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† 難波 英嗣† 竹澤 寿幸†

†広島市立大学大学院 情報科学研究科 〒731-3194 広島県広島市安佐南区大塚東 3-4-1

E-mail: †{fukuda, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp

あらまし 本研究では，科学研究費助成事業データベース(KAKEN)を対象に，KAKEN の分類体系に基づいて，学術論文を自動的に分類する手法を提案する．自然言語処理の分野において，文書分類は代表的な研究課題の一つであり，機械学習に基づいた手法が数多く提案されているが，学術論文固有の特徴に焦点を当てた分類手法を提案する．一般に，論文中には，新しい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．このような要素技術とその効果の対が，その論文を特徴付けていると考えることができる．本研究では，要素技術とその効果を示す表現を自動的に抽出し，素性として特徴付けることで，対象の論文に該当する研究分野を効率的に解析することを実現する．

キーワード 文書分類，情報抽出，学術論文

1. まえがき

本稿では，様々な研究分野における学術論文を効率的に検索するために，学術論文に対して特定の分類体系に基づく分類コードを自動的に付与する手法を提案する．近年，研究者数の増加，学問分野の専門分化によって学術情報量が爆発的に増加している．Web上で利用可能な学術情報データベースの整備に伴い，研究者は，短時間で膨大な学術情報にアクセスすることが可能になってきているが，限られた時間の中で特定分野の情報を網羅的に収集することが益々困難になってきつつある．しかし，学術論文に分類コードを付与すれば，論文検索がより効率的に行えるようになる．実際，特許やACM Digital Libraryなどの文献データベースでは様々な分類体系が考案されており，個々の文献に付与された分類コードを用いた検索が日常的に行われている．一方で，これらの分類体系は一定期間ごとに改訂が行われるが，改訂の度に，過去の文献に対して人手で新しい分類コードを改めて付与しなおすのは非常にコストがかかる．このため，多くの場合，一度文献に付与された分類コードは，分類体系が改訂された後もそのままになっている場合が少なくない．しかしながら，論文に分類コードを自動付与するシステムが実現できれば,このような問題も容易に解決できる．

その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)を対象に,KAKENで定められている分類体系に基づき，国立情報学研究所が運営する学術論文情報データベースであるCiNii articleに収録されている学術論文を自動的に分類することを行う．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・公開している過去に採択された研究課題を検索できるデータベースである．研究者が効率的に過去の研究課題を調べることができるよう，個々の研究課題には，人手で分野コードが付与されている．KAKENの分類体系は，工学，医学，社会科学などほぼ全ての学術領域を対象に作られているため，研究課題だけでなく論文の分類にも適した体系であると考えられる．

自然言語処理の分野において，文書分類は代表的な研究課題の一つであり，機械学習に基づいた手法が数多く提案されているが，本研究では学術論文を対象とし,学術論文固有の特徴を用いた分類手法を提案する．一般に，学術論文中には，新しい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．このような技術とその効果の対が，その論文を特徴付けていると考えることができる．さらに，特定の研究課題で，ある技術が有効であることが確認された時，その技術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が少なくない．例えば，古くは，Hidden Markov Model(HMM)が音声認識で有効であることが1980 年代終わりに確認されると，90年代に入って形態素解析に応用されている．90年代後半には決定木学習の代表的な手法であるC4.5が，2000年以降はSupport Vector Machine(SVM)などの機械学習手法が，画像処理や自然言語処理を含む知能情報学分野の多くの研究課題で利用され，従来の分類精度を大幅に向上できることが確認されている．一方，医学分野では，In Situ Hybridization(ISH)法やPolymerase Chain Reaction(PCR)法などの技術が広く用いられており，その効果について，「細胞数の減少」や「DNAの増幅」など，知能情報学の分野とは大きく異なる表記が論文中で見られる．

このように，論文に記述されている要素技術やそれによって得られる効果は，特定の研究分野の特徴を表すための重要な手掛かりとなり，学術論文の分類に対して有効であると考えられる．本研究では，要素技術とその効果を自動的に抽出し，素性として特徴付けることで,論文への研究分野の効率的な解析を実現する．

