1L-04

複数研究室の研究情報集約システムの提案

杉嵜　諒吾†　　打矢　隆弘‡　　内匠　逸‡

†名古屋工業大学 工学部 情報工学科
〒 446-8555 愛知県 名古屋市 昭和区 御器所町

‡名古屋工業大学 大学院 工学研究科
〒 446-8555 愛知県 名古屋市 昭和区 御器所町

1　はじめに

研究室の情報は，主に各研究室が運営する Web サイトによってインターネット上で公開されている．また，学術論文情報を検索できる CiNii[1] や，研究者のプロフィールを管理，公開する researchmap[2] といったインターネット上のサービスが充実し，これらを利用することで研究室での研究内容をより詳細に得ることができる．情報を求めている学生や企業はインターネット上で Web サイトにアクセスして情報を得るが，情報を得る初期の頃は複数の研究室の情報を調べて比較を行うことが多いと考えられる．その場合，複数の研究室を比較するために多くの Web サイトを閲覧する必要があり，大きな手間となる．そこで本研究では，特定の大学の学科に属する研究室に関する情報を解析，集約，表示するシステムを提案する．これにより，研究室情報の効率的な提供を実現する．

2　先行研究

加藤ら [3] は，研究者の論文をラベル付きのクラスタに分類し，クラスター一つを一つの研究内容とすることによって，研究履歴を自動生成する手法を提案している．その中ではクラスタリング手法として *K*-means 法が用いられている．この手法は，クラスタのアイテム数が等しいことを暗に仮定しているが，実際は研究を行っている期間の差などにより，各研究テーマに対する論文の数に大きな差があることが多いため，高いクラスタリング精度が期待されない．また，*K*-means 法ではクラスタの数をクラスタリング前に定める必要があるが，この手順は自動化されていない．

3　提案機構

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 1).

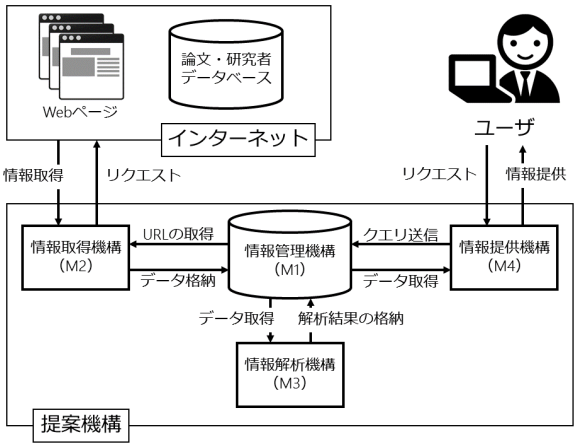

図 1: システム概要

(M1) 情報管理機構
収集したインターネット上の情報，及び提案機構内での解析結果を保存する

(M2) 情報取得機構
インターネット上の Web サイトやデータベースから定期的に情報を取得し，機構 M1 に格納する

(M3) 情報解析機構
収集した Web ページ，論文等の解析を行い，研究室の特色や研究内容などの研究室情報を抽出する

(M4) 情報提供機構
提案機構にて保持している研究室情報を，Web ページとしてユーザに提供する

4　論文解析機能

情報解析機構 (M3) の一機能として，研究室に所属する研究者の論文を解析，集約することにより研究室の情報を得る機能を導入する．

4.1　研究テーマの自動抽出機能

各研究室から公表された論文を研究室毎にクラスタリングし，クラスタ毎にラベル付けを行う．これによ

Proposal of Summarization for Laboratories' Research Information
†Ryoa SUGISAKI　‡Takahiro UCHIYA　‡Ichi TAKUMI
†School of Engineering, Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya, Aichi,466-8555 Japan
‡Graduate School of Engineering, Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya, Aichi,466-8555 Japan

り，各研究室が扱っている研究テーマをラベルとして取り出す．また，研究テーマごとに論文を分類分けする．

4.2 類似した研究を行う研究室の推薦

前項の機能で生成した論文のクラスタを利用し，類似した研究を行う研究室を推薦する機能である．ユーザに対し，特定の研究室の情報を提供する際に併せて表示し，研究内容の比較を支援する．また，類似したクラスタのラベルを提示することで，推薦における高い説明性の実現を図る．

