複数研究室の研究情報集約システムの実装

杉嵜 諒吾[1]　打矢 隆弘[1]　内匠 逸[1]

概要：近年のインターネット技術の普及，及びオープンサイエンスの推進により，研究成果をインターネット上から集めることが容易になりつつある．それに伴い，多くの研究成果を収集・管理する手間が問題となる．本研究では，研究成果を研究室単位にて自動で収集・解析し，ユーザが少ない手間でこれらの情報を得るシステムの開発を目指す．

Implementation of Summarization for Laboratories' Research Information

RYOA SUGISAKI[1]　TAKAHIRO UCHIYA[1]　ICHI TAKUMI[1]

1. はじめに

近年，インターネット技術の普及により，Web ブラウザを介した情報収集を多くの人が手軽に行えるようになった．検索サービスを提供する Web サイトにアクセスし，キーワードを入力することで，キーワードに関する情報を含む Web サイトを容易に見つけることができる．検索サービスを提供している Google[1] は，自社の検索エンジンが処理している 1 日の検索数が約 30 億クエリ (2015 年 1 月時点) であることを明らかにしており [2]，人々が日常的に情報検索を行っていることがわかる．また，ブログサイトや Social Networking Service(SNS) 等を利用することで，情報発信を個人で行うことも容易となった．

インターネットでの収集，発信が容易となった情報の一つに学術情報がある．昨今は多くの研究室が Web サイトを運営している上，論文や書籍の情報を提供する CiNii[3]，研究者情報を提供する researchmap[4] などの Web サービスが充実してきており，これらを利用することで人々は学術情報を詳細に得ることができる．また，研究成果のオープンアクセス (OA) の動きが進んでおり，米国や英国では公的資金を使った研究を対象に論文の OA 義務化を段階的に進めている．日本も 2016 年 1 月 22 日に閣議決定された第 5 期科学技術基本計画にてオープンサイエンスの推進を掲げており [5][6]，今後は今まで以上に多くの学術情報の提供が期待される．

一方で，多くの情報の取得や比較・管理をする手間が，情報量の増加に伴って増大している．特に，目的の情報が細かく定まっていない場合は情報を幅広く取得せざるを得ず，利用者自身の意図に合う情報を見つけ出すことが困難になってしまう．また，複数の Web サイトにて情報が別々に提供されているため，それらの情報を網羅的に得るための手間がかかってしまう．

2. 基礎知識

本章では，本研究に関連する基礎知識について述べる．

2.1 テキストマイニング

テキストマイニングとは，テキストデータを対象としたデータマイニングのことである．つまり，統計学やパターン認識などのデータ解析技法を大量のテキストデータに適用することで知識を取り出す技術である．自由記述のアンケートや書籍，SNS の投稿など，自然文の形で記述されたテキストデータは数多く存在しているため，解析により有益な知見を得ることが期待できる．テキストデータは数値化・定型化することが難しかったが，自然言語処理の発展により，実用的な水準での解析が可能となった．

[1] 名古屋工業大学 大学院 工学研究科 情報工学専攻
〒 446-8555 愛知県 名古屋市 昭和区 御器所町

2.1.1 形態素解析

形態素解析とは，自然言語で書かれた文章を形態素[*1]に分割し，品詞を判別することである．狭義ではこれを計算機において自動で処理する技術のことを指す．英語の場合は予め単語間が明確に区切られているため，形態素への分割が容易である．しかし，日本語文法における明確な区切りは句読点の部分的な挿入のみであるため，形態素への分割が困難である．日本語文の分割を目的として，形態素解析の技術が用いられることが多い．テキストマイニングにおいても，形態素解析は主に統計的処理の前処理として用いられる．

2.1.2 tf・idf 法

tf・idf 法とは文書に出現する単語に重み付けを行う手法であり，テキストマイニングの他，情報検索分野における索引語の重み付け手法として利用される．tf・idf 法は，文書を特徴付ける単語はその文章に多く登場し，他の文章にはあまり登場しないものであるという考えに基づいている．テキストマイニングの分野では各文書に登場する単語の重み付け値 (以下 tf・idf 値) をベクトル化し，文書の特徴を表す値として利用される．

