71

ビッグデータ分析技術を応用した ソフトウェア不具合の分析実施事例[1]

1. 概要

大規模なソフトウェア開発プロジェクトの現場では、テストにより不具合が数千件摘出されることもある。従来は不具合の発生傾向や改善が必要な箇所を把握するために「欠陥分類法」を用い、不具合情報に「重要度」「現象分類」「原因分類」「作り込み工程」「摘出すべき工程」「未発見理由」などの定型的な分類情報[2]を付与し、それらの分類情報を機能やプログラムごとに集計することで品質状況の分析を行ってきた（図 71-1 参照）。

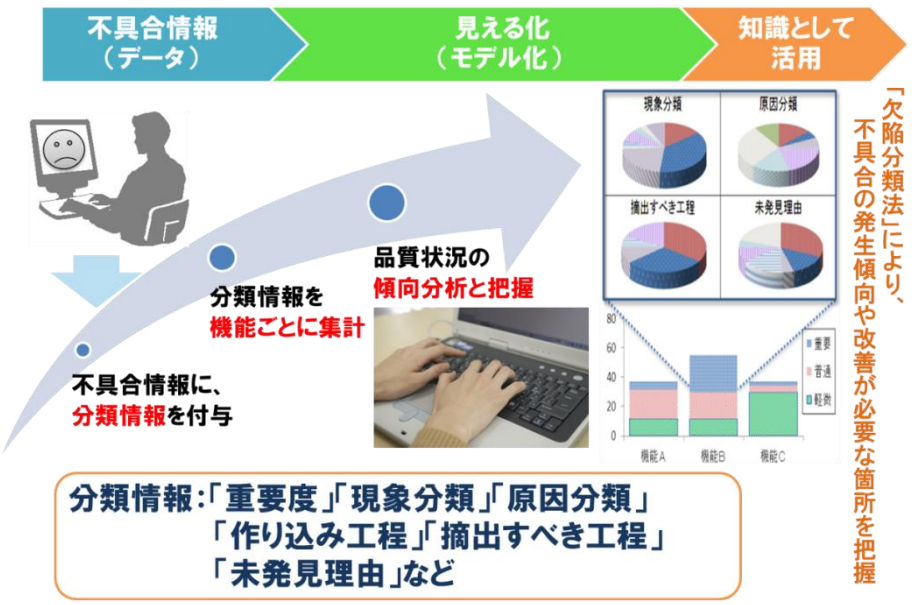


図 71-1　欠陥分類法による品質分析の概要

しかし、ソフトウェア開発プロジェクトで扱う問題は多岐にわたり、分析する情報の全てが定型的に分類されているとは限らない。また、分類されていたとしても、情報の欠落や属人的な分類判断により、しばしば精度に欠けた情報で分析を行わざるを得ないのが実状である。プロジェクトによっては分析の目的／情報／期間は様々で、必ずしも上述の方法で品質状況や問題を的確に分析できるとは言えず、効率的な分析手法を確立する必要があった。

このような現状を踏まえて、分類前の不定形な問題情報（テキスト情報）を分析するために、

[1] 事例提供: 株式会社日立ソリューションズ ソリューション品質保証本部 角口 勝隆 氏

ビッグデータ分析技術の適用を図った。これにより、大量に蓄積された問題情報を効率良く分析することが可能となった。本編では、ビッグデータ分析技術の１つである「テキストマイニング」を活用した問題発生状況の把握、および改善が必要な箇所のキーワードを抽出する分析事例を紹介する。

2. 取り組みの目的

2.1. 解決すべき課題

コンピュータの誤動作が社会へ与える影響が年々高まっている中、金額計算結果不正などの重要障害を撲滅させるための取り組みが当社内で立ち上がった。その取り組みの一環として、当社の「障害管理システム」に蓄積された不具合情報を分析し、共通かつ潜在的な障害発生要因（障害のポテンシャル）を把握することが求められた。

作業開始当初は、従来のソフトウェア品質分析のように欠陥分類法を用いて、定型的に分類された情報を元に障害管理情報の分析を試みた。しかし障害管理システムはお客様状況の把握や対策完了までのステータス管理を主目的としており、ソフトウェア不具合票のように細分化された分類情報は付与されていない。この状況で分析に欠陥分類法を用いる場合の隘路事項を以下に挙げる。

（1） 記載された文章情報（テキスト）から「現象」「原因」等の分類種別ごとにキーワードを抽出し、分類・整理をしなければ、分析を行うことができない。

（2） 人手で大量データを分類・整理するには、結果を得られるまで時間を要する。

（3） 複数人で分類を行うと、属人的な分類判断により分類結果が異なり、分析精度が悪くなることがある。

（4） 最適な分類種別を定めることが困難である。分類種別が抽象的であれば属人的差異が低減し分析所要時間が早くなる反面、具体的事象を把握しにくい。分類種別を具体化すれば事象を具体的に把握できるが、分類作業が煩雑となり、属人的差異が生じやすくなる。（図 71-2 参照）

図 71-2　分類種別の抽象度による差異

解決すべき課題は、文章（テキスト）として記述された「現象」や「原因」などの情報から、共通的な要因を効率良く探し出すことである。

2.2. 分析作業の目的

問題情報を分析する上での目的を表 71-1 に示す。分析のために与えられた期間は８日間であり、期間内に実効性のある施策を提案資料としてまとめる必要があった。

本編では「①現状把握」および「②傾向分析」の作業を効率良く実施した事例を紹介する。

表 71-1　分析作業の目的

分析作業の目的	
①現状把握	問題の根本原因を対策するために、問題の背景にある要因を掘り下げて、正しく現状を把握すること
②傾向分析	効率的に対策効果を上げるために、問題の発生傾向や共通的な要因を分析し、問題の発生しやすい部分（ウィークポイント）を洗い出すこと
③改善施策	洗い出したウィークポイントから、再発防止／未然防止に対し実効性のある施策を提案すること

本編の対象範囲

2.3. 課題解決のための手段

課題を解決するためには、大量に蓄積された問題情報の分析における分類作業の自動化・効率化が必要である。そこで、テキスト情報を自動的に分類・可視化する「テキストマイニング」の活用を図った。なお、テキストマイニングを適用するに至った経緯として、システム稼働後の問合せ内容を要約するためにテキストマイニングを活用しており、この分析方法はソフトウェアの不具合分析にも応用可能ではないかと考えたことが挙げられる。

