週報のテキストマイニングによるリスク対応キーワード抽出

野々村　琢人
(株)東芝　インダストリアルICTソリューション社
takuto.nonomura@toshiba.co.jp

安村　明子
(株)東芝　インダストリアルICTソリューション社
akiko.yasumura@toshiba.co.jp

弓倉　陽介
(株)東芝　インフラシステムソリューション社
yosuke.yumikura@toshiba.co.jp

要旨

ソフトウェアの研究開発管理において，週報などの報告書による情報は重要であるが，以下の課題がある．部門内のチームリーダによる集約された週報では情報が欠落してしまう．逆に部門メンバー全員の週報は膨大で冗長な表現も多く，短時間での確認が困難である．

今回，メンバー個人の週報および部門全員の週報に対してテキストマイニングを行い，リスク対応に必要なキーワードを抽出できることを確認した．

1. はじめに

ソフトウェアの研究開発において，担当者は週報（間隔によって，日報，半月報，月報などもある）を管理者に提出する．管理者は報告会などで直接担当者から情報を得る以外に，提出された週報から進捗やリスクを確認する．

部門メンバーが多い場合には，個人の週報ではなくチーム毎にチームリーダが纏めた週報を管理者は読む．しかし，チーム内の複数案件の情報を報告するため，チームリーダが適切に情報を集約しないと重要な情報が欠落する可能性がある．さらに週報にはフォーム（形式）を決めにくい，あるいは決めても守られないという傾向があり，個人ごとの週報の場合，そこから課題やリスクに繋がる情報を確実に読み取るには，時間がかかる．

そこで，主なトピックスや概要はチームリーダ週報で把握し，気になる／気にすべきトピックスを必要に応じて担当者個人から直接ヒアリングしたいという管理者の希望がある．

近年，アンケートなどの自由記述の回答データの分析にはテキストマイニングが利用されている．規則性のない非構造化データである「週報」を分析するのに，このテキストマイニング技術が有効であると考え，試行実験を行ったので，それを報告する．

なお本稿では「リスク」とは「それが発生（顕在化）すると影響を与える不確実な事象・状態」[1]とする．

2. 本試行実験の方法

リスクを発見し，その問題を事前に解決するため，次のようなステップを考える（図1）．Step1　週報を入力としたテキストマイニングによりリスク発見に繋がるキーワードを抽出，Step2　抽出したキーワードを元に担当者へのヒアリング，Step3　ヒアリングに基づく課題対応．

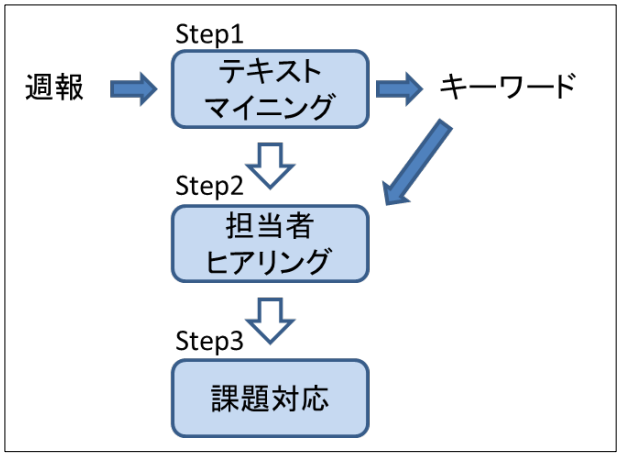

図1：リスク対応ステップ内のテキストマイニング

今回この Step2 の担当者ヒアリングに利用可能なキーワードを Step1 で効果的に抽出できることを確認するため，次の2つの実験を行った．

実験1：個人週報からキーワードを抽出

週報などの報告資料ではリスクを早めに検知し回避することが重要である．しかし悪い報告は遅れがちであり，特にトラブルかどうかの判断が難しい情報を担当者やチ

ームリーダがトラブルでないという情報としてしまうこともある．そこでチームリーダが集約したチーム週報と，担当者の個人週報をテキストマイニングで分析する．それにより管理者からみてリスクかもしれない状況を発見するためのキーワード抽出にどの程度効果があるかを確認する．

実験2：部門全員の週報からキーワードを抽出

部門内にメンバーが多い場合，管理者が全員の週報を確認する時間を確保することは容易でなく，部門を分割したチームの週報で把握することも多い．チームリーダが自チームの週報を集約することで消えてしまった「リスク発見に繋がるキーワード」を全員の週報から発見すること，及び全員の週報を俯瞰的にみて，チーム間での調整が必要なことなどに関する新たな表現に気づくことが可能かを確認する．

