テキストマイニングによるトラブル事例情報の有効活用

Effective Use of Trouble Information by Text-mining Method

東北大学大学院　高橋　信	Makoto Takahashi	Member
東北大学大学院　内松　洋輔	Yousuke Utimatu	Non-Member
東北大学大学院　加須屋秀彰	Hideaki Kasuya	Non-Member
東北大学大学院　北村正晴	Masaharu Kitamura	Member

Abstract

The method of text-mining has been applied to accumulated documents on anomaly for nuclear power plant to represent them in systematic way. The twelve event reports from the public database of the nuclear power plant anomalies have been taken as examples. It has shown that the similarity and singularity of the event can be successfully described in the proposed framework. It has also shown that the important key words and the contextual structure of the report have been successfully extracted by the proposed methods.

Keywords:Text mining, Principal Component Analysis,

Email:makoto.tahakashi@qse.tohoku.ac.jp

1. 背景

近年のシステムの大規模化に伴い、故障やトラブルの影響も甚大なものになりつつある。このような背景からシステムにはより一層の信頼性が要求されている。大規模システムの高信頼性化に関しては、設計・施工における信頼性の向上も重要な要素ではあるが、現実問題としては長い供用期間中における保全活動が中心的役割を果たす。保全に関しては、状態をより詳細に監視しながら状況に応じた保全を行うという「状態監視保全」の考え方が積極的に導入されつつあるが、別の方向からの考え方として過去の故障事例を有効に活用する考え方の重要性が近年指摘されている。通常、原子力プラントに代表される高度な信頼性が要求されるシステムにおいて異常が発見された場合、水平展開という形で故障を起こした機器と同様の機器を利用している他のプラントで再点検が行われる。このような事例を活用した保全活動の場合、蓄積されている事例をどのように有効に活用するかが重要なポイントとなる。計算機技術の発展により、文書情報として故障事例を蓄積することは容易になってはいるが、問題は蓄積された情報をどのように有効活用するかという点である。原子力プラントに関しては、異常事例を記述した報告書は、大量の電子データとして蓄積され一般に公開されているが、表現形式や詳細度の違いなどで統一的に整理されているとは言い難いのが現状である。他の産業分野においても、社内や業界内で独自の故障事例データベースを持つ場合もあるが、それらも同様に表現形式や詳細度は統一されていない。

このような現状に対して、表現形式を統一したデータベースを作成する方向での対処法も考えられるが、現実問題としてこれまで蓄積された大量の事例をデータベース化することは困難である。本研究グループでは、過去の事例から故障生起に関する汎用的知識を導出する研究[1]や、導出された知識のデータベース化に関する研究[2]を行ってきた。本研究では、近年大量に蓄積された文書データから構造化された知識を抽出する手法として注目されているテキストマイニングの手法[3,4]の適用を試みた。テキストマイニングの適用により、異常事例に関する知識を抽出し、事例のその共通性、特異性を把握することが可能であると考えられる。

2. 目的

本研究では原子力プラントにおける異常事例に対しテキストマイニング技術を適用し、原子炉プラントにおける異常事例からの知識抽出、故障生起知識の構造化の２点を行い、その有効性を確認することを目的とする。

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. 前処理とキーワード抽出

文書からキーワードを取得するため、処理の第一段階として形態素解析を行った。形態素解析は、文章を文法的に解析することで、品詞分解を行う技術であり、本研究では、形態素解析ソフトウェア「茶筅」[5]を使用する。

3.2. 単語の重み判別

本研究では文書中における有意な単語を選択するために、以下の二つの観点から単語の重要度、すなわち重みを決定した。

A)単語の頻度による重み付け

単語を文章集合における頻度を尺度として重み付けをする手法であり、基本的に頻度が高いものが優先される。

B)確率的見地による重み付け

文書中の単語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、内容語は従わないと仮定し、その違いから重みを決定する手法。以上の二つである。

