研究者ディレクトリデータベースからのキーワード抽出による分野間の関連分析

○西澤正己、孫媛、矢野正晴

A Study on relationship between the Information Science and other fields based on keyword analysis

○Masaki NISHIZAWA, Yuan SUN, Masaharu YANO

Abstract

In this study, relationships between the Information Science and other fields were investigated in terms of keywords extracted from research themes of each researcher. First, procedure for extracting keywords from the database of Directory of Researchers was presented. As an index measuring the relationship, we introduced a coefficient of normalized complete congruence and further applied the Correspondence Analysis method to analyze the relations between the three items of the so-called Information Science and other related research fields at Japanese universities.

1. 概要

近年、応用分野が急速に広がりつつある情報科学分野の研究と他分野間の関連を、キーワードを抽出する方法を用いて調べた。まず、研究者ディレクトリデータベースから、各研究者の研究課題からキーワードを抽出し(研究課題キーワード)、次に、比較的狭い意味で研究分野の特徴を把握するために、各研究者が記入したキーワードそのものからの抽出を行なった(分野特徴キーワード)。これら2種類のキーワードの一致度を、平均値により規格化した後、完全一致と部分一致の両方調べたが、ここでは、情報科学の3細目（計算機科学、知能情報学、情報システム学）それぞれに対して、少なくともいずれかの分野で規格化完全一致数が 30 以上であった細目分野（33 細目）について、対応分析を行なった。その結果、情報科学の3細目のそれぞれと関連が深い他の細目と、その距離関係が示された。

2. 研究者ディレクトリデータベース

国立情報学研究所およびその前身である学術情報センターでは、平成 4 年度より研究者の研究活動に関するデータベースである「研究者ディレクトリ」[1)]を作成している。調査対象は国・公・私立大学等の高等教育機関と文部省および文化庁並びにそれらの施設等機関、文部省所轄民間学術研究機関に所属する常勤の教職員・研究者である。今回の分析には平成10年度(1998 年度)における4年制大学の調査結果（総数 179,605 人のうち、回答があったのは 128,650 人）[2)]を用いることとする。

本分析には研究者ディレクトリデータベースより、研究者が記入した、現在の研究課題とその内容を最もよく表すキーワード(研究課題に付き3つまで)、またその研究課題が密接に関連している専門分野コード(研究課題に付き3つまで)をデータとして用いる。なお、専門分野コードは複数ある場合重要な順に記入してもらっているので、本研究では、第1番目の分野コードのみを用いて、分析を行った。 ここで言う専門分野コードは、科学研究費補助金の「系・部・分科・細目表」の細目を基に付与されたものである。

3. 分析およびキーワード抽出手法

研究分野間の研究の関連を定量的に調べるために、まずその分野の研究者が実際に行なっている研究課題からキーワードを抽出することにより(研究課題キーワード)、応用範囲をも含めたより広い関連分野の情報も含んだキーワードが抽出できるものと考える。さらにある特定の分野を特徴付けるキーワード(分野特徴キーワード)を以下の手法で選び出すことにより、分野に密接に関連したキーワードを得る。この2種類のキーワード群を比較することにより、分野間の関連を分析する。

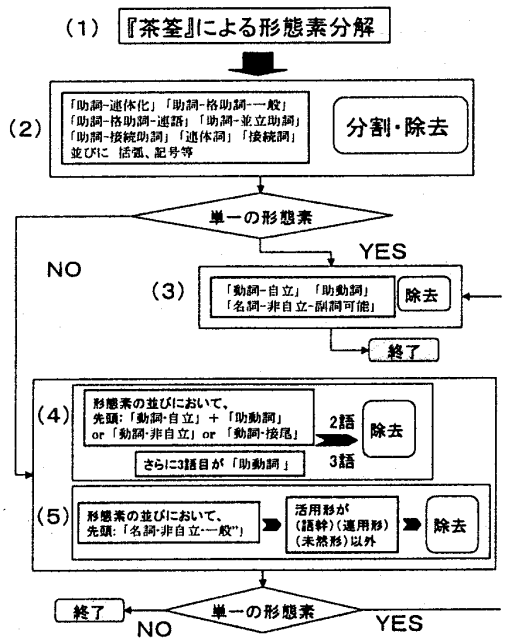
最終的にはすべての細目分野について分野間の関連を調べる予定であるが、これまでの分析[3]との比較によりさらに手法が洗練できる事と、近年応用分野が急速に広がりつつある分野であることを考慮し、始めに情報科学分科の三細目(計算機科学、知能情報学、情報システム学)と他の細目分野の関連を調べることとした。

3.1 研究課題キーワード

現在の研究課題からのキーワード抽出に当たっては、始めに日本語形態素解析システム『茶筌』(version 2.0b6)を用い形態素分解をおこなった。その後の処理手順を図1に示す。図中の(2)の過程では助詞等の品詞および品詞細分類を調べて文の分割を行い、(4)および(5)では(2)の過程で分割したことによる問題点を処理している。この処理により、「用いた遺伝子解析」("動詞-自立"+"助動詞"+ "名詞-一般"+"名詞-サ変接続")等は「遺伝子解析」("名詞-一般"+"名詞-サ変接続")というキーワードになる。