3. 個人週報からキーワードを抽出（実験1）

今回の実験では計量テキスト分析やテキストマイニング専用のフリーソフトウェア「KH Coder」[2][3]を利用した．これは「茶筅」[4]などの形態素解析システムにより語を抽出し，語の繋がりの強さなどを統計的に解析するソフトウェアである．アンケートの自由回答項目やインタビューデータなどのテキスト分析に利用されている．

3.1. 実験内容

・利用データ：次2つのデータを利用

(1) 担当者（1 人）の個人週報
［6 ヶ月間 22 回分 543 行］

(2) チームリーダ(1 人)のチーム週報
［6 ヶ月間 22 回 2007 行］

・分析手法：共起ネットワーク分析（Graph 理論に基づき，関係の深さを頂点と辺により表現するもの．共起パターンを使うことで語の繋がりを判断する）及び，文書検索機能

・実験手順：

(1) 個人週報を入力とし，共起ネットワーク分析を実施
(2) チーム週報を入力とし，共起ネットワーク分析を実施
(3) 個人週報の共起ネットワーク図からヒアリングに利用したいキーワードを管理者が視覚的に選定
(4) 選定したキーワードが，チーム週報の共起ネットワーク図に抽出されているかを確認

・KH coder の設定：

最低出現回数の設定＝ (1)チーム週報の時 3 回及び 2 回 (2)個人週報の時 3 回
（最低出現回数の設定値が高すぎると汎用的な語しか抽出できず，低すぎると抽出される語が多すぎ特徴をつかめない）

3.2. 実験結果

共起ネットワーク分析により出現頻度が高い語として抽出された語数は表 1 のようになった．また，個人週報の共起ネットワーク分析結果は 68 個がグループ化され，図 2 のように表示される．この図は出現頻度が高い語と，それらの語間の繋がりの強さを示している．

表1：共起ネットワーク分析での出現語数

	最低出現回数設定	グループ数	語数
チーム週報	3	２１	１１６
	2	３９	３７１
個人週報	3	１５	６８

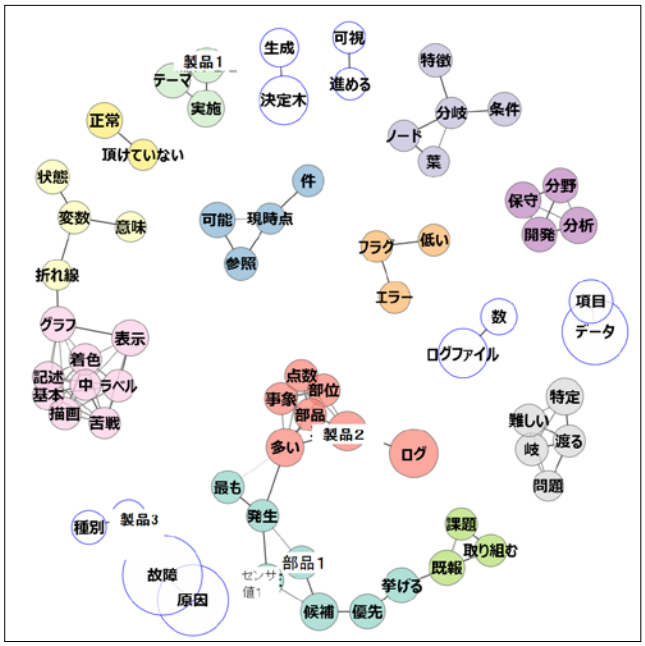

図 2：個人週報の共起ネットワーク分析

この共起ネットワーク図から個人週報のヒアリングに利用可能なキーワード（センテンス）として視覚的判断したものは以下の 3 件である．

(1)「（まだ）正常な（データを）頂けていない」
(2)「ラベル描画に苦戦中」
(3)「問題が多岐にわたり特定が難しい」

また，「故障」「原因」という語も頻出しているが，今回．

これらの語を含むセンテンスはヒアリングのキーワードとして選ばなかった．その理由は，この個人週報のテーマの一つが「故障原因の候補を早く見つける手法」であり，テーマ名として，この2語が頻出しているためである．

