講演番号：B05

テキストマイニングによるプラント異常事例の構造化表現

○加須屋秀彰（東北大学）　高橋信（東北大学）北村正晴（東北大学）

Application of Text Mining Method to Structurized Representation of Plant Anomaly Events

*H.Kasuya (Tohoku University) and M..Takahashi (Tohoku University) and M..Kitamura(Tohoku University)

Abstract- The method of text-mining has been applied to the trouble case reports of the nuclear power plant. The reports written in natural language have been represented in the form of structurized representation and the essential cause-consequence relationships have been extracted. It has been confirmed through the application to the example case reports that the present method can be used for the extracting the “concepts”, which explicitly represent the nature of the troubles.

Keyword　Textminig, Knowledge management

1 背景

近年の IT 技術の急速な進展により、様々な情報が大量に蓄積されるようになりつつあるが、蓄積された情報を有効に利用する手段に関してはまだ研究開発途上である。データマイニング手法はこのような要請に応えるための一つの手段であり、その中でも、テキストマイニング手法は、文書情報中から有効な情報を抽出することを目的としており、様々な分野で実用的に利用され始めている。

本研究では、原子力発電所の異常事例の文書情報を対象とする。発電所におけるトラブル事例は、法令により報告が義務づけられており、一般に閲覧可能な情報として入手可能であるが、この報告書は組織ごとの相違、詳細度の相違などがあり、統一的な記述とは言い難く、これをもとにして知識を統一的に活用することは現状では非常な労力を必要とする仕事となる。将来のトラブ防止のために有効な情報を含むと考えられる過去の事例を有効に活用するために、電子媒体化されたプラントトラブル事例に対して、テキストマイニング技術を応用することで、各事例の統一的な表現が可能になると考えられる。このようにして得られた表現を基に、異常事例を効率的に把握し、その共通性、特異性を見出すことで、大規模システムの長期的な安全運用を可能にすることができると考えられる。

2．目的

本研究では、自然言語で記述されたプラントとラブル事例に対しテキストマイニング技術を適用し、その適用可能性について検討を行うことを目的とする。具体的には、コンセプト抽出を用いたプラントトラブル事例に対する構造的な表現を行う。また、形態素を基準とし、コンセプトの適切な重み付けを行うための手法の選択に関しても検討を行う。

3．手法

本研究では、故障事例から知識を抽出するための手法として以下の方法を適用した。

3.1 前処理

まず始めに、文書をコンセプト群に分別する必要があるため、文章を品詞別に分類する形態素解析を行った。形態素とは品詞別に見た文章の最小構成要素である。形態素解析を行うにあたっては、代表的なソフトウェアである「茶筅」を用いた。さらに形態素を組み合わせて、情報量の多いコンセプト（以下コンセプトと呼ぶ）を抽出した。これを行うにあたっては、市販ソフトウェアである Textminig for Clementine を用いた。

3.2．牽引語の判別

本研究では、文書中を表現するのに有意なコンセプト（牽引語と呼ぶ）を選択するために、以下の二つの観点からコンセプトの重み付けを行った。

A)コンセプトの頻度による重み付け

文書をj、コンセプトをiとする。コンセプトの文書内頻度l_{ij},文書集合における頻度g_i,文書を正規化するためのn_jを用いて以下の式のようなコンセプト頻度d_{ij}による重み付けを行った。

$$d_{ij} = \frac{l_{ij} g_i}{n_j} \tag{3.1}$$

l_{ij}, g_i, n_jはそれぞれ以下の式より導出される。

$$l_{ij} = \begin{cases} 0.5 + \dfrac{f_{ij}}{\max f_{ij}} & (f_{ij} > 0) \\ 0 & (f_{ij} = 0) \end{cases} \tag{3.2}$$

f_{ij}：文書D_jにおける単語w_iの出現頻度

$$g_i = \log \frac{N}{n_i} \qquad (3.3)$$

N：総文章数、n_i：単語w_iを含む文書数

$$n_j = \sqrt{\sum_{i=0}^{m} (l_{ij} g_i)^2} \qquad (3.4)$$

B)確率的見地による重み付け

文書中のコンセプトは文章の意味を決定づける内容語とそれ以外の一般語に分類され、一般語はその文章中での出現がポアソン分布に従い、牽引語は従わないと仮定する。その違いによる重みは RIDF（Residual inverse document frequency）と呼ばれ、確率的見地からの重み付けにおいて一般的な手法である。本研究では、以下の式より RIDF を算出した。

$$\begin{aligned} RIDF_i &= \log \frac{N}{n_i} - \log(1 - p(0;\lambda_i)) \\ &= g_i - \log\left(1 - \exp\left(-\frac{F_i}{N}\right)\right) \end{aligned} \qquad (3.5)$$