文書 j における単語 i の tf・idf 値 $w_{i,j}$ は，文書 j における単語の出現頻度を $tf_{i,j}$，文書集合 D の総数を $|D|$，D のうち単語 i を含む文書の総数を df_i とすると式 (1) で表現できる．

$$w_{i,j} = tf_{i,j} \cdot \log \frac{|D|}{df_i} \tag{1}$$

2.1.3 コサイン類似度

コサイン類似度とは，ベクトルの類似性を示す指標の一つである．この指標は 2 つのベクトルのなす角 θ の余弦を類似度としているため，値が 1 に近づくほど θ の値が小さく，類似しているベクトルであると言える．テキストマイニングではコサイン類似度を 2.1.2 項で述べた tf・idf 値のベクトルに適用することで，文章の類似性の有無を示すことができる．コサイン類似度の計算式は，文書 A, B の tf・idf 値ベクトル v_A, v_B を用いると式 (2) で表現できる．

$$\cos(v_A, v_B) = v_A \cdot v_B = \frac{\sum_i w_{iA} w_{iB}}{|v_A||v_B|} \tag{2}$$

2.2 クラスタリング

クラスタリングとは，多変量データを自動的に分類する教師なし学習の手法，およびアルゴリズムである．クラスタリングにより，対象となるデータ集合は類似したデータ同士で構成される部分集合 (クラスタ) に分けられる．

2.2.1 凝集型階層的クラスタリング

凝集型階層的クラスタリングとは，類似性の高い二つのクラスタを逐次的に併合し大きなクラスタを生成する，クラスタリングの一手法である．分析者は二つのクラスタの非類似性を表すクラスタ間距離を定義し，その値が小さいクラスタ同士から順に併合する．併合の過程はデンドログラムによって表される (図 1)．分析者はこれを基に任意のクラスタ間距離でクラスタを分割することで複数のクラスタを得ることができる (図 2)．

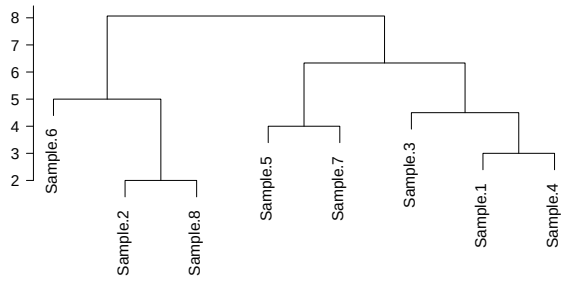

図 1 デンドログラム

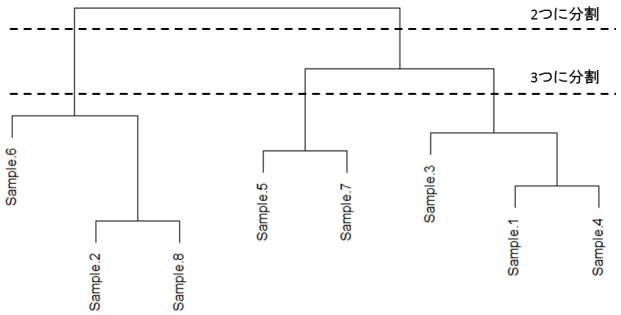

図 2 クラスタの分割

3. 関連研究

論文をクラスタリングする手法を用いた研究として，論文のメタ情報を利用した研究履歴自動抽出・可視化システム [7] がある．この研究は研究履歴の自動生成をするために，メタ情報を基に論文をクラスタリングし，各クラスタ毎に研究テーマを抽出している．システムの概要を図 3 に示す．

3.1 メタ情報収集モジュール

メタ情報収集モジュールは，NII 論文情報ナビゲータ CiNii[3] と科学研究費補助金データベース KAKEN[8] から著者の論文メタ情報を取得するモジュールである．CiNii