次にチーム週報を同様に分析した結果，最低出現回数設定＝3 回のとき，個人週報で抽出したキーワードのうち(1)のみ存在し，(2)と(3)は表示されなかった．(表 2)

さらにチーム週報の最低出現回数設定＝2 回の場合，キーワードは出現したが，共起ネットワーク図に表れる語数が371となり，図としても円が重なり判別が非常に困難であった．これはチーム週報においては汎用的な技術用語が多くなり，リスクに繋がるキーワード出現数が相対的に下がるためと考えられる．

表 2 個人週報とチーム週報でのキーワード出現

管理者指摘のキーワード	個人週報	チーム週報 [設定 3 回]	チーム週報 [設定 2 回]
正常－頂けていない	○有り	○有り	○有り
ラベル－描画－苦戦中	○有り	×無し	○有り
問題－多岐－特定－難しい	○有り	×無し	○有り
判別可能な図か？	YES	YES	NO

このように，担当者の個人週報に対する共起ネットワーク分析は特徴的な語を抽出することができ，チーム週報の共起ネットワーク分析よりも効果的であることを確認できた．

3.3. 追加実験（テキスト目視との比較）

共起ネットワーク分析を利用せず，週報全体を全部読んだ場合との比較はどうなるかを，追加実験として行った．

対象データは(1)個人週報の共起ネットワーク図 6 枚，(2)チーム週報（リーダによる集約版），(3)個人週報 6 人分である．これらの所要時間とキーワード出現の有無を確認した（表 3）．共起ネットワーク図は事前準備に時間を取られるが，作成後の俯瞰は短時間ですむ（3.4 の文書検索による詳細確認を除く）．個人週報全員分の読み込みは時間がかかるだけでなく，キーワードが記載されていても読み落とすことも発生した．チーム週報全体では集約により，共起ネットワーク図で見つけたキーワードは含まれていなかった．その意味でチーム週報全体を閲覧することよりも，共起ネットワーク分析の利用は効果的であった．

表 3 追加実験（テキスト目視確認との比較）

対象のデータ	準備時間	確認時間	所定キーワードは出現したか
個人週報の共起ネットワーク図×6 枚（6 人分）	35 分	2 分	○出現
チーム週報全体テキスト（2007 行）	0 分	40 分	×非出現
個人週報 6 人分テキスト（3696 行）	0 分	60 分	○出現

3.4. 適用上の留意点

この試行実験結果を適用する場合の留意点の一つは，そのキーワードに関するヒアリングを至急行うべきかどうかの判断を行う前に，次のような 5W1H の情報を把握する必要があることである．

(a) どの程度の頻度（回数）で出現しているのか
(b) いつの週報に出現しているのか（また，最後に出現した週報はいつか）
(c) どの文脈で（どのセンテンスの中で，どの意味で）出現しているのか

これにより例えば「まだデータを頂けていない」という状態について，それがヒアリングに値するかどうかを判断する．

この確認には KH coder の機能の一つである「文書検索機能」を利用した．抽出されたキーワードを選ぶと，それを含むコンテンツ（該当するセンテンスや行）を表示する機能である．それにより今回のケースでは「データを頂けていない」という状態は 4 週間続き，それが現在は収束していることが分かった．

4. 部門全員の週報の分析（実験 2）

4.1. 実験内容

・実験データ:
(1) 部門全員(4 チーム計 34 人)の個人週報（マージ版）
［1 回分，1230 行］
(2) 各チームのチーム週報（4 チーム分）
［1 回分，平均 52 行／チーム］
・分析手法：対応分析，特徴度抽出（Jaccard 類似性）
・実験手順:

(1) 個人週報(マージ版)を入力とし対応分析を実施
(2) 分析出力の分布図から，ヒアリングに利用したいキーワードを管理者が視覚的に選定
(3) 各チームのチーム週報に，選定したキーワードが含まれるかを検索で確認
(4) チームの特徴語を Jaccard 類似性測度で抽出し，チーム週報に含まれるかを確認
(5) 分布図と上位出現リスト表(30 語)の比較

・KH coder の設定： 最低出現回数の設定＝18 回

4.2. 実験結果

部門全体の週報から 98 個のキーワードが取り出され，4 チームの特徴が分布図で視覚化された(図 3)

