとし　広島市立大学大学院情報科学研究科
[*2] なんば　ひでつぐ　広島市立大学大学院情報科学研究科
[*3] たけざわ　としゆき　広島市立大学大学院情報科学研究科

Automatic Classification of Research Papers Focusing on Elemental Technologies and Their Effects

Satoshi FUKUDA
Hiroshima City University
Hidetsugu NANBA
Hiroshima City University
Toshiyuki TAKEZAWA
Hiroshima City University

We propose a method for the automatic classification of research papers in terms of the KAKEN classification index using a machine learning method. This classification index was originally devised to classify reports for the KAKEN research fund in Japan, and it is organized as a three-level hierarchy: Area, Discipline, and Research Field. Traditionally, researcher and conference names are used as cue phrases to classify the research paper efficiently. In addition to these cue phrases, we focus on elemental technologies and their effects, as discussed in each research paper. Examining the use of elemental technology terms used in each research paper and their effects is important for characterizing the research field to which a given research paper belongs. Therefore, we use elemental technology terms and their effects as additional cue phrases for machine-learning-based text classification. To investigate the effectiveness of our method, we conducted some experiments using the KAKEN and CiNii article data. From the experimental results, we obtained average precision scores of 0.853, 0.712, and 0.615 for the Area, Discipline, and Research Field levels in the KAKEN classification index, respectively. These scores are higher than those for the method not using elemental technologies and their effects. From these results, we confirmed the effectiveness of using elemental technology terms and their effects as cue phrases.

本研究では，科学研究費助成事業データベース(KAKEN)の分類体系に基づいて，学術論文を機械学習により横断的に分類する手法を提案する。これまで，論文を効率的に分類するための情報として，研究者名や雑誌名などが用いられてきた。我々は，これらの情報に加え，論文固有の特徴表現である要素技術とその効果に着目する。一般に，論文には，新しい技術(要素技術)を用いて得られた新たな研究成果(効果)が記述されている。このような研究動向を示す表現は，特定の研究分野の特徴を表す重要な手掛かりになる。提案手法の有効性を検証するために，KAKEN の研究課題データと CiNii article の論文データを用いて実験を行った。そして，KAKEN の分類体系である「分野・分科・細目表」を対象とした時，それぞれ平均 0.853，0.712，0.615 の分類精度が得られた。これらの値は，要素技術とその効果に関する表現を用いない場合より高いことから，本手法の有効性が確認された。

目　　次

1. はじめに

研究領域全般を横断した学術論文の分類は，網羅的かつ効率的な論文検索や技術動向分析などの支援を可能にする。一部の学術論文データベースでは，特定の研究分野を対象にした分類体系が考案されており，この分類体系に基づいてデータベース内の論文を人手で分類している。しかし，これから発表されていく論文や未分類のすべての論文を人手で分類することは，非常にコストがかかる。また，対象とする分類体系が改訂された時，改めて人手で論文を分類しなおすのは現実的でない。そこで本研究では，特定の分類体系に基づいた学術論文の自動分類手法を提案する。

文書分類は，自然言語処理などのデータ解析の分野における代表的な研究課題の一つであり，事前に与えられたデータ集合に基づき未知のデータを自動的に予測・分類する機械学習手法が多く提案されている[1)-4)]。本研究では，学術論文固有の特徴を用いた機械学習による分類手法を提案する。

理工系などの分野における多くの論文中では，研究課題に対して提案された新しい技術や既存技術を応用した技術，研究課題を解決するための手段などが記述されている。また，これらの技術や手段などを用いて得られた知見を，研究課題に対する成果として述べている場合が多い。そして，特定の研究課題において有用とされている技術や手段が確認された時，それらは同一あるいは近い分野の他の研究課題にも利用されることも少なくない。このような研究動向を示す技術や手段(要素技術)とそれらにより得られた知見(効果)に関する情報は，研究分野の特徴を表す重要な手掛かりになると考えられる。本研究では，要素技術とその効果を自動的に抽出し，論文を分類するための手掛かりとして利用することで，その有用性を示す。また，人文社会系のような要素技術や効果に関する情報があまり見られないような研究分野に対しても，本手法が有効に機能するかどうかを検証する。

特定の研究領域で考案された分類体系に基づく学術論文の分類に関しては，これまでにもいくつかの研究が行われている[5), 6)]。しかし本研究では，すべての研究領域を網羅した学術論文の分類を目

指している。これを実現するための第一歩として，本研究では，科学研究費助成事業データベース(KAKEN)[7]の分類体系を用いる。KAKENとは，国立情報学研究所が文部科学省，日本学術振興会と協力して作成・公開しているデータベースであり，過去に採択された67万件以上の研究課題を検索することができる。KAKENの分類体系は，理工系，人文社会系，生物系といったほぼすべての研究分野を網羅しており，研究領域によって「系・分野・分科・細目表」と呼ばれる4種類の階層に構造化されている。また，文部科学省において，分類体系の審議･改訂が年度ごとに行われており，研究分野の新設や統廃合，細分化などが実施されている。このように，KAKENの分類体系は，最新の研究領域の動向を考慮した横断的な学術論文の分類に適している。

2. 関連研究

本節では，「学術論文の自動分類」と「技術動向分析」に関する関連研究について述べる。

2.1 学術論文の自動分類

学術論文の自動分類に関して，これまでにいくつかの先行研究がある。国立情報学研究所が主催した第7回[8]および第8回[9]NTCIRワークショップ・特許マイニングタスクで実施された学術論文分類サブタスクでは，国際特許分類(International Patent Classification: IPC)と呼ばれる分類体系に基づき，学術論文を自動分類するという課題が設定された。このタスクにおいて，シャオ(Xiao, Tong)ら[10]は，文書分類における機械学習手法のひとつであるk-NN (k-Nearest Neighbor)法を用いて，任意の論文抄録に対して候補となるIPCコードリストを作成した後， IPCコードリストをリランキングする手法を提案した。

アクリタイディス(Akritidis, Leonidas)ら[11]は，計算機科学・情報技術分野を対象とした電子ジャーナルサービス，ACM Digital Library[12]で考案された分類体系(ACM CCS (Computing Classification System))[13]に基づき，学術論文を自動的に分類する手法を提案した。このタスクにおいて彼らは，研究者欄，収録刊行物欄，キーワ

ード欄といった論文のメタ情報に着目し，SVM (Support Vector Machine)を用いて分類を行った。

今井ら[14]は，岩波情報科学辞典と呼ばれる，情報科学の分野に特化した索引用語辞典に基づく学術論文の分類手法を提案した。彼らの手法は，論文の表題構造解析に基づいており，「標準化」と「コード割当て」という2つの処理から構成されている。「標準化」では，文字列処理による不要部分の削除・分割を行い，木構造を変形する。その後，単語処理による不要部分の削除・分割を行う。この処理を繰り返し，論文表題をいくつかの部分要素に分割する。「コード割当て」では，各部分要素内の専門用語を抽出し，その用語を岩波情報科学辞典の分類コードと対応付け，論文を分類する。上記で述べた研究で使用された分類体系とその研究領域，カテゴリ数，分類手法，論文項目および分類手法で用いられた手掛かり語を表1にまとめる。

表 1

アクリタイディスらの研究では，論文への候補となる研究分野は，その論文を発表した研究者または学会・出版雑誌が扱う研究分野になる可能性が高いという考えに基づいて，研究者名や学会・雑誌名が手掛かり語として用いている。このようなメタ情報を用いた文書分類に関する研究はこれまでにもいくつか行われている[15), 16)]。また，言語横断による論文の自動推薦[17)]や異種・同種データコレクションからなるデータ群からのトピックの発見[18)]など，文書分類以外の分野においても，メタ情報は有用な手掛かりとして活用されている。

しかし近年では，工学分野と農学分野の研究者による農業用ロボットの研究開発など，研究内容が大きく異なる分野間での共同研究が盛んに行われており，特定の研究分野を対象とした学術会議や論文雑誌においても様々な分野の研究課題が扱われることが多くなっている。このような専攻分野が異なる研究者により発表された論文を適切な研究分野に分類する場合，研究者名やその論文を発表した学会・雑誌名だけでは手掛かり語として不十分な可能性がある。アクリタイディスらは，論文に付与されているキーワードを手掛かり語として用いているが，その数や粒度は論文を執筆した研究者によって異なる。また，キーワードそのものが付与されていない論文も多く存在する。そ