こうして得られたキーワードを研究課題に付けられた研究分野コード(第1位)によって纏め、それぞれの研究分野(細目)に対する「研究課題キーワード」とした。

図1 研究課題キーワードの抽出手法

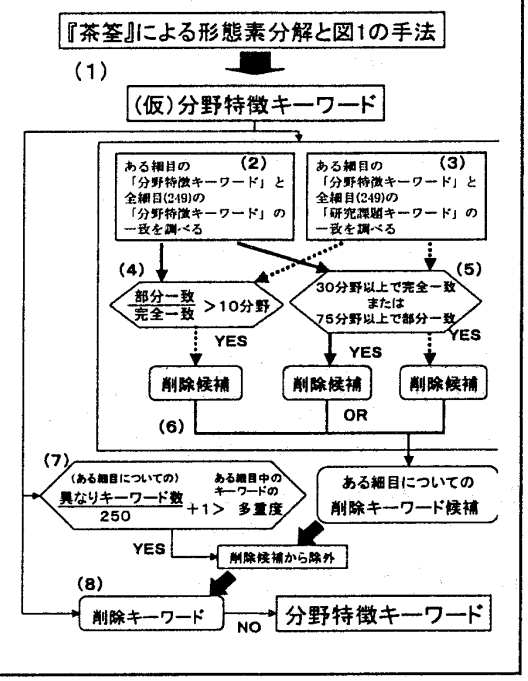


3.2 分野特徴キーワード

「分野特徴キーワード」としては比較的狭い意味で研究分野の特徴を示すキーワードが必要である。本来はこのような目的で収集されたキーワードが必要であるが、ここでは研究者が記入した研究課題の内容を表すキーワードが、研究課題に付けられた第1位の研究分野コードに即していると推測し、このキーワードを分野特徴キーワードとした。しかし、キーワードの記述が短文となっているものもあり、「研究課題キーワード」と一貫性を保つ目的からも、前節で説明した形態素分割による手法を同様に適用している。

このようにして抽出したキーワードには意味の広いもの、どの分野でも使われるキーワードも含まれている。

図2 分野特徴キーワードの抽出手法



よって、このようなキーワードを除くために、すべての 249 細目分野に対して「分野特徴キーワード」を抽出した後に図 2 の手順で意味の広いキーワードの除去をおこなった。図中の(4)では部分一致の多さに注目し“特性”や“最適化”などのように分野に特徴的でない複合語の一部となるキーワードの除外を目的とした。また、手順(5)では分野を跨いだ完全一致、部分一致の数を見ることにより分野に特徴的でない一般的なキーワードを除外対象としている。さらに、これら分野の中で重要な頻出するキーワードが削除されないように(7)の手法により配慮をおこなった。(4)、(5)、(7)で用いたパラメータについては、今回は結果を見比べながら決定したが最適化の検討は不十分であり、今後の課題である。

4. 解析結果

3章で述べた手法によって情報科学分野3細目、「計算機科学」、「知能情報学」、「情報システム学」について「分野特徴キーワード」を抽出し、全細目分野の「研究課題キーワード」との一致を調べた。一致はワードの完全一致または部分一致(「分野特徴キーワード」が「研究課題キーワード」に包含される:完全一致を含む)を調べている。

ここで細目ごとに「研究課題キーワード」の数(延べ数) Tk が違うため、基準として「研究課題キーワード」数の平均値 $<Tk>$ により規格化を行なった。「研究課題キーワード」数の平均値は 1,416 であり、標準偏差は 1,149 と分野間のばらつきはかなり大きい。次に「計算機科学」、「知能情報学」、「情報システム学」に対しての「分野特徴キーワード」の数 Fn の違いを補正することにより、「計算機科学」と「システム工学」のつながりの強さと「知能情報学」と「システム工学」のつながりの強さ等を比較できるようになる。この情報科学の 3 細目それぞれに対する「分野特徴キーワード」の異なり数は、1,222、965、722 であり、全細目分野に対する平均の「分野特徴キーワード」の異なり数 $<Fn>$ は 561.6 であった。ここでは全分野におけるそれぞれの「分野特徴キーワード」異なり数の平均を基準として規格化する。 よって、「分野特徴キーワード」と「研究課題キーワードの細目分野における数の違いを規格化するための式は次のようになる。

$$Cn_i = C_i \times \frac{<Tk>}{Tk_i} \times \frac{<Fn>}{Fn_i}, \quad Pn_i = P_i \times \frac{<Tk>}{Tk_i} \times \frac{<Fn>}{Fn_i}$$

ここで、Cn_i, C_i は i-細目の規格化後完全一致数および、規格化前完全一致数、Pn_i, P_i は i-細目の規格化後部分一致数および、規格化前部分一致数、Tk_i, Fn_i は i-細目の「研究課題キーワード」の延べ数と「分野特徴キーワード」の異なり数である。