[*1] 言語において意味を持つ最小の単位

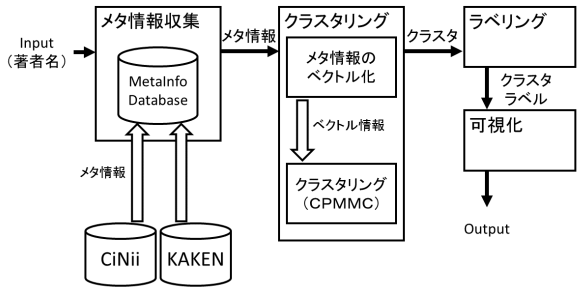

図 3 システムの処理フロー

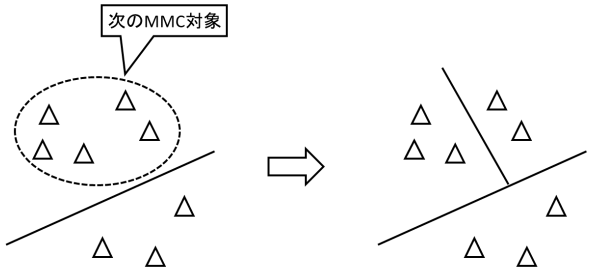

図 4 MMC の繰り返し

からは著者情報，出版年，キーワード*2，引用情報を取得する．KAKEN からは関連プロジェクト情報を取得する．CiNii，KAKEN はいずれもウェブサービスであるため，これらのサイトにリクエストを送信してレスポンスメッセージを受信し，メッセージの解析によりメタ情報を得る．また，このモジュールはデータベースシステムを有しており，収集したメタデータをデータベースに格納している．

3.2 クラスタリングモジュール

クラスタリングモジュールは，論文のクラスタリングを行うモジュールである．メタ情報収集モジュールのデータベースからメタ情報を取得し，各論文のメタ情報をベクトル化する．そして，ベクトル情報に基づいて論文のクラスタリングを行う．

論文のクラスタリングには，マージン最大化クラスタリング (MMC) を高速化した CPMMC[9] を利用する．MMC はサポートベクターマシン (SVM) を教師なし学習に拡張したものであり，クラスタを分割する超平面に加えて全データへのラベル割り当ても変化させてマージン最大化を行い，データが属するクラスタの決定をするクラスタリングである．1 回の MMC で論文集合は 2 つのクラスタに分割されるが，この研究では論文数が最も多いクラスタに対して MMC を繰り返し行うことで 3 つ以上のクラスタを生成している (図 4)．

3.3 ラベリングモジュール

ラベリングモジュールは，クラスタリングモジュールにて生成された各論文クラスタに対して研究テーマ名を決定するモジュールである．クラスタ毎に論文のキーワードを参照し，出現頻度が最も高いキーワードを研究テーマ名とする．

3.4 可視化モジュール

可視化モジュールは，抽出された研究テーマとその研究期間を可視化するモジュールである．研究期間は，論文の出版年を基に計算する．可視化には Java の SWT(Standard Widget Toolkit) と Jface Toolkit を利用し，研究テーマと研究期間をブロックで表現する (図 5)．

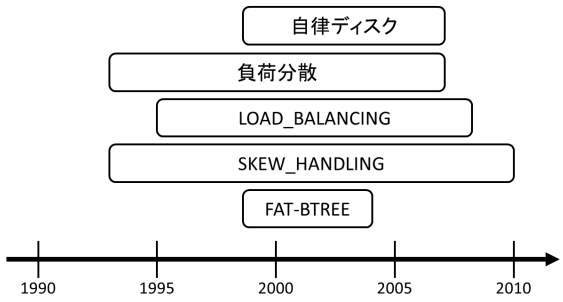

図 5 研究履歴の可視化 (イメージ)

3.5 考察

この研究では，論文のメタ情報の解析によって研究履歴を自動的に抽出するシステムを提案している．しかし，この研究で行われているクラスタリングでは，クラスタ数の設定に明確な基準が設けられていない．そのため，論文集合の過度な分割や分割不足によるクラスタリング精度の低下に繋がる恐れがある．また，論文数の多いクラスタを再分割する手法を取っているため，このシステムで生成されるクラスタは論文数が同程度になりやすいが，実際は研究が行われている年数の差により，研究テーマごとの論文数に差があるため，クラスタリング精度の低下に繋がる．

4. 提案機構

4.1 提案機構の設計

提案機構はインターネット上の情報を解析，集約し，複数の Web サイトの情報を単一の Web サイトで提供する．これにより，ユーザは研究分野や研究テーマといった研究室情報を効率よく得ることが可能になる．提案機構は以下の 4 つの内部機構で構成される (図 6)．

