学会メーリングリストからの専門用語および人名の抽出と関連付け
Extraction and Association of Technical Terms and Personal Names from Academic Mailing Lists

不動 雄樹
Yuki Fudo
広島工業大学情報学部知的情報システム学科
Email: ba09218@cc.it-hiroshima.ac.jp

松本 慎平
Shimpei Matsumoto
広島工業大学情報学部知的情報システム学科
Email: s.matsumoto.gk@cc.it-hiroshima.ac.jp

***Abstract*—Using Internet information is effective for learning about new developments on a certain research field. Everyone can find research papers what they want with PDF search service of large search engines and Internet search services of academic articles such as CiNii, J-Stage, and Google Scholar, however it relies on one's search skill. Recently, since there have been many reports with similar techniques or objectives even though they belong to different academic societies, the work of survey that encompasses various academic fields is indispensable. Based on the background, until now, there have been some reports for supporting research survey: a system which can extract the information of persons from Web documents, and a system which can show researchers related with technical terms. This paper focuses on academic mailing list sending information including research meetings for the purpose of managing the latest trend of research. The work of survey is supported with structured information including the date of publication while associating technical terms and personal names.**

I. はじめに

研究領域の最新動向を調査する上で，インターネットの活用は既に一般的なものとなった．先行研究の成果を調査する際，大手検索エンジンの PDF 検索や，CiNii，J-Stage，Google Scholar といったサービスはもちろんのこと，学会大会のプログラム，Slide Share，論文 Relation といった論文検索支援サービスを活用すれば，検索スキルがある限り自分が求める情報をある程度得ることができる．最近では，所属学会が異なっていても類似の技術や目的での成果が数多く報告されているため，学会や分野を横断した検索が不可欠である．以上の背景を踏まえて，研究領域や論文関係，著者関係のネットワーク構造構築に向けての取り組みは近年多数報告されるようになった．また，得られた知見を活用して，Web 文書からの人物情報抽出するシステムや，専門用語に強く関係する研究者を提示するシステムなどの開発事例が報告されている．ここに報告年月という時間的要因を含めて，学術関係情報を構造的に管理する意義は大きいのではないかと考えられる．研究会発表や全国大会などのプログラムや研究活動状況を活用すれば，研究会への参加機会も含めて，より最新の研究活動状況を把握できるようになるのではないかと考えられる．

本稿では，研究の最新動向を把握することを目的として，研究会情報を配信している学会メーリングリストに着眼した．専門用語と人名を関連付けながら，報告年月を含めて情報を構造化管理することで，最新の研究動向調査を支援可能な著者データベース構築を目的とする．著者の名前と技術的専門用語を著者に対して属性値として関連付けることで，情報の効率的な管理と応用システム開発を目指す．著者と専門用語関係を明確に定義し，その運用として，著者名を検索エンジンのクエリとしての活用までを想定している．著者名で PDF 検索をすれば，論文を入手することができるためであり，また，著者が Web サイトを公開していれば，公開されている業績情報や研究室活動の情報から，学術論文や関連する研究，現在に至るまでの成果を順に辿っていくことができる．以上見解に基づき，本稿は，第一段階として，まず人名と技術用語が含まれるメールの抽出に取り組んだ．具体的には，人名と研究題目が含まれる研究会プログラムのメールを抽出するため，人工知能学会等のメーリングリストのデータに対して，単純ベイズ分類器を用いてメールの内容に基づく分類を行った．

II. 関連研究

本研究は，関連研究の調査支援を目指したものである．通常，調査の際には当該研究領域のキーワードか，第一人者の名前が重要となる．これらのどちらかかあるいは双方を利用して，論文や資料を入手する．よって，自動的な調査支援まで進めるためには，研究者と専門用語との関係をまず明確に定義する必要がある．例えば，研究者の関係については，木實らは，所属，共著した論文の有無，論文誌や書誌の共同編集の有無，同じ会議への論文投稿の有無で評価されると述べている [1]．専門用語の関係については，同一 Web ページ内での用語の共起頻度 [2] や，あるいは学術論文で与えられる著者キーワード中での共起情報の利用 [3] などによって通常定義される．

論文検索支援の開発事例としては，高久らは，具体的な語彙を適切に選択することが難しい状況にあっても，比較的容易な検索により，様々な関連分野の専門的な論文を発見することが可能なシステムを開発した [4]．鉢木らは，学術論文中の専門用語に対して解説等の有用なページへのリ

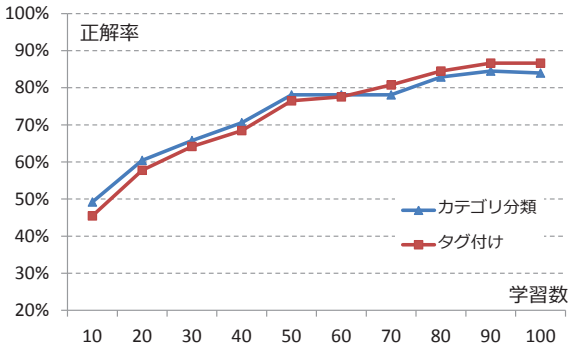

図 1.　研究報告資料における単語関係のネットワーク可視化

ンクを提供することで論文の閲覧を支援するシステムを開発した [5]．Web ページのリンク解析アルゴリズムで論文をランク付けし，論文閲覧時にその論文に関連する論文を推薦することで，論文閲覧を支援している．増田らは，入力された専門用語に関係する分野に詳しい研究者を CiNii から検索して，さらにその人物の関連情報を Web から抽出し，プロファイルを自動生成するシステムを提案した [6]．