3.2.1.単語の頻度による重み付け

重み付けの手法には様々な提案がなされているが、以下に述べる重み付けを考慮し単語頻度による重みを決定する。

(1)文書内頻度に基づく重み l_{ij}

文書 D_j における単語 w_i の出現頻度 f_{ij} に基づいて計算され、以下の式で示す。

$$l_{ij} = \begin{cases} 0.5 + \dfrac{f_{ij}}{\max f_{ij}} & (f_{ij} > 0) \\ 0 & (f_{ij} = 0) \end{cases} \quad (1)$$

(2)大域重み g_i

文書集合に対する単語 w_i の頻度に基づいて計算され、以下の式で示す。

$$g_i = \log \frac{N}{n_i} \quad (2)$$

ただし、Nは総文章数、n_i は単語 w_i を含む文書数

(3)文書正規化係数 n_j

文書が長ければそれだけ単語の数も増加し、重みが大きくなってしまう。そこで文書の長さによる影響を排除することを目的とし、以下の式で計算される。

$$n_j = \sqrt{\sum_{i=0}^{m} (l_{ij} g_i)^2} \quad (3)$$

これら三つの指標に基づく単語頻度による重み付けは以下の式で表すことができる。

$$d_{ij} = \frac{l_{ij} g_i}{n_j} \quad (3.4)$$

3.2.2.ポアソン分布を利用した重み付け

ポアソン分布は一定の時間や範囲において、ランダムな事象の発生回数を確率的に表現するモデルであり、以下の式で表される。

$$P(x;\lambda) = \exp(-\lambda)\frac{\lambda^x}{x!} \quad (5)$$

そこで、ある文書における単語の出現頻度分布をポアソン分布として近似する。ここでλは単語 w_i が1文書中に出現する回数の期待値となり、文書集合の総数をN、単語 w_i の大域頻度を F_i とすると、$\lambda_i = F_i/N$ となる。これより、単語 w_i が文書中に出現する確率は $P(x;\lambda_i)$ となる。よって、実際の大域重みによる単語の出現頻度と上の式から計算された出現確率の差を用いた重み RIDF (Residual inverse document frequency) を計算することによって、ある単語における頻度分布の偏りが把握できる[6,7]。以下にその式を示す。

$$\begin{aligned} RIDF_i &= \log \frac{N}{n_i} - \log(1 - p(0;\lambda_i)) \\ &= g_i - \log\left(1 - \exp\left(-\frac{F_i}{N}\right)\right) \end{aligned} \quad (6)$$

3.2.3.総合指標

本研究では、以下述べた、単語頻度による重み付けとポアソン分布を利用した重み付けを組み合わせた総合指標を利用する。その式を以下に示す。

$$v_{ij} = \frac{d_{ij}}{\max d_{ij}} \cdot \frac{RIDF_i}{\max RIDF_i} \quad (7)$$

この総合指標を用いて牽引語の選択を行う手法を tf･idf 法と言う。この指標は文書中の単語出現頻度のランダム性と文書に依存した偏りを総合的に取り入れた指標であると考えることができる。

3.3. 共起度

共起度とは文書中における単語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。本研究では単語間距離と前節で示した総合指標に注目した式を使用する。単語間距離を $l_{w1,w2}$ 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。

$$co(w1,w2) = \begin{cases} (v_{w1D} + v_{w2D}) * \dfrac{L_i - l_{w1,w2}}{L_i} & (d_{w1}, d_{w2} > 0) \\ 0 & (d_{w1}, d_{w2} = 0) \end{cases} \quad (8)$$