のため，一般的には論文内容から判断する必要があるが，表題や概要においても，要素技術やその効果といったその論文を特徴付ける重要な手掛かり語が存在する。本研究では，論文内容から要素技術とその効果に関する表現を自動的に発見し，手掛かり語として用いることを行う。そして，研究者名や学会・雑誌名に加えた，文書分類に対する新たな手掛かり語としての有用性を示す。

論文中から要素技術とその効果に関する表現を発見する場合，シャオらが用いたBag-of-Wordsによる方法や，今井らの手法では困難である。Bag-of-Wordsは，単語の並び方や係り受け関係などは考慮せず，文書をモデル化する方法である。しかし，文書中のある単語が要素技術または効果を表すものかどうかを判断することはできない。さらに，効果に関する表現は論文表題では記述されないため，今井らの手法を用いて効果表現を発見することはできない。そこで本研究では，筆者ら[19]が提案した機械学習による抽出手法を適用する。次節で詳細を述べる。

2.2 技術動向分析

研究動向の解析に対していくつか先行研究があり，文内の語句の意味役割を手掛かりとしたアプローチが多く提案されている。グプタ(Gupta, Sonal)ら[20]は，研究アイディアの発展過程を調べるために,「FOCUS」「TECHNIQUE」「DOMAIN」という3種類のカテゴリに該当する語句を論文概要から自動的に抽出する手法を提案した。彼女らの手法はパターンマッチに基づいており,例えば,動詞「propose」の直後に出現する直接目的語を「FOCUS」を表す語句として抽出している。

Tateishiら[21]は，論文中に出現するモノとモノの意味関係を同定するための手法を提案した。まず，エンティティを示す語句に対して「TEAM」「OBJECT」「MEASURE」のいずれかのタグを付与し，エンティティ間に「PERFORM（動作主体)」や「CONDITION（実験条件)」など16種類の有向関係を付与したタグ付けコーパスを構築した。そして，作成したコーパスを用いてSVMによる関係抽出器を作成し，その結果に基づいて論文中の文を解析している。

難波ら[22]は，ある研究分野において，「どのよ

うな要素技術がいつ頃から使われているのか」という情報を網羅的に収集するために，論文表題から「RESTRICT」「GOAL」「METHOD」に対応する語句を抽出する手法を提案した。難波らは，各構造タグに対応する手掛かり語表現をいくつか用意し，表題中で一致する手掛かり語を構造タグに置き換えることで解析を行っている。

技術動向分析に関するこのほかの研究として，第8回NTCIRワークショップ・特許マイニングタスク[23]で実施された技術動向マップ作成サブタスクがある。これは，「要素技術」と「その効果」という観点から，論文と特許を分類した技術動向マップを作成することを目的とした研究プロジェクトである。このようなマップを作成するツールは，先行技術調査や無効資料調査の支援ツールとして利用することができる。そして，技術動向マップを自動的に作成するために，技術動向マップ作成サブタスクでは，論文や特許から要素技術とその効果を表す表現を自動的に抽出するという課題を設定している。

筆者ら[24]は，技術動向マップ作成サブタスクにおいて，機械学習に基づいた手法により，日本語論文および日本語特許から要素技術とその効果に関する表現を自動的に抽出している。そして，「論文の表題と概要に，要素技術とその効果を示すタグを付与する」という系列ラベリング問題として捉え，SVM を用いて，以下に示すタグの自動付与を行っている。

- **TECHNOLOGY**: 要素技術(新しく提案された技術，既存技術を応用した技術，研究課題を解決するための手段など(例: 協調フィルタリング，SVM))
- **EFFECT**: 効果(新しい機能の追加，新しく得られた物質，精度などの数値または増加・減少など)。EFFECT タグには，ATTRIBUTE タグと VALUE タグを含む。
- **ATTRIBUTE, VALUE**: 「速度(ATTRIBUTE)が向上(VALUE)」のように，要素技術に対する効果は「属性(ATTRIBUTE)」と「属性値(VALUE)」の対で表現する。

表題解析ではTECHNOLOGYタグのみの付与を行っている。これは，前節で述べたように，論文表題には要素技術を用いて得られた効果に関す

る記述はほとんどされないためである。以下に，「SVMを用いたChunk同定」という論文表題に<TECHNOLOGY>タグを付与した例を示す。

<TECHNOLOGY>SVM</TECHNOLOGY>を用いたChunk同定

概要の解析では，上記で示したすべてのタグを付与する。以下に，「英語の単名詞句とその他の句の同定問題にSVMを適用し，実際のタグ付けデータを用いて解析を行ったところ，従来手法に比べて高い精度を示した」という論文概要にタグを付与した例を示す。

英語の単名詞句とその他の句の同定問題に<TECHNOLOGY>SVM</TECHNOLOGY>を適用し，実際のタグ付けデータを用いて解析を行ったところ，従来手法に比べて<EFFECT><VALUE>高い</VALUE><ATTRIBUTE>精度</ATTRIBUTE></EFFECT>を示した。

上記で示したタグを自動的に付与するために，筆者らは，機械学習に用いる素性として，単語，品詞，要素技術を示す手掛かり語表現(例：を用いた，を適用し)，効果を示す手掛かり語表現(例：を示した，が可能になる)の有無，および属性，属性値を表す語句の有無を使用している。ここで，要素技術および効果を示す手掛かり語表現の収集について，これらの表現には分野依存性がそれほどないため，人手で収集することが可能である。一方で，属性や属性値を表す語句は，研究分野や研究内容によってそれぞれ異なるため，人手で収集することは容易ではない。そこで筆者らは，係り受け関係や分布類似度などの手法を用いて半自動的に収集した。さらに，論文の解析を行う際に，ドメイン適応手法を用いている。そして，第8回NTCIRワークショップ・特許マイニングタスクで提供されたデータを用いて実験を行った結果，部分一致による正解も含め，0.276 の再現率および0.539 の精度を獲得している。本研究では，このシステムを用いて解析を行う。

3. 要素技術とその効果を用いた分類手法

3.1 提案システム

本研究のタスクは，表2に示すようなKAKENの分類体系に基づき，各階層に対して論文を適切な研究分野に自動分類することである。KAKENの分類体系は，表1で示したIPCやACM CSS，岩波情報科学辞典のような特定の研究・技術分野を対象とした分類体系とは異なり，人文学，社会科学，生物学，農学，医歯薬学など，幅広い研究領域をカバーしている。なお，以降では，分類体系における「分野」を第1階層，「分科」を第2階層，「細目表」を第3階層と呼ぶ[25]。また，この分類体系への分類対象として，国立情報学研究所(NII)[26]が運営するCiNii article[27]の論文データを用いる。CiNii articleでは，幅広い研究領域で発表された論文を500万件以上収録している。

表 2

これまでの文書分類タスクでは，事前に与えられたデータ集合に基づき未知のデータを予測して分類する機械学習手法が多く提案されている。本研究でも同様に，機械学習に基づいた手法を適用し，論文を特徴付けるための表現として，研究者名，学会・雑誌名，および論文表題・概要から形態素解析により得られる単語を用いる。そして，論文表題および概要から得られた単語が要素技術または効果(属性，属性値)を表す表現かどうかを判別し，一般的な単語との区別を行うことで，研究者名，学会・雑誌名に加えた要素技術とその効果による特徴表現の有効性を示す。

ここで，論文の適切な研究分野への自動分類に対する手掛かり語表現の有用度は，それぞれ異なると考えられる。例えば，主に情報科学の分野では，「SVM」が要素技術として用いられることが多いが，「精度が向上」という効果表現は，情報科学の他に工学などの分野でも使用される場合がある。そのため，論文の自動分類において，要素技術は効果より有用性が高いと考えられる。しかし，人文社会系の分野の論文では「精度が向上」という効果表現はほとんど使用されない傾向にあるため，一定の有用性はあると考えられる。そこで本研究では，要素技術，属性，属性値を表す表現を収集した重み付き手掛かり語リストを作成し，形態素解析により得られた単語の有用度を判別する。次節で詳細を述べる。