この図の中心にある語は特徴の少ない語であり，横軸の左右または縦軸の上下に寄った語ほど，特定のチームを特徴付ける語となる．

例えば「実装，機能，モジュール」などの語はどのチームの週報にも表れ，他の語との関連性も弱いことが分かる．また，左上のチーム 1 を特徴づける語は「環境，ビルド」などであることが分かる．このように特定のチームに出現する語の例は表 4 のようになる．

表 4 抽出された語が含まれる分布図のチームエリア

語	チーム 1	チーム 2	チーム 3	チーム 4
1.実装	○	○	○	○
2.機能	○	○	○	○
3.モジュール	○	○	○	○
4.環境	◎			
5.ボード	◎			
6.設備		◎		
7.アダプタ		◎		
8.特許			◎	
9.閾値			◎	
10.ヒアリング				◎
11.バックアップ				◎

(○:全チームで出現，◎:特定チームで出現)

この分布図を俯瞰し，ヒアリングに利用したいキーワードとしては以下を選定した． チーム 1:「ライセンス」， チーム 2:「用語 3」， チーム 3:「特許」， チーム 4:「ヒアリング」．

キーワードとして選定したのは次のような理由である，チーム 1 の「ライセンス」は，ライセンス契約で気になることがあるのか？の確認を行いたい．チーム 2 の技術用語「用語 3」はこれまで報告のなかった用語でその発生元を含め確認したい．チーム 3 の「特許」は特許出願を準備中か他社特許の調査を行っていると推測できるが，特許関連で何を行っているのか？問題はないか？などを確実にフォローしたい．またチーム 4 では「ヒアリング」というキーワードがある．もしヒアリングが既に終わっている予定であれば，担当の今週の週報にあがってくるのは，まだ何かが続いている可能性があり要注意である．あるいは単に過去のヒアリング結果に対する分析なのか，などを確認する必要がある．

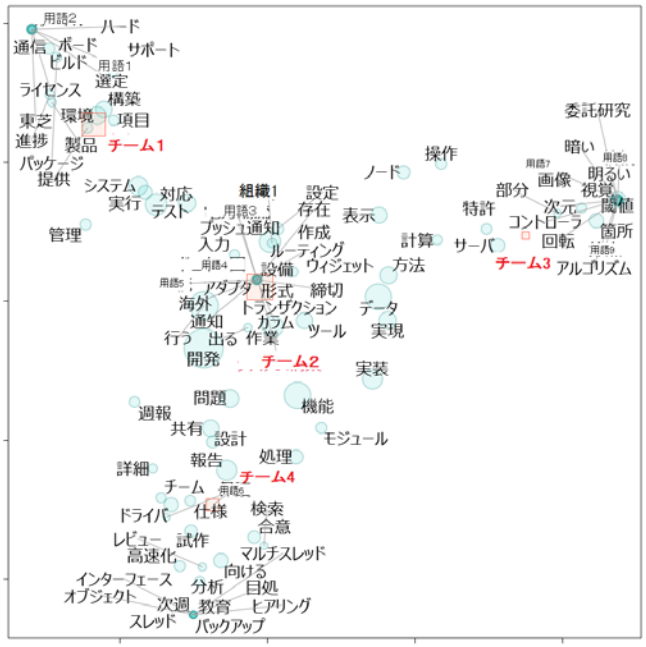

図 3 部門全体の週報(個人週報全員分)の対応分析

図 3 を確認して選定した 4 つのキーワードが，各チームのチーム週報に記載されているかどうかも確認した，その結果は次の通りである．

チーム 1 週報:「ライセンス」・・×(なし)
チーム 2 週報:「用語 3」・・・・・○(あり)
チーム 3 週報:「特許」・・・・・・×(なし)
チーム 4 週報:「ヒアリング」・・○(あり)

また，各チームを特徴づける語として，「Jaccard の類似性測度」(以下測度)を計算し各チーム 6 語をリストアップした．これはデータ全体に比べ，それぞれのチーム週報において特に高い確率で出現している度合いを示す数値である．更にその特徴的な語が各チームの週報内に出現するかを確認し，ヒット率を計算した(表 5)．