- 情報管理機構
  情報管理機構は，インターネット上から取得した情報

*2 論文冒頭に研究分野や利用した手法等を記載し，タイトルの情報を補うもの

や情報解析機構での解析結果を保管し管理する機構である．他の機構から送られる情報を保管し，クエリを受け取った際は要求された情報を返す．

- 情報取得機構
  情報取得機構は，インターネット上の Web サイトやデータベースから研究室に関する情報を取得する機構である．この機構は情報管理機構から URL 情報を取得し，一定間隔ごとにリクエストを送信してレスポンスを得る．取得した情報は情報管理機構に格納する．この機構で取得する情報 (ファイル) を以下に記す．
  – 学術論文
    研究内容について書かれた論文を，論文書誌を扱う CiNii 等のデータベースにアクセスし取得する．論文は pdf ファイルフォーマットのものを扱う．
  – 研究室のホームページ
    研究室が運営しているホームページから html ファイルを取得する．取得した Web ページに同一サイト内の別のページへの URL が含まれていた場合，その URL を情報管理機構に格納することで新たなアクセス先とする．
- 情報解析機構
  情報解析機構は，情報管理機構に格納された情報を解析し有益な情報を取り出す機構である．インターネット上から取得した情報はそのままでは情報量が多すぎるため，この機構での解析によって情報の抽出や要約を行う．
- 情報提供機構
  情報提供機構は，収集や解析によって得られた情報をユーザに提供する機構である．ユーザは Web ブラウザを通じて提案機構にアクセスすることで，研究室に関する情報を Web ページとして得ることができる．

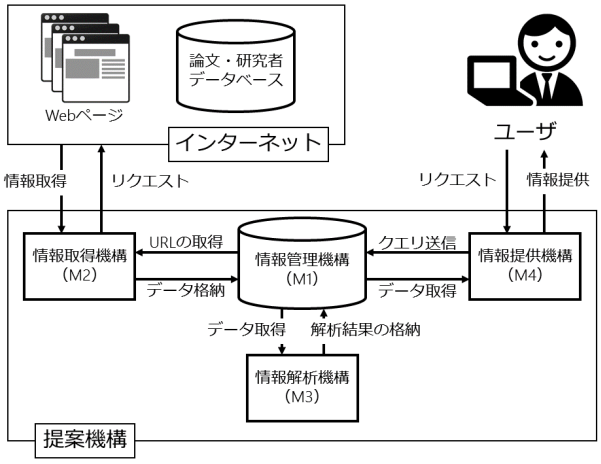

図 6 システム概要

4.2 プロトタイプシステムの設計・実装

情報解析機構のプロトタイプシステムとして，論文の解析によって研究室の情報を抽出するシステムを実装した．本システムは，ローカルに保存した pdf ファイルからテキスト情報を抽出し解析することで，各研究室が扱っている研究テーマの抽出を行う．また，論文の解析結果を他の研究室のものと比較することで，似た研究を行う研究室を探索する．プロトタイプシステムの実装には，PDF ビューアの Xpdf[10] で用いられているプログラム，及び形態素解析を行う MeCab[11]，計 2 つのツールを用いた．

本システムは最初に論文のクラスタリングを行う．クラスタリングには凝集型階層的クラスタリングを用いる．この手法を用いることで，研究室毎のクラスタ数やクラスタ毎の論文数の適正化が可能であり，クラスタリング精度の向上を図ることができる．

5. 実験と評価

本章では，プロトタイプシステムを用いた評価実験を行う．

5.1 概要

評価実験では，クラスタリングと研究テーマ抽出の精度の評価を行う．データセットには本学の卒業論文発表会，及び修士論文審査会の予稿集を用いる．クラスタリングや研究テーマ抽出の精度の評価は，データセットをプロトタイプシステムにて解析した結果を，人手によって分類，研究テーマ抽出を行ったものと比較することで行う．

5.2 評価用データセット

本実験では評価用データセットとして，名古屋工業大学における「平成 26 年度 情報工学科知能系プログラム 卒業研究発表会」の予稿 50 稿，及び名古屋工業大学大学院工学研究科における「平成 26 年度 情報工学専攻（知能科学分野） 修士論文審査会」の予稿 20 稿[*3]の pdf ファイルを用いた．これらの 71 稿を，著者が所属している研究室ごとに分けて用いた (表 1)．また，プロトタイプシステムの評価を行うために，論文を人手によって分類し，人手で研究テーマを付与した (表 2)．以降の評価では，これを正解分類結果とする．