2. システムの動作例

本章では，学術論文の検索結果を研究分野ごとに分類し，さらに任意の研究分野における技術動向を分析した結果を提示するシステム，CiNii Miningの動作例および仕組みについて説明する．CiNii articleデータベースには約1,500万件の学術論文が収録されており，キーワードや著者名を手掛かりに，CiNii論文検索APIから取得した論文を分類している．図1は，「音声認識」をクエリとしてシステムに入力した時の検索結果を示している．図1において，「音声認識」という語句を含んだ論文が，知能情報学，教育工学，日本語教育など，様々な研究分野に分類されていることが分かる．これらのカテゴリは，KAKENで定められている「系・分野・分科・細目表」を用いており，「分野・分科・細目表」から研究分野の分類レベルを選択することができる．また，図1における，「この研究分野の技術動向を見る」という箇所をクリックすると，選択した研究分野における技術動向を示したページに移動する(図2)．

図2では，クエリ「音声認識」に対する，知能情報学の分野における技術動向マップを示している．図2において，左側に「音声認識」に関する，知能情報学の分野の各論文中で使われている要素技術を列挙している．また，中央には，各技術が使われた年を示しており，右側には，各技術を用いて得られた効果を示している．例えば，知能情報学の分野における「音声認識」に対して，2004年に「GMM (Gaussian Mixture Model)」を要素技術として用いており，「認識性能を大幅に改善」という効果を得ていることが，図2の技術動向マップから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，情報抽出手法[1]を用いて自動的に抽出を行っている．

図1: クエリ「音声認識」に対する検索結果画面

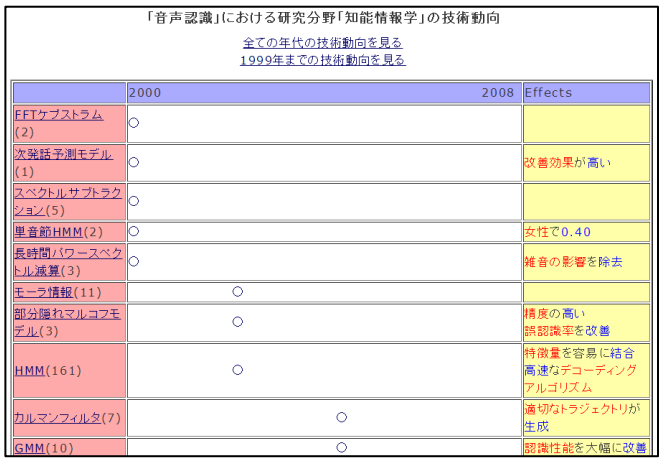

図2: 知能情報学の分野における，「音声認識」に対する技術動向マップ画面

3. 関連研究

本章では，「学術論文の自動分類」と「要素技術とその効果の抽出」に関する関連研究について述べる．

3.1. 学術論文の自動分類

学術論文の自動分類には，いくつかの先行研究があるが，その多くは，ある特定の分野に対してのみに焦点が当てられている[2,3]．また，近年では，学術論文をACM Digital Libraryのカテゴリに自動的に分類するという研究が行われている[4]．ACM Digital Libraryでは，計算機科学，情報技術の分野に特化した分類体系を構築している．一方，KAKENでは，上記で述べたような研究分野だけでなく，人文学や社会科学などといった全ての学術領域をカバーしており，このような分類体系は我々の知る限り存在しない．そのため，本研究の分類アプローチは，学術論文を全ての研究分野に対して網羅的に分類することができるといえる．また，これまでにもKAKENとDBLP(Digital Bibliography & Library Project)における著者名の同定[5]など，KAKENを対象とした研究がいくつか存在するが，KAKENの分類体系を用いて学術論文の自動分類を行うという研究は我々が初めてである．

一方で，KAKENと同様に，様々な技術分野を網羅している分類体系の一つに，国際特許分類(International Patent Classification: IPC)がある．IPCは国際的に統一されて用いられている分類体系であり，特許文献の技術内容によって，5階層から構成・分類されている．このような特許分類体系と学術論文を対象にした研究は，これまでにも数多く存在している．その一つに，国立情報学研究所が主催した第7回NTCIRワークショップ

(NTCIR-7)特許マイニングタスク[6]がある．これは，特許と論文を対象にした検索や動向分析など，様々な目的に利用可能な言語処理技術の開発を目的とした研究プロジェクトであり，その第一歩として，学術論文を国際特許分類に自動分類するタスクを設定している．