テキストマイニングによる分析作業の概要を図 71-3 に示す。

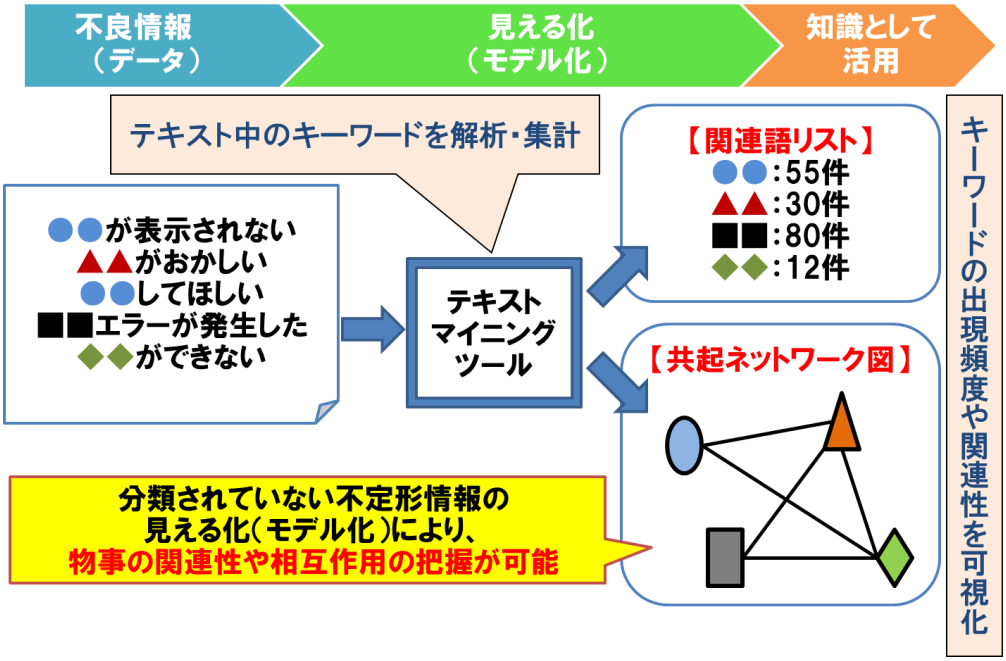


図 71-3　テキストマイニングによる分析作業の概要

3. 適用ツール・分析技法

3.1　取り組みの対象

本編での分析事例の対象は、当社の「障害管理システム」に蓄積された 1182 件の不具合情報である。

3.2　適用ツール

本編ではテキストマイニングツールとして著名な、オープンソースの「KH Coder（作：立命館大学産業社会学部 樋口耕一准教授）[4]」を用いる。商用ツールではなく、オープンソースを採用した理由は以下のとおりである。

- 初期導入費用を抑えることが可能
- 簡易な分析目的であれば、十分実用に耐え得る
- 無償ツールを扱うため、分析ノウハウを一般展開しやすい

KH Coder では、Excel や CSV 形式データの特定列に記述されたテキスト情報を元に、様々な分析を行うことができる。図 71-4 に、KH Coder の分析機能による出力結果の一例を示す。

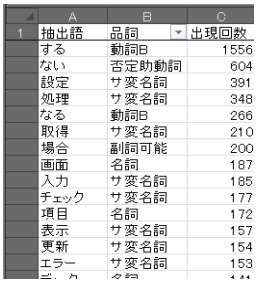
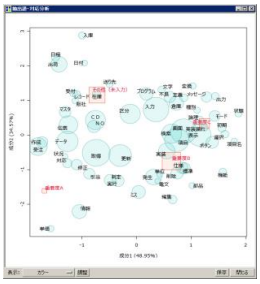
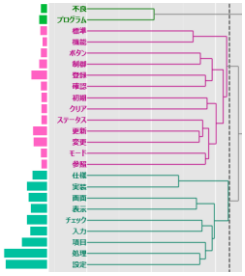
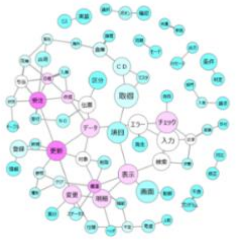
<table>
<tr>
<td>抽出語リスト
 
<table>
<tr><th></th><th>A</th><th>B</th><th>C</th></tr>
<tr><td>1</td><td>抽出語</td><td>品詞</td><td>出現回数</td></tr>
<tr><td></td><td>する</td><td>動詞B</td><td>1556</td></tr>
<tr><td></td><td>ない</td><td>否定助動詞</td><td>604</td></tr>
<tr><td></td><td>設定</td><td>サ変名詞</td><td>391</td></tr>
<tr><td></td><td>処理</td><td>サ変名詞</td><td>348</td></tr>
<tr><td></td><td>なる</td><td>動詞B</td><td>266</td></tr>
<tr><td></td><td>取得</td><td>サ変名詞</td><td>210</td></tr>
<tr><td></td><td>場合</td><td>副詞可能</td><td>200</td></tr>
<tr><td></td><td>画面</td><td>名詞</td><td>187</td></tr>
<tr><td></td><td>入力</td><td>サ変名詞</td><td>185</td></tr>
<tr><td></td><td>チェック</td><td>サ変名詞</td><td>177</td></tr>
<tr><td></td><td>項目</td><td>名詞</td><td>172</td></tr>
<tr><td></td><td>表示</td><td>サ変名詞</td><td>157</td></tr>
<tr><td></td><td>更新</td><td>サ変名詞</td><td>154</td></tr>
<tr><td></td><td>エラー</td><td>サ変名詞</td><td>153</td></tr>
</table>
</td>
<td>対応分析
  </td>
<td>階層クラスター分析
  </td>
<td>共起ネットワーク
  </td>
</tr>
<tr>
<td>語の出現頻度を、品詞別に一覧表示したもの。
品詞ごとの特徴や傾向を確認することが可能。</td>
<td>似たような語を、関連する項目と組み合わせて配置した図。
項目ごとに、語の出現傾向を確認することが可能。</td>
<td>類似する語を階層構造で分類した図。
不具合パターンの把握や類似見直し観点の考察時に有効。</td>
<td>語と語の繋がりをネットワーク図化したもの。
感覚的に、どのような事象が頻出しているのかを確認することが可能。</td>
</tr>
</table>

図 71-4　KH Coder の分析機能による出力結果の一例

3.3　分析技法

図 71-4 以外にも KH Coder の分析機能は多々あるが、本編では不定形な問題情報（テキスト情報）から物事の関連性や相互作用を可視化するために有用な「共起ネットワーク」を用いた分析事例を取り上げる