（ヒット率＝ヒットした語の測度合計／6 語の測度合計）

表 5 各チームの特徴語

	特徴語	測度	出現有無	ヒット率
チーム1	環境	0.072	○	
	構築	0.059	×	
	システム	0.047	○	
	ビルド	0.042	×	
	ボード	0.028	○	
	パッケージ	0.028	×	
	チーム1ヒット率			53.3%
チーム2	ツール	0.029	×	
	通知	0.026	○	
	設備	0.019	×	
	ウイジェット	0.019	×	
	アダプタ	0.017	×	
	用語4	0.017	×	
	チーム2ヒット率			20.5%
チーム3	視覚	0.088	×	
	閾値	0.044	○	
	画像	0.027	×	
	明るい	0.035	×	
	次元	0.031	×	
	特許	0.023	×	
	チーム3ヒット率			17.7%
チーム4	報告	0.059	×	
	ドライバ	0.035	×	
	試作	0.03	○	
	高速化	0.028	○	
	仕様	0.02	○	
	バックアップ	0.018	○	
	チーム4ヒット率			50.5%

（出現有無 ○:有り,×:無し）

このように，この部門のチーム週報はトピックスの一部のみを重点的に記載する傾向があるためか，特徴語は17.7%~53.3%しか含まれていなかった．

4.3. その他の確認結果

上位出現リストとの比較は，個人週報（全員分）の上位 30 語の出現リストの確認により行った（表 6）．リストで 84 回出現し 12 位に入っている「組織 1（固有名詞）」は部門としては関わりが深い組織となるが，分布図（図 2）で確認するとチーム 2 のみが関わっている組織であることが分かる．このように単純な出現リストより，対応分析の分布図の方が状況をより適切に示していると言う意味で，有効であると考えられる．

チーム間での調整などに関する新たな事実に気づくことが可能か？に関しては，同じ問題に遭遇しているケースや似たものを無駄に二重開発しているケースを想定した．

しかし今回，部門間での調整などが必要と思われる情報は特に抽出できなかった．これは当該部門の各チームの業務や特徴が大きく異なり，似た課題に遭遇するなどの状況が表れにくいことも影響していると考えられる．

表 6 実験 2 の抽出語と出現回数 上位 30 語

抽出語	出現回数	抽出語	出現回数
開発	207	対応	69
画像	153	確認	65
作業	144	組織 2	65
アルゴリズム	135	技術 1	64
部品	124	技術 2	63
ファイル	110	担当	61
データ	105	処理	60
作成	98	必要	58
ログ	95	実施	57
検出	93	最適	56
案件	88	検討	56
故障原因	84	適用	50
行う	84	報告	48
組織名1	84	予定	48
機能	73	可能	47

5. 課題と工夫点

5.1. 一般的なテキストマイニングでの課題と対応

本手法での課題とは別に，一般的なテキストマイニングでの前処理なども実際の運用では必要である．

（1）データクレンジング

テキストデータに対して以下を実施しておくことが分析精度に大きな影響を与えるので不可欠である．

a)誤記，誤字の修正

b)機種依存文字の削除，変更

c)半角，全角の統一

d)表記の揺れの修正

今回の試行では分析対象のテキストデータが週報のため，a)誤記・誤字 や b)機種依存文字 は殆ど無く，対応がほぼ不要であった．しかし c)半角・全角 および d)

表記の揺れは多く出現しており表現の統一を行った.

(2) 強制抽出(専門用語の指定)

自然言語処理の形態素解析では, 専門用語が分割されて認識されることがある. 例えば「ログファイル」が「ログ」と「ファイル」, 「ハードディスク」が「ハード」と「ディスク」などである. 用語を正しく抽出するために, 分割されてしまう専門用語を事前にピックアップし, KH coder の「強制抽出」設定を行った.

これにより本来意識したい用語を適切に抽出し, 分析の精度をあげることができる.

指定用語としては「プラットフォーム」「通信インターフェース」「ソースコード」などのソフトウェア関連の用語の他にも, プロジェクト内で使われる製品ドメイン固有の技術用語も指定した.

5.2. 本手法での課題と対策

先の 5.1(2)の強制抽出のための専門用語の指定は, 一般に抽出結果を見て専門用語の分割に気づき, 用語を指定する手順の繰り返しとなる. そのためソフトウェアの研究開発のように専門用語が多い週報では, 指定すべき用語を試行錯誤で決めるオーバヘッドが非常に大きい.

そこでソフトウェアの研究開発に関連する用語を Web 上の IT 用語辞典(e-Words[5]など)からピックアップして, 強制抽出したい用語の指定に利用した.

これらの IT 用語辞典では, 製品ドメイン固有の技術用語はカバーされないが, 一般的なソフトウェア関係の技術用語はほぼカバーされている. 語数は, プログラミング用語が 1068 件, ソフトウェア用語が 912 件, セキュリティ用語が 649 件などである. (e-Words の場合)

また, 分野をより絞ったカテゴリ, 例えば「ソフトウェアテスト」114 件などもあり, 対象とするチームの特性を考慮してカテゴリを選んだ.

