

データマイニング手法を用いた研究動向分析手法の一考察

A Discussion of Text Mining Methods for Analyzing Trends on Articles in Academic Journals

華山 宣胤[†] 山本 樹^{††} 山崎 航^{†††}

Nobutane Hanayama,[†] Tatsuki Yamayamoto,^{††} Wataru Yamazaki^{†††}

^{†,††,†††}尚美学園大学 芸術情報学部 情報表現学科

^{†,††,†††}Department of Information Technology, Shobi University

概要：本研究では、学術雑誌に掲載された論文記事をテキスト・マイニングと多変量解析手法を用いて分析し、当該分野の研究動向を明らかにするための1手法を提案する。提案手法は、単にキーワード間の関連を見るだけではなく、学術論文に特有の文法ルールに注目した解析を行うことにより、各論文を内容に従って比較・分類する点に特徴がある。また、日本テレワーク学会誌に掲載された論文記事を実際に分析することにより、提案手法の有用性を検討する。

1. はじめに

学術論文誌に掲載されている研究動向を把握することは、新しい研究分野に取り組もうとする研究者および学会誌の企画・編集者にとって重要である。そこで本研究では、1999年に設立された「日本テレワーク学会」の発行する学会誌を例題として取り上げ、その研究動向分析の手順を提案するとともに、実際の分析結果から提案手順の有用性を議論する。

2. 分析手順

提案する分析手順は下記のとおりである。

- (1) 形態素解析によるキーワード抽出
- (2) 度数分布からの主要キーワード選出
- (3) ダミー変数による因子分析
- (4) 因子得点によるクラスター分析

提案手順の特徴は(3)と(4)にある。

2.1. 形態素解析によるキーワード抽出

形態素解析とは、コンピュータ等の計算機を用いた自然言語処理の基礎技術の一つである。対象言語の文法の知識や辞書を情報源として用い、自然言語で書かれた文を形態素の列に分割し、それぞれの品詞を判別する作業を指す。実際の分析には、フリーソフトの『茶筌 (奈良先)』を使用した。

2.2. 論文の構文特徴によるキーワードの選出

茶筌を使用した形態素解析では、文章が単語や分節に分解され、品詞別に表示される。この中で下記の2つの条件を満たす名詞をキーワードとして抽出する。

- 「本論文/研究」を含む文中のサ変接続名詞
 - 転換の接続詞がある場合はその後位置する
- そして、適当な数(本研究では4とした)以上の論文に含まれているキーワードを分析に用いるキーワードに採択する。表1に各論文含まれるキーワードを示す。

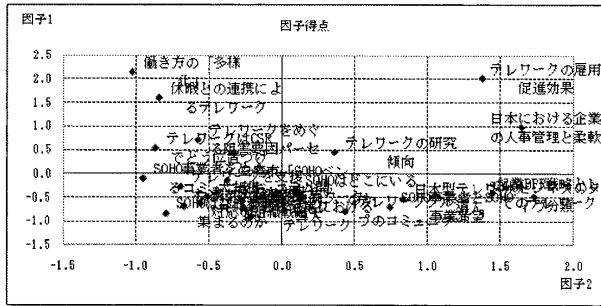
図表 1

選定された上位のキーワード			
研究	調査	考察	雇用
実施	導入	分析	検討
提案	就業	組織	継続
整理	提示		

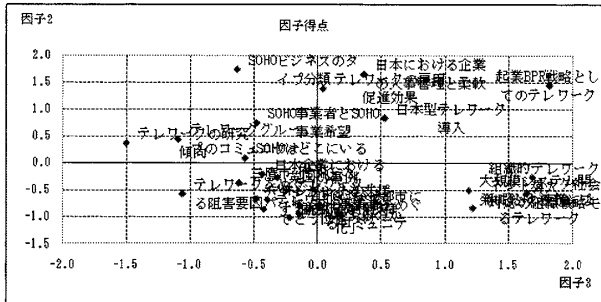
2.3. ダミー変数による因子分析

各キーワードを含むか否かを各論文の属性項目(含む=1, 含まない=0)と考える。そして、論文間の類似性を全体的に把握するために、ダミー変数を属性用いた因子分析を行う。

図表 2. 因子得点のプロット (第 1 × 第 2)