3.2 手掛かり語の収集方法

本節では，要素技術，属性および属性値を表す手掛かり語の収集方法について述べる。まず，筆者ら[28]が提案した技術動向分析システムを用いて，KAKENの研究課題672,397件の表題および概要に対して<TECHNOLOGY>，<ATTRIBUTE>，<VALUE>タグを付与する[29]。その後，タグ付けされた語句をそれぞれ抽出し，要素技術，属性，属性値リストとしてリスト化する。

上記で述べた要素技術，属性，属性値リストと同様に，本研究では，研究者名および学会・雑誌名を収集したリストの作成を行う。これは，2.1節でも述べたように，近年では研究内容が大きく異なる分野間での共同研究が盛んに行われており，特定の研究領域を対象とした学術会議や論文雑誌においても様々な分野の研究が扱われていることを考慮している。論文の適切な研究分野への分類において，ある特定の研究領域を専門としている研究者や学術会議・論文雑誌は，多岐にわたる研究領域を対象としているものと比べて，有用な手掛かりになると考えられる。本研究では，研究者名および学会・雑誌名に付随する研究分野の数が少ない場合，分野数が多いものより重要な手掛かり語になると考え，研究分野数に閾値を設ける。そして，研究者名および学会・雑誌名において，それぞれ閾値による 2 種類のリストを構築する。そして，閾値以下の手掛かり語により構築されたリストに対する重みを，閾値より大きい値の手掛かり語から構成されたリストより高くすることで，各研究者名および学会・雑誌名に対する手掛かり語としての有用度を表現する。まず，研究課題における研究者欄から代表者，研究分担者，連携研究者を含むすべての研究者名と，研究課題の発表文献欄における各文献が掲載されたすべての学会・雑誌名を正規表現により抽出する。そして，研究者名または学会・雑誌名と，それらを抽出した研究課題に付与されている(細目表に位置する)研究分野を対応付ける。その後，手掛かり語に付随する研究分野数に対して閾値を設ける。閾値は，4.1 節で述べるチューニング方法を用いて 1 から設定し，最も高い性能を示した時の値を用いる。この時，閾値より高い値を持つ手掛かり語は重み 1 を与えることで調整する。これにより本研究で

は，2 分野以下に属する研究者名から研究者リスト 1 を構築し，それ以外は研究者リスト 2 として構築した。同様に，9 分野以下に属する学会・雑誌名から学会・雑誌リスト 1 を作成し，それ以外は学会・雑誌リスト 2 として作成した。

7種類のリストにおける手掛かり語の例，収集した語句の数，各リストに対して与える重みを表3に示す。各リストの重みは，1から50までの範囲で決定し，4.1節で述べるチューニング方法により，1または50から1ずつ重みを変更していき，最も性能が高くなるように人手で調整を行った。

表 3

さらに本研究では，共同研究者リストを作成する。これは，2.1節で述べたような研究者間における共同研究の状況に加えて，一般的に，研究者は同じ分野の研究者と共同研究を行う機会が多いことを考慮している。共同研究により発表された研究課題や論文が多いほど，その研究者らは同一の研究分野を専攻している可能性が高く，その分野を特徴付けるための重要な手掛かりとなるといえる[30), 31)]。そこで，特定の研究者間において発表された研究課題および論文の数に閾値を設け，閾値以上の値を持つ研究者間を収集したリストを作成する。この共同研究者リストは，論文の適切な研究分野への分類において，論文中の研究者欄に記述されている研究者名と関連する研究者名を手掛かり語として追加する時に用いる。まず，KAKENの研究者欄における代表者，研究分担者，連携研究者から，各研究者間における発表された研究課題の数を数え，その後，一定の閾値を設定する。閾値は，4.1節で述べるチューニング方法を用いて1から設定し，最も性能を示した時の値を用いる。この時，閾値未満の値を持つ手掛かり語はリストから除外を行うことで調整する。その結果，研究課題の本数が1本以上ある共同研究者を対象にリストの作成を行う。また，本研究では，KAKENに加えてCiNii articleからの共同研究者リストの作成を行う。CiNii articleでは，共著論文の数が5本以上ある共同研究者を対象とする。本研究では，KAKENおよびCiNii articleからそれぞれ1,094,510対および3,268,625対を獲得した。

3.3 システム構成

図1に本システムの構成を示す。提案システム

は,「索引作成モジュール」と「文書分類モジュール」から構成される。以下では，各モジュールについてそれぞれ説明する。

図 1

3.3.1 索引作成モジュール

索引作成モジュールでは，入力論文からクエリファイル$\vec{q}$を作成する。$\vec{q}$は，論文内における接頭詞を含む名詞$(w_{q1},\dots,w_{ql})$，研究者名$(a_{q1},\dots,a_{qm})$，学会・雑誌名$(p_{q1},\dots,p_{qn})$を用いてベクトル化したデータを格納したものである$(\vec{q}=(w_{q1},\dots,w_{ql},a_{q1},\dots,a_{qm},p_{q1},\dots,p_{qn}))$。この時，各ベクトルに対して，3.2 節で作成した手掛かり語リストにより重み付けを行う。

まず，名詞への重み付けについて説明する。論文概要に対して形態素解析を行い，抽出した名詞が，表 3 で示した要素技術，属性，属性値リストのいずれかに存在していれば，各リストに対応した重みを与える。もし，どのリスト内にも存在していなければ重み 1 を与える。その後，表題に対して形態素解析を行い，抽出した語句が表 3 の要素技術リストに存在していれば重み 17 を，存在していなければ 1 を与える。これは，語句が出現する文書内の項目に応じて重みを変えることは有効であることが報告されているためである [32), 33)]。なお，重みの決定は，4.1 節で述べるチューニング方法により，要素技術リストの重みから 1 ずつ値を変更し,最も性能の高くなるように調整した。

次に，研究者名および学会・雑誌名への重み付けについて述べる。まず，論文の研究者欄から研究者名を，収録刊行物欄から学会・雑誌名を正規表現により抽出する。その後，抽出した研究者名と関連する研究者名を，KAKEN および CiNii article の共同研究者リストから抽出する。そして，研究者欄および共同研究者リストから抽出した研究者名が表 3 で示した研究者リスト 1 または 2 に含まれていれば,リストに対応した重みを与える。また，抽出した学会・雑誌名が学会・雑誌リスト 1 または 2 に含まれていれば，リストに対応した重みを与える。抽出した研究者名または学会・雑誌名がリストに存在しない場合，その論文の特徴を表す手掛かり語ではないとみなし，クエリファイルの作成に用いない。

KAKEN の研究課題からインデックスファイル

$\vec{d}$を作成する際も，上記で述べた手順を適用する。この時，学会・雑誌名は発表文献欄における各文献から抽出する。また，名詞を抽出する時，表題，概要のほかにキーワードも対象とする。研究課題では，それに関連する技術用語(ツール，モデル)などがキーワードとして選定されており，要素技術となる手掛かり語が多く含まれていると考えられる。まず，各キーワードに対して形態素解析を行い，抽出した名詞が要素技術リストに存在していれば，表 3 に従い重み 14 を，そうでなければ重み 1 を与える。また，共同研究者リストは用いない。これは，インデックスファイルの作成において，共同研究者リストを用いた場合と用いない場合よる予備実験の結果から判断した。

3.3.2 文書分類モジュール

文書分類モジュールでは，文書分類タスクにおいて一般的に適用されており，また，表 1 で述べた既存研究でも使用されている k-NN 法および SVM 手法という 2 種類の分類手法を用いる。各分類手法について，以下で詳細を述べる。

k-NN 法

k-NN 法とは，訓練用データに含まれる情報を用いて，入力文書がどのようなカテゴリに属するのかを自動的に予測するための機械学習アルゴリズムである。k-NN 法は，「類似度計算」と「ランキング」という 2 つのステップから構成される。まず，入力論文と訓練用データ内の研究課題との類似度を計算する。その後，類似度に基づき，候補となる上位k件の研究課題を選択する。そして，それらに付与されている研究分野をランキング手法により選択することで，入力論文に対する適切な研究分野を決定する。以下では，類似度計算およびランキング手法について詳細を述べる。

- 類似度計算

まず，汎用連想計算エンジン GETA[34]を用いて，入力論文と研究課題間の類似度を測定する。その後，類似度が高い順に研究課題をソートし，上位k件を選択する。類似度計算には，GETA のライブラリで提供されている SMART[35]を用いる。