特許マイニングタスクでは，論文に付与するIPCコードの数が30,855件と非常に多く，さらに，訓練用データが350万～450万件と膨大である．このため，ほとんどの参加グループは，k-Nearest Neighbor(k-NN)法を用いていた．一方，KAKENに登録されている採択課題の件数および研究分野(細目表)の数は，2011年度においてそれぞれ，672,397件，297分野と，特許マイニングタスクで提供されたものほど多くない．そのため本研究では，k-NN法に基づいた分類手法と，自然言語処理の分野で一般的に使用されるSVMを用いて研究分野の自動付与を行い，それぞれの手法における分類精度の比較を行う．

NTCIR-7特許マイニングタスクにおいて，Xiaoら[7]は，k-NN法を用いて，任意の英語論文抄録に対する候補となるIPCコードのリストを作成した後，リランキング手法を用いてリスト内のIPCコードを並べ替えるという手法を提案している．k-NN法を用いてIPCコードの候補リストを作成する際，Xiaoらは，5種類の類似度計算手法と5種類のランキング手法を組み合わせ，最もMAP(Mean Average Precision)値が高かった手法を採用している．本研究でもXiaoらと同様に，k-NN法に基づく様々なランキング手法を用いて対象の学術論文に対する研究分野の候補リストを作成する．

3.2. 要素技術とその効果の抽出

NTCIR-8特許マイニングタスク[8]における学術論文分類サブタスクでは，要素技術とその効果を示す表現を特許や論文から自動的に抽出することを目的としている．また，特許と論文を要素技術とその効果という観点から分類した技術動向マップを作成することを目指している．

福田ら[1]は，学術論文分類サブタスクに基づき，訓練用データである論文と特許文書集合の両方を機械学習に用いるというドメイン適応を用いて日本語論文と特許から要素技術と効果を抽出しており，その有効性を示している．本研究では，このシステムを用いて，KAKENの採択課題を解析する．4章で詳細を述べる．

4. 要素技術とその効果を用いた学術論文の自動分類

本章では，要素技術とその効果を用いた，KAKEN分類体系の観点における，CiNii articleの学術論文を自動的に分類する手法を述べる．4.1節では，要素技術とその効果をどのように抽出し，使用するのかについて述べ，4.2節では，本システムの構成について述べる．

4.1. 要素技術とその効果の抽出・リスト化

4.1.1. 要素技術とその効果の抽出

1 章でも述べたように，要素技術とその効果を用いることは，各分野を特徴づけるために有用であると考えられる．本研究では，福田らの情報抽出手法を用いて，要素技術とその効果の抽出を行う．この手法では，「論文の表題と概要において，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考え，SVM を用いた機械学習によって，以下に示すようなタグの自動付与を行っている．

- **TECHNOLOGY**: 要素技術(例: SVM, HMM)
- **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，問題点の抑制や解決したこと)．EFFECT タグには，ATTRIBUTE タグと VALUE タグを含む．
- **ATTRIBUTE, VALUE**: 「処理速度(ATTRIBUTE)が向上(VALUE)」のように，要素技術に対する効果部は「属性(ATTRIBUTE)」と「属性値(VALUE)」の対で表現する．

概要に上記のタグを付与した例を以下に示す．

<TECHNOLOGY>CRF</TECHNOLOGY>を用いた手法では，<EFFECT><VALUE>0.935</VALUE>の<ATTRIBUTE>精度</ATTRIBUTE></EFFECT>が得られた．

なお，表題解析では TECHNOLOGY タグのみが付与される．これは，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からである．

4.1.2. 手がかり語の抽出・リスト化

本研究では，KAKEN に収録されている採択課題から，要素技術とその効果を示す表現をそれぞれ抽出し，リスト化を行う．まず，福田らの情報抽出手法を用いて，672,397 件の採択課題の表題および概要の解析を行い，<TECHNOLOGY>, <ATTRIBUTE>, <VALUE>タグを付与する．その後，タグ付けされた語句をそれぞれ抽出し，要素技術リスト，属性リスト，属性値リストを作成する．

上記で作成した3種類のリストに加えて，本研究では，採択課題における研究者欄と発表文献欄から，正規表現を用いて著者リストと学会･出版リストを作成する．これは，著者名や学会名などのデータフィールドを利用することで，異種・同種のデータコレクションからなるデータ群からトピックを発見できることや[9]，言語横断的に学術論文を推薦することができる[10]など，データマイニングや文書分類の分野において有用な手掛かり語として活用されており，本研究においても，対象の論文に対して最も適切な研究分野を付与するための有効な手掛かり語として用いることができると考えられるためである．ここで，特定の分野を専門とし

