CiNii データベースを用いた研究動向分析システムの構築

福田悟志[1]　難波英嗣[1]　竹澤寿幸[1]　武田英明[2]
相澤彰子[2]　大向一輝[2]　宮尾祐介[2]　内山清子[2]

[1]広島市立大学大学院　[2]国立情報学研究所

1.はじめに

論文検索では同一キーワードを入力しても，ユーザの研究分野によって求める論文が異なる場合がある．例えば「SNS」というキーワードに対し，「データベース」を専門としているユーザの場合，SNS 上に蓄積されたデータを用いたテキストマイニングに関する論文を，一方で「社会学」を専門としたユーザは，SNS を利用した新たな地域交流の創設に関する論文を検索したい場合などがそれに当たる．

本研究では，国立情報学研究所(NII)の論文情報ナビゲータである CiNii[1]を対象に，検索結果を研究分野毎に自動的に分類し，さらに分野毎の動向情報分析結果を提示するシステムの開発を目指す．CiNii とは，論文や図書・雑誌などの学術情報で検索できるデータベース・サービスである．従来の CiNii では検索結果として，出版年が新しい順で論文が表示される．これらを研究分野毎に分類して表示させることで，ユーザが求めているキーワードの意味を含んだ論文を検索することが容易になり，またユーザが専門としている研究分野以外の様々な分野における動向情報を表示することで，キーワードの持つ様々な側面をユーザに提供することが出来るなど，より有益な検索結果を提供出来ると考えられる．

2.CiNii Mining

本節では，CiNii のデータベースを対象に，検索結果を研究分野毎に分類し，さらに分野毎の技術動向分析を提示するシステム“CiNii Mining”の仕様について説明する．

2.1.研究分野の自動分類

2.1.1.外部仕様

画面例を図 1 に示す．トップ画面でキーワードと著者名の入力を行い，表示件数を選択する．その後，search ボタンを押すと，学術論文の情報が研究分野毎に振り分けて表示される．最大で 100 件までの論文情報を検索することが出来る．また，分野名をクリックすることで該当の検索結果までスクロールされ，ユーザの専門分野の論文情報を容易に検索出来る点が特徴である．論文情報のタイトル名をクリックすると，CiNii の論文詳細表示画面へ移動することが出来る．論文情報の表示件数を変更したい場合，図 1 の件数を選択し直し，再度 search ボタンを押すことで可能となる．また，キーワードや著者名を新たに検索したい場合も同様に，図 1 上の入力フォームにキーワードまたは著者名を入力し，search ボタンを押すことで再度検索を行う事が出来る．

図 1：CiNii Mining の検索画面

2.1.2.内部仕様

CiNii Mining では，CiNii 論文検索 API を用いて CiNii データベースから学術論文の情報を取得している．CiNii 論文検索 API とは，CiNii にある 1,500 万件以上の学術論文に関する情報を，キーワードや著者名で検索出来る API のことである．本システムでは，入力キーワードを含んだ学術論文の情報を取得する際に，検索画面で選択された表示件数に関わらず，内部では 200 件の論文情報を取得する．そしてその 200 件のうち，ユーザが選択した件数の論文情報を対象に，研究分野の自動分類を行う．

2.2.技術動向分析

2.2.1.外部仕様

図 1 の「この分野の技術動向を見る」という箇所をクリックすることで，研究分野毎における入力キーワードの技術動向を閲覧できるページへ移動することが出来る．画面例を図 2 に示す．このページでは，振り分けられた論文情報を自動的に解析し，論文情報内の「要素技術」と「その効果(属性と属性値)」を色つき文字で強調して表示している．本システムでは，要素技術に該当する語句を青字で，属性に該当する語句を赤字で，属性値に該当する語句を緑字で強調して表示している．また，図 2 の「もっと見る」という部分をクリックすると，選択している

[1]http://ci.nii.ac.jp/

研究分野に該当する，入力キーワードを含んだ論文情報をさらに取得し，自動解析して表示する．

図 2：技術動向ページ

2.2.2.内部仕様

CiNii Mining では，福田ら[福田 2011]が作成した技術動向分析システムを組み込むことで，選択した研究分野に振り分けられた論文情報のタイトルと概要を自動的に分析する．福田らは，特定分野の技術動向を効率的に把握するために，学術論文から要素技術とその効果を示す表現を自動的に抽出し，「要素技術」と「その効果(属性と属性値)」という 2 つの観点で分類を行った．例えば，「CRF を用いた場合，88%の精度が得られた．」という文の場合，「CRF」が要素技術，「精度」が属性，「88%」が属性値に該当する．

図 1 の「この分野の技術動向を見る」がクリックされた時，選択した研究分野に振り分けられている論文情報を対象に動向分析を行う．また，図 2 の「もっと見る」がクリックされた時，2.1.2 節で取得した 200 件の論文情報の内，まだ分類されていない論文情報を対象に研究分野の自動分類を行う．そして，選択している研究分野に該当する論文情報を取得し，動向分析を行う．