- ランキング

まず，類似度計算により選択された上位k件の

研究課題に基づく研究分野のスコア$Score(c)$を計算する。$Score(c)$は，論文に研究分野cが付与される可能性の尺度を表す。そして，$Score(c)$が高い順に研究分野をソートし，研究分野リストとして出力する。$Score(c)$の算出に対して，本研究では，シャオらの研究[36]で用いられた4種類のランキング手法を適用する。そして，本システムにおいて有用なランキング手法を調べる。

(1) Naïve

Naïve は，最も類似度が高い研究課題に付与されている研究分野から順に決定していくランキング手法である。

(2) Sum

Sum は，研究分野cにおける，入力論文と研究課題間の類似度の総和を算出し，そのスコアに基づいて候補となる研究分野を決定するランキング手法である。以下の(1)式を用いて算出する。

$$Score_{Sum}(c) = \sum_{i=1}^{k} occur(c, d_i)\, Sim(q, d_i) \quad (1)$$

$occur(c, d_i)$は，研究分野cが研究課題d_iに付与されているかどうかを示す関数を表す。もし，付与されていれば 1，付与されていなければ 0 となる。$Sim(q, d_i)$は，入力論文qとd_i間の類似度を表す。

(3) Listweak (List)

Listweak は，上記で述べた Sum 手法に基づくランキング手法であるが，この手法では，入力論文との類似度が低い研究課題はノイズであると仮定し，より類似度の高い研究課題を強調する。以下に述べる(2)式により算出される。

$$Score_{Listweak}(c) = \sum_{i=1}^{k} occur(c, d_i)\, Sim(q, d_i) r_1^i \quad (2)$$

r_1は，より類似度の低い研究課題に対してペナルティを与えるパラメータを表す$(0 < r_1 < 1)$。本研究では，シャオらの研究に従い，0.95 と設定する。

(4) Weak

k-NN法の欠点として，訓練用データ内の各研究分野が持つ研究課題の数に偏りが大きいほど，入力論文に対する研究分野の予測において，その研究分野が選ばれやすくなることが挙げられる。Weakは，このような分野間の偏りを考慮するラ

ンキング手法である。スコアは(3)式で算出される。

$$Score_{Weak}(c) = \sum_{i=1}^{k} occur(c_i, d_i) Sim(q, d_i) r_2^{crank(c,i)\times\frac{size(c)}{k}} \quad (3)$$

$size(c)$は，上位k件の研究課題集合におけるcの数を表し，$crank(c,i)$は，その研究課題集合における上位$i-1$件までのcの出現頻度を表す。r_2は，より類似度の低い研究課題に対してペナルティを与えるパラメータを表す$(0 < r_2 < 1)$。本研究では，シャオらの研究に従い，0.90と設定する。

SVM 手法

次に，SVMを用いた分類手法について述べる。SVM は，2 値分類のための教師あり学習アルゴリズムである。マージン最大化による識別平面の決定により高い汎化性能を持ち，様々なソースデータをモデリングする場合において，その柔軟性が優れていることなどから，パターン認識の分野をはじめ,様々な分野で広く用いられてきている。

まず,3.3.1 節で作成したインデックスファイル集合を用いて，各階層内の研究分野を表す分類器を SVM により作成する。この時，すべてのインデックスファイル内に存在する名詞，研究者名，学会・雑誌名およびそれらに付随する重みを，SVM で用いる表現および重みとして使用する。次に，各分類器に対して，3.3.1 節で作成した入力論文を表すクエリファイルを適用する。この時，クエリファイル内の名詞，研究者名，学会・雑誌名およびそれらに付随する重みを,SVM で用いる表現および重みとして使用する。そして，各分類器において入力論文との超平面間の距離を測り，その距離の値が最も小さかった分類器が表す研究分野から順に出力する。

4. 実験

提案手法の有効性を調べるための実験を行った。4.1 節で実験データについて，4.2 節で評価尺度について，4.3 節で比較手法について，それぞれ説明する。また，4.4 節で実験結果を報告し，4.5 節で結果を考察する。

4.1 実験データ

KAKEN

本研究では,KAKENの研究課題を訓練用データとして用いる。この時，表題，研究概要，キーワード，研究者欄，発表文献欄，および2011年度に使用可能な分類体系(表2)において，第3階層に位置する研究分野が付与されている研究課題を対象とする。また，第3階層の各研究分野におけるデータ数の偏りを無くすために，1種類の研究分野に対して200件の研究課題を用いる。そして，第3階層に位置する研究分野に基づき，その上位となる第1，第2階層の研究分野を各研究課題に付与する。例えば，表2の分類体系に基づき，知能情報学が付与されている研究課題に対して,総合領域，情報学という研究分野を新たに付与する。その結果，28,400件の研究課題および各階層における研究分野のうち，第1階層では10分野，第2階層では44分野，第3階層では142分野を本実験で使用する。ここで，第1，第2階層における各研究分野の研究課題の数には偏りがあることに注意する。

3.2節で述べた研究者リスト,学会・雑誌リスト,共同研究者リストの構築では，本実験で対象とする研究分野が付与されている283,686件の研究課題データを用いた。また，3章で述べた手掛かり語リストへの重み付けや閾値の決定について，本研究では，訓練用データを除く2,000件の研究課題をチューニング用データセットとして作成し，3.3.2節で述べたKNN(Sum)手法に基づき，精度@1(4.2節で述べる)において最も性能が高くなるように調整した。

CiNii article

CiNii articleは,2012年までに5,924,679件の論文データを収録しており，それらは主に，表題，概要，研究者欄，収録刊行物欄から構成されている。このうち，概要を含む1,000件の論文データ(Abstデータセット),および概要を含まない1,000件の論文データ(Titleデータセット)を評価用データとして使用する[37]。なお，訓練用データ内の発表文献欄に存在する論文データは用いない。また，評価用データで扱う研究分野は訓練用データで用いているものを対象とし，第3階層における1種類の研究分野に対して20件の論文データ(Abst: 10件，Title: 10件)を用いる。そして，各データに付

与されている研究分野の上位となる第1，第2階層の研究分野も付与する。その結果，評価用データでは，第1，第2，第3階層において10分野，39分野，100分野を対象とする。3.2節で述べた共同研究者リストの構築では，すべての論文データ(5,924,679件)を用いた。

4.2 評価尺度

システムが評価用データに自動付与した研究分野と評価用データに付与されている研究分野が一致した時，正解とする。本研究では，これを精度として評価する((4)式)。

$$\text{精度} = \frac{1}{K}\sum_{i=1}^{K} t_i \qquad (4)$$

ここで，Kは，評価用データの数を表しており，t_iでは，システムにより付与された上位i件の研究分野のうち，正解となる研究分野が出現していればt_i=1，出現しなければt_i=0を表す。

また，MRR (Mean Reciprocal Rank)による評価も行う。そして，評価用データがそれに付与されている研究分野に適切に分類されているかを上位i件までの研究分野候補を見ることで判断する。

本研究では，評価用データに自動付与された研究分野のうち，上位1件(@1)，2件(@2)，3件(@3)までのものを対象として精度を算出する。また，MRR値の算出では，システムが出力した上位3件までの研究分野を対象とする。

4.3 比較手法

以下に述べる 5 種類の提案手法と 3 種類の比較手法を用いて実験を行った。なお，比較手法における k-NN 法では，実験を通して全体的に精度が高かった B_KNN(List)および B_KNN(Weak)手法を記載する。また，k-NN 法における上位k件の決定について，チューニングデータセットを用いてkの値を 1 から 50 まで 1 刻みで設定し，各実験条件において最も精度の高かった時の値を用いた。なお，形態素解析には MeCab[38]を用いた。また，SVM による機械学習パッケージは TinySVM[39]を用い，カーネル関数は線形カーネルを使用した。