5 論文解析機能の実装

5.1 研究テーマの自動抽出機能

論文の特徴量として，tf-idf 法による単語への重み付け値（以下 tf-idf 値）を定義する．そして，論文間の距離に tf-idf 値ベクトルのコサイン類似度の逆数を用いて階層型クラスタリングを行う．また，各クラスタの論文において，tf-idf 値が高い単語をクラスタの研究テーマとする．

tf-idf 法は，文書に出現する単語に重み付けを行う手法である．この手法では，「文書を特徴付ける単語はその文書に多く登場し，他の文書にはあまり登場しない」という考えに基づき単語に重み付けが行われる．tf-idf 法によって文書 j の単語 i に与えられる tf-idf 値 $w_{i,j}$ は，文書 j における出現頻度 $tf_{i,j}$，文書の集合 D の総数を $|D|$，D のうち単語 i を含む文書を df_i とすると式 (1) で表される．

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

ここで，本機能によって特定の研究室の論文を解析する手順を以下に示す (図 2).

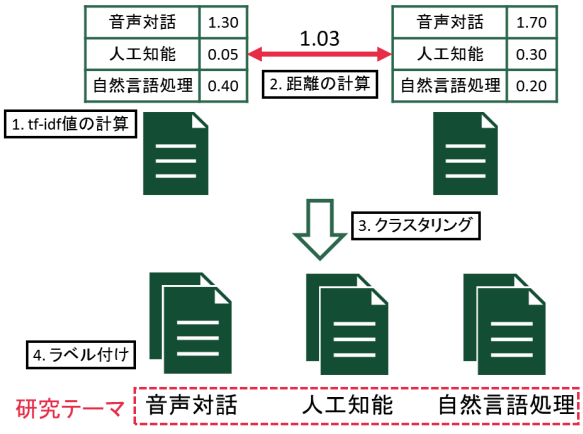

図 2: 研究テーマの自動抽出機能

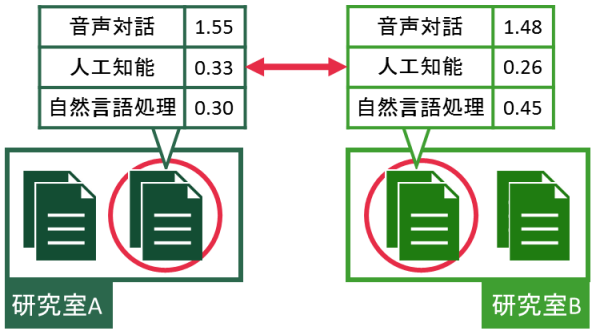

図 3: 研究テーマの比較

1. 論文内における名詞の tf-idf 値を計算し，論文ごとにベクトルとして保持する
2. 論文が保持するベクトルを用いて，全ての論文同士の距離を計算する
3. 階層型クラスタリングを行い，クラスタ間距離を基にクラスタ数を決定する
4. クラスタごとに論文の tf-idf 値の平均を計算し，値が最も高い名詞をクラスタの研究テーマとする

5.2 類似した研究を行う研究室の推薦

論文のクラスタの特徴量として，論文の tf-idf 値の平均値を用いる．他のクラスタとのコサイン類似度を計算し，類似度が高いクラスタ同士は内容が似た研究テーマとみなす．実際にユーザへ推薦を行う際は，特定の研究室が保持しているクラスタと類似度が高い上位 N 個のクラスタを保持している研究室を推薦する．

6 まとめと今後の予定

本稿では研究室情報を効率よく提供する手法として，複数の研究室に関する情報を解析，集約，表示するシステムを提案した．また，システムの一機能として，論文の解析を行う機能を提案した．今後は論文解析機能の実装，評価を行う．

参考文献

[1] “CiNii Articles”, http://ci.nii.ac.jp/

[2] “researchmap”, http://researchmap.jp/

[3] 加藤 大智, NGUYENMANH CUONG, 橋本 奏一, 横田 治夫, “論文のラベル付きクラスタリングのための情報利得を用いたキーワード選定”, DEIM Forum 2012 E10-1, 2012.