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

5U—02

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。前報[1]では、幾つかの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た結果を報告した。

本稿では、それぞれの技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の比較と考察を報告する。更に、今後の研究に導入する予定の新たな手法についても述べる。

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。具体的な分析方法の内容やその利点等は前報[1]を参照されたい。

今回の分析方法も、それ自体は前回と同じである。

Historical Analysis of Information Science Studies using Natural Language Processing to Article Titles
Akira WAKATSUKI　Noritaka KATATANI
Faculty of Engineering , Ymanashi University

2．2　本研究で用いる技法

本研究で用いている自然言語処理の技法は、以下の２つである。

1．辞書マッチングと形態素解析による簡易な語切り出し

2．自然言語処理ツールＪUMANの利用

技法１：

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の技法を用いている。前回の発表で用いていた技法である。

技法２：

京都大学長尾研究室が中心となって開発された形態素解析ツールＪUMANを利用する。ＪUMANは使用者によって、文法（辞書）の定義、単語間の接続関係の定義が容易に行える。現時点では、対象が論文タイトルという特殊な例の為、デフォルトのものを利用している。

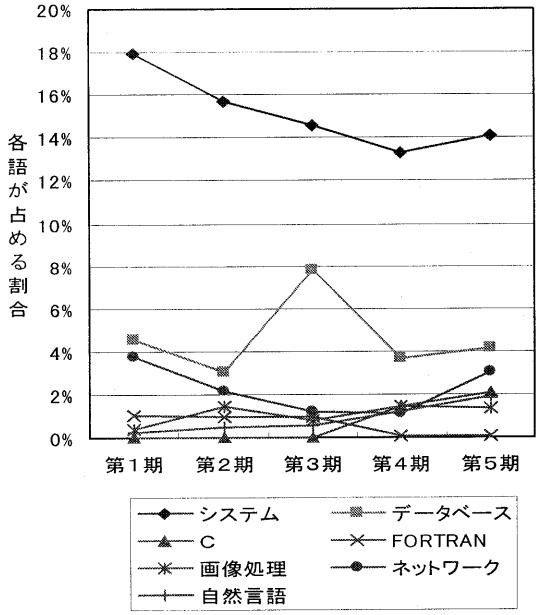
3．　分析とその結果の比較

3．1　分析対象

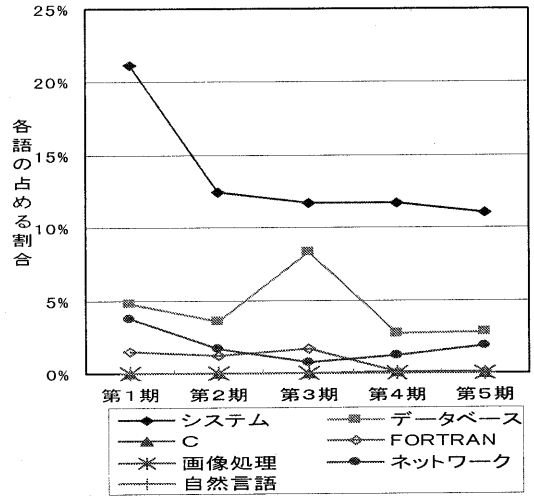
情報科学研究の歴史的分析として用いる学術雑誌として、「情報処理」学会誌を選択した。分析期間は 1975～1989 年とし、時系列の比較を行う為、３年毎の５区間に分割し分析を行う。辞書ファイルとしては、コンピュータ用語辞典に掲載されている単語 2619 語をファイル化し、使用している。

3．2　分析結果

２つの技法をそれぞれ用いた結果を比較すると、技法１では、「情報処理」学会誌に関わりの深い分野の用語集である「コンピュータ用語辞典」からマッチングに使う辞書ファイルを作成し、また辞書登録単語のカウント結果は別に出力する為、技法２より多くの専門用語を抽出していることがまず分かる。



技法１を用いた場合の主要な語の時系列変化



技法２を用いた場合の主要な語の時系列変化

技法２の解析結果の中でも、主要な語の内、上位に位置する「システム」、「データベース」（常に上位に位置していることから、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）や「ネットワーク」（情報化やインターネットの発達に伴った増加）はほぼ同じ変化をしている。一方で「画像処理」、「自然言語」は解析結果の中に全く表れておらず、共に「画像」と「処理」、「自然」と「言語」に分割されてカウントされていた。従来の方法では辞書マッチングで抽出されていた単語が、JUMAN ではより小さい形態素に分割されて抽出されていることが考察される。一方、「C」に注目してみると、「画像処理」等と同様に全くカウントされていない。原因としては、辞書マッチングが、辞書が表す意味の「C」とは異なる、任意の語の部分的な C を拾い出してしまっているという従来の技法が持つ問題点が考えられる。このように、辞書マッチングによって部分的に先に切り出された為に、従来の方法では異なる形で抽出された単語が、JUMAN では正しい形で抽出されているケースも中には見られた。

４．　今後新たに導入する手法

分析を行うにあたり、単語ごとの出現論文数のみで重み付けを行うだけでは、評価を行う指標としては不十分である。そこで現在、切り出した単語の重み付けを行う過程中において、情報検索で広く用いられているｔｆ・ｉｄｆ法の導入を考えている。

ｔｆ・ｉｄｆ法を導入することにより、新たに可能になり得る機能として、以下の２つが挙げられる。

１：新たな重みを導入することによる、従来の手法では見逃していた重要語の抽出

２：共起情報（１つのセグメント中に存在する語の組み合わせ）の利用によるセグメント・単語間の関連付け

今後の研究において、期待できるのは２である。１に関しては、問題点が存在する。セグメントであるタイトルは１つの文でしかなく、１つのセグメント内の情報量（形態素数）が圧倒的に少ない為、どの語もｔｆ値が同じになってしまう。そこで、セグメント単位の変更、ｔｆ・ｉｄｆの計算方法に独自の考えに基づく改良を加えることを考えている。

参考文献

[1]若月　玲，片谷　教孝：「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」　情報処理学会第 59 回全国大会講演論文集Ⅱ，pp373-374

[2]高橋　三雄：「コンピュータ用語辞典」　ナツメ社　1991

[3]大森　信行，岡村　潤，森　辰則，中川　裕志：「tf･idf 法を用いた関連マニュアル群のハイパーテキスト化」　http://www.forest.dnj.ynu.ac.jp/~ohmori/Paper/NL121/paper1.html