論文タイトルの自然言語処理による科学研究の歴史的分析

8M—05

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，分析に専門知識を要するという問題が生じることになる。

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]，[2]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。

今回、より詳細な分析を行う為、上記の技法に新たな語の重み付けを行う２つの改良を加えた。本稿では、前報とは異なる分野において、これらの技法を用いた分析結果とその考察を報告する。

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[2]を参照されたい。

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。

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。

2．3　上記の技法の拡張

複合語の集約化

上記の技法によって抽出された単語に対し、半角英数字を除く２～３文字までの単語全てを「構成語」（複合語を構成する語を、ここではこう呼ぶことにする）とする。これらの構成語と他の切り出された単語全てとのパタンマッチングにより、構成語ごとにその構成語を含む複合語を集約化する。

共起情報の抽出

1．タイトル毎に、その中の全ての共起単語の組合せを抽出し出現頻度をカウント

2．区間内の各単語の出現頻度をカウント

3．１と２で求めた頻度を基に共起の得点付け

具体的な得点付けの基準として、区間内の全単語の出現頻度において、少数のタイトルに出現してる語から成る共起の得点が高くなるようにした。その為、計算式は次のように設定した。

$$\log((M * con) / (left * right))$$

M：区間内の全単語の出現頻度の合計

con：区間内の各共起の出現頻度

left：共起中の左に位置する単語の出現頻度

right：共起中の右に位置する単語の出現頻度

3．　分析とその考察

3．1　分析対象

科学研究の歴史的分析として用いる学術雑誌とし

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

Akira WAKATSUKI　Noritaka KATATANI

Faculty of Engineering , Ymanashi University

て、本稿では「安全工学」学会誌を選択した。分析期間は1962～1988年とし、時系列の比較を行う為、7年毎の4区間に分割し分析を行う。今回は辞書作成時間の都合上、辞書ファイルは使用していない。

3．2　分析結果

図1から、全体を通した流れとしては、やはり「ガス」や「爆発」といった語が顕著に出現している。このことから、これらの語が常に事故に密接に関わってくるキーワードだということが分かる。

時系列で見てみると、第2期には「大気汚染」、「有害物質」という2つの語が着目すべき点として挙げられる。これらの語から、この区間では単に安全問題を取り扱うだけではなく、環境問題についても論じられた機会が多かったことが伺える。また、第4期からは、安全対策や事故に関わる語以外に、「シュミレーター」や「シュミレーション」のような語が出現している。これらからは、計算機等の発達に伴い、机上で論じるだけではなく、仮想実験も行われるようになったということが読み取れる。

また全体を通してみて、他に「関連」・「関係」といった語が出現しており、安全工学では事象（事故や原因等）の関連付けが基本であることを窺い知ることが出来る。

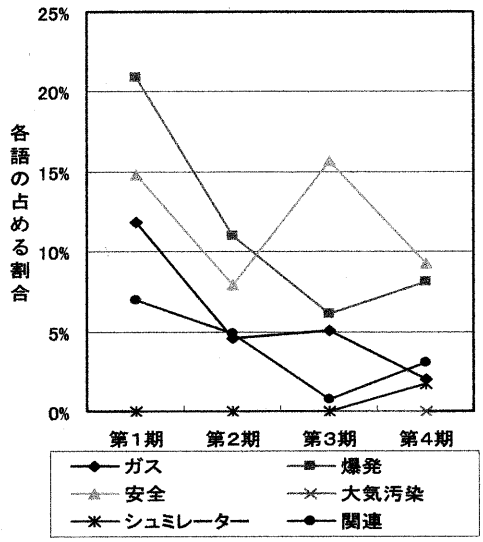

図1　技法2を用いた主要な語の時系列変化

図2でも、「安全」や「爆発」といった単語はどの期間でも大抵上位に位置していることから、図1と同じ考察が出来る。更に、構成語の詳細から、「安全」

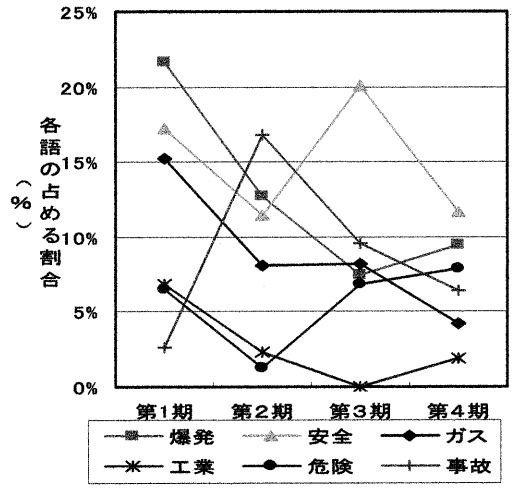

図2　主要な複合語の時系列変化

では（安全な状態の「管理」，危険に備えた「対策」，安全性を保つ「装置」，etc）といった「安全」における研究の対象が、「爆発」では（爆発の限界，爆発の危険性，爆発がもたらす災害，etc）といった「爆発」において注目すべき部分が読み取れる。

共起情報の抽出結果においては、注目すべき語の共起が多数存在していた。当然のことだが、助詞（特に一文字から成る平仮名）との共起をカウントする必要はないということも分かった。タイトル中に数多く存在する「の」等の助詞は、その語を挟んで出現している2つの単語の共起により大きな重みをつけるといった利用方法を行うべきだと考えられる。また、タイトル中の最後に位置する「～の分析・研究」といった語の共起はカウントしないことも解決策として挙げられる。

4．　まとめ

本稿では、科学研究の歴史的分析を行う上で、前報までの自然言語処理の技法に加え、語の重み付けにおいて2つの改良を加えた。その結果、より詳細な分析を行うことが可能になった。

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]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60回全国大会講演論文集Ⅲ，