4S-2

キーワードの相関性に基づく研究者，研究機関ごとの特徴抽出

安藤志宙† 堀幸雄† 今井慈郎†
香川大学 工学部†

１．はじめに

現在，インターネットを通じて，膨大な数の論文を閲覧することが可能となっている．しかし，その中から全体を把握し，有益な情報を得るのは困難であり，それらをいち早く得るためには早い段階で個別の研究の特色を見抜くことが重要である．研究トピックに対応するような文書群を分析する方法として，文書群中に特徴的に現れるキーワードを抽出しリストアップする方法などが多く利用される．しかしながら，このような方法では全ての文書に共通なキーワードが現れる．

長尾らは全文献をいくつかの分野に分けて，分野ごとの単語の頻度を数え，ある単語が各分野に偏りなく出現すれば一般語，少数の分野に偏って出現すれば重要語と呼んでいる[1]．

そこで，本研究では，各研究者の出版した論文から，研究者，研究機関ごとのキーワードの相関性に着目し，個々の研究者，研究機関ごとの特色を抽出する手法を提案し，抽出結果の可視化を行う．

２．関連研究との比較

現在に至るまでにも，ある文書から，相関性に基づき特定の情報を抽出する研究，手法は多々行われている[2,3,4]．しかし，これらの手法では，一般的な語を抽出してしまう．そこで，本研究では，新しい手法として AIC(赤池情報量基準)[5]を提案し，各キーワード手法による比較，検討を行い，抽出結果の可視化を行う．

また，学術論文を対象にした，要素技術を抽出する研究[6]は過去に行われているが，研究者や研究機関そのものを対象にした研究は著者が知る限りでは前例がない．そこで，本研究では分野全体と，研究者，研究機関をキーワードの相関性に基づき比較し，研究者，研究機関の特色を抽出する．

３．提案手法

提案手法における研究者，研究機関ごとの特色のあるキーワードの抽出手法の概要を図１に示す．研究者，研究機関の発表している論文に偏って出現するようなキーワードを統計的な基準を用いて自動的に抽出する．

ここで，本研究で使用する研究者，研究機関のデータは下記の流れで収集した．

(1)中四国の研究機関リストを取得
(2)研究機関リストから，所属する研究者のデータを研究者 DB である J-GLOBAL より取得
(3)研究機関名，研究者名から，各研究者が発行した論文を，論文 DB である CiNii より取得
(4)得られた論文を分野ごとに分類する

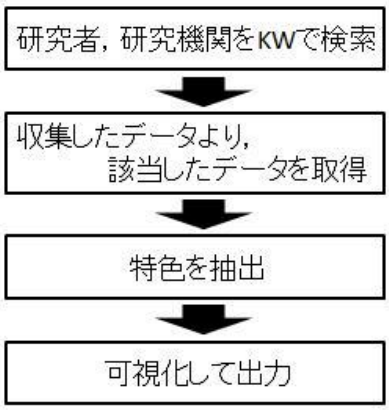

図１．処理フロー

３．１ 特色のあるキーワードの抽出

ある研究者，研究機関の発表した論文に偏って出現するキーワードを抽出する手法として，AIC を用いる．ここで，あるキーワード k が出現するある研究者，研究機関の発表した論文数 N_{11} と該当分野における論文数 N_{21}，キーワード k が出現しない発表した論文数 N_{12} と該当分野における論文数 N_{22} のある研究者，研究機関の発表した論文に含まれるキーワードについて求める(表１)．キーワード k が研究者，研究機関に偏って出現する度合い E(k)を AIC の独立モデルに対する値 AIC_IM および従属モデルに対する値 AIC_DM を用いて次のように定義する(式(1),(2))．

表１．AIC 算出に用いるキーワード k の出現回数

	k が出現	k が非出現	合計
対象	N_{11}	N_{12}	N_p
分野全体	N_{21}	N_{22}	N_n
合計	N(k)	N(¬k)	N

The Character Extraction of Researcher and Research Affiliation using Correlations Between Keyword Occurrences.
†Yukihiro Ando：Kagawa University

$$N_{11}(k) / N(k) > N_{12}(k) / N(\neg k) \text{ のとき}$$
$$E(k) = AIC\_IM - AIC\_DM$$
$$N_{11}(k) / N(k) \leq N_{12}(k) / N(\neg k) \text{ のとき}$$
$$E(k) = AIC\_DM - AIC\_IM \quad (1)$$