3.関連研究

2 章では CiNii Mining のシステムについて述べたが，本システムにおいて，より高精度な研究分野の自動分類を目指すために，本研究では学術論文情報から詳細な情報を抽出する「情報抽出」という形式を採用した．本章では，情報抽出を用いた研究について紹介する．

3.1.Gupta らの研究

Gupta ら[Gupta 2011]は，研究アイディアの発展過程や研究分野の動態性などを把握するために，学術論文から「FOCUS」，「TECHNIQUE」，「DOMAIN」という 3 つのカテゴリに該当する語句を抽出する手法を提案した．そしてこれらのカテゴリ句を用いて，ある手法やツールが様々な研究分野に対してどのような影響を及ぼしているかについて研究を行い，また，ある研究分野における過去 20 年間の影響力に関するタイムラインも示した．本研究では，福田ら[福田 2011]が作成した技術動向分析システムを用いて，論文情報から「要素技術」，「属性」，「属性値」の 3 つのカテゴリに該当する語句を抽出する．

3.2.内堀らの研究

内堀[内堀 2003]は，「著者名」，「出版者名」を手掛かりとして組み入れた図書の自動分類について研究を行った．その結果，著者名による分類，出版者名による分類いずれにおいても，913(小説，物語)をはじめとする 9 類文学で高い正解率を得た．内堀はこの結果から，「書名中の語句からカテゴリを判断することは困難である」とされていた 9 類の文学について，著者名と出版者名を用いれば分類が可能であることを示した．本研究では，内堀が図書の自動分類に用いた著者名に着目し，論文情報から著者名の抽出を行い，手掛かりとして組み込む．

4.研究分野の自動分類手法

4.1.訓練用データと研究分野の分類指標

CiNii の論文情報には研究分野は付与されていない．そのため CiNii Mining では，研究分野が付与されている科学研究費補助金データベース(KAKEN)[2]の採択課題データを訓練用データとして用いることで，CiNii 論文検索 API から取得した論文情報に対して自動的に研究分野の付与を行なっている．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る．KAKEN では全ての学問分野に渡って幅広く取り扱っているので，全分野の最新の研究情報を網羅することが出来る．

また，本研究では KAKEN で定められている研究分野の「系・分野・文科・細目表」と時限付き細目の内，2011 年で使用可能な細目表，分野に属している「広領域」，及び時限付き細目とする 291 の研究分野を分類指標として用いる．

表 1：KAKEN における「系・分野・文科・細目表」

系	分野	文科	細目名
総合・新領域系	総合領域	情報学	情報学基礎 ソフトウェア
		生活科学	生活科学一般 食生活学
人文社会系	人文学	言語学	言語学 英語学
	社会科学	法学	刑事法学 民事法学

4.2.GETA を用いた分類手法

汎用連想計算エンジン GETA[3]は，文書検索における頻度付き索引データを典型とする大規模かつ疎な行列(Word-Article-Matrix;WAM)を対象として，行と行または列と列の類似度を内積型メジャーで高速計算するツールである．CiNii Mining では，GETA における WAM を生かして，検索システムとして有用で高速な自動分類を目指す．

4.3.頻度付き索引ファイル及びクエリファイルの作成

頻度付き索引ファイルの作成には，KAKEN 内の採択課題において，タイトル，キーワード，概要，著者名，及び 2011 年で使用可能である研究分野が含まれている 242,772 件の採択課題を用い，クエリファイルの作成には，CiNii 論文検索 API から取得した論文情報の内，タイトルと著者名，またはタイトル，概要，及び著者名を含む論文情報を用いる．本研究では，以下で述べる手法

[2]http://kaken.nii.ac.jp/
[3]http://geta.ex.nii.ac.jp/geta.html

を用いて頻度付き索引ファイル及びクエリファイルを作成することで，より高精度な自動分類を目指す．

4.3.1.リストの作成

本研究ではまず，福田ら[福田 2011]の作成したシステムを用いて，KAKEN における 639,341 件の採択課題に記載されているタイトルと概要に対して技術動向分析を行い，要素技術，属性，属性値に該当する語句をそれぞれ抽出し，要素技術リスト，属性リスト，及び属性値リストを作成する．抽出された語句の総数は以下となった．

要素技術:306,505 属性:162,167 属性値:41,391

収集した語句の例を以下の表に記載する．

表 2:各リストに属する語句の例

要素技術	属性	属性値
rt-pcr法	温度	向上
分子動力学法	分解速度	短縮
触覚センサ	エネルギー効率	10%程度
遺伝子欠損マウス	磁気転移温度	約10%
dna修復酵素遺伝子欠損マウス	周波数利用効率	克服