C)総合指標

本研究では、コンセプト頻度とポアソン分布を利用した重み付けを組み合わせた$value_{ij}$を用いて、牽引語を選択する。この手法はtf・RIDF法と呼ばれ牽引語を抽出する方法の１つである。$value_{ij}$は以下の式より導出する。

$$value_{ij} = \frac{d_{ij}}{\max d_{ij}} \cdot \frac{RIDF_i}{\max RIDF_i} \qquad (3.6)$$

従って、$value_{ij}$>0.1 となるときのコンセプトを牽引語として選択する。

3.3 コンセプトの統計的な性質

tf・RIDF 法により牽引語を抽出したが、コンセプトは形態素を組み合わせるものであるので、統計的な性質がどう変化するのかを検証する。それによって、文書構造化にとって有効な重み付けを決定する。

A)RIDF 値による形態素とコンセプトの比較

RIDF 値を用いて形態素とコンセプトの比較を行った。ここで、コンセプト及び形態素の出現が特定の文書に偏るほどポアソン分布とのずれが大きくなり、RIDF 値は大きくなる。コンセプトにおいて 1 事例につき約 50%が事例独自のものとなっており、形態素の結果に比べて RIDF 値が約 30%増加している。これより、コンセプトの性質として、事例独自のものが多いことが確認できた。しかしながら、表現されるコンセプトが各事例ごとで独自性が強いために、他の事例と共通するコンセプトの数は少なくなる傾向が見られた。

B)頻度による比較

コンセプトと形態素を比較した場合、コンセプトはより詳細な概念を表現していることになる。このとき、コンセプトの頻度は形態素の頻度より有効性が高くなることが示唆される。

C)検証

以上より、コンセプト抽出では、より詳細なコンセプトの選択が可能であることが示されたが、それによって、コンセプトの持つ意味が狭められ、形態素と比較して他事例と共有する数が減少する。これは事例集合を用いる重み付けが不適切となる可能性を示唆している。また、コンセプト抽出の性質を考慮するとき、その頻度はより重要な指標となりうることが示された。このような結果は、事例の数が 10 事例と少ないことに起因すると考えられるが、本研究の枠組みでは、この範囲内で効率的な事例の構造化を行うために、一般的な手法である tf・RIDF 法による牽引語の選択を行った後、コンセプト頻度に基づく文書行列の作成、事例文書の構造化を行うこととした。以下、その内容について述べる。

3.4 共起度

共起度とは文書中におけるコンセプト間の結びつきの強さを表すもので、情報検索分野において一般的な概念である。本研究では、3.3 の結果を受けて、共起度の計算式にはコンセプト間距離とコンセプトの頻度重みに注目した式を使用する。計算に用いる頻度重みはコンセプト頻度d_wとその最大値の商で表す。そしてコンセプト間距離を$l_{w1,w2}$とし、文書D_iにおける全体の長さをL_iとすると、共起度は以下の式で計算される。

$$co(w1,w2) = \begin{cases} \left(\frac{d_{w1}}{\max d_{w1}} + \frac{d_{w2}}{\max d_{w2}}\right) * \frac{L_i - l_{w1,w2}}{L_i} & (d_{w1}, d_{w2} > 0) \\ 0 & (d_{w1}, d_{w2} = 0) \end{cases} \qquad (3.7)$$

3.5 文書行列の作成

以上より、各事例における共起度を利用した文書行列を以下に示す。

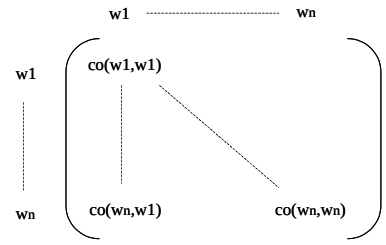

Fig.1 The document procession by the degree of coincidence

ただし、コンセプト$W=(w_1,\cdots,w_n)$は１事例のコンセプト群である。

3.6 事例構造化

本研究の目的として、プラントトラブル事例の構造化が挙げられている。今まで述べた手法において事例文書内のコンセプトの共起関係は得られたが、その関係は構造化された形式では表現されていない。重要なコンセプトとその共起関係だけでは、その事例が持つ本質的な知識構造を十分に表現しているとはいえない。そのため、それらのコンセプトがお互いにどのような関係を持っているのかを表現する必要がある。