ている研究者や学会・出版社は，正しい研究分野の判定において，特に有効であると考えられる．そこで本研究では，研究者名および学会・出版名に付随する研究分野の種類を調べ，それぞれの種類数に対して一定の閾値を設定する．そして，閾値以下および閾値より上回っている手掛かり語集合からリストをそれぞれ構築する．予備実験の結果から，2分野以下の研究分野に属する研究者から著者リスト1を収集・構築し，それ以外の研究者名は著者リスト2 として収集・構築する．同様に，9分野以下の研究分野に属する学会・出版名から学会・出版リスト1を作成し，それ以外の学会・出版名は学会・出版リスト2として作成する．それぞれの閾値は，人手で設定を行い，最も精度の高かったものを採用している．上記で作成した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，各リストの手掛かり語に対して与える重みを表1に示す．各リストにおける重みは，予備実験に基づいて人手で設定を行い，最も精度の高かったものを用いている．

また，CiNii articleの論文データにおける概要は非常に短いものが多く，中には概要が存在しないものも数多く存在する．そのため，構成するクエリファイルの情報量不足により，類似度を十分に算出することが出来ず，研究分野の付与が行えない場合がある．これを解決するために，本研究では，KAKENおよびCiNii articleから共著者リストをそれぞれ作成し，抽出した著者名と関係のある著者名をクエリファイルに追加することを行う．これは，特定の分野における研究者は通常，同じ分野の研究者と共同研究を行うため，その分野を特徴付けるための重要な手掛かりとなると考えられる[11]．しかし近年では，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いる．そのため本研究では，各著者に対するこれまでの共同研究の回数を数え，一定の閾値を設定する．予備実験の結果から，KAKENから作成する場合，1回以上のものを対象とし，CiNii articleから作成する場合，5回以上のものを対象とする．これにより，本研究では，KAKENから1,094,510対の共著関係およびCiNii articleから3,268,625対の共著関係を獲得した．

表1: 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
著者リスト1	荒牧英治	144,108	50
著者リスト2	川原稔	15,567	1
学会・出版リスト1	日本教育情報学会	125,232	40
学会・出版リスト2	情報処理学会	4,352	4
要素技術リスト	rt-pcr法，有限要素法	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

4.2. システム構成

図 3 に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」から構成されている．以下に各モジュールについて説明する．

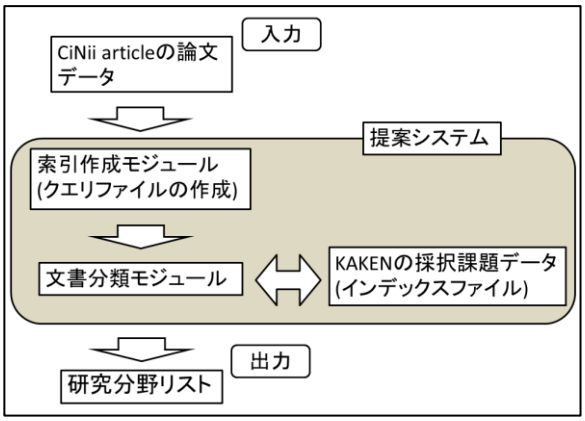

図 3: システム概要

4.2.1. 索引作成モジュール

索引作成モジュールでは，4.1.2 節で作成した手掛かり語リストを用いて，CiNii article の論文データからクエリファイルを作成する．この時，本研究では，抽出する項目によって与える重みを変更する．これは，文書内の各項目に対して出現する単語の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[12]．

まず，概要に対して形態素解析を行い，接頭詞を含む名詞を抽出する．次に，抽出した語句が，要素技術リスト，属性リスト，属性値リストのど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，次に，抽出した名詞が要素技術リストに存在していれば，重み17を与える．存在していない場合，重み1を与える．これらの重みは，予備実験に基づいて人手で設定を行い，最も精度の高かったものを用いている．

最後に，正規表現を用いて，著者欄から著者名を，発表文献から学会・出版名をそれぞれ抽出する．その後，抽出した著者名と関連する著者名を，4.1.2節で作成したKAKENおよびCiNii articleの共著者リストから抽出する．そして，抽出した著者名が著者リスト1または著者リスト2に含まれていれば，各リストに対応した重みを与える．また，抽出した学会・出版名が学会・出版リスト1または学会・出版リスト2 に含まれているかどうかを調べ，各リストに対応した重みを与える．抽出した著者名または学会・出版名がどのリストにも存在しない場合，クエリファイルの作成には用いない．

