インシデントレポートの自動分類とその分析環境の構築

Categorization of Incident Reports and Construction of Analysis Environment

稗方和夫[1]　大和裕幸[1]　中村覚[1]　岡田伊策[2]　齋藤稔[2]　安藤峻[3]

Kazuo HIEKATA[1], Hiroyuki YAMATO[1], Satoru NAKAMURA[1],
Isaac OKADA[2], Minoru SAITO[2], and Takashi ANDO[3]

[1] 東京大学大学院新領域創成科学研究科
[1]Graduate School of Frontier Sciences, THE UNIVERSITY OF TOKYO.
[2] 富士通株式会社 SI 技術サポート本部
[2]SYSTEM INTEGRATION TECHNOLOGY SUPPORT UNIT, FUJITSU LIMITED.
[3] 株式会社ウニークス
[3] UNICUS Co., Ltd.

アブストラクト: 情報システム企業では情報システム製品の顧客環境での運用時のインシデント情報を収集している。本研究は、新しいインシデントを効率的に解決するために既存のインシデントレポートの知識を再利用することを目的とする。分類対象に依存したテキスト処理以外の共通部分をプラットホーム化することで横展開可能な文書自動分類プログラムを開発し、特定の文書群について自動分類を行い、その有用性の検証を行う。

1 はじめに

1.1 背景

情報システム企業 A 社では情報システム製品の顧客環境での運用時等のインシデント情報をインシデントレポートとして蓄積し、新規のインシデントが発生した際に過去の類似インシデントに関するインシデントレポートを参照することでインシデントの解決に役立てている。

現在の問題点として、既存レポートを参照する際には全文検索が用いられているが、蓄積された膨大な数のレポートを十分に絞り込めないケースが存在する。そこで先述した知識抽出という観点から、レポートをカテゴリ毎に分類することで検索能力を向上させ、目的とするレポートへのアクセスを容易にする試みがなされている。しかし現在は人手でカテゴリが付与されており、コストや労力がかかる上、各担当者の経験の差異や主観によって客観的な分類が難しい等の問題がある。

本研究ではインシデントレポートに対して自然言語処理技術、機械学習技術を用いてカテゴリの自動付与を行う。テキストデータからの知識獲得の試みは数多く行われている[1][2][3]。しかし従来の自然言語処理の研究対象は新聞記事や論文、特許文書等の十分に推敲された文書を扱っている。一方、本研究で対象とするインシデントレポートは企業・組織内でのみ参照される情報のため、不特定多数の読者を想定しておらず、結果として誤字・脱字・省略等を含む点が特徴として挙げられる。このような文章を対象とした関連研究として、那須川[4]による既存研究が挙げられる。那須川はコールセンターのテキストデータからの知識獲得を目的とし、テキストに出現する用語にカテゴリを与えて表記の統一化を図る意味辞書の作成や、文法的に係り受け関係にある名詞概念や述語概念とのペアを抽出することで、テキストデータからさまざまな情報を抽出している。

1.2 目的

本研究では蓄積されたインシデントレポートに対して自然言語処理技術、機械学習技術を用いてカテゴリを自動付与することを目的とする。

また分類対象に依存したテキスト処理以外の共通部分をプラットホーム化することで横展開可能な文書自動分類プログラムを提案する。さらに情報システム企業 A 社のデータに対して、本プログラム上でカテゴリの自動付与を行い、その有用性の検証を行う。

2 インシデントレポート

2.1 インシデントレポート

インシデントレポートの例を図 1 に示す。

ID:a1234-5678　**日付**:2013/1/1　**OS**: Windows　**製品**:製品 B

1. 質問概要
製品 A の起動シェルを置き換え後、OS 再起動を実施しましたが製品 B が起動しなくなりました。

2. 回答概要
D コマンドを実行していないのが原因です。マシンブート時の製品 B の自動起動設定がされているか確認してください。

3. ヒアリング
・OS 再起動の時間　・システム等の変更点はないか　・復旧（製品 A 起動）方法

4. 原因要約
D コマンドを実行していないのが原因です。

5. 処理要約
マシンブート時の製品 B の自動起動設定がされているか確認してください。

6. 参考情報
・マニュアル　製品 B 運用ガイド
付録 C 製品 B 統合コマンドによる運用操作
＞C.3　製品 B の起動
C.3.5　マシンブート時の製品 B の自動起動

・過去事例　a3456-9876

図 1: インシデントレポート例.