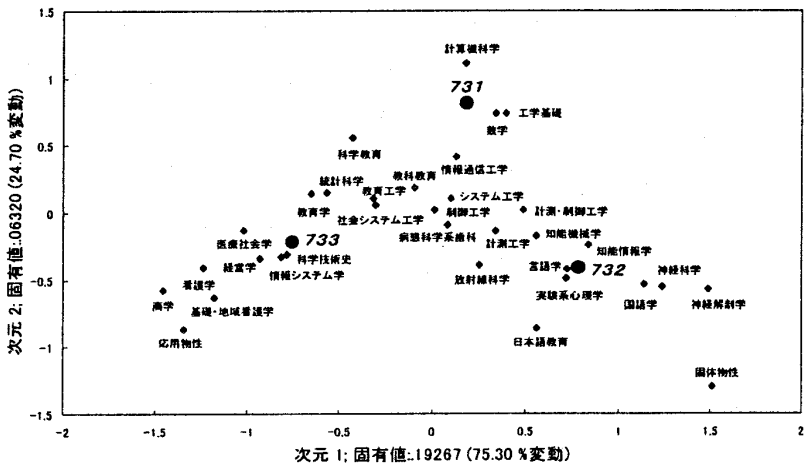
以上の手法によって、「計算機科学」、「知能情報学」、「情報システム学」に対しての規格化完全一致数および規格化部分一致数が得られた。完全一致と部分一致では分野間のつながりの相対的強度は若干異なっている。今回は、まず完全一致の結果のみを使ってこれら情報科学 3 分野と他分野との関連を対応分析によって調べた。

対応分析に用いたデータは、「計算機科学」(731)、「知能情報学」(732)、「情報システム学」(733)それぞれに対して、少なくとも何れかの分野で規格化完全一致数が30以上の値を得た細目分野(33 細目)のものである。これら33×3の 2 元データに対応分析 [4]を適用して2次元プロット図を描いたのが図 3 である。

まず、計算機科学(731)、知能情報学(732)と情報システム学(733)は、互いに距離を保ち、細目レベルの分野は、この 3 つを核として分散していることが図に示されている。情報通信、システム工学、

計測・制御工学、数学、工学基礎などが計算機科学と、知能機械学、言語学、実験系心理学、神経科学などが知能情報学と、そして、統計学、教育学、経営学、医療社会、看護学などが情報システム学と関連が強いことが示された。

図3 情報科学関連分野の2次元図示



5. 考察およびまとめ

この分析によって、情報科学の 3 細目と関連が深い他の細目とその距離関係が読み取れる。中央部は 3 分野に対して平均的に関連が深い分野、731、732、733の方向にそれぞれやや偏って関連のある分野が集まっている。「放射性科学」や「病態科学系歯科」が中央付近にあるが、これは“画像診断”や“画像解析”といったキーワードが寄与している。また、「固体物性」は“磁性”と言うキーワードが「知能情報学」に含まれたことで関連付けられており、「応用物性」は“評価”と言うキーワードで「情報システム学」に関連付けられことが、個々のキーワードの分析で分かっている。このように特定の1つの分野のみに関連がやや強く現れた場合は731、732, 733の方向の外側に現れてくる。

この分析ではある意味で関連分野の実態を顕著に表しており、新たな知見が得られることが分かるが、対応分析に選んだデータの抜き出し方によるバイアスも現れることに注意が必要である。さらに、分野特徴キーワードの抽出方法もさらなる考慮が必要で、アンケートによるキーワードの調査なども必要であろう。また、分野の大きさによる Tk の違いが規格化時に誤差を拡大する可能性があり、この影響も分析の上で考慮が必要であろう。今後は、これらの課題や他の問題点を明らかにし、他領域についても分析をおこなう予定である。

参考文献

1) 太田和良幸、柿沼澄男、西澤正己、 孫媛、山下泰弘, 我が国における学術研究活動の状況 - 「平成 7 年度学術研究活動に関する調査」結果概要 - , 情報管理, Vol. 40, No. 9, pp770-789, 1997
2) 「我が国における学術研究活動の状況」- 平成10年度学術研究活動に関する調査結果- , 学術情報センター, 平成 12 年 3 月
3) 西澤正己; 孫媛; 矢野正晴, 「情報科学研究の分野分類に関する調査研究」, 学術情報センター紀要, 第 12 号, pp.121-128, 1999
4) Gifi,A. (1990) Nonlinear Multivariate Analysis. John Wiley & Sons.

西澤 正己 国立情報学研究所(〒101-8430 東京都千代田区一ツ橋 2-1-2)
孫 媛 同上
矢野 正晴 同上
Masaki Nishizawa (nisizawa@nii.ac.jp) National Institute of Informatics
Yuan Sun (yuan@nii.ac.jp) National Institute of Informatics
Masaharu Yano (yano@nii.ac.jp) National Institute of Informatics.