「共起ネットワーク」の「共起」とは、任意の文書や文において“ある文字列と他のある文字列が同時に出現する（＝共に起こる）こと”である。「共起ネットワーク図」とは、文字列間の共起性をリンクとして表したものである。図 71-5 に、共起ネットワーク図の作成概念を示す。

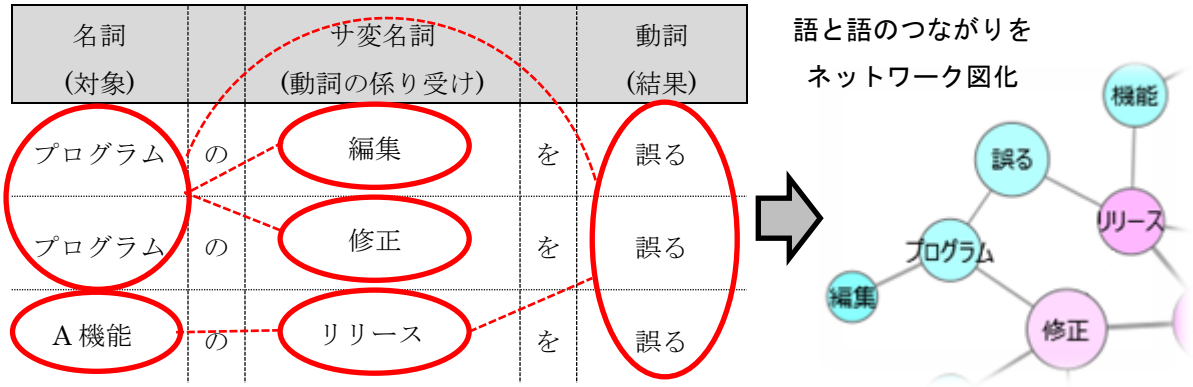


図 71-5　共起ネットワーク図の作成概念

ネットワーク図を用いた分析方法として、Linton.C.Freeman[5]が提唱する「中心性（Centrality）」という指標が存在する。「中心性」とは、ネットワークを構成する各要素が、ネットワーク内でどの程度中心的な位置にあるかを示す指標である。例えば社会ネットワーク分析では、情報を効率良く拡散させたい場合は、媒介的な位置づけとなる人物を特定して情報を伝達している。問題分析においては、複数要因の中から関連性の強い要因を特定し除去することで、全体的な問題発生頻度を抑止できると捉える。

図 71-6 に、社会ネットワーク分析を例とした、問題分析への応用例を示す。

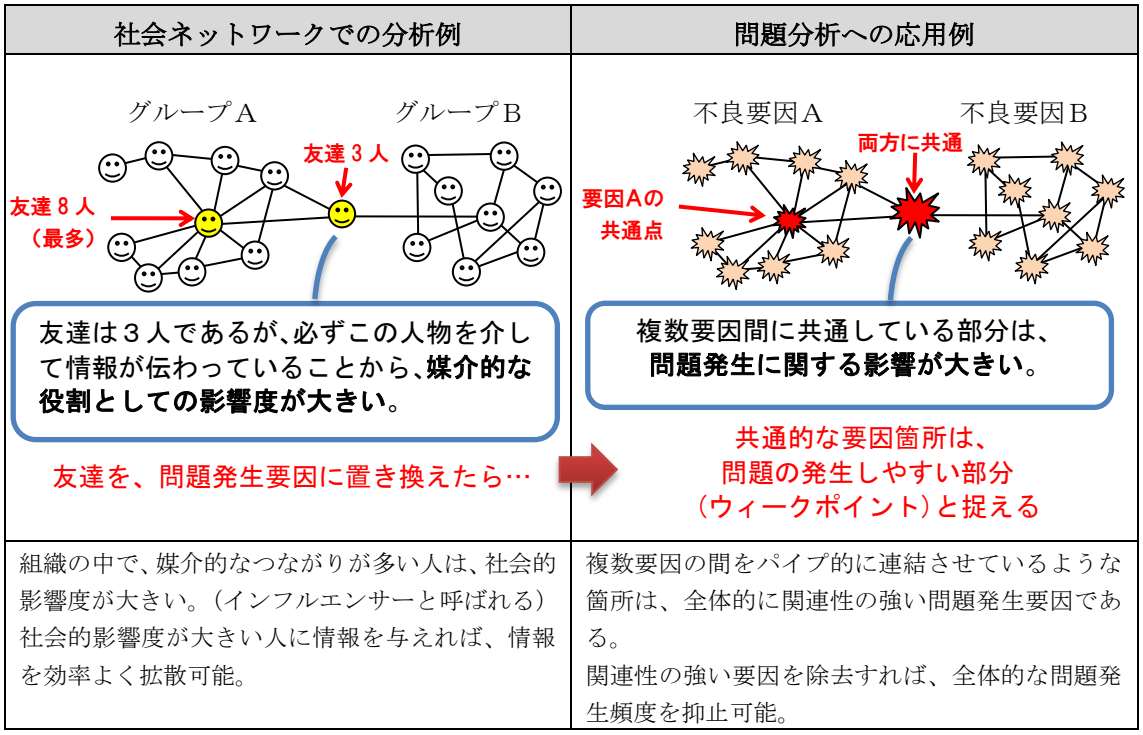
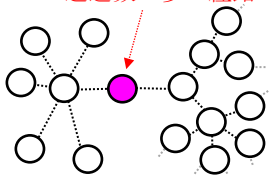
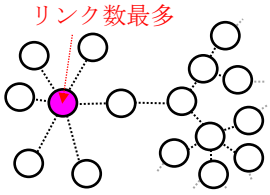
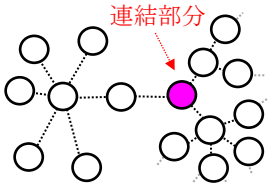


図 71-6　社会ネットワーク分析の問題分析への応用例

代表的な中心性の指標と、問題分析での応用例を表 71-2 に示す。

表 71-2　ネットワーク分析での中心性指標と問題分析での応用例

中心性の指標	意味と、社会ネットワーク分析での指標例	問題分析での応用例
媒介中心性 (betweenness centrality) 通過数の多い経路 	各要素を最短経路で結んだ場合に、経路が要素を通過する回数の多さを示す。 集団内において、他のメンバーをつなぐパイプとしての役割を示す指標。 ※ネットワーク機器のブリッジみたいなもの	<u>全体影響が大きい</u> 問題要素と捉える。
次数中心性 (degree centrality) リンク数最多 	他の要素へのリンクの多さを示す。 より多くのメンバーと仲間関係を持っていることを示す指標。 ※ネットワーク機器のハブみたいなもの	<u>共通性のある</u> 問題要素と捉える。
固有ベクトル中心性 (eigenvector centrality) 他と関連の多い要素の連結部分 	他と関連の多い要素が、多く連結しているかを示す。 仲間関係の多い他者とつながることによる、集団での影響力を示す指標。例えるなら、直接の知人ではなくとも関連先に豊かな人脈を持つ人物。 ※ネットワーク機器のルータや、ゲートウェイみたいなもの	問題要素を芋づる式に把握する際の、 <u>起点になる</u>と捉える。