いま、文書行列Dがあり、その要素をco_{nm}として、コンセプト間の関係を以下のように段階を経て構造化を行う。

1. 文書行列の各行に対して、自分自身co_iとの

$$S_i = co_i \cdot co_i$$

内積S_iをとる

2. 最も内積S_iが大きかったもの、つまりコンセプト群に対して最も影響力のあるコンセプトを選択する（これをkeywordとする）
3. keyword との共起が最も大きいコンセプトを 3 つ選択する(これを secword とする)
4. 2.と同様のことを secword についても行う。
5. 選択したコンセプトを基に共起の方向性を考慮し、keyword、secword を結びつける。

以上の処理を通じて、事例中での重要なコンセプトとその間の因果関係を構造化することが可能となる。

Fig2 に示すように、Level1 が文書中で最も重要度の高いコンセプト。Level2 がそれを支えるコンセプト群であり、Level3 はさらに詳細な説明を与えるコンセプト群と位置付けることができる。

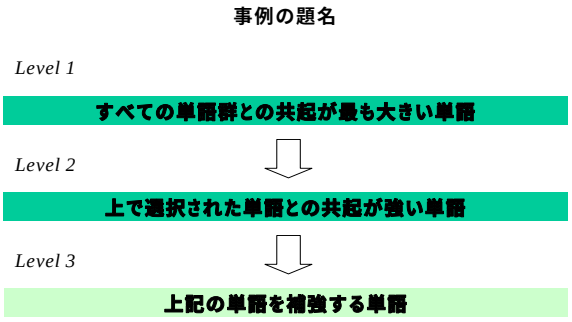

Fig.2 the result of the structured representation of report

4.適用結果

手法で述べた事例構造化手法の有効性を検証するため、旧通商産業省が公開していた国内の原子力故障事例[9]の中からTable1に示す10件の故障事例を対象にして手法を適用した。

No	プラント名	件名	発生日時
1	福島第二原子力発電所3号機	ジェットポンプ内装置振動による手動原子炉停止について	平成5年5月29日
2	福島第一原子力発電所5号機	シュラウド中間部リングひびについて	平成5年5月29日
3	伊方発電所1号機	蒸気発生器伝熱管の損傷について	平成5年5月29日
4	柏崎刈羽原子力発電所5号機	タービン制御油漏えいによる手動原子炉停止について	平成5年7月13日
5	美浜発電所3号機	格納容器サンプ水位上昇による手動原子炉停止について	平成5年10月13日
6	高浜発電所3号機	第A高圧注入ポンプ軸封部からの漏えいについて	平成5年11月14日
7	伊方発電所1号機	湿分分離加熱器逃し弁の損傷について	平成5年11月5日
8	高浜発電所3号機	B-主給水制御弁点検による手動原子炉停止について	平成5年11月26日
9	福島第一原子力発電所3号機	ジェットポンプ入口配管のひびについて	平成5年12月24日
10	敦賀発電所1号機	化学体積制御系配管からの漏えいによる手動原子炉停止について	平成6年10月13日

Table1. the 10 reports of nuclear plant accident

次に、故障生起知識構造化の有効性を検証するため、本研究の手法を Table1 の事例群に適用した。例として事例 No.1 への適用結果を Fig3 に示す。No.1 はジェットポンプビーム部の折損事故である。事故の原因はジェットポンプビーム取り付け時にわずかな位置ずれを生じ、応力腐食割れを起こしたためである。事例解析結果では、事例のキーワードである「ビーム」は、「漏えい」、「応力腐食割れ」及び取り付け時のずれを表す「応力」によって支えられていることが示されている。また、「ビーム端部」、「インレットミキサ」など事象が発生した箇所を捉えており、現象面との関係も構造化されている。以上から、この事例に関しては、適切な構造化が行われていると考えることができる。