提案手法

- **KNN(Naive)**: 入力論文との類似度が高かっ

た研究課題に付与されている研究分野から順に決定する。

- **KNN(Sum)**: 各研究分野における，入力論文と研究課題間の類似度の総和を算出し，そのスコアに基づいて候補となる研究分野を決定する。
- **KNN(List)**: 各研究分野における,入力論文と研究課題間の類似度の総和を算出し，最もスコアが高かった研究分野から順に決定する。この時，入力論文との類似度が低い研究課題にペナルティを与える。
- **KNN(Weak)**: KNN(List)に基づく手法であるが，訓練用データセットにおける研究分野間の文書数の偏りを考慮する。
- **SVM**: 各分類器において，入力論文に対する超平面の距離を測定し，最も距離が小さい結果を示した分類器が表す研究分野から順に決定する。

比較手法

- **B_KNN(List)**: KNN(List)手法において，要素技術とその効果(属性，属性値)に対応する表現を用いない。
- **B_KNN(Weak)**: KNN(Weak)手法において，要素技術とその効果に対応する表現を用いない。
- **B_SVM**: SVM 手法において，要素技術とその効果に対応する表現を用いない。

4.4 実験結果

第1，第2，第3階層の研究分野を対象にした時の実験結果を表4，表5，表6にそれぞれ示す。Ave.は，各データセット(Title，Abst)を用いて算出した評価値を平均した値(マクロ平均)を示している。なお，すべての実験条件において，合計2,000件のデータセットに対して1種類以上の研究分野が付与された。表4から表6の結果から，第1階層から第3階層において，出力結果の上位1件までを正解とした場合，KNN(List)手法により，平均で最大0.853，0.712，0.615の精度が得られ，MRRでは，0.909，0.800，0.711の平均値を示した。また，k-NN法およびAbstデータセットを用いた時のSVMの結果から，研究者名および学会・雑誌名のみを手掛かり語として用いた場合では正しく分

表 4
表 5
表 6

類できなかった論文を，要素技術とその効果を手掛かり語として加えることで改善できたことが分かった。

また，提案手法と比較手法において全体的に性能が高かったKNN(List)手法とB_KNN(Weak)手法に対して，t検定による統計的有意差検定を行ったところ，Abstデータセットを対象とした時，第3階層におけるすべての条件において有意水準1%で有意差があることが確認された。さらに，第2階層における上位2件の研究分野を正解対象とした時，有意水準5%で有意差が確認された。これらの結果から，本手法における要素技術とその効果を用いることの有効性を示せたといえる。

4.5 考察

4.5.1 各研究分野に対する要素技術とその効果の有効性

本節では，各研究分野において，要素技術とその効果に関する表現を手掛かり語として用いることで，精度がどのくらい向上したのかについて述べる。ここでは，最も一般的な研究内容を扱う第1階層の10分野を対象に，表4における上位1件でのAve.が最も高かったKNN(Weak)手法と，同様のランキング手法を用いているB_KNN(Weak)手法の実験結果を調べた。各手法における詳細結果を表7に示す。表7では，評価用データにおける各研究分野の論文数および正解件数を示している。

表 7

表7から，工学や化学において，精度が向上していることが分かる。特に工学では，Abstデータセットにおける正解件数が231件から245件へと大幅に向上している。これは，本研究で用いた技術動向分析システム[40]は元々，理工系の分野を対象に構築されており，研究課題から各分野の特徴となる要素技術とその効果に関する表現を多く抽出できたためと考えられる。

次に，人文学と社会科学の分野に対する正解件数を比較する。表7を見ると，要素技術とその効果に関する表現を手掛かり語として用いることで，人文学ではTitleおよびAbstデータセットにおいて正解件数がそれぞれ4件，2件増えており，同様に，社会科学では正解件数がそれぞれ6件増えていることが分かる。ここで，特に正解件数が増加した社会科学において，どのような語句が要素技術ま

たは効果であると判断されているのかについて調べた。その結果，「質問紙法」や「情報公開法」などが要素技術，「教育水準」や「回収率」などが効果とみなされていることが分かった。実際にこれらの要素技術が抽出された論文を見ると，主に，研究課題に対する問題を調査・解決するための手段として用いられていた。これらの結果から，本手法は，理工系だけでなく，人文社会系の分野においても，一定の効果があると考えられる。

また，総合領域と複合新領域における実験結果について考察する。総合領域および複合新領域には，人文社会系，理工系，生物系のうち 2 つ以上の系をまたがる学際的な研究分野が含まれている。表 7 を見ると，複合新領域では，比較手法と提案手法では性能に変化がなかった。しかし，総合領域では，Title および Abst データセットにおいて，正解件数がそれぞれ 1 件，3 件増えている。これは，複合新領域は，2003 年度の分類体系の大幅な改訂において新設された比較的新しい研究分野であり，複合新領域に対する特徴的な要素技術や効果が少なかったためと考えられる。一方で，総合領域では，2003 年度以前の分類体系において，複合領域と呼ばれる分野に含まれていた研究分野が多く扱われている。そのため，KAKEN における長い研究期間において多くの確立された技術や手法が開発され，学際的な研究領域に対する特徴的な知見が得られていたためと考えられる。

要素技術とその効果を用いることの有効性をさらに確かめるために，本研究では，より専門的な研究内容を扱う第3階層の研究分野を対象に，精度がどのように変化したのか調べた。ここでは，表6における上位1件でのAve.が最も高かったKNN(List)手法とそれに対応する比較手法であるB_KNN(List)手法の実験結果を比較した。その結果，本手法を用いることで精度が向上した研究分野は42分野であり，精度が低下した分野は16分野であることが分かった[41]。また，表8に，各研究分野に対する詳細結果の一部を示す。上段では，比較手法より精度が向上した研究分野の例を示し，中段では，比較手法より精度が低下した例を示している。下段では，KNN(List)手法を用いた時，最も精度が低かった研究分野の例を示している。

表 8

ここで，B_KNN(List)手法より精度が低下した

研究分野(無機材料・物性，基礎獣医学・基礎畜産学，細菌学(含真菌学))およびKNN(List)手法において最も精度が低かった研究分野(応用物理学一般，神経科学一般，生物系薬学) について詳しく見ていく。本研究では，論文に対してシステムが誤って付与した研究分野の傾向について調べた。上記の研究分野が付与されている評価用データに対して，KNN(List)手法により誤って付与された研究分野の例を表9に示す。括弧内の数値は，システムが誤って付与した研究分野における論文の数を表す。

表 9

まず，無機材料・物性の結果を見ると，金属生産工学，金属物性，材料加工・処理など，工学関連の様々な研究分野が誤って付与されていることが分かる。同様に，細菌学(含真菌学)では耳鼻咽喉科学や病態検査学，生物系薬学では呼吸器内科学や循環器内科学など，それぞれ医歯薬学に関連する研究分野が誤って付与されていることが分かる。一方で，基礎獣医学・基礎畜産学の論文に誤って付与された研究分野を見ると，主に応用獣医学であると判断されていることが分かる。同様に，応用物理学一般の論文には，応用光学・量子光工学が誤って付与されていることが分かる。なお，基礎獣医学・基礎畜産学と応用獣医学は，分類体系の第2階層における畜産学・獣医学に属しており，応用物理学一般と応用光学・量子光工学は，第2階層の応用物理学・工学基礎に属している。また，各分野間の研究内容は比較的類似している。

本研究では，基礎獣医学・基礎畜産学，応用獣医学，応用物理学一般および応用光学・量子光工学において，訓練用データ内でどのような語句が要素技術または効果として用いられているのか調べた。表10に，4種類の研究分野における要素技術とその効果の例を示す。括弧内の数値は，訓練用データ内において要素技術または効果が出現した研究課題数を示す。表10から，基礎獣医学・基礎畜産学と応用獣医学では，要素技術を表す用語として「マウス」や「ウイルス」が，効果を表す用語として，「細胞」や「活性」がそれぞれ主に用いられていることが分かった。また，応用物理学一般と応用光学・量子光工学では，「レーザー」や「レンズ」が要素技術として，「特性」や「周波数」が効果として頻出していることが分かった。

これらの結果から，研究内容が類似している分野間の特徴をより詳細に捉えるために，各研究分野における要素技術と効果の出現傾向を考慮した類似度計算を考案するといった新たな枠組みが必要と考えられる。