Table 1 対象事例

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A系流量変動に伴う原子炉手動停止について	平成6年5月29日
2	福島第一原子力発電所2号機	シュラウド中間部リングひびについて	平成6年6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7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伴う原子炉手動停止について	平成7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伴う原子炉手動停止について	平成7年10月13日
6	高浜発電所2号機	第6A高圧給水加熱器細管漏えいについて	平成8年1月14日
7	伊方発電所3号機	湿分分離加熱器逃し弁の損傷について	平成8年1月6日
8	高浜発電所1号機	B-主給水制御弁点検に伴う原子炉手動停止について	平成8年11月26日
9	福島第一原子力発電所1号機	ジョットポンプ入り口配管のひびについて	平成8年12月24日
10	敦賀発電所2号機	化学体積制御系配管からの漏えいに伴う原子炉手動停止について	平成9年10月13日
11	福島第一原子力発電所4号機	中性子計測ハウジングのひびについて	平成9年10月13日
12	高浜発電所2号機	昇圧変圧器保護継電装置の動作による原子炉自動停止について	平成8年3月15日

以上より、各事例における共起度を要素とした文書行列を作成した。次に文書行列群からの知識抽出について述べる。行列表現された事例群から同じ行をそれぞれ抜き出し新たな行列を作成し、その行列に対して主成分分析を用いて知識抽出（知識ベクトルと呼ぶ）を各行ごとに行った。そして、知識ベクトルを足し合わせた知識行列を用いて復元行列を求める。復元行列は以下の分散・共分散行列の主成分分析の性質を利用し作成した。ある文書行列Dにおける、ランクRのi行目をD_{Ri}、i行目に対する知識ベクトルをD'_{Ri}とするとき、その内積S_{Ri}には次の関係が存在する。

$$D_{Ri} \cong S_{Ri} \cdot D'_{Ri} \qquad (9)$$

これによって知識ベクトルから文書行列を復元することができる。このとき複数の固有ベクトルを求める基準として、累積寄与率を80%としている。

3.4. 構造化手法

ここでは得られた知識をどのように構造化し、提示するかについて述べる。ある文書行列Dの復元行列D'を用いて、自分自身との内積を求め最上位にある行を選択し、その中で値の大きい上位3つのキーワード（w_p、w_q、w_r）を選択する。さらに、（w_p、w_q、w_r）以外でw_pとの共起度が上位3つのキーワード（w_s、w_t、w_u）を選択する。（w_p、w_q、w_r、w_s、w_t、w_u）以外でw_sとの共起度が上位3つのキーワード（w_x、w_w、w_v）を選択した。

4. 適用結果

4.1. 異常事例文書からの知識抽出

3．述べた知識抽出方法の有効性を確認するために、適用事例として以下の原子力プラントトラブル事例12事例を用いた。

このとき既存事例をNo1-No10まで、新規事例をNo11,12とした。No11は既存事例の中に類似事例が存在しておらず、No12は類似事例が存在している。このとき知識行列の有効性の確認方法として、文書行列と復元行列とのコサイン尺度（情報の損失なく知識ベクトルを作成しているかどうか）、知識ベクトルと新規事例の文書行列とのコサイン尺度（これまで得られた知識で新規事例をどの程度説明できるかどうか）を用いた。表4.2に適用結果を示す。

10事例を既存事例とした場合、10事例中に類似事例が存在しないNo.11のコサイン尺度が小さくなっており、これはこの事例がこれまでの事例とは異なる新規事例であることを示している。

Table 2 抽出知識の有効性検証結果

既存事例総数	既存事例のコサイン尺度	事例No11のコサイン尺度	事例No12のコサイン尺度
10事例	0.92	0.18	0.41
12事例（No11,12を含む）	0.93	0.94	0.94

10事例を既存事例とした場合、10事例中に類似事例が存在しないNo.11のコサイン尺度が小さくなっており、これはこの事例がこれまでの事例とは異なる新規事例であることを示している。これに対して、No.12のコサイン尺度はNo.11に比べると高いことから、既存事例中に類似の事例が存在していること示唆している。12事例を既存事例とした場合は、No.11,No12に対するコサイン尺度はともに高くこれらが既存事例中に含まれることを明確に示している。この結果から、提唱手法は事例の意味的な類似度を良好に判別していることを示している。

