4N－1

論文タイトルの自然言語処理による情報科学研究の歴史的分析

若月　玲　　片谷　教孝

山梨大学　工学部

1．　はじめに

科学の任意の部門における歴史的な流れを分析することは、当該分野の研究者にとっては研究の方向づけの参考となり、歴史学者に対しては史学的な関心の対象となる。しかし、この歴史的分析を行う為には、通常 100 年以上もの期間の文献を地道に調べることが必要とされ、結果非常に手間のかかる作業を行わなければならない。更にこの方法では、分析結果に分析者の主観が入ることは避けられず、また同時に、分析に専門知識を要するという問題も生じることになる。

そこで本研究では、前述した手法とは全く異なる手法を使用している。概要を説明すると、自然言語処理の技法を適用することにより、分析対象となる分野において、比較的短期間且つ簡潔な歴史的分析を行っている。

本稿では、この手法の詳細と実際に行った情報科学研究の歴史的分析結果を報告している。

2．　手法の詳細

2．1　論文タイトル

論文タイトルはその論文の最も短い要約であり、分析を行う上で非常に重要であると言える。本研究ではこの考えに基き、定期刊行の学術雑誌に焦点を当て、論文のタイトルとその中の単語群に注目している。具体的には、一定期間内の学術雑誌における論文タイトルを集めたデータに対し、情報処理の分野で著しい進歩を遂げている自然言語処理の技法を適用することにより、単語毎に出現している論文数をカウントする。この値が高い程、その期間における重要度が高いことを意味する。更に、時系列での比較を行う為、期間内の総論文数に対して各単語が占める割合を百分率で求める。

このように自然言語処理の技法を用いて機械的に解析を行うことにより、誰もが容易に客観的な分析を行うことが可能になると考えられる。しかしながら、分析対象が継続刊行の学術雑誌の為、100 年単位の長期的な分析は困難という欠点も挙げられる。

2．2　自然言語処理

本研究では自然言語処理の手法自体の研究を目的とするのではなく、あくまで分析を行う為の２次的なツールとして、既存の自然言語処理手法を用いている。自然言語処理には辞書、シソーラス、構文解析、意味解析等の様々な手法が存在するが、ここでは辞書ファイルマッチングによる切り出しと、文を幾つかの構成要素（語）に分解する形態素解析のみを用いる。また、形態素解析においては、文法的な規則は無視し、２文字以上の漢字や仮名で構成される名詞を切り出すというヒューリスティクスのみを採用している。

2．3　語切り出しの手順

マッチングに用いる辞書ファイルは、分析する学術雑誌の分野に関する分野に関する用語集から語を抽出して作成している。また、２文字以上の文字列のみを選定するのは、語数の増大

Historical Analysis of Information Science Studies using Natural Language Processing to Article Titles

Akira WAKATSUKI　Noritaka KATATANI

Faculty of Engineering , Ymanashi University

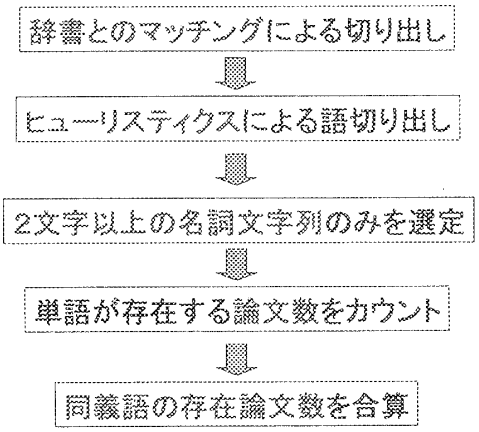

図１　語の抽出とカウントの流れ

防止と、１文字の語は大半が助詞等の意味を持たない語であるという理由からである。１つのタイトルの中に同じ語が重複して存在した時の二重カウントを防ぐ為に、語の総数ではなく論文数をカウントする。同義語の合算では、表記の揺れ等の明らかに同義と分かるものを、手作業で合算する。

３．　分析

３．１　分析対象

情報科学研究の歴史的分析として用いる学術雑誌として、「情報処理」学会誌を選択した。分析期間は 1975～1989 年とし、時系列の比較を行う為、３年毎の５区間に分割し分析を行う。辞書ファイルとしては、コンピュータ用語辞典に掲載されている単語2619語をファイル化し、使用している。

３．２　分析結果と考察

主要な語の時系列変化を図２に示す。全区間を通してみてまず目につくのは、「システム」、「データベース」の２つの語が常に上位に位置していることである。この２つは、情報処理研究という分野の中において、常に中心に存在するテーマだということが分かる。またほぼ全区間を通して、「データベース管理システム」等のように、両者が情報処理研究の中で互いに密接に関係しているのを読み取ることが出来る語が幾つも出現している。

他に、コンピュータ言語の「C」や「ＦＯＲＴＲＡＮ」に注目してみると、３期を境にＦＯＲＴＲＡＮやＬＩＳＰに代わり、移植性に富み且つ、開発が容易なＣが受け入れられていった

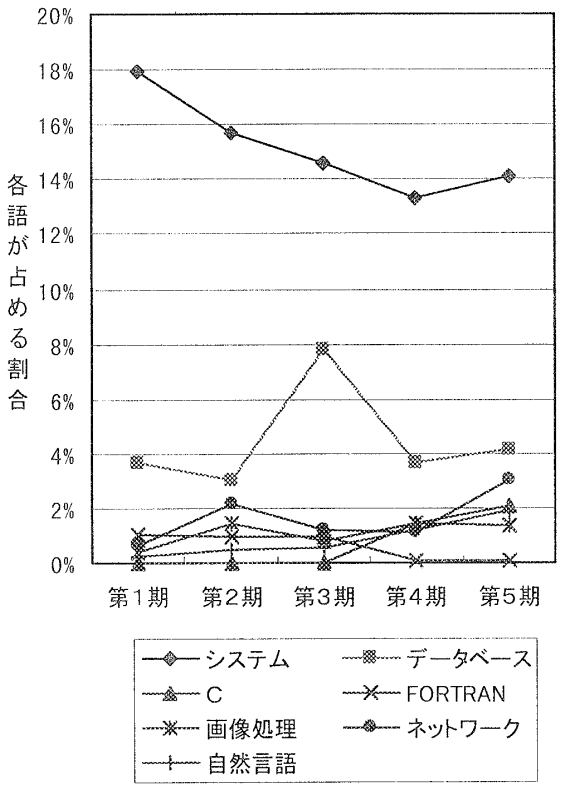

図２　　主要な語の時系列変化

様子が分かる。また、計算機の性能の向上に伴う「画像処理」（他に「コンピュータグラフィックス」）が次第に注目され始めている。情報化やインターネットの発達に伴う「ネットワーク」の増加（「ワークステーション」等も同様）も注目すべき点である。

４．　おわりに

今回、情報処理研究の歴史的な分析を行う上で、自然言語処理の技法を用いた新たな手法を提案し、分析を行った。その結果、おおまかではあるが歴史的流れや傾向を見出すことができた。

参考文献

高橋　三雄：「コンピュータ用語辞典」
ナツメ社　1991