KAKENのデータからインデックスファイルを作成する際も，上記で述べた手法を用いる．この時，本研究で訓練用データとして用いるKAKENデータには全て研究概要が含まれており，情報量の不足という観点

による文書間の類似度の測定に対する問題は考慮しなくても良いため，共著者リストは用いない．

4.2.2. 文書分類モジュール

本システムでは，k-NN 法および SVM の 2 種類の分類手法を用いて，研究分野の付与を行う．

k-NN 法

- 類似度計算

k-NN 法に基づく分類器において，入力文書と訓練用データの各文書間の類似度を計算するために，本研究では，情報検索システムの分野で幅広く用いられている SMART[13]を採用する．

- ランキング

本研究では，Xiao らの研究[6]で用いられたランキングのうち，以下で述べる 2 種類の手法を用いて，対象の文書に適合する研究分野のランキングを行う．まず，計算された類似度と共に，上位 k 件の文書を抽出する．次に，抽出された文書の研究分野に対するスコア値 $Score(c)$を計算する．ここで，$Score(c)$は，入力クエリにラベル c が付与される可能性の尺度である．そして，各研究分野における $Score(c)$が高い順のソートし，上位 n 件までの研究分野を，入力された文書に付与する．本研究で適用する各ランキング手法を以下に述べる．

(1) Listweak

上記で述べた Sum 手法に基づいたランキング手法であり，抽出された文書集合において，より類似度の高い文書を強調する手法である．

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

ここで，$occur(c, d_i)$は，ラベル c が文書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，そうでなければ 0 となる．$Sim(q, d_i)$は，入力クエリ q と文書d_i間の類似度を表す．r_1は，抽出された文書集合において，より類似度の低い文書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，デフォルト値である 0.95 と設定する．

(2) Weak

k-NN法の欠点として，訓練用データ内の文書に付与されているラベルの偏りが大きいほど，入力クエリに対してそのラベルが付与されやすい傾向にあることが挙げられる．本実験で用いる訓練用データセット(5章で詳細を述べる)では，第3階層における研究分野の偏りがないように作成しているが，第1および第2階層の研究分野に置き換えた場合，文書数の偏りが発生する．例えば，第1 階層に位置する研究分野に置き換えた時，工学の分野が付与されている文書は6,000 件であるのに対し，人文学の分野が付与されている文書は1,600 件である．Weak手法では，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．

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

ここで，$size(c)$は，抽出された文書内におけるラベル c の数を表し，$crank(c, i)$は，上位 i-1 件の文書におけるラベル c の出現頻度を表す．本研究では，デフォルト値である 0.90 と設定する．

SVM 手法

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ける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．本研究では，KAKENデータベースの規模から，分類器の作成に要する計算コストを考慮し，カーネル関数として線形カーネルを用いる．以下に，SVMを用いた分類手法について述べる．

まず，4.2.1節により作成されたKAKENデータからのインデックスファイルを用いて，各階層における研究分野に対する分類器を作成する．次に，各分類器に対して，4.2.1節により各単語への重み付けが行われた入力クエリ(CiNii articleの論文データから作成されたクエリファイル)を適用する．そして，正例であると判断された場合，その分類器を表す研究分野を出力する．ここで，本研究のタスクでは，入力クエリに対して少なくとも1つの研究分野を付与しなければならないと定めている．しかし，全ての分類器が入力クエリに対して負例を示した場合，候補となる研究分野が存在しないことになる．そのため本研究では，入力クエリに対する各分類器の超平面の距離を用い，最も計算結果が高かった分類器の研究分野を入力クエリに付与する．

5. 実験

5.1. 実験方法

5.1.1. 実験データ

KAKEN データベース

KAKEN の分類体系は，「系・分野・分科・細目表」で構成されており，年度ごとに文部科学省の科学技術・学術審議会学術分科会科学研究費補助金審議部会で審議が行われ，改訂が実施される．このうち，2011 年度に使用可能な分野(第1階層; 10分野)・分科(第2階層; 69分野)・細目表(第3階層; 297分野) を対象とする．KAKENの分類体系の一部を表2に示す．

CiNii articleの論文データには研究分野が付与されていない．そのため本研究では，KAKENデータベースの採択課題データ(KAKENデータ)を訓練用データとして用いる．また，各KAKENデータには発表文献欄が記載されており，そのリストの中には，CiNii articleとリンクしている論文データが含まれている．しかし，KAKENの発表文献と実際にリンクしている論文の割