4.2. 故障知識の構造化

次に、故障生起知識構造化の有効性を検証するため、提唱手法を表4.1の事例群に適用した。例として事例No.1への適用結果を以下に示す。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。図1より、「ビーム」、「小片」

などの現象面での構造化されており、「ビーム」から「応力」「割れ」など異常事例がどのように起こったか、その原因の関係が構造化されていることが示されている。

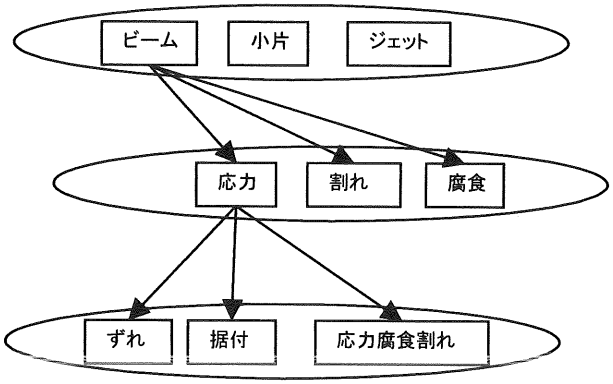

Fig.1　Cause-Consequence Relationship of Failure Event No.1

5. 考察

本研究では、キーワードの重要度を2つの相反する統計的手法を用いることにより表現することで、特異性と一般性の間でのバランスのとれた重み付けが可能となった。また、本研究において異常事例をキーワード間の関連度である共起度を用いて行列化することを提案しているが、これにより異常文書を共通の次元で評価でき、コサイン尺度による文書間の意味的比較も可能となる手法であることを確認した。

知識抽出によって得られる知識は人間によって理解・把握されるべき知識であり、この知識をどうやって人間に伝えるかが重要な問題として挙げられる。人間にとって理解しやすいように、故障生起知識の構造化においては、キーワードを9個に制限しているが、人間が見過ごしているような因果関係を見いだすためには、更に広範囲に表現を行う必要があると考えられる。

6. 結論

大規模複雑システムの保全活動における既存知識の有効活用のため、蓄積された異常事例文書に対してテキストマイニング技術を適用し、以下の知見を得た。

・原子炉プラントにおける異常事例からの知識獲得

本研究で提案した知識抽出手法は事例間の類似度・特異度の判定、新事例からの新知識の発見などの点において、有効であることを確認した。

・故障生起知識の構造化

本研究で示した構造化手法は、異常事例から少数のキーワードだけで故障知識を構造化しうる可能性を示した。

今後は、より広範囲の事例に対して本手法を適用しその有効性を更に検討する予定である。最終的には、多くの産業分野の事例を集積しそこから構造化された故障生起に関する知識を統合的に抽出し、広範囲に異常発生の可能性を検討する方策に結びつけていく予定である。

7. 参考文献

[1]高橋信，“原子炉運転員支援高度化のための知識処理技術の開発”(1991)
[2]尾暮拓也，“プラント診断知識導出のための故障事例データベース構築の基礎研究”，(1997)
[3]砂山渉，大津幸生，谷内田正彦，“KeyGraph キーワノード抽出ツールから発見ツールへの展開”，発見科学とデータマイニング，共立出版，2000，pp.45－pp.53
[4]豊田正，芝山悦哉，“ズーミング技術を用いた対話的情報検索インタフェース”，発見科学とデータマイニング，共立出版，2000，pp.262－pp.271
[5]茶筅，“http://chasen.aist0nara.ac.jp/index.html.ja”
[6]北研二，津田和彦，獅々堀正幹著，“情報検索アルゴリズム”，共立出版（2002）
[7]松倉健志，“文書の話題構造と文書間の意味的関連の発見にもとづく Web 検索に関する研究”，(2001)