表 10

4.5.2 本手法の分類精度に対する要素技術とその効果の有効性

次に，要素技術とその効果に関する表現をそれぞれ単独に用いた場合，どのように分類精度が変化するのか調べる。ここでは，表4から表6において，全般的に高い精度とMRRを示しているKNN(List)手法を対象に調査する。B_KNN(List)手法に対して，表題中の要素技術，概要中の要素技術および概要中の効果を表す表現をそれぞれ単独に加えた時の結果を表11に示す。表11では，Abstデータセットを用いて実験を行い，第1，第2，第3階層の研究分野を対象とした時の結果を記載している。また，KNN(List)およびB_KNN(List)手法の結果も記載する。表11から，概要中の効果を表す表現のみを用いた場合，ベースライン手法と比べて精度およびMRR値は上回っているが，一方で，要素技術を表す表現のみを用いた場合，ベースライン手法とほとんど変わらない結果を示す場合があることが分かる。しかし，要素技術とその効果を表す表現を手掛かり語としてすべて用いることで，効果表現のみを用いた場合の性能から，さらに上回る結果を示すことが分かった。

表 11

本研究では，表題および概要中の語句が要素技術，属性または属性値であるかどうかを決定するために，技術動向分析システム[42)]を用いて<TECHNOLOGY>，<ATTIBUTE>，<VALUE>タグをKAKENの研究課題に付与し，タグ付けされた語句の抽出・リスト化を行っている。そして，その語句が各リストのいずれかに存在するかどうかを調べることで手掛かり語として判定している。しかし，2.2節でも述べたように，本研究で用いた分析システムの精度は0.5393と決して高いとはいえない。そのため，要素技術または効果でない語句を誤って手掛かり語として抽出してしまい，その結果，誤った重みを与えた可能性がある。

本研究で構築した各リストを調べると，要素技術リストでは，「導入」「解明」「再構築」など，

属性または属性値である方がふさわしいと考えられる語句や，「予定」「観点」「こと」など，手掛かり語としてふさわしくない語句も含まれていた。同様に，属性リストでも，「ナノコンポジット」「支援」「それぞれ」など，要素技術，属性値またはそれらのいずれにも属さないと考えられる語句が含まれていた。属性値リストにおいても，「ニュース」「レシピエント」「スイッチ」「ビジュアル」など，属性値とは考えにくい用語が含まれていた。しかし，各リスト内における，これらの語句をタグ付けしている文書の数は低い傾向にある。実際，要素技術リストにおいて，「熱処理」という語句は870文書から抽出されているのに対し，「予定」や「再構築」といった要素技術とは考えにくい用語は，それぞれ3文書，13文書のみから抽出されていた。この問題は，技術動向分析システムの精度向上や，各リスト内の手掛かり語に対して一定の閾値を設定するといった処理により改善すると考えられる。

5. システムの動作例

本章では，提案手法を用いて構築したシステム，CiNii Miningついて説明する。本システムでは，CiNii論文検索API[43]から取得した論文データを対象に自動分類している。

図2は，「音声認識」をクエリとして入力した時の検索結果を示している。この時，本手法を用いて計算されたスコアが最も高かった研究分野に基づいて学術論文を分類している。また，近年において，KAKENの分類体系では網羅されていないような学際的な研究が増加していることを考慮し，最もスコアが高かった研究分野の他に，スコアが2番目および3番目に高かった研究分野をその他の研究分野候補として提示している。図2において，「音声認識」というキーワードで検索された論文が，知能情報学や教育工学など，様々な研究分野に分類されていることが分かる。また，本システムは，KAKENの分類体系に従い，論文の分類レベルを選択することができる。

図 2

図3は，クエリ「音声認識」に対する，知能情報学の技術動向マップを示している。図3の左側では，「音声認識」に関する，論文内で使用され

た要素技術を列挙している。また，中央では，各技術が使われた年を示しており，右側には，要素技術を用いて得られた効果を示している。このような技術動向マップから，例えば，2004年に「GMM (Gaussian Mixture Model)」が要素技術として用いられており,「認識性能を大幅に改善」という効果を得ていることを読み取ることができる。これらの要素技術や効果に関する表現は，CiNii articleの各論文に対して，技術動向分析システム[44]を用いて自動的に抽出を行っている。また，図中の○にカーソルを重ねることで，その論文の書誌情報がポップアップウィンドウ内に表示される。さらに，○をクリックすることで，その論文におけるCiNiiのページにアクセスすることができる。

図 3

6. おわりに

本稿では，研究領域全般を横断した学術論文の自動分類手法を提案した。本手法は機械学習に基づいており，研究者名や学会・雑誌名に加え，論文中で記述されている要素技術とその効果に関する表現を手掛かり語として用いた。そして，KAKENの分類体系である「分野・分科・細目表」を対象に評価実験を行った結果，各階層における上位1件の結果に対して，KNN(List)手法により，それぞれ平均0.853，0.712，0.615の精度が得られた。また，上位3件までの出力結果に対して，同様の手法により，平均で0.909，0.800，0.711のMRR値が得られた。これらの結果は，要素技術とその効果に関する表現を手掛かり語として用いない場合より高い値を示していることから，本手法の有効性が確認された。

注・引用文献

1) Akritidis, Leonidas. and Panayiotis Bozanis. “A Supervised Machine Learning Classification Algorithm for Research Articles,” *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, 2013, p. 115-120.
2) McCallum, Andrew, Kamal Nigam, Jason Rennie, and Kristie Seymore. “A Machine Learning Approach to Building

Domain-Specific Search Engines," *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, 1999, p. 662-667.

3) 宮田洋輔，石田栄美，神門典子，上田修一「NDCの階層構造を利用した図書の自動分類の試み」『日本図書館情報学会春季研究集会発表要綱』2006, p. 51-54.

4) Rocha, Leonardo, Fernando Mourão, Hilton Mota, Thiago Salles, Marcos André GonçAlves, and Wagner Meira Jr. "Temporal Contexts: Effective Text Classification in Evolving Document Collections," *Information Systems*, Vol. 38, No. 3, 2013, p. 388-409.

5) 前掲1)

6) Xiao, Tong, Feifei Cao, Tianning Li, Guolong Song, Ke Zhou, Jingbo Zhu, and Huizhen Wang. "KNN and Re-ranking Models for English Patent Mining at NTCIR-7," *Proceedings of the 7th NTCIR Workshop Meeting*, 2008, p. 333-340.

7) KAKEN - 科学研究費助成事業データベース. http://kaken.nii.ac.jp/

8) Nanba, Hidetsugu, Atsushi Fujii, Makoto Iwayama, and Taiichi Hashimoto. "Overview of the Patent Mining Task at the NTCIR-7 Workshop," *Proceedings of the 7th NTCIR Workshop Meeting*, 2008, p. 325-332.

9) Nanba, Hidetsugu, Atsushi Fujii, Makoto Iwayama, and Taiichi Hashimoto. "Overview of the Patent Mining Task at the NTCIR-8 Workshop," *Proceedings of the 8th NTCIR Workshop Meeting*, 2010, p. 293-302.

10) 前掲6)

11) 前掲5)

12) ACM Digital Library. http://dl.acm.org/

13) ACM Computing Classification System ToC. http://www.acm.org/about/class/

14) 今井俊，佐藤理史「表題解析による科学技術論文の自動分類」『情報処理学会第 57 回全国大会講演論文集』1998, p. 211-212.

15) 前掲2)

16) 前掲3)

17) Uchiyama, Kiyoko, Hidetsugu Nanba, Akiko Aizawa, and Takeshi Sagara.