※グラフ理論では「要素」は「ノード・点」「リンク」は「エッジ・枝」と呼ばれているが、本編では表現をわかりやすいものへ置き換えている。
※個々の問題に対しどの中心性指標を用いるべきかは一概に言えないが、各々の中心性の特徴から検討した問題分析での応用方法を踏まえて、影響の大きい問題要素を特定すればよい。

分類前の不定形な問題情報（テキスト情報）の中から、関連性が強い要素を「共起ネットワーク図」で可視化し、「中心性」指標に基づいて影響の大きい要素を特定することで、問題の発生傾向や改善が必要な箇所を把握できる。次に、当社の障害管理システムに蓄積された情報を対象にテキストマイニングを実施した事例を紹介する。

4. 取り組みの実施、及び実施上の留意事項、対処・工夫

4.1　実施内容

共起ネットワーク図を用いた分析作業の流れを図 71-7 に示す。分析を実施する前に、障害管理システムから分析対象データを CSV 形式で取り出し、KH Coder で解析が可能なようにデータを整備する。データ整備後、KH Coder で共起ネットワーク図を作成し、共起ネットワーク図の中からウィークポイントとなるキーワードや、改善を図るための観点を抽出する。

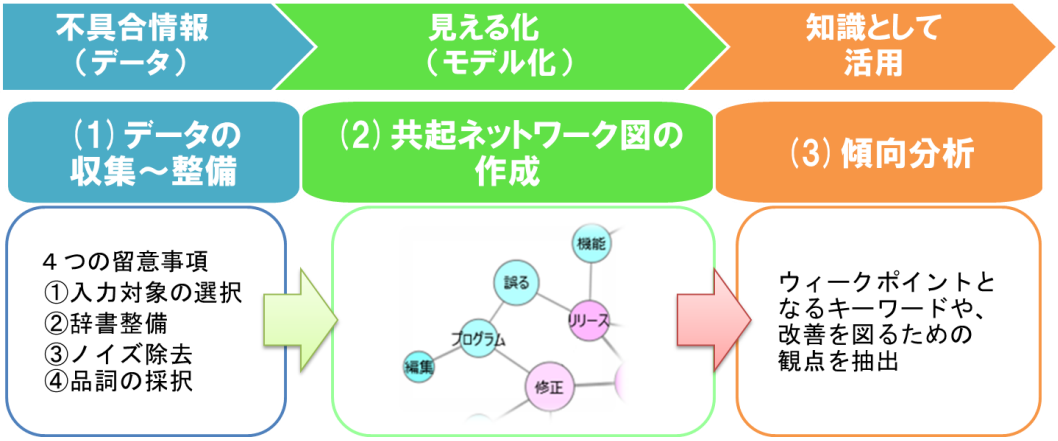


図 71-7　共起ネットワーク図を用いた分析作業の流れ

4.2　具体的な取り組み内容

4.2.1. データの収集～整備

共起ネットワーク図から意味のある分析結果を得るためには、「入力対象の選択／辞書整備／ノイズ除去／品詞の採択」といった前準備を実施する必要がある。共起ネットワーク図を作成する上で留意すべき事項を、表 71-3 に示す。

表 71-3　共起ネットワーク図を作成する上での留意事項

	留意事項
①	最適な入力対象を選択する
②	よく使われる単語を辞書としてツールに登録する
③	「ノイズ（不要な語）」を除去する
④	分析対象とする品詞を採択する

上記留意事項の詳細を、次に解説する。

① 最適な入力対象を選択する

KH Coder で入力対象とするテキスト情報は、「現象欄」や「原因欄」のように意味のある内容で分離されていることが望ましい（図 71-8 参照）。これは、記載内容の意味や目的が混在したものを入力対象とした場合、得られる結果も混在したものとなるためである。また、内容が十分に記載されている方が、具体的な分析結果を得やすくなる。

その他、「現象欄」や「原因欄」で分離されていれば以下の副次的効果が得られる。

・「現象欄」を解析した場合　⇒　テストケースやレビュー観点となるヒントを得やすい

・「原因欄」を解析した場合　⇒　失敗事例集のネタになる情報を得やすい

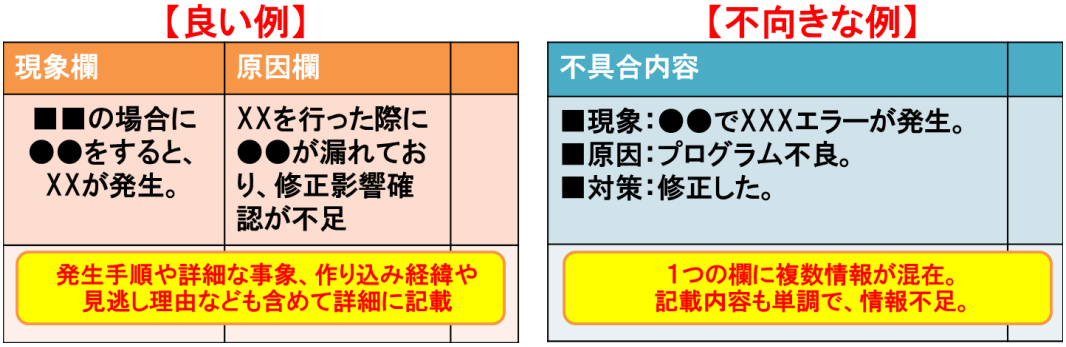


図 71-8　最適な入力対象の例

② よく使われる単語を辞書としてツールに登録する

KH Coder は各種研究機関が公開している辞書を同梱しており、日常生活程度の単語は正しく分析することができる。しかし特殊な専門用語が存在すると、語が適切に分割されず分析精度が悪くなることがある（図 71-9 参照）。このような場合、KH Coder では専門用語を「強制抽出語」として登録しておくことで、語の分割が適切に行われる。