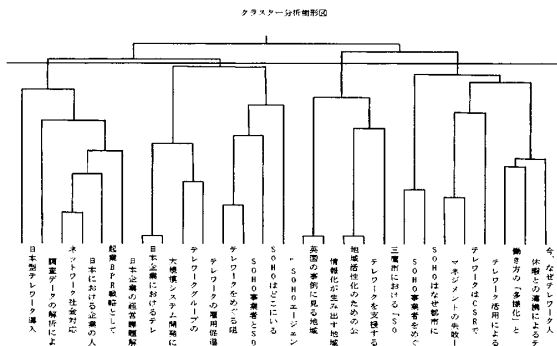
図表 3. 因子得点のプロット (第 2 × 第 3)



2. 4. 因子得点によるクラスター分析

因子得点に基づいてクラスター分析を行う。

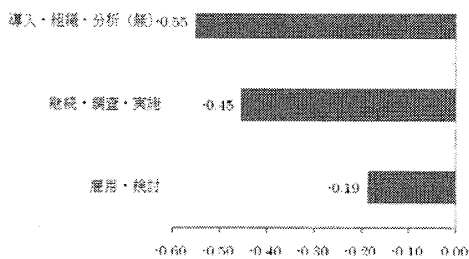
図表 4. 因子得点によるクラスター分析



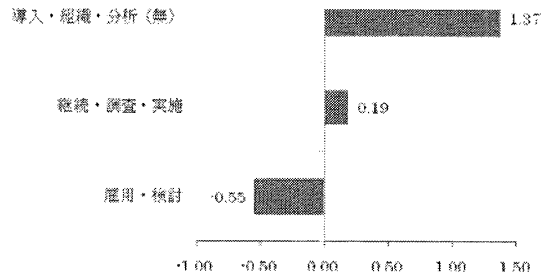
3. 分析結果の検討

前節で分類した論文のクラスター別の特徴を因子得点の平均値から見てみよう。

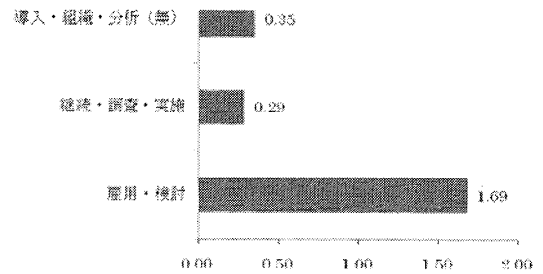
図表 5. 第 1 クラスターの因子得点平均値



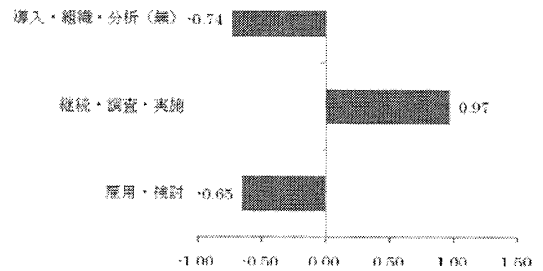
図表 6. 第 2 クラスターの因子得点平均値



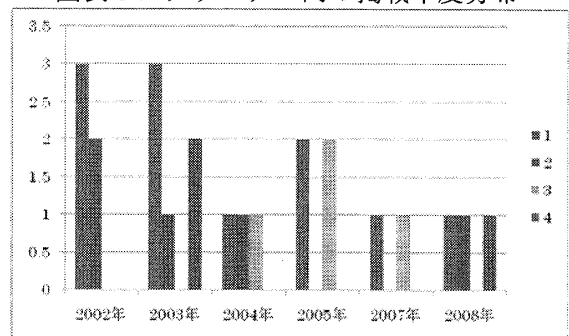
図表 7. 第 3 クラスターの因子得点平均値



図表 8. 第 4 クラスターの因子得点平均値



図表 9. クラスター内の掲載年度分布



参考文献

- [1] 浅川 雅美, 岡野 雅雄, テレビCM に対する視聴者反応の分析—自由回答文のテキストマイニング, 文教大学女子短期大学部研究紀要 48 集, 1-6, 2005
- [2] 林 俊克, Excel で学ぶテキストマイニング入門, オーム, 2002