"OSUSUME: Cross-lingual Recommender System for Research Papers," *Proceedings of the 2011 Workshop on Context-awareness in Retrieval and Recommendation*, 2011, p. 39-42.
18) 坂本剛彦, ホーツーバオ「データコレクション間の類似度を利用したトピック発見手法の提案」『電子情報通信学会第19 回データ工学ワークショップ』2008.
19) 福田悟志, 難波英嗣, 竹澤寿幸「論文と特許からの技術動向情報の抽出と可視化」『情報処理学会論文誌データベース』Vol. 6, No. 2, 2013, p. 16-29.
20) Gupta, Sonal. and Christopher D. Manning. "Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers," *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, 2011.
21) Tateisi, Yuka, Yo Shidahara, Yusuke Miyao, and Akiko Aizawa. "Annotation of Computer Science Papers for Semantic Relation Extraction," *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014, p. 26-31.
22) 難波英嗣, 谷口裕子「学術論文データベースからの研究動向情報の抽出と可視化」『言語処理学会 第12回年次大会 併設ワークショップ－言語処理と情報可視化の接点－』2006, p. 35-38.
23) 前掲9)
24) 前掲19)
25) 本実験では, 2011年度のKAKENの分類体系を用いるが, この分類体系は2013年度から大幅な改定が実施されており, その構造は大きく変化している。また, 2011年度の分類体系(「系・分野・分科・細目表」)における「系」では, 総合・新領域系, 人文社会系, 理工系, 生物系の4分野しか扱われていないため, 本実験では分類対象としなかった。
26) 国立情報学研究所. http://www.nii.ac.jp/
27) CiNii article. http://ci.nii.ac.jp/
28) 前掲24)
29) KAKENの研究課題および学術論文における表題および概要の記述形式は類似しており, 技術動向分析システムによる抽出性能はほぼ同等であると判断した。
30) Hoche, Sumanne. and Peter Flach.

"Predicting Topics of Scientific Papers from Co-Authorship Graphs: a Case Study," *Proceedings of the 2006 UK Workshop on Computational Intelligence (UKCI2006)*, 2006, p. 215-222.
31) Zhang, Xiaodan, Xiaohua Hu, and Xiaohua Zhou. "A Comparative Evaluation of Different Link Types on Enhancing Document Clustering," *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2008, p. 555-562.
32) Larkey, Leah S. "Some Issues in the Automatic Classification of U.S. Patents," *Working Notes for the AAAI-98 Workshop on Learning for Text Categorization*, 1998, p. 87-90.
33) 間瀬久雄，辻洋，絹川博之，石原正博「特許テーマ分類方式の提案とその評価実験」『情報処理学会論文誌』Vol. 39, No. 7, 1998, p. 2207-2216.
34) 汎用連想計算エンジン(GETA) 公開HP. http://geta.ex.nii.ac.jp/geta.html
35) Salton, Gerard. "The SMART Retrieval System – Experiments in Automatic Document Processing," *Prentice-Hall, Inc., Upper Saddle River, NJ*, 1971.
36) 前掲10)
37) CiNii articleには，概要が存在しない論文データが約21%含まれており，表題のみを含む論文を対象とした性能評価は必要であると考えられる。
38) MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html?sess=3f6a4f9896295ef2480fa2482de521f6
39) TinySVM: Support Vector Machines. http://chasen.org/~taku/software/TinySVM/
40) 前掲28)
41) これらの結果は，TitleおよびAbstデータセットにおける各手法の比較結果を統合したものから判断している。
42) 前掲40)
43) CiNii article - メタデータ・API - CiNii article 論文検索の OpenSearch . http://support.nii.ac.jp/ja/cia/api/a_opensearch
44) 前掲42)

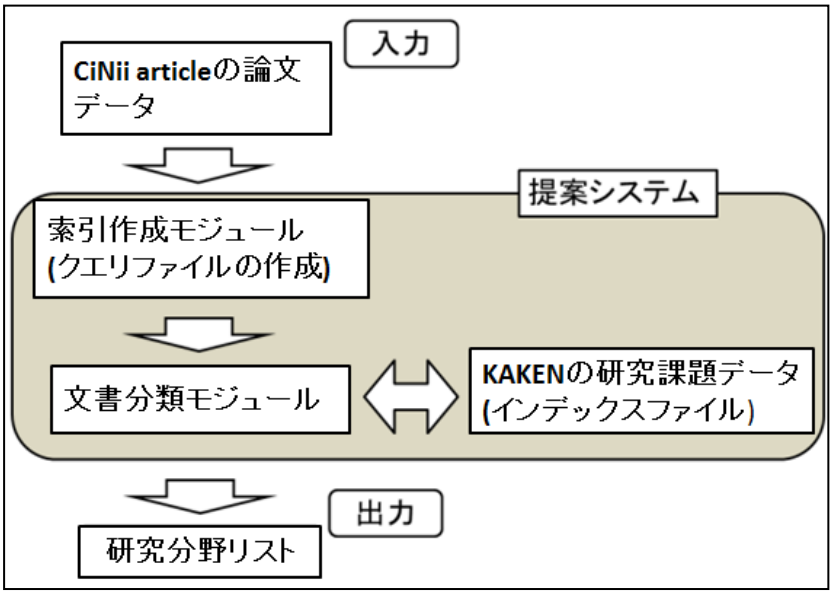

図１　システム構成

図2 クエリ「音声認識」に対する検索結果画面

図3　「音声認識」に対する技術動向マップ画面

表 1　学術論文の自動分類における既存研究で使用された分類体系，研究領域，カテゴリ数，分類手法，論文項目，手掛かり語

	分類体系	研究領域	カテゴリ数	分類手法	論文項目	手掛かり語
シャオら	IPC	生活必需品 処理操作 科学 繊維 機械工学 物理学	30,885件 (第5階層)	k-NN リランキング	表題 概要	Bag-of-Words
アクリタイディスら	ACM CSS	計算機科学 情報技術	276件 (第3階層)	SVM	研究者欄 収録刊行物欄 キーワード欄	研究者名 学会・雑誌名 キーワード
今井ら	岩波情報科学辞典	情報科学	約4,500件	標準化 コード割当て	表題	表題中の専門用語

表 2　KAKEN の分類体系(2011 年度)の例

分野 (第 1 階層)	分科 (第 2 階層)	細目表 (第 3 階層)
総合領域	情報学	知能情報学, ソフトウェア, 図書館情報学・ 人文社会情報学
	教育工学・ 科学教育	教育工学, 科学教育
人文学	言語学	言語学, 日本語教育
	人文地理学	人文地理学
工学	機械工学	機械力学・制御, 知能機械学･機械システム
	電気電子工学	システム工学, 計測工学, 制御工学
	総合工学	航空宇宙工学, 原子力学

表 3　各リストに属する手掛かり語の例

	手掛かり語の例	手掛かり語の数	重み
研究者リスト1	荒牧英治	144,108	50
研究者リスト2	川原稔	15,567	1
学会・雑誌リスト1	教育情報学会	125,232	40
学会・雑誌リスト2	情報処理学会	4,352	4
要素技術リスト	SVM，CRF	430,920	14
属性リスト	精度，安定性	296,152	6
属性値リスト	向上，抑制	70566	3

表 4　第 1 階層における研究分野を対象にした場合の精度および MRR の結果

		精度									MRR		
		@1			@2			@3					
		Title	Abst	Ave.	Title	Abst	Ave.	Title	Abst	Ave.	Title	Abst	Ave.
提案手法	KNN(Naïve)	0.778	0.832	0.805	0.913	0.953	0.933	0.952	0.976	0.964	0.859	0.900	0.880
	KNN(Sum)	0.822	**0.878**	0.850	**0.930**	0.964	**0.947**	0.957	**0.986**	0.972	0.888	0.927	0.908
	KNN(List)	0.827	0.877	0.852	0.925	**0.966**	0.946	0.962	0.984	**0.973**	**0.889**	**0.928**	**0.909**
	KNN(Weak)	**0.828**	0.877	**0.853**	0.927	0.964	0.946	0.958	0.985	0.972	0.888	0.927	0.908
	SVM	0.737	0.815	0.776	0.835	0.892	0.864	0.873	0.938	0.906	0.799	0.869	0.834
比較手法	B_KNN(List)	0.815	0.852	0.834	0.921	0.944	0.933	0.958	0.976	0.967	0.880	0.909	0.895
	B_KNN(Weak)	0.813	0.853	0.833	0.924	0.944	0.934	**0.964**	0.980	0.972	0.882	0.911	0.897
	B_SVM	0.750	0.777	0.764	0.854	0.878	0.866	0.892	0.910	0.901	0.815	0.838	0.827