情報システム企業 A 社では情報システム製品の顧客環境での運用時のインシデント情報をインシデントレポートとして蓄積している。「OS」や「製品名」等の項目は選択形式になっているが、他の「質問概要」、「回答概要」、「参考情報」等の項目は自由記述形式になっている。

なお、本研究で扱うインシデントレポートは XML 形式で記述されており、上記の各項目が構造化されている。

2.2 既存インシデントレポートの検索方式

新規のインシデントが発生した際に過去の類似インシデントに関するインシデントレポートを参照することでインシデントの解決に役立てている。具体的には新規のインシデントで発生した際、その事象から検索キーワードを想起し、既存のインシデントレポート群に対して全文検索を行う。得られた検索結果について「質問概要」項目に、現在のインシデント事象と類似した事象が記述されているインシデントレポートを選択し、「回答概要」や「参考情報」といった項目を参照しながら新規のインシデントへの対応を行う。しかし現在の検索方式である全文検索では検索結果が十分に絞り込めず、目的とするインシデントレポートを得るのに多大な時間を要するといったケースが見られる。

2.3 分類カテゴリ

そこで情報システム企業 A 社はインシデントレポートをカテゴリ毎に分類し、検索能力を向上させることによって目的とするレポートへのアクセスを容易にする試みを行っている。カテゴリは「検索方式」-「製品」-「インシデントの症状」-「インシデントの発生箇所」に代表される階層構造を持つ。この例を図 2 に示す。

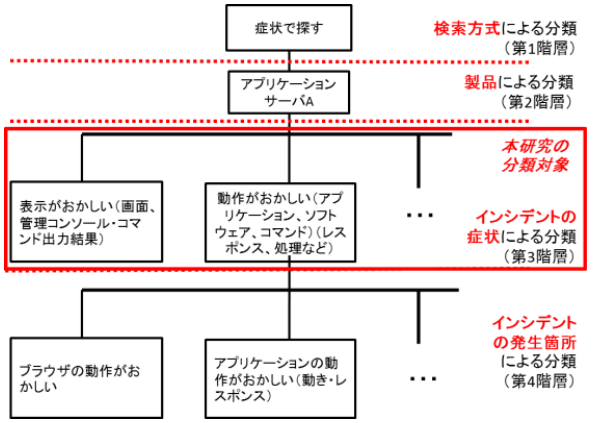

図 2: 階層構造を持つカテゴリ例.

本研究では情報システム企業 A 社の製品の一つであるアプリケーションサーバ A に関するインシデントレポートについて、その症状で分類することを目的とする。製品 A に関するインシデントレポートは図 3 に示す 11 カテゴリに分類される。

図 3: インシデントの症状に関する分類.

なお本研究では「カテゴリに分類する」ことを、「インシデントの症状」というメタデータフィールドに対して、そのメタデータバリューを図 3 のカテゴリから重複を許して付与する問題へと帰着させ、以下「カテゴリ付与」と表現し論を進める。

3 分析プラットホーム「Kashiwade」

3.1 概要

本研究では分類対象に依存したテキスト処理以外の共通部分をプラットホームとして提供する「Kashiwade」（以下、Kashiwade）を開発した。その

概要を図 4 に示す。このプラットホームは Web ブラウザから使用する Web アプリケーションとして構築されている。次節より Kashiwade が保有する各機能について説明する。

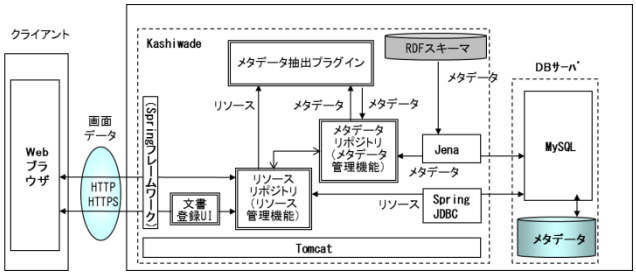

図 4: Kashiwade 概要.

3.2 リソース管理機能

リソースは文書登録 UI を通じて Kashiwade に登録される。リソースが持つ情報は MySQL によって構築されたリソースリポジトリとメタデータリポジトリに保存される。リソースリポジトリには文書名やバイナリデータ、URI が保存され、これらは基本的に変更されない。一方、メタデータリポジトリにはメタデータとして編集可能な文書名やグループ名、URI が保存される。これらリポジトリは URI によって結合される。

登録されたリソースは図 5 に示すリソース一覧画面に表示される。本画面ではメタデータ検索によるリソースの絞り込みや、各リソースのダウンロードや更新、削除を行うことができる。

図 5: リソース一覧画面.

3.3 メタデータ管理機能

特定のリソースについてのメタデータフィールドとメタデータバリューは、図 6 に示すメタデータ一覧画面に表示され、メタデータバリューの更新を行うことができる。

図 6: メタデーター覧画面.

3.4 メタデータ抽出プラグイン管理機能

本プラットホームはプラグインを読み込み・実行する機能を備えている。プラグインはプラットホームに保存された文書名やバイナリデータといったリソース情報と、リソースに付与されたメタデータ（フィールドとバリューのセット）を読み込み、既存メタデータフィールドのバリューの更新や新規メタデータの追加を行う。

プラグインは図 7 に示す管理画面に一覧表示され、選択したメタデータフィールドのバリュー別、リソースの拡張子別にプラグインを実行できる。

図 7: プラグイン管理画面.

4 カテゴリ自動付与プログラム

4.1 カテゴリ自動付与プログラム

本研究で提案するカテゴリ自動付与プログラムについて説明する。本提案手法は以下の 3 プロセスで構成される。

① Kashiwade を利用した分類対象に依存しないテ

キスト処理に関する前処理
② インシデントレポートの特徴に基づいた特徴ベクトルの作成
③ Support Vector Machines を用いたカテゴリ付与

4.2 Kashiwade を利用した前処理

分類対