4.3.2.重み付けの検討

訓練用データのタイトル，キーワード，概要，及び CiNii 論文情報のタイトル，概要に対して形態素解析を行う．もし得られた語句が 4.3.1 節で作成した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，重みとして 1 を与える．形態素解析には MeCab を用い，形態素解析して得る語句は名詞または(接頭詞を含んだ)名詞句のみとする．本研究では，各リストにおける重みは以下のように設定した．

要素技術リスト	:重み 14
属性リスト	:重み 6
属性値リスト	:重み 3
上記のリスト以外	:重み 1

本研究では著者名への重み付けも行う．訓練用データ及び CiNii の論文情報に記載されている著者名を抽出し，重みを付与する．本研究では重みを 50 と設定した．

4.4.検索インタフェースの作成

4.3 節の手法を用いて作成された頻度付き索引ファイルのインデックス(WAM)と，クエリファイルをつけ合わせることで各文書間の類似度を計り，検索インタフェースを作成する．類似性尺度には SMART[Salton 1971]を用いる．その後，各クエリデータにおける上位 k 件までの訓練用データを対象に，同名の研究分野をまとめ上げ，研究分野毎の類似度の総和を算出する(k-NN 法)．そして，最も類似度の高い研究分野を対象の論文情報に付与する．

5.評価実験

5.1.実験データ，ツール，及び正解判定

本実験では，科学研究費補助金データベース(KAKEN)の採択課題データにおいて，タイトル，キーワード，概要，著者名，及び 2011 年において使用可能な研究分野(291 件)が記載されている 242,772 件の採択課題を用いる．この内，240,722 件を訓練用データ，2,000 件を評価用データとして用いる．タイトル，キーワード，概要の形態素解析には MeCab を用いる．

本実験は CiNii Mining における CiNii 論文情報に対して研究分野の自動付与を想定している．しかし CiNii 論文検索 API では，KAKEN のようなキーワード群は取得できないため，本実験ではクエリファイル作成において，評価用データからキーワードを取り除いた．

正解判定として本実験では，評価用データへ自動付与した研究分野名と評価用データに元々記載されている研究分野名が一致した場合に対して正解と判断する．また，検索インタフェースの作成において，k-NN 法における k の値について，本実験では 1 件ずつ値を変えていき，1~15 件までの k の値に対する正解率を示す．

5.2.ベースライン

本実験では，頻度付き索引ファイル及びクエリファイルの作成において，要素技術リスト，属性リスト，属性値リストを用いない場合をベースラインとして定義する．以下では，ベースラインで使用するリストについて説明する．まず，KAKEN 内に付与されているキーワードを収集してリスト化する．そして，訓練データのタイトル，キーワード，概要に対して形態素解析を行い，もし得られた語句が上記で作成したキーワードリストに存在していれば，重みとして 14 を与える．ここで，形態素解析して得る語句は名詞または(接頭詞を含んだ)名詞句のみとする．抽出された語句の総数は 1,415,482 個であった．

5.3.比較手法

頻度付き索引ファイル及びクエリファイルの作成の際に用いる比較手法を述べる．

BASELINE:4.3 節で述べたキーワードリスト，及びキーワードリストに属さない名詞または名詞句への重み付けを用いる．

METHOD:4.3 節で述べた要素技術リスト，属性リスト，属性値リスト，著者名，及びこれらのリストに属さない名詞または名詞句への重み付けを用いる．

5.4.実験結果

まず，タイトル，概要，著者名を含んだ論文情報に対して研究分野の自動付与を行った場合の結果について述べる．比較手法の評価結果を図 3 に示す．横軸は k-NN 法における k の値，縦軸は各 k の値における正解率を示す．

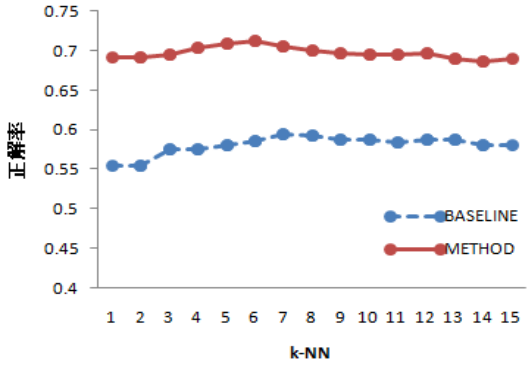

図 3:実験結果(1)

図 3 から，提案手法の性能はベースラインの性能を全体的に上回っていることから有効であったといえる．そして，提案手法において k=6 の時の性能が最も高く，正解率は 71.25%であり，正解件数は 1,425 件であった．