合は，全体の約6.10%である．本研究では，KAKENデータとリンクしているCiNii articleには，そのKAKENデータの研究分野が付与されているとみなし，評価用データとして用いる．

KAKENデータベースには，672,397件(1965-2011年)の採択課題が含まれている．このうち，本研究では，表題，研究概要，キーワード，研究者，発表文献欄，研究分野が存在する採択課題を訓練用データとして用いる．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つの研究分野に対して200件の採択課題を用いるように訓練用データを作成する．その結果，28,400件の採択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分野を訓練用データとして用いる．ここで，第1，第2階層における各研究分野の採択課題の数には偏りがあることに注意する．また，4.1.2節で述べた著者リスト，学会・出版リスト，共著者リストについて本実験で対象としている研究分野が付与されている283,686件のデータを用いている．

表 2: KAKEN 分類体系(2011 年度)の例

分野(第 1 階層; 10 分野)	分科(第 2 階層; 69 分野)	細目表(第 3 階層; 297 分野)
総合領域	情報学	知能情報学 ソフトウェア
	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, 流体工学
	電気電子工学	システム工学, 計測工学

CiNii article データベース

CiNii articleデータベースの論文データは主に，ID, 表題，概要，著者，発表文献から構成されている．このうち本研究では，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用いる．これらのデータは，KAKENデータの発表文献欄内の論文データとリンクしており，各評価用データには，リンク先のKAKENデータの研究分野が付与されている．なお，評価用データで扱う研究分野は，訓練用データで対象としているもののみを用いており，第3階層の研究分野におけるデータ数の偏りを無くすために，1つの研究分野に対して20件の評価用データ(Abstデータセット:10件，Titleデータセット:10件)を用いている．このため，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．

また，4.1.2節で述べた共著者リストについて，著者名を含んでいる5,924,669件のデータを用いている．

5.1.2. 評価尺度

評価尺度には，システムが評価用データに自動的に付与した研究分野と，評価用データに元々付与されている研究分野が一致した場合に対して正解と判断する精度およびMRR(Mean Reciprocal Rank)を用いる．

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3件までの研究分野を正解対象とする．また，訓練用データおよび評価用データに付与されている研究分野を，第1階層，第2階層，第3階層それぞれに置き換えた場合における精度およびMRR値を示す．

5.2. 比較手法

以下で述べる 3 種類の提案手法と 3 種類のベースラインを用いて実験を行った[1]．形態素解析には MeCab を，SVM における機械学習には TinySVM を用いた．

提案手法

- **KNN(Listweak)**: 任意の研究分野が付与されている，抽出された全ての文書のスコア値の総和を算出し，最も値が高かった研究分野から順に付与する．この時，より低いランクに位置する文書のスコア値に対してペナルティを与える．
- **KNN(Weak)**: KNN(Listweak)に基づいたスコア値の計算を行うが，訓練用データセット内の研究分野間における文書数の偏りを考慮する．
- **SVM**: 入力クエリに対する各分類器の超平面の距離を用い，最も計算結果が高かった分類器を表す研究分野から順に付与する．

ベースライン手法

- **BASE_KNN(Listweak)**: KNN(Listweak)手法において，要素技術とその効果(属性，属性値)に対応する素性を用いない．
- **BASE_KNN(Weak)**: KNN(Weak)手法において，要素技術とその効果に対応する素性を用いない．
- **BASE_SVM**: SVM 手法において，要素技術とその効果に対応する素性を用いない．

5.3. 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それぞれ表3，表4，表5に示す[2]．表3から表5における各手法のMRR値を見ると，KNN(Listweak)手法が全体的に高い値を示していることが分かった．また，第1，第2階層における本システムの出力結果の上

[1] 4章で述べた各手掛かり語の重みや閾値について，KAKENデータベースから，(訓練用データ内のデータを除く)2,000件のAbstデータセットを作成し，チューニングを行うことで決定した．

[2] 提案手法およびベースライン手法におけるk-NN 法では，チューニング用データセットを用いて閾値kを1から50までの1刻みの範囲で設定を行い，各手法や実験条件におけるそれぞれの閾値による結果において，最も精度の高かった時の値を採用している．