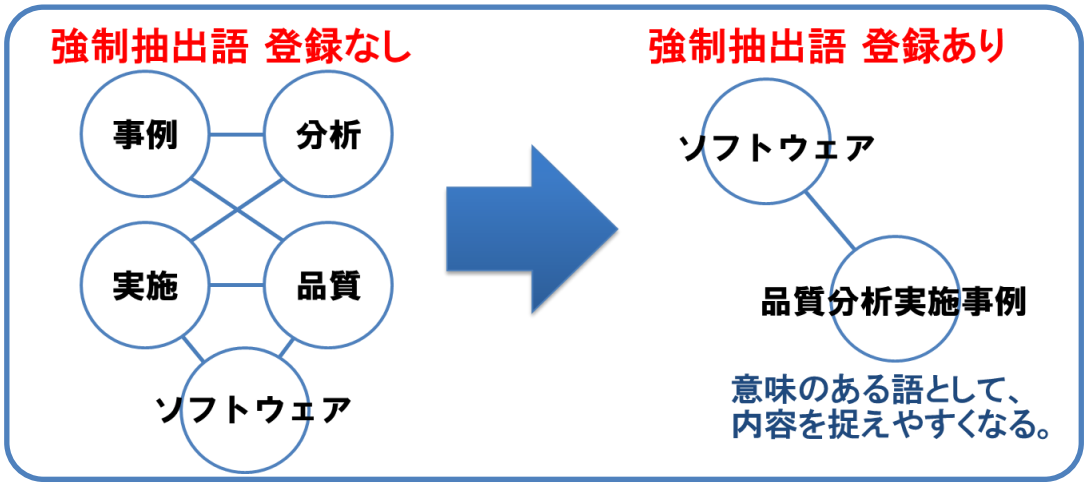


図 71-9　強制抽出語を登録した結果の例

③　「ノイズ（不要な語）」を除去する

分析精度を上げるためには、「ノイズ（不要な語）」を除去しておく必要がある。特に、図 71-10 に示すとおり、「プログラム」や「不良」のような出現頻度は高いが分析しても意味の無いような語は分析対象から除外しておくとよい。

テキストマイニングの要点は、余分な情報（枝葉）を剪定し、本質的な部分（根幹）を掴むことである。

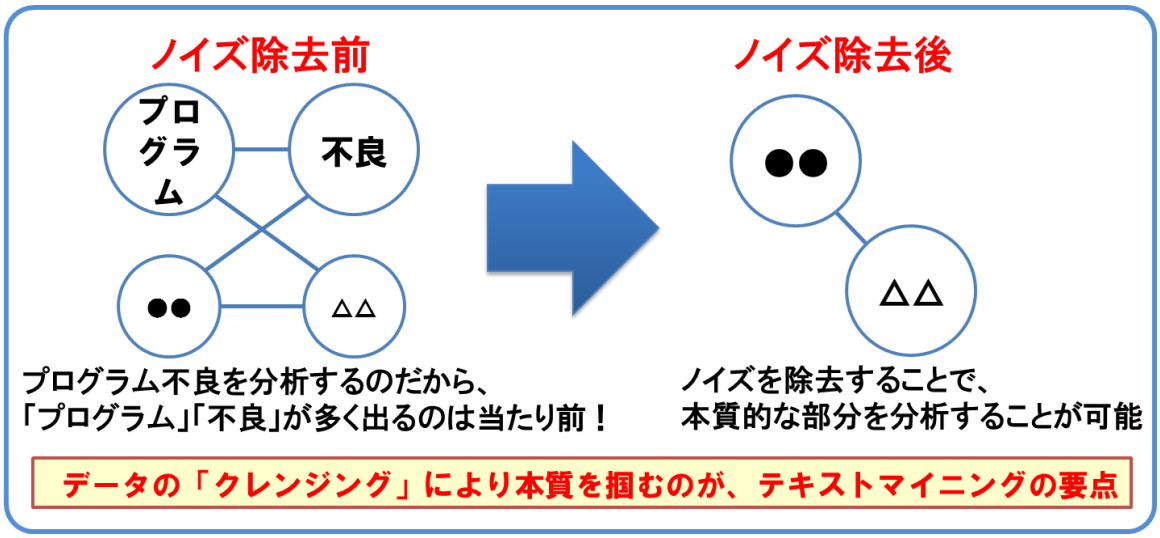


図 71-10　ノイズ除去実施後の例

④　分析対象とする品詞を採択する

分析対象とする品詞を絞り込むことにより、文脈を要約／抽象化した状態で意味を捉えやすくなる。図 71-11 に示すように、「名詞」「サ変名詞」「動詞」を主として採択すると良い。共起ネットワーク図を作成する際にこれらの品詞を採択することで、不具合内容の文脈を抽象化して捉えることができる。

対象　顕在契機（トリガー）　結果

名詞（対象）		サ変名詞（動詞の係り受け）		動詞（結果）
プログラム	の	編集	を	誤る
プログラム	の	修正	を	誤る
A機能	の	リリース	を	誤る

※これにより、必要な要素のみ分析可能となる

図 71-11　抽象化して文脈を捉えるために品詞を指定した例

4.2.2. 共起ネットワーク図の作成

KH Coder で処理が可能なようにデータを整備した後、KH Coder のツール機能で共起ネットワーク図を作成し、共起ネットワーク図の中からウィークポイントとなるキーワードや、改善を図るための観点を抽出する。図 71-12 は、障害管理システム情報から出力した情報の「原因欄」を入力対象として作成した共起ネットワーク図である。