$$AIC\_IM(k) = -2 \times MLL\_IM + 2 \times 2$$
$$MLL\_IM = N_p(k)\log N_p(k) + N(k)\log N(k) + N_n(k)\log N_n(k) + N(\neg k)\log N(\neg k) - 2N\log N$$
$$AIC\_DM(k) = -2 \times MLL\_DM + 2 \times 3$$
$$MLL\_DM = N_{11}(k)\log N_{11}(k) + N_{12}(k)\log N_{12}(k) + N_{21}(k)\log N_{21}(k) + N_{22}(k)\log N_{22}(k) - N\log N \quad (2)$$

４．実験

表２に AIC および，tf-idf，出現確率比，χ^2 値によるキーワード抽出結果を示す．ここで，特色の抽出対象として，著者らの大学に所属する富永浩之氏を選出した．

結果としては，出現確率比が最も特色を抽出できていたのではないかと思う．提案手法である AIC は，研究者固有のキーワードを抽出することはできたが，それが特色と言えるものではなかった．

また，抽出結果を可視化したものを図２に示す．横軸は出現確率比を表し，縦軸は出現頻度を表す．研究者の特色と判断したキーワードを赤色で左側に，分野全体によく現れると判断したキーワードを青色で右側に表示する．キーワードが書かれたバブルをクリックすることで，ページ下部に，そのキーワードで論文検索をした結果を表示する．

５．おわりに

本研究では，AIC を用いて研究者，研究機関ごとの特色を抽出する手法を提案した．今後は，得られた研究者，研究機関の特色の分析，評価を行なっていきたいと考えている．

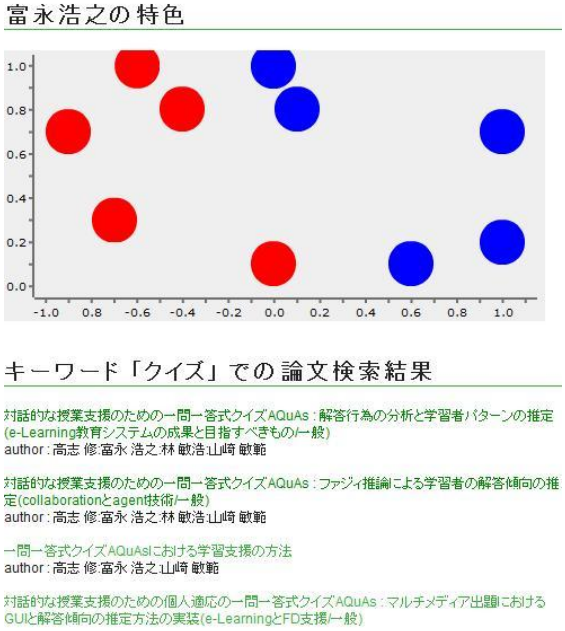

図２．出力結果

参考文献

[1] 長尾，水谷，池田：日本語文献における重要語の自動抽出，情報処理，Vol.17，No.2，pp.110-117,1976.
[2] 吉岡真治: 多言語ニュースの対照分析のための Wikipedia 活用手法の研究，人工知能学会 第23 回全国大会，2009.
[3] 中崎寛之，川場真理子，宇津呂武仁，福原知宏: Wikipedia エントリを知識源とする日英ブログからの文化間差異発見支援，第 23 回人工知能学会全国大会，2009.
[4] 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，人口知能学会，Vol.17，No.3，pp217-223，2002.
[5] 鈴木義一郎：情報量基準による統計解析入門，講談社，1995.
[6] 難波英嗣，谷口裕子: 学術論文データベースからの研究動向情報の抽出と可視化，言語処理学会 第 12 回年次大会，2006.

表２．特色キーワードの抽出結果

分野全体(tf-idf)	tf-idf	出現確率比	χ^2 値	AIC
システム	解答	注意事項	システム	作業
データ	システム	ネットワーク環境	開発	報告
学習	学習	VR 技術	汎用	作成
作業	漢字	学習目的	学習	可能
コマンド	操作	道具	効果	構築
作成	試験	行為支援	利用	構成
支援	字形	問題演習	注意事項	ベース
使用	出題	文字装飾	日本	比較
開発	解答方式	漢字検索表示システム	タブレット	基礎
アイコン	試験システム	クイズ	難解	本稿