表 5　第 2 階層における研究分野を対象にした場合の精度および MRR の結果

		精度									MRR		
		@1			@2			@3					
		Title	Abst	Ave.	Title	Abst	Ave.	Title	Abst	Ave.	Title	Abst	Ave.
提案手法	KNN(Naïve)	0.623	0.688	0.656	0.782	0.849	0.816	0.852	0.905	0.879	0.726	0.787	0.757
	KNN(Sum)	0.667	0.753	0.710	0.811	0.871	0.841	0.879	0.918	**0.899**	0.760	0.824	0.792
	KNN(List)	0.676	0.744	0.710	**0.828**	**0.880**	**0.854**	0.872	**0.925**	**0.899**	**0.767**	**0.832**	**0.800**
	KNN(Weak)	0.670	**0.754**	**0.712**	0.810	0.872	0.841	**0.880**	0.916	0.898	0.763	0.829	0.796
	SVM	0.572	0.685	0.629	0.663	0.781	0.722	0.709	0.822	0.766	0.633	0.747	0.690
比較手法	B_KNN(List)	**0.677**	0.715	0.696	0.815	0.844	0.830	0.866	0.902	0.884	0.763	0.800	0.782
	B_KNN(Weak)	0.672	0.717	0.695	0.812	0.853	0.833	0.874	0.911	0.893	0.764	0.805	0.785
	B_SVM	0.591	0.637	0.614	0.710	0.753	0.732	0.764	0.803	0.784	0.669	0.712	0.691

表 6　第 3 階層における研究分野を対象にした場合の精度および MRR の結果

		精度									MRR		
		@1			@2			@3					
		Title	Abst	Ave.	Title	Abst	Ave.	Title	Abst	Ave.	Title	Abst	Ave.
提案手法	KNN(Naïve)	0.523	0.583	0.553	0.693	0.747	0.720	0.758	0.811	0.785	0.630	0.686	0.658
	KNN(Sum)	0.569	0.636	0.603	0.725	0.790	0.758	0.787	0.837	0.812	0.670	0.732	0.701
	KNN(List)	**0.588**	0.641	**0.615**	**0.744**	**0.800**	**0.772**	**0.790**	**0.852**	**0.821**	**0.683**	**0.738**	**0.711**
	KNN(Weak)	0.570	**0.642**	0.606	0.730	0.790	0.760	0.789	0.845	0.817	0.671	0.732	0.702
	SVM	0.527	0.607	0.567	0.634	0.708	0.671	0.666	0.765	0.716	0.591	0.677	0.634
比較手法	B_KNN(List)	0.574	0.603	0.589	0.724	0.733	0.729	0.776	0.795	0.786	0.667	0.690	0.679
	B_KNN(Weak)	0.572	0.607	0.590	0.730	0.739	0.735	**0.790**	0.809	0.800	0.671	0.697	0.684
	B_SVM	0.502	0.561	0.532	0.634	0.654	0.644	0.686	0.715	0.701	0.585	0.628	0.607

表 7　第 1 階層における研究分野ごとの正解件数(上位 1 件)

	工学		社会科学		総合領域		人文学		農学	
	Title	Abst	Title	Abst	Title	Abst	Title	Abst	Title	Abst
KNN(Weak)	**218**/260	**245**/260	**37**/60	**54**/60	**40**/70	**41**/70	**16**/20	**17**/20	68/80	**74**/80
B_KNN(Weak)	213/260	231/260	31/60	48/60	39/70	38/70	12/20	15/20	**71**/80	70/80
	医歯薬学		化学		複合新領域		数物系科学		生物学	
	Title	Abst	Title	Abst	Title	Abst	Title	Abst	Title	Abst
KNN(Weak)	**342**/360	333/360	**26**/30	**21**/30	7/20	12/20	67/80	70/80	7/20	**12**/20
B_KNN(Weak)	339/360	**340**/360	23/30	19/30	7/20	12/20	**68**/80	70/80	**10**/20	10/20

表8 第3階層における研究分野毎の精度(上位1件)

	KNN(List)手法とB_KNN(List)手法を比較して精度が向上した研究分野の例					
	応用光学・量子光工学		衛生学		会計学	
	Title	Abst	Title	Abst	Title	Abst
KNN (List)	**0.60**	**1.00**	**0.70**	**0.60**	**0.80**	**0.80**
B_KNN (List)	0.40	0.70	0.50	0.30	0.50	0.70

	KNN(List)手法とB_KNN(List)手法を比較して精度が低下した研究分野の例					
	無機材料・物性		基礎獣医学 基礎畜産学		細菌学(含真菌学)	
	Title	Abst	Title	Abst	Title	Abst
KNN (List)	0.40	0.20	0.10	0.40	0.50	0.60
B_KNN (List)	**0.60**	**0.30**	**0.30**	0.40	**0.60**	**0.70**

	KNN(List)手法において最も精度が低かった研究分野の例					
	応用物理学一般		神経科学一般		生物系薬学	
	Title	Abst	Title	Abst	Title	Abst
KNN (List)	0.10	0.30	0.00	**0.30**	0.00	**0.20**
B_KNN (List)	0.10	0.30	0.00	0.20	0.00	0.10

表 9　KNN(List)手法においてシステムが誤って付与した研究分野の例

Correct field	B_KNN(List)と比較して精度が低下した研究分野（表 8 中段）	
	Title	Abst
無機材料・物性	● 金属物性（1） ● 材料加工・処理（1） ● 応用物性・結晶工学（1）	● 金属生産工学（2） ● 金属物性（1） ● 材料加工・処理（1）
基礎獣医学・基礎畜産学	● 応用獣医学（4） ● 応用動物科学（2） ● 解剖学一般（1）	● 応用獣医学（4） ● 畜産学・草地学（1） ● 農業土木学（1）
細菌学(含真菌学)	● 耳鼻咽喉科学（2） ● 応用獣医学（1） ● 無機材料・物性（1）	● 病態検査学（2） ● ウイルス学（1） ● 土木環境システム（1）
Correct field	KNN(List)手法において最も精度が低かった研究分野（表 8 下段）	
	Title	Abst
応用物理学一般	● 原子力学（2） ● 機械力学・制御（2） ● 流体工学（1）	● 応用光学・量子光工学（3） ● 原子力学（1） ● 航空宇宙工学（1）
神経科学一般	● 計測工学（4） ● 耳鼻咽喉科学（1） ● 麻酔・蘇生学（1）	● 環境生理学（2） ● 神経・筋肉生理学（2） ● 脳神経外科学（1）
生物系薬学	● 呼吸器内科学（2） ● 生理学一般（1） ● 薬理学一般（1）	● 循環器内科学（2） ● 構造生物化学（2） ● 薬理学一般（2）

表10　第3階層の研究分野において抽出された要素技術とその効果の例

基礎獣医学・基礎畜産学		応用獣医学	
要素技術	効果	要素技術	効果
ラット（65）	細胞（101）	血清（79）	細胞（106）
マウス（60）	活性（87）	マウス（66）	反応（93）
アミノ酸（38）	遺伝子（87）	リンパ（45）	成績（80）
ウイルス（37）	濃度（52）	ウイルス（43）	活性（70）
応用物理学一般		応用光学・量子光工学	
要素技術	効果	要素技術	効果
顕微鏡（58）	特性（88）	レーザー（87）	特性（114）
レーザー（40）	成果（83）	半導体（73）	波長（105）
半導体（37）	試料（79）	ファイバ（34）	成果（79）
レンズ（21）	周波数（58）	レンズ（31）	周波数（57）

表 11　Abst データセットにおける表題中の要素技術，概要中の要素技術および概要中の効果表現を単独で加えた場合の精度および MRR の結果

KNN(List)で用いる素性	第1層				第2階層				第3階層			
	精度			MRR	精度			MRR	精度			MRR
	@1	@2	@3		@1	@2	@3		@1	@2	@3	
ALL (KNN(List))	**0.877**	**0.966**	0.984	**0.928**	**0.744**	**0.880**	**0.925**	**0.832**	**0.641**	**0.800**	**0.852**	**0.738**
TITLE_TECHNOLOGY	0.856	0.939	0.977	0.910	0.719	0.843	0.895	0.800	0.613	0.741	0.802	0.698
ABST_TECHNOLOGY	0.855	0.958	0.979	0.912	0.720	0.865	0.914	0.811	0.609	0.761	0.820	0.705
ABST_EFFECT (ATTRIBUTE and VALUE)	0.870	0.960	**0.985**	0.923	0.737	0.875	0.917	0.820	0.631	0.775	0.830	0.724
NOTHING (B_KNN(List))	0.852	0.944	0.976	0.909	0.715	0.844	0.902	0.800	0.603	0.733	0.795	0.690