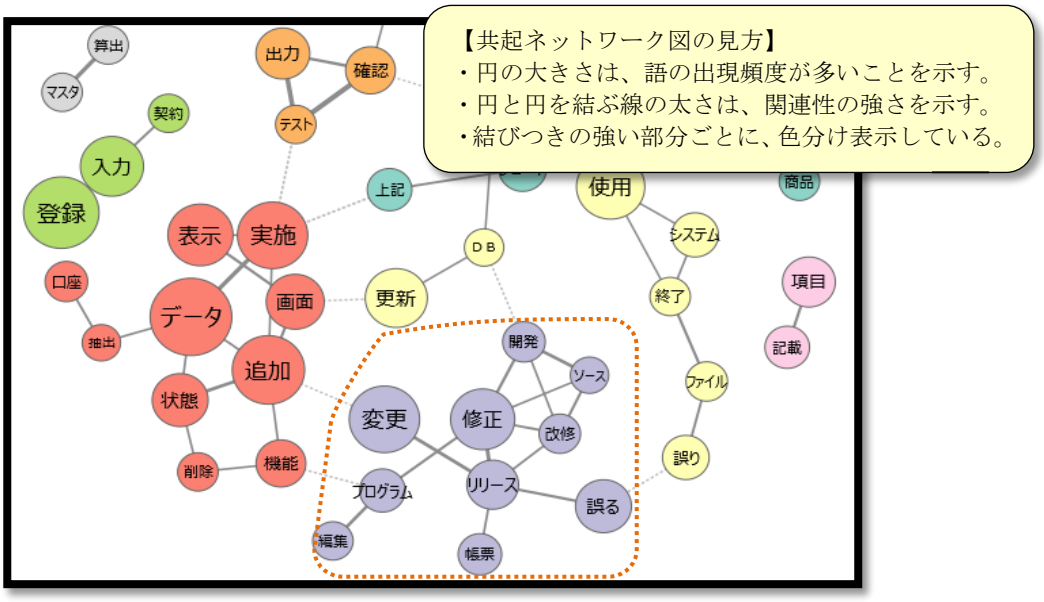


図 71-12　障害管理システムから出力した情報の「原因欄」を入力対象として作成

上記の共起ネットワーク図では、円の大きさは語の出現頻度が多いことを示し、円と円を結ぶ線の太さは関連性の強さを示している。また、結びつきの強い部分ごとに、色分け表示を行っている。図 71-12 の下部（点線で囲った範囲）では「変更」と「修正」の出現頻度が高く、そこから「リリース」「誤る」との関連性が強いことから、修正／変更作業とリリース時の作業に誤りがあったと、感覚的に捉えることができる。

4.2.3. 傾向分析

共起ネットワーク図を用いれば問題発生状況を抽象化した状態で可視化できる。ただし、ここで留意しておくべき事項は、捉えた内容は抽象化された情報であるため、あくまでも推定ということである。より具体的な特徴や要因を分析した上で事実確認をしなければ、現場の改善を図ることはできない。掴んだ傾向や特徴を掘り下げて要因を特定する作業（ドリルダウン）、および掴んだ特徴や要因から元データを確認する作業（ドリルスルー）を実施し、より正確に情報を捉える必要がある。具体的な分析作業の流れを、図 71-13 に示す。

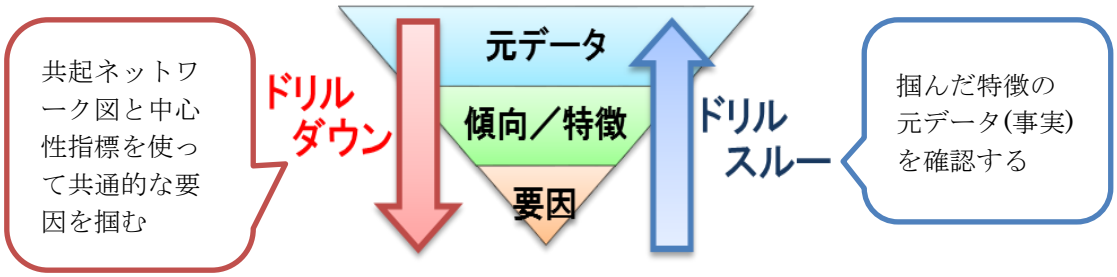


図 71-13　具体的な分析作業の流れ

(1)　ドリルダウンの実施：ウィークポイントとなるキーワードの分析

ウィークポイント（共通的な問題要素）を探索するために、ネットワーク図の中で中心性の強い問題要素を下記 1)、2)の手順で分析する。

1)　「固有ベクトル中心性」を用いて、問題要素を芋づる式に把握する際の起点を探す

2)　1)で探索した起点を基に、他の中心性から共通的な問題要素や影響の大きい問題要素を探す

KH Coder では、中心性が強い要素をピンク色で示すことが可能である。問題要素の起点を捉えるために、図 71-12 の共起ネットワーク図に「固有ベクトル中心性」を指定した結果を図 71-14 に示す。「固有ベクトル中心性」を指定した結果、「修正」が最も濃いピンク色となった。このことから、何らかの「修正」を起点として、不具合が発生していると捉えることができる。

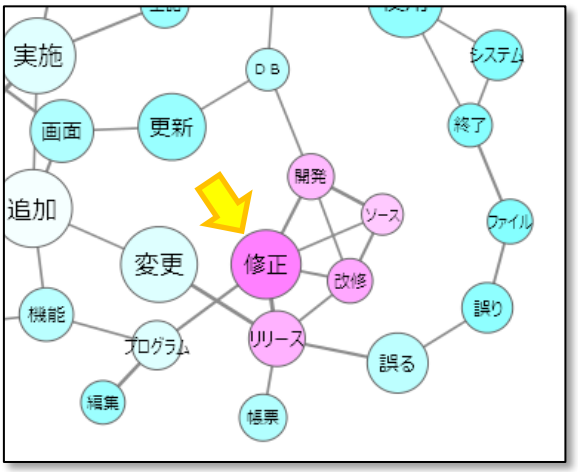


図 71-14　固有ベクトル中心性の指定結果

続いて、「修正」を行った時の関係性を掘り下げるため、「修正」に関連する語で共起ネットワーク図を作成した結果を図 71-15 に示す。なお、修正時における共通事項を確認するために、図 71-15 では「次数中心性」を指定している。

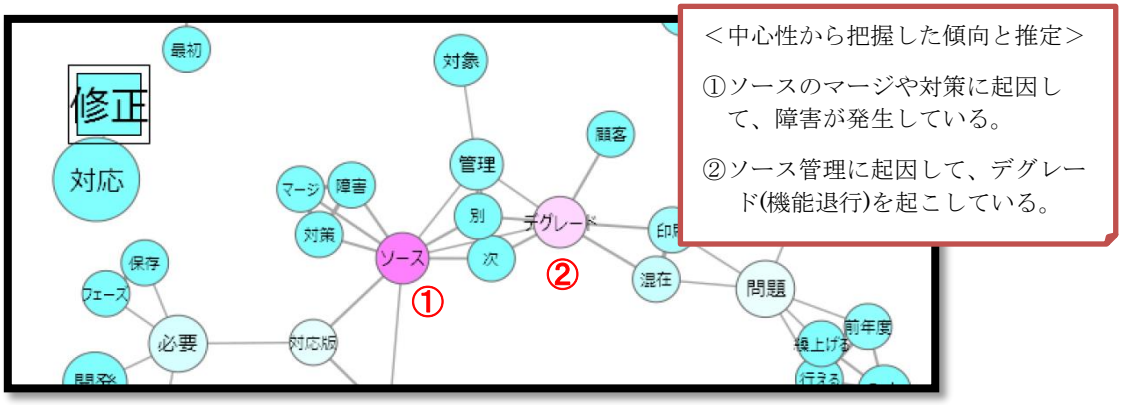


図 71-15　「修正」に絡む語をネットワーク図化し、次数中心性を求めた結果

図 71-15 では「ソース」、「デグレード」で次数中心性が高くなっている。その周辺の関連語から、ソースのマージや対策に起因して、障害が発生している可能性が推察される。また、ソース管理に起因して、デグレード（機能退行）が発生している可能性が推察される。