専門用語と研究者を対応付けた成果としては，まず，橋本らの提案した研究者逆引きデータベースシステムがある [7]．また，森らは，Web 上の情報を用いて，研究者の情報をキーワードとして自動的に抽出する手法を提案した [8]．小林らは，Web ページに含まれる技術的概念の詳細やそれらの間の関係を表す手法を示した [9]．

III. 実験及び結果

実験では，人工知能学会メーリングリストを中心とした約 3000 通の電子メールに含まれるテキストデータを利用した．本研究では，日本語の電子メールのみを解析の対象とした．また，キーワードを多く含む研究会案内や発表プログラムの告知メールを抽出することを第一の課題とした．日本語文章であるため，まず KAKASI により分かち書きを行った後，単純ベイズ分類器によりメールのカテゴリ分けを試みた．カテゴリとして，“論文誌案内”，“学会研究会案内”，“教員・学生募集”，“ツール紹介”，“広告” を用意し，いくつかのメールを任意選出し学習データとして利用した．分類の方法は 2 種類用意した．一つはカテゴリ分類法であり，これは各メールに対してどれか一つのカテゴリを設定する方法である．もう一つはタグ付け法である．これは各カテゴリをタグとして見なした手法であり，各タグが付くか否かを各メール毎に学習させる方法である．よって，タグ付け法では，最大 5 つのタグが設定されることもある．本稿では，まずキーワード抽出の可能性を検証する必要があると判断し，3000 通の中から 300 通を任意に抽出し，単純ベイズ分類器の精度を実験的に確認した．

実験結果を図 1 に示す．実験のとおり，ある一定数の学習を行うことで，十分な精度で分類ができることがわかった．本稿での正解率は完全一致を表しているため，部分一致も成果と見なすと，100 通の学習で 95%程度の正解率であった．なお，両手法のカテゴリ分類一致率は，約 92%～約 98%であり，大きな相違は確認されなかった．ただ，タグ付け法の場合は分類の自由度が高いため，今後はタグ付け法を基準に研究を進める計画である．

次に，Web スクリプトを開発し，キーワード抽出に向けての基盤構築を行った．ここでは，MeCab を用いて，メール文書の形態素解析を試みた．MeCab のデフォルトの辞書だけでは精度が不十分であるため，IT 用語辞典 e-Words と Wikipedia の項目名を用いて辞書を作成し，また，CiNii API を利用して，人名辞書を作成した．これら辞書を MeCab に追加して，メール本文に対して形態素解析を施した．著者名を取得する方法として，まず，MeCab で形態素解析を行った結果出力される [名詞, 固有名詞, 人名] に注目した．これらキーワードを全て取得し，繰り返し処理で取得したキーワードを順に処理した．処理内容は以下の通りである．実験の結果，一般的な姓名を持つ著者名の取得には成功したが，正常に取得できない場合も確認された．

1) 姓→名の順で出た場合，それを結合して著者名とする
2) 姓→姓→名の順で出た場合，最初の姓を捨て，残りの姓名を結合して著者名とする
3) 姓→名→名の場合，3 番目の名前はを捨て，残りの姓名を結合して著者名とする

IV. おわりに

本稿では，研究の最新動向を調査する際人名が重要なキーワードとなることに着眼し，学会メーリングリストのテキストデータを用いて，専門用語と関連付けながら人名を抽出することを試みた．とりわけ本稿では，キーワードを多く含む研究会開催告知や発表プログラムの選別を第一の課題として取り組んだ．単純ベイズ分類器を用いて，メーリングリストのデータからメールの内容に応じたメールの分類を試みた結果，高い精度での分類に成功した．次に，IT 用語辞典 e-Words と Wikipedia を用いて専門用語辞書を構築し，また，CiNii のデータを用いて人名辞書を構築した．実験により，これら辞書を用いたキーワードの抽出に成功した．今後は，TermExtract の利用やパターン認識手法の適用などにより，キーワード抽出の精度を向上させたい．

参考文献

[1] 木實新一，井上創造，小林隆志，土田正士，喜連川優, 学術会議における参加者関係発見のためのネットワーク表示システムの利用, DEWS2006, 4A-o6, 2006.
[2] 松尾豊, 石塚満, 語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム, 人工知能学会誌 17(3), pp.217-233, 2002.
[3] 相澤彰子, 影浦峡, 著者キーワード中での共起に基づく専門用語間の関連度計算法, 電子情報通信学会論文誌, J83-D-I(11), pp.1154-1162, 2000.
[4] 高久雅生, 江草由佳, セレンディピティを促す論文検索ツール「ふわっと関連検索」, ディジタル図書館, No.38, pp.pp.35-41, 2010.
[5] 鉢木稔浩, 太田学, 高須淳宏, 学術論文閲覧支援システムのための関連論文推薦, DEIM Forum 2011, F9-4, 2011.
[6] 増田浩司, 太田学, CiNii を利用したエキスパートサーチシステム, DEIM Forum 2011, F6-4, 2011.
[7] 橋本泰一, 乾孝司, 内海和夫, 石川正道, 研究者逆引きデータベースシステムの構築, 2009 年度人工知能学会全国大会 (第 23 回), 2L1-4 ,2009.
[8] 森純一郎，松尾豊，石塚満, Web からの人物に関するキーワード抽出, 人工知能学会論文誌, Vol.20, No.5, pp.337-345, 2005.
[9] 小林慎一, 白井康之, 比屋根一雄, 粂野文洋, 犬島浩, 山内規義, インターネットリソースを用いた技術動向の時系列的分析, 電気学会論文誌 C, Vol.125, No.5, pp.720-729, 2005.

問い合わせ先
〒 731-5193
広島県広島市佐伯区三宅 2 丁目 1-1
広島工業大学情報学部知的情報システム学科
不動 雄樹