位1件までを正解とした場合，わずかであるがKNN(Weak)手法が最も高い値を示している．これらの結果から，「分野・分科・細目表」を対象に2種類のデータセットを用いて実験を行った結果,本手法により，出力結果の上位1件までを正解とした場合,平均で最大0.8525，0.7120，0.6145の精度が得られることが分かった．また，MRRにおいて，平均0.9083，0.7995，0.7102の値を示すことが分かった．さらに，対象とする階層やデータセットなどを考慮して適用するランキング手法を変更することが重要であることが分かった．

次に，提案手法およびベースライン手法において全体的に性能が高かったKNN(Listweak)手法とBASE_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階層における全ての条件において有意水準1%で有意な差になった．また，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のAbstデータセットに対して，有意水準5%で有意な差になった．これらから，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．

5.4. 考察

本節では，各研究分野に対して，要素技術とその効果を素性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調べる．ここでは，より一般的な研究内容を扱っている第1階層に属する10分野を対象に調べる．表3において，上位1件での平均精度が最も高かったKNN(Weak)手法とBASE_KNN(Weak)手法を用いて研究分野を自動付与した時における，研究分野ごとの上位1件に対する精度を表6に示す．また，評価用データセット内における各研究分野が付与されている文書数および正解件数も示している．

表6から，理工系の分野を扱う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，2種類の評価用データセット内において，工学の分野を表す，合計520件の文書のうち,正解件数が444件から463件へと大幅に向上していることが分かる．これは，本研究で用いた技術動向分析システムは，主に理工系の分野を対象としたシステムであり,KAKENデータからそれらの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することができたためであると考えられる．

また，人文社会系に属する人文学と社会科学の分野に対しても，要素技術とその効果を示す表現は有効であることが分かる．特に，社会科学では，合計120文書のうち，正解件数が79件から91件へと増えている．ここで，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」などが効果表現とみなされていることが分かった.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対しても有効であることが分かる．

6. おわりに

本研究では，学術論文を科学研究費補助金データベースにおける研究分野の分類体系に基づいて分類を行う手法を提案した．本手法では学術論文情報から，要素技術とその効果を示す表現を抽出し，学習に用いる際の素性として用いた．そして，分類指標である「分野・分科・細目表」における研究分野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，それぞれ平均0.8525，0.7120，0.6145の精度が得られた．また，上位3件までの結果を正解対象とした時，平均で0.9083，0.7995，0.7102のMRR値が得られた．これらの結果は，要素技術とその効果を用いない手法による実験結果よりも上回っていることから，本研究における提案手法の有効性が確認された．

参 考 文 献

[1] 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, pp.16-29 (2013)

[2] 今井 俊, 佐藤理史: 表題解析による科学技術論文の詳細分類，情報処理学会第 57 回全国大会講演論文集，pp.211-212 (1998)

[3] McCallum, A., Nigam, K., Rennie, J. and Seymore, K.: A Machine Learning Approach to Building Domain-Specific Search Engines, Proc. of the 16th International Joint Conference on Artificial Intelligence, pp.662-667 (1999)

[4] Rocha, L., Mourão , F., Mota, H., Salles, T., Goncalves, M.A. and Meira, W.Jr.: Temporal Contexts: Effective Text Classification in Evolving Document Collections, Information Systems, pp.388-409 (2012)

[5] Tran, D.H., Takeda H., Kurakawa, K. and Tran, M.T.: Combining Topic Model and Co-author Network for KAKEN and DBLP Linking, Proc. of the 4th Asian Conference on Intelligent Information and Database Systems, pp.396-404 (2008)

[6] Nanba, H., Fujii, A., Iwayama, M. and Hashimoto, T.: Overview of the Patent Mining Task at the NTCIR-7 Workshop, Proc. of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, pp.325-332 (2008)

[7] Xiao, T., Cao, F., Li, T., Song, G., Zhou K., Zhu, J. and Wang, H.: KNN and Re-ranking Models for English Patent Mining at NTCIR-7, Proc. of the 7th NTCIR Workshop Meeting, pp.333-340 (2008)

[8] Nanba, H., Fujii, A., Iwayama, M. and Hashimoto, T.: Overview of the Patent Mining Task at the NTCIR-8 Workshop, Proc. of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, pp.293-302 (2010)

[9] 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)