(2)　ドリルスルーの実施：捉えた特徴の詳細を確認

KH Coder では、「KWIC コンコーダンス」機能で特定のキーワードに着目して、前後の文脈を確認することが可能である。図 71-16 に、「ソース」をキーワードとして抽出し、前後の文脈を確認した例を示す。これにより、対象を絞った上で要点を確認することが可能である。その結果、図 71-15 での推定と、詳細確認結果は一致していた。

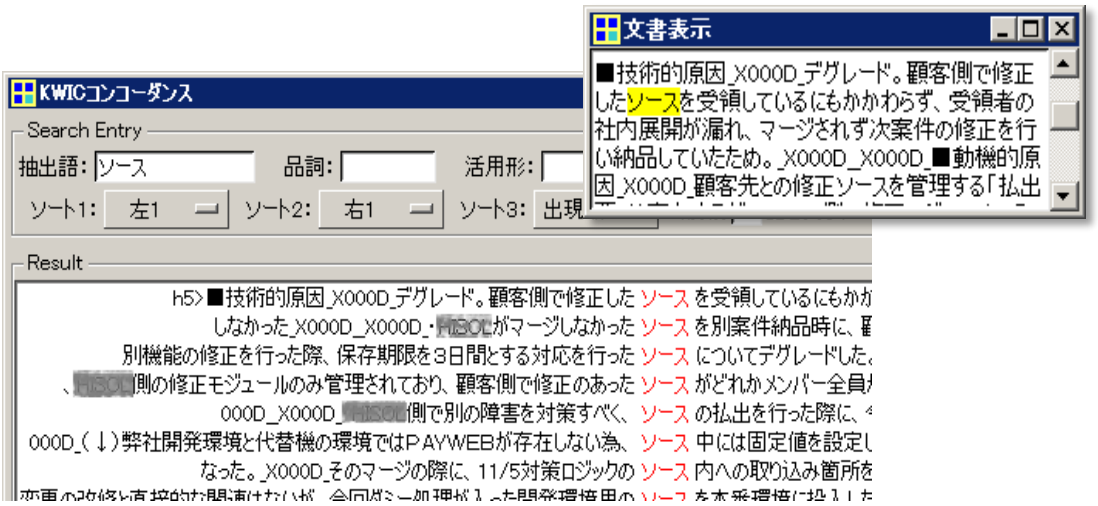


図 71-16　捉えた特徴の詳細確認を実施した例

4.3　分析結果と考察

ネットワーク図と中心性指標を用いて分析した結果、ソース管理や、ソースのマージ作業で障害が発生していたことが確認できた。このことから、プロジェクトメンバーが入れ替わった際、ソース管理ルールなどの躾が引き継がれていないのではないかという仮説が考えられた。この仮説に基づき事実を確認し、さらに要因を掘り下げて根本原因を対策することで有効な再発防止／未然防止を図ることが可能である、として分析結果の見解をまとめた。

5. 達成度の評価、取り組みの結果

表 71-1 の「分析作業の目的」に対する達成状況を表 71-4 に示す。テキストマイニングを採用した場合でも、分析作業の目的を達成できている。

表 71-4　分析作業の目的に対する達成状況

分析作業の目的		達成状況	評価
① 状況把握	問題の根本原因を対策するために、問題の背景にある要因を掘り下げて、正しく現状を把握すること	問題の背景を掘り下げるための仮説（ヒント）を得やすい	○
② 傾向分析	効率的に対策効果を上げるために、問題の発生傾向や共通的な要因を分析し、問題の発生しやすい部分（ウィークポイント）を洗い出すこと	ネットワークの中心性から、問題が発生しやすい部分の把握が可能	○

欠陥分類法を用いた場合の所要時間（想定）と、テキストマイニングを用いた場合の分析所要時間の比較を図 71-17 に示す。与えられた分析期間は８日間であったが、テキストマイニングを活用することで、より短期間（約３日短縮）で分析結果を提示することができている。

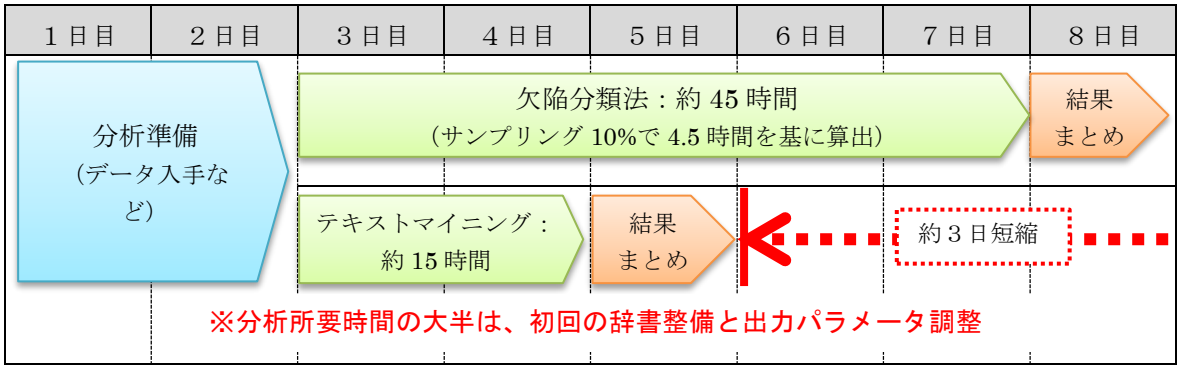


図 71-17　分析所要時間の比較

分析にテキストマイニングを活用することで、表 71-5 に示す効果が得られる。しかし、テキストマイニングでは期待できない事項もある。これらのメリット／デメリットを把握した上で、適切に分析技法を使い分ける必要がある。