これにより, 当初は強制抽出する専門用語の指定に 50 分程度かかっていたものが, 10 分程度に短縮することができた.

6. 評価と考察

6.1. 個人週報の分析(実験 1)

個人週報の共起ネットワーク分析では, データとしてある程度のボリューム(何週間分か)が必要であるが. 次のような効果を期待できる.

1.毎週続けて記載した(訴えた)情報はヒアリング用キーワードとして抽出が可能, 2.リスク分析には特に文意が重要であるが, KH coder 内のリンク機能で個別週報に戻らずに直ぐに内容を確認できる.

実験 1 の結果に関するアンケート(対象数 36 人)では以下のような評価を得た.

(1)使えそう／使ってみたい	14%
(2)(改善すれば)ある程度使えそう	59%
(3)あまり使えない	19%
(4)まったく使えない	08%

6.2. 部門全員の週報の分析(実験 2)

チームごとの特徴的キーワード抽出では, 部門の人数が多く, 生の週報を全部確認できない場合に利用すると次のような効果を期待できる. 1.確認やフォローの切り口を容易に見つけられる, 2.チーム固有の活動状況や課題を見つけられる(事例では特許, ヒアリング)

実験 2 の結果に関するアンケート(対象数 36 人)では以下のような評価を得た.

(1) 使えそう／使ってみたい	11%
(2)(改善すれば)ある程度使えそう	50%
(3)あまり使えない	28%
(4)まったく使えない	11%

6.3. 評価アンケートコメント

改善を前提に使えそう／ある程度使えそうという回答(上記の(1)と(2))が合計 61%～73%と, 一定の評価は得られた.

しかし「まったく使えない」という強いネガティブ評価も 8%～11%あり, 活用にはまだまだ改善が必要と思われる.

なお代表的なコメントは, 以下のようなものである.

[ポジティブコメント]

(1)週報によっては, 明るいニュースを優先し, 現在も残っている過去のリスクに触れないことがある. そのようなケースでもプロジェクトの不安から自身の週報に記載している個人週報がある. そこから見つけることを期待したい. (2)部門が 20 人, 30 人となった時には時間的制約から, チームリーダの週報のみを見ていたが, 全員の週報確認であってもこの分布図のボリュームなら使ってみたい.

[ネガティブコメント]

(1)クレンジングなどのオーバヘッド削減の見込みがないと, 使いにくい.

全体として. 特に大量の週報に対する俯瞰的分析が

期待をされている．しかしクレンジング作業のオーバヘッドがまだ大きく準備に時間を取られるため，その改善が利用の前提となる．

6.4. 考察

リスクに繋がるキーワード抽出以外の活用として，チーム間での情報交換や部門長と担当者のコミュニケーションにも効果があると思われる．（他チームの膨大な週報を読むことはあまり行われないため）

また，ソフトウェアの研究開発におけるリスク管理では進捗管理がよく利用されるが，進捗遅れが出る前にその原因となるリスクを見つけることが本来望ましい．本手法では「遅れた」と明記していない段階で気になるキーワードを見つけてヒアリングに繋げることが可能である．これがソフトウェアの研究開発において，週報のテキストマイニングを利用するメリットの一つである．

7. 課題と改善方針

7.1. 課題

今回の実験により，以下のような課題もリストアップされた．

1) 判断に関して
   ・肯定か否定か（成功か失敗）の判断には，動詞を含む詳細な分析が必要
   ・実験1では過去 6 ヶ月分の週報を用いたが，この分析でリスクに気づいた時には既に手遅れとなっている可能性がある．
2) 準備や整備に関して
   ・同じ用語で表現が異なる（例：インターフェース，インタフェース，Interfaceなど）と抽出に誤差が出るので，週報作成時に表現を統一することが望ましい
   ・専門用語が形態素解析で分割されることがないように，強制抽出設定を行う（5.1(2)）など，分析前のデータ整備の工数削減が必要
3) 運用に関して
   ・リスクを隠したいメンバーは該当するキーワードをNGワードとして利用しなくなる可能性もある