次に，タイトル，著者名のみを含んだ論文情報に対して研究分野の自動付与を行った場合の結果について述べる．比較手法の評価結果を図 4 に示す．

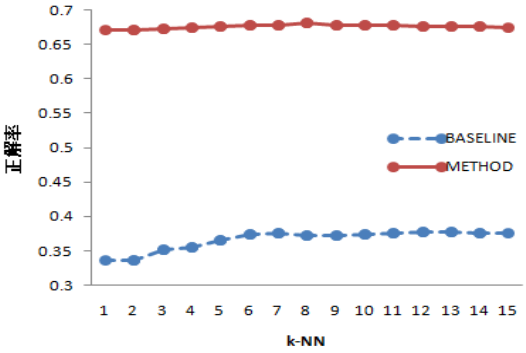

図 4：実験結果(2)

図 4 から，提案手法の性能はベースラインの性能を全体的に上回っていることから有効であったといえる．そして，提案手法において k=8 の時の性能が最も高く，正解率は 68.33%であり，正解件数は 1,366 件であった．また，概要が存在しないため，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，分野の付与が行えなかったデータが発生した．ベースライン手法では 20 件，提案手法では 1 件発生した．

5.5.考察

本節では主に，5.4 節の図 3 において最も正解率が高かった k=6 の結果を用いて考察を行う．

まず，誤りであった評価用データ 575 件内の研究分野を見ていく．以下の表に，上位 6 件の結果を示す．

表 3：誤りの評価用データ内の研究分野(細目)と件数

研究分野(細目)	件数	研究分野(細目)	件数
社会学	14	環境技術・環境材料	8
医用生体工学・生体材料学	12	経済政策	8
機能物質化学	9	社会福祉学	7

次に，評価用データへ自動付与した研究分野名(細目)と評価用データに元々記載されている研究分野名(細目)それぞれに対して，1 つ上の階層に位置する研究分野名(文科)に置き換えて正解判定を行った．その結果，正解率は 79.9%となった．この時の誤りであった評価用データ 402 件における研究分野(文科)の上位 6 件の内訳を以下の表に示す．

表 4：誤りの評価用データ内の研究分野(文科)と件数

研究分野(文科)	件数	研究分野(文科)	件数
社会学	21	複合化学	16
内科系臨床学	18	人間医工学	15
基礎医学	16	環境学	14

表 3，表 4 から，誤りであった評価用データの文科と細目，及びその文科に属している細目数の関係を調べると，表 5 のような関係が分かった．

表 5：誤りであった評価用データの文科と細目の関係

研究分野(文科)	研究分野(細目)	細目数
社会学	社会学　社会福祉学	2
複合化学	機能物質化学	6
人間医工学	医用生体工学・生体材料学	3
環境学	環境技術・環境材料	4

表 5 から，表 3 の「社会学」，「社会福祉学」などの誤分類は細目レベルでなく，文科レベルから発生していることが分かる．この結果から，文科レベルにおける分野の明確な特徴や違いを学習させ，他の文科レベルでの誤分類を無くすことで，正解率の向上に繋がると考えられる．

また，表 3 の「経済政策」について，「経済政策」の 1 つ上の階層に位置する「経済学」に属している細目数を調べると，「理論経済学」や「経済史」など 10 件の分野が存在した．このように，文科に属する細目数が多い場合，文科内における細目毎の明確な特徴を表す語句を新たに用意して学習に加えることで，細目毎の差別化が図れ，正解率の向上に繋がると考えられる．また，図 4 の場合にも同様に適用させることで，クエリデータの情報量が豊富になり，類似度の計算を十分に行うことが出来ると考えられる．

6.おわりに

本稿では，国立情報学研究所の論文情報ナビゲータ CiNii のデータベースを対象に，検索結果を研究分野毎に自動的に分類し，さらに分野毎の動向分析結果を提示する CiNii Mining のシステム概要について紹介した．そして，CiNii Mining における研究分野の自動分類の精度を向上させるために，学術論文情報から「要素技術」，「属性」，「属性値」，及び「著者名」を抽出し，汎用連想計算エンジン GETA を用いて高速に実行する手法を提案した．その結果，比較実験において全体で 7 割程度の分類精度が得られた．

今後の展望として，ユーザによって発生する表記揺れに頑健なシステムや，入力キーワードをシステムの内部で翻訳し，日英両方の論文情報を取得できるシステムを CiNii Mining に組み込むことで，より高度な検索結果画面の提供を目指す．

参考文献

[福田 2011] 福田 他:"技術文書からの動向情報の抽出と可視化". 言語処理学会第 17 回年次大会, 2011.

[Gupta 2011] Gupta et al.:"Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers". Proceeding of IJCNLP 2011, pages 1-9, 2011.

[内堀 2003] 内堀:"著者名・出版者名を組み入れた図書の自動分類". 平成 15 年度慶応義塾大学卒業論文, 2003.

[Salton 1971] Salton:"The SMART Retrieval System". Experiments in Automatic Document Processing, Prentice Hall Inc., 1971.