表 71-5　ネットワーク分析による効果（メリット／デメリット）

メリット：期待できる効果	デメリット：期待できない事項
・問題発生傾向の概要を素早く把握 不定形な問題情報（テキスト情報）を、共起ネットワーク図で可視化することにより、現状の問題発生傾向を素早く把握することが可能。 **・改善が必要な箇所を素早く把握** 中心性指標により、ネットワーク図から改善が必要な箇所を素早く把握することが可能。 **・大量データを分析可能** 人手による分類や集計作業は不要。 大量データの分析で、特に効果を発揮する。	**・定量的な判断ができない** 分析結果は相関関係であり、定量的な良し悪しは判断できない。定量データと併せた分析・評価が必要。 **・対象データが少ないと効果を得られない** 目安として 100 件以上のデータが必要。 **・分析結果は分析者のスキルに依存** 全体的な特徴が抽象化された状態で可視化されるため、テキストとして特徴のない情報は把握できずに見逃してしまう可能性がある。 また、抽象化された情報となるため、具体的に物事を捉えるには、現場経験や、問題背景の理解が必要となる。

6. 今後の取り組み

6.1　今後の取り組み

(1)　効果を上げるために必要な要素

共起ネットワーク図による問題情報の可視化は、現状の問題発生傾向や、改善が必要そうな箇所を素早く把握する用途には有効である。ただし、この分析方法が効果を上げるかどうかは、適用する上で必要となる要素が揃っているかどうかに大きく依存する。図 71-18 に、共起ネットワーク分析で必要となる要素を示す。

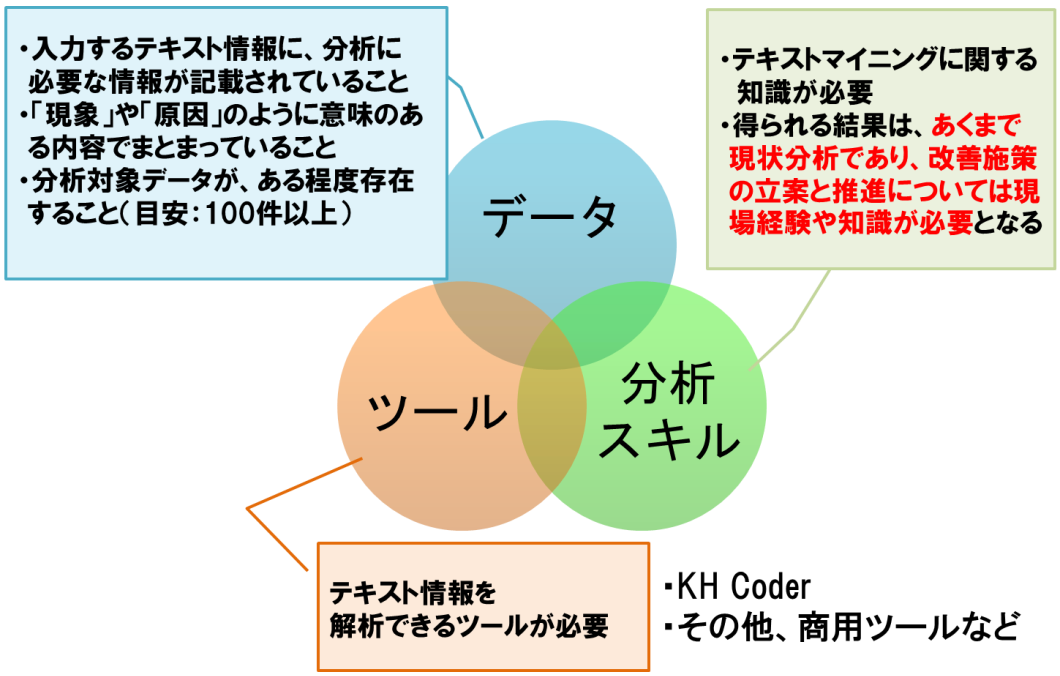


図 71-18　共起ネットワーク分析で必要となる 3 つの要素

(2)　今後の課題

より分析効果を上げるためには、以下の取り組みが必要である。

①　データの効率的な収集方法の検討

入力フォームのテンプレート化や、入力ガイダンス／必須入力チェックにより、必要な情報をより確実に収集できる仕組み／仕掛けを用意する。

②　分析スキルの一般化

分析スキルを一般化するために、優れた個人ノウハウの収集を図るとともに、標準的な分析作業プロセスを明確にする。

6.2　考察

共起ネットワークによる分析は、大量の不定形な問題情報（テキスト情報）から、共通的な問題要素を分析する用途に有効である。本編で紹介した事例に留まらず、以下のような活用方法が期待できる。

(1)　より多角的な分析が可能

従来の定性的評価（欠陥分類法による分析） および 定量的評価（生産規模あたりのチェックリスト件数や、不良発生件数）と組み合わせて分析することで、より多角的に問題を捉えることができる。例えば、欠陥分類法により重要度の高い問題に対象を絞った上で原因傾向を分析したり、生産規模あたりのチェックリスト件数は低いが不良発生件数が多い場合の現象内容（テキスト情報）からレビュー観点やテスト観点を抽出しフィードバックするなどの活用方法が挙げられる。

(2)　あらゆる現場での改善に活用可能

テキストマイニングを活用すれば、ソフトウェア開発プロジェクトに限らずあらゆる現場で改善を図ることができる。例えばシステム維持保守作業での作業ミスや、システム利用者からの問合せなど、様々な情報を分析し改善が必要なポイントを探すことが期待できる。

(3)　新たな発見への期待

共起ネットワーク図は、現在の状態をモデル化したものであり、現場が暗黙知的に捉えている意識を可視化できる効果がある。しかし「それは当たり前」、「だから何？」という結果で分析作業が終わってしまうこともある。このような事態を避けるためには、ありきたりのデータだけを分析するのではなく、新たな分析視点を考察したり、他の要素と組み合わせて分析を行うとよい。（分析視点例：部位／カテゴリ別、経年変化／定点観測等）

近年は IoT、M2M など、大量かつ様々な情報を収集するための基盤が整ってきた。これらのデータと組合せてテキスト情報を分析することで、新たな発見が期待できる。

参考文献

[1] 角口勝隆、ネットワーク型データモデルを用いた問題点の可視化と問題分析への応用例、ソフトウェア品質シンポジウム 2015、A2-1 セッション、2015

[2] 保田勝通、奈良隆正、ソフトウェア品質保証入門 高品質を実現する考え方とマネジメントの要点、日科技連出版社、2008

[3] 那須川哲哉、テキストマイニングを使う技術／作る技術―基礎技術と適用事例から導く本質と活用法、東京電機大学出版局、2006

[4] 樋口耕一、社会調査のための計量テキスト分析―内容分析の継承と発展を目指して、ナカニシヤ出版、2014

[5] Linton.C.Freeman、A Set of Measures of Centrality Based on Betweenness、American Sociological Association、1977

独立行政法人情報処理推進機構 技術本部 ソフトウェア高信頼化センター（IPA/SEC）