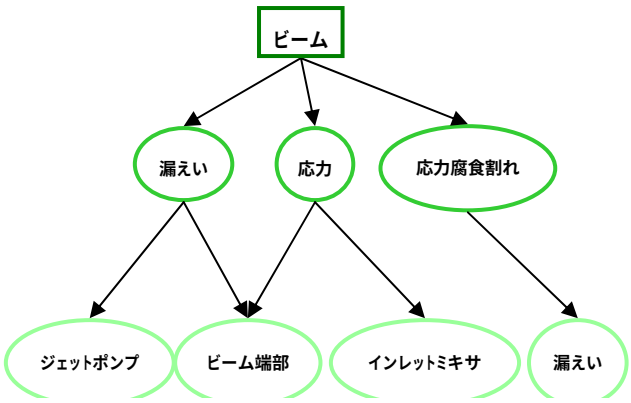

Fig.3 the result of the structured representation of report No1

5．考察

A)形態素とコンセプトの比較

一般に、原子炉機器、現象面の一部などは複数の形態素を組み合わせるものが多いが、コンセプトを用いることによって、1 つのコンセプトとして扱うことが可能となり、より詳細な事例文書の構造化が進んだと考えられる。しかし、コンセプト抽出によって事例同士の共有するコンセプトが減少したこと

で、RIDF 値が実際の値より大きくなってしまい、事例集合における検証が難しくなったことも事実である。

B)頻度レベルによる統計手法的なコンセプトの重み付け

本研究において、牽引語選択はコンセプト頻度と RIDF 値の積を用いて行った。牽引語選択の際に重視されるのは牽引語漏れであり、RIDF 値が大きい場合、重要度の低いコンセプトも含めてしまうが牽引語の抽出漏れの可能性が低いと考えたからである。また、事例固有のコンセプトを抽出するという観点から、広く事例に共有されるコンセプトを削除するという点も考慮したためである。しかし、RIDF 値はポアソン分布とのずれという仮想的な値をとり、サンプル数に著しく影響を受ける。本研究においては事例数が 10 事例という少ない数であり、さらに内容的な偏りも存在している。そのため、RIDF 値に基づくコンセプトの重み付けは文書行列の作成にとって不適切であると判断した。

C)事例の構造化手法

本研究では、事例文書に対してコンセプト抽出によるコンセプト頻度レベルでの共起度を用いた文書行列の作成することを提案した。これにより、事例の構造化結果は従来研究と比較してより本質的な知識を反映しており、適切なコンセプトの抽出を行うことができた。また、事例の構造化に使用するコンセプトは 10 個以内と制限したことで、人間にとってより理解しやすい表現が実現され、選択されたコンセプト間の関連性の把握を容易にした。

6．結論

本研究では、テキストマイニング技術を応用し、自然言語で表現された事例に含まれる知識の構造化を、コンセプトという単位で行う手法を提案し、その有効性を示した。

1. 原子力発電所におけるプラントトラブル事例の構造化
   本研究において提案した構造化手法は、構造化結果の正確性、可視化の観点から有効な手法である事を確認した。

2. コンセプト抽出の有効性の確認
   コンセプト抽出は、情報量の大きい単語の選択を可能にし、事例の構造化にあたって有効な手法となることを確認した。

以上から、本研究で提案した手法はプラントトラブル事例の視覚的な構造化を可能にするものであり、大規模複雑システムにおけるトラブル事例の効率的な利用に対して有効な手法となりうることを示した。

7．参考文献

[1]内松洋輔、“故障生起知識の構造化のための文書情報処理に関する研究”、(2002)
[2]http://aoki2.si.gunma-u.ac.jp/
[3]長尾真、“自然言語処理”、岩波書店(1996)
[4]新田義彦、“正規表現とテキスト・マイニング”、明石書店(2003)
[5]金明哲ら、“言語と心理の統計　ことばと行動の確率モデルによる分析”、岩波書店(2003)
[6]マイケル J.A.ベリー/ゴードン・リノフ、“マスタリング・データマイニング”、海文堂(2002)
[7]林俊克、“Excel で学ぶテキストマイニング入門”、オーム社(2002)
[8]北研二、津田和彦、獅々堀正幹、“情報検索アルゴリズム”、共立出版(2002)
[9]http://www2.jnes.go.jp/atom-db/jp/index.html
[10]http://chasen.aist-nara.ac.jp/