7.2. 改善方針

1) キーワード選定のためのデータセット準備

今回，分布図や対応図からキーワードを抽出する方法は，管理者が経験から気になるキーワードを選定する方法であった．しかし事前に「ソフトウェア研究開発の管理において注意すべきワード」をデータセットで用意しておけば，それを利用した抽出が可能になる．特にソフトウェアの研究開発管理という視点での抽出が実現する．

これはまた，管理者の経験に依存せず（異なる領域部署の管理者として門外漢が新たに着任した場合など）リスクに繋がるキーワードを容易にピックアップすることが可能になると期待される．

2) 強制抽出用の専門用語のデータセット準備

今回，強制抽出のための専門用語の指定では，IT 用語辞典を利用する工夫を行った．かなりの工数削減を実現できたが，上記同様にデータセットとして準備することで，精度を更に向上できる．すなわち研究開発対象のドメインの技術用語や製品用語を知識継承的にストックしておき利用する．

3) 語の出現回数の最適な設定

KH coder では語の抽出に関して「出現数」と「文書数」という設定がある．「出現数」は語がデータ全体を通して出現した回数を示し，「文書数」は語がいくつの文書中に出現したかを示す．出現数が極端に少ない語はキーワードとも言いがたく，また統計的分析に利用しにくい．そこで最小出現数を設定し，それ以下の語がノイズとして出現しないようにする．本実験 1 では 2 回 or3 回，実験 2 では 18 回の指定である．また，同じ出現数でも，文書数が多い語はどのコンテキスト（文脈）でも現れる語（例えば「ソフトウェア」）と見なせる．これに対しては，満遍なく出現する「キーワードとは言えない一般的な語」を除外するために，最大文書数を設定し出現数を制限する．これらの設定値は試行錯誤で行うことが多くオーバヘッドが大きくなる．

そこで，語の出現回数の最適な設定値を，部門の人数や週報のボリューム（総行数）から，デフォルト値として決めたい．つまり，部門 X（人数 n 人）で，その週報の記載の仕方（トレンド）から，最大・最小出現数の最適値を割り出しておく．

4) 週報表現の統一方法の工夫

週報の記載に際して，そのフォームや，専門用語（テクニカルターム）の書き方を統一することで，より精度の向上とオーバヘッド時間の短縮が期待できる．

8. 終わりに

8.1. まとめ

テキストマイニングはアンケート結果やコールセンターのクレーム内容などの分析に利用が進んでいる．またソ

フトウェアの研究・開発への適用としては LDA（Latent Dirichlet Allocation：潜在的ディレクトリ配分法）を利用したバグローカライゼーション分析などがある[6].

今回ソフトウェアの研究開発の報告書の一つである週報に対して，KH coder を使うテキストマイニングを用いて状況把握，リスク発見に繋がるキーワード抽出を行うことを試みた．頻出リストよりも見やすくセンテンスの特徴を表現する共起ネットワーク図や対応分析図がキーワード抽出に効果的であることを確認した．またアンケートでも条件付きながら 61%～73%の評価を得られた．

部門全員の週報分析（実験 2）では，6.評価と考察で示した効果の他に，今回は検証にいたらなかったが次の三点も期待される．

1) チーム間での不整合や対立の発見
2) 共通に出現する表現から，複数のチームでの二重開発やツールの二重購入などのムダの発見
3) 複数チームでの共通の課題（品質，コスト，納期など）の早期発見と横断的な解決

勿論，正確な情報分析を行うには，直接担当者に口頭やメールで確認するプロセスも必要である．

その問い合わせや確認のための「切り口」「キーワード」をテキストマイニングからスピーディーに得られれば，より担当者への確認が適切なものになる．

今後は，ソフトウェアの研究開発固有のデータセットの収集や，用語表現の統一などのオーバヘッド削減に関して引き続き検討し，ソフトウェア研究開発における他のドキュメントへの適用なども検証していきたい．

参考文献

[1] PMBOK®ガイド第 5 版
[2] 樋口耕一, KH Coder <http://khc.sourceforge.net/>
[3] 樋口耕一，社会調査のための計量テキスト分析，ナカニシヤ出版
[4] NAIST，形態素解析システム茶筅 http://chasen.naist.jp/hiki/ChaSen/
[5] IT 用語辞典 e-Words http://e-words.jp/
[6] S. K. Lukins, N. A. Kraft, and L. H. Etzkorn, “Bug localization using Latent Dirichlet Allocation,”Information and Software Technology vol. 52, no. 9, pp. 972–990, 2010.