表 3: 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	Title	Abst	Title	Abst	Title	Abst
提案手法	KNN(Listweak)	0.8270	**0.8770**	0.9250	**0.9660**	0.9620	0.9840	**0.8885**	**0.9280**
	KNN(Weak)	**0.8280**	**0.8770**	**0.9270**	0.9640	0.9580	**0.9850**	0.8880	0.9270
	SVM	0.7370	0.8150	0.8350	0.8920	0.8730	0.9380	0.7987	0.8689
ベースライン	BASE_KNN(Listweak)	0.8150	0.8520	0.9210	0.9440	0.9580	0.9760	0.8803	0.9090
	BASE_KNN(Weak)	0.8130	0.8530	0.9240	0.9440	**0.9640**	0.9800	0.8828	0.9113
	BASE_SVM	0.7500	0.7770	0.8540	0.8780	0.8920	0.9100	0.8147	0.8382

表 4: 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	Title	Abst	Title	Abst	Title	Abst
提案手法	KNN(Listweak)	0.6760	0.7440	**0.8280**	**0.8800**	0.8720	**0.9250**	**0.7670**	**0.8320**
	KNN(Weak)	0.6700	**0.7540**	0.8100	0.8720	**0.8800**	0.9160	0.7628	0.8292
	SVM	0.5720	0.6850	0.6630	0.7810	0.7090	0.8220	0.6328	0.7467
ベースライン	BASE_KNN(Listweak)	**0.6770**	0.7150	0.8150	0.8440	0.8660	0.9020	0.7630	0.7996
	BASE_KNN(Weak)	0.6720	0.7170	0.8120	0.8530	0.8740	0.9110	0.7635	0.8053
	BASE_SVM	0.5910	0.6370	0.7100	0.7530	0.7640	0.8030	0.6685	0.7117

表 5: 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	Title	Abst	Title	Abst	Title	Abst
提案手法	KNN(Listweak)	**0.5880**	0.6410	**0.7440**	**0.8000**	**0.7900**	**0.8520**	**0.6826**	**0.7378**
	KNN(Weak)	0.5700	**0.6420**	0.7300	0.7900	0.7890	0.8450	0.6706	0.7321
	SVM	0.5270	0.6070	0.6340	0.7080	0.6660	0.7650	0.5912	0.6765
ベースライン	BASE_KNN(Listweak)	0.5740	0.6030	0.7240	0.7330	0.7760	0.7950	0.6670	0.6896
	BASE_KNN(Weak)	0.5720	0.6070	0.7300	0.7390	**0.7900**	0.8090	0.6711	0.6971
	BASE_SVM	0.5020	0.5610	0.6340	0.6540	0.6860	0.7150	0.5853	0.6278

表 6: 第 1 階層における研究分野ごとの精度(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)	**0.839** (218/260)	**0.942** (245/260)	**0.617** (37/60)	**0.900** (54/60)	**0.571** (40/70)	**0.600** (41/70)	**0.800** (16/20)	**0.850** (17/20)	0.850 (68/80)	**0.925** (74/80)
BASE_KNN (Weak)	0.819 (213/260)	0.889 (231/260)	0.517 (31/60)	0.800 (48/60)	0.557 (39/70)	0.543 (38/70)	0.600 (12/20)	0.750 (15/20)	**0.888** (71/80)	0.875 (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)	**0.950** (342/360)	0.925 (333/360)	**0.867** (26/30)	**0.700** (21/30)	0.350 (7/20)	0.600 (12/20)	0.838 (67/80)	0.875 (70/80)	0.350 (7/20)	**0.600** (12/20)
BASE_KNN (Weak)	0.942 (339/360)	**0.944** (340/360)	0.767 (23/30)	0.633 (19/30)	0.350 (7/20)	0.600 (12/20)	**0.850** (68/80)	0.875 (70/80)	**0.500** (10/20)	0.500 (10/20)

[10] Uchiyama, K., Aizawa, A., Nanba, H. and Sagara, T.: OSUSUME: Cross-lingual Recommender System for Research Papers, Proc. of the 2011 Workshop on context-awareness in Retrieval and Recommendation, pp.39-42 (2011)

[11] Zhang, X., Hu, X. and Zhou, X.: A Comparative Evaluation of Different Link Types on Enhancing Document Clustering, Proc. of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.555-562 (2008)

[12] Fall, C.J., Torcsvari, A., Benzineb, K. and Karetka, G.: Automated Categorization in the International Patent Classification, Proc. of the ACM SIGIR Forum, pp.10-25 (2003)

[13] Salton, G.: The SMART Retrieval System - Experiments in Automatic Document Processing, Prentice-Hall, Inc., Upper Saddle River, NJ (1971)