5.3 実験 1：クラスタリング精度の評価

プロトタイプシステムによるクラスタリング精度を評価するために，評価用データセットをシステムにて分類し正解分類結果と比較することで分類の妥当性を検証した．評価には Rand 尺度 [12] を用いた．

[*3] 全 21 稿のうち全て英語で書かれた 1 稿を除外している

表 1 評価用データセット

	卒業研究発表会	修士論文審査会	合計
伊藤研究室	6	0	6
犬塚研究室	4	1	5
打矢研究室	5	4	9
加藤研究室	6	8	14
新谷・大園研究室	5	3	8
世木研究室	3	0	3
内匠研究室	5	1	6
竹内研究室	6	0	6
中村研究室	4	3	7
山岸研究室	1	0	1
和田山研究室	5	0	5

表 2 正解分類結果

研究室	研究テーマ	論文数
伊藤研究室	エージェント	2
	行動戦略	2
	機械学習	1
	学習支援	1
犬塚研究室	データマイニング	3
	オントロジー	1
	エージェント	1
打矢研究室	エージェント	6
	MMDAgent	2
	情報推薦	1
加藤研究室	機械制御	5
	運動推定	4
	感情推定	3
新谷・大園研究室	PowerPoint	3
	Web	2
	情報共有	2
	介護支援	1
世木研究室	データマイニング	3
内匠研究室	脳波	3
	MMDAgent	2
	位置推定	1
竹内研究室	機械学習	4
	秘密計算	1
	個別化医療	1
中村研究室	オノマトペ	4
	CG	2
	ヒューマンインタフェース	1
山岸研究室	幾何学	1
和田山研究室	符号化技術	5

5.3.1 Rand 尺度

Rand 尺度とは，クラスタリングによるシステム分類結果を正解分類結果と比較することで分類の妥当性を評価する尺度の一つである．データ集合中の全ての対 $x_1, x_2 \in X (x_1 \neq x_2)$ がシステム分類結果・正解分類結果において同じクラスタにあるか否かを調べ，2 つの分類結果で一致した対の割合を求める．Rand 尺度は 0 以上 1 以下の値をとり，1 に近いほど良いとされる．対の数を M，2 つの分類結果の両方で同じクラスタにあった対の数を a_{11}，両方で別のクラスタにあった対の数を a_{00} とすると，Rand 尺度は式 3 で表される．

$$Rand = \frac{a_{11} + a_{00}}{M} \tag{3}$$

5.3.2 評価方法

評価用データセットの論文を研究室ごとにクラスタリングし，Rand 尺度の平均値を算出した．正解分類結果は全体で 29 の研究テーマに分類されたため，システムの階層的クラスタリングではクラスタ数が 26〜32 となるように分割の閾値を設定し，それぞれの結果に対して評価を行った．また比較のために，論文間の類似度に 0 以上 1 未満の一様分布に従う乱数を与えてクラスタ数が 26〜32 となるようにクラスタリングした場合，各研究室において全ての対が同じクラスタである場合 (クラスタ数 11)，全ての対が異なるクラスタである場合 (クラスタ数 70) において同様の評価を行った．

5.3.3 評価結果

クラスタリング精度の評価結果を表 3 に示す．なお，山岸研究室は論文が 1 つであり Rand 尺度の算出ができないため除外している．クラスタ数 26〜32 における評価値の平均は 0.6131 となった．クラスタ数 26〜32 のいずれにおいても，論文の類似度をコサイン類似度で定義したものの方が評価値は高くなっており，クラスタ数 11,70 における評価値を上回っている．このことから，今回使用したデータセットのクラスタリングには一定の妥当性があることが示された．

表 3 クラスタリング精度の評価結果

	類似度	
クラスタ数	コサイン類似度	乱数
26	0.6223	0.4727
27	0.6168	0.4799
28	0.5968	0.4999
29	0.6039	0.5106
30	0.6061	0.5128
31	0.6218	0.5150
32	0.6328	0.5364
11	0.4476	-
70	0.5524	-

5.4 実験 2：研究テーマ抽出精度の評価

プロトタイプシステムにおける研究テーマ自動抽出の精度を評価するために，システムによって抽出された研究テーマを正解分類結果の研究テーマと比較した．

5.4.1 評価方法

プロトタイプシステムにより研究テーマ抽出を行い，正解分類結果の研究テーマがいくつ含まれるかを数えた．システムによるクラスタリングではクラスタ数が 32 となる

ように分割の閾値を設定した．なお，研究テーマが文字列として一致していなくても，指し示す内容が同一であると判断される場合は同一の研究テーマとした．

5.4.2 評価結果

評価結果を表 4 に示す．正解分類結果における研究テーマ 29 個に対し，抽出できたのは 8 個であった．また正解分類結果での論文数に着目したところ，論文数が 3 稿以上ある研究テーマは 11 個中 6 個抽出されたのに対し，2 稿以下の研究テーマは 18 個中 2 個のみ抽出された．以上のことから，現在のプロトタイプシステムでは抽出されない研究テーマが存在し，論文が少ない研究テーマほど抽出される可能性が低くなると考えられる．

表 4 抽出された研究テーマ

研究室	研究テーマ	一致
伊藤研究室	エージェント 試作 PV-CBOW アーム	○
犬塚研究室	super 論点 候補パターン 属性	
打矢研究室	エージェント 重複排除	○
加藤研究室	制御 自己主張力 推定 推計 BV-MDPs	
新谷研究室	被介護者 オブジェクト 検索結果 操作 登山計画書	
世木研究室	マイニング 次元削除	○
内匠研究室	MMDAgent 運動 速度	○
竹内研究室	学習 秘密計算 組み合わせ	○ ○
中村研究室	オノマトペ	○
山岸研究室	フラクタル図形	
和田山研究室	符号化 2 次元	○

5.5 総括

クラスタリング精度の評価結果によって，今回使用したデータセットのクラスタリングに一定の妥当性があることを示した．このデータセットは情報工学分野の研究内容に関して記述された論文集合であるため，情報工学分野を研究している他の研究室の論文の解析にも高い精度が期待できる．一方で研究テーマ抽出精度の評価結果を見ると，論文の解析により一部の研究テーマの抽出が可能であるものの，現在のプロトタイプシステムでは十分な精度で抽出しているとは言い難く，改善の余地があると考えられる．

6. まとめ

本研究は，本学の研究室情報の収集，管理に伴うユーザの手間を解消する事が目的である．本論ではインターネット上から研究室情報を自動で取集し解析，要約したものを単一の Web サイトとして提供するシステムを提案した．また，提案機構の一機能として，論文の解析による研究テーマの抽出，及び類似した研究を行う研究室の探索を行う機能を設計し，プロトタイプシステムを実装した．さらに，プロトタイプシステムを用いて評価実験を行った．評価実験の結果から，提案手法の有効性を確認することができた．今後は，情報提供機構のプロトタイプを作成し，ユーザに対する有効性を調査していく．

参考文献

[1] “Google”, https://www.google.co.jp/.
[2] “How Google Search Dealt With Mobile”, https://medium.com/backchannel/how-google-search-dealt-with-mobile-33bc09852dc9/.
[3] “CiNii Articles”, http://ci.nii.ac.jp/.
[4] “researchmap”, http://researchmap.jp/.
[5] “平成 28 年 1 月 22 日 (金) 定例閣議案件”, http://www.kantei.go.jp/jp/kakugi/2016/kakugi-2016012201.html/.
[6] “第 5 期科学技術基本計画に向けた中間取りまとめ” http://www8.cao.go.jp/cstp/tyousakai/kihon5/chukan/index.html/.
[7] Nguyen Manh Cuong, 加藤大智, 橋本泰一, 横田治夫, “論文のメタ情報を利用した研究履歴自動抽出・可視化システム”, DEIM Forum 2011 C2-2, 2011.
[8] National Institute of Information. KAKEN. http://kaken.nii.ac.jp/.
[9] Bin Zeao , Fei Wang, and Changshui Zang, “Efficient maximum margin clustering via cutting plane algorithm”, SIAM International Conference on Data Mining, pp.751-762, 2008.
[10] Foo labs, “Xpdf”, http://www.foolabs.com/xpdf/.
[11] 工藤拓, “MeCeb”, http://taku910.github.io/mecab/.
[12] William M Rand, “Objective Criteria for the Evaluation of Clustering Methods”, Journal of The American Statistical Association, Vol.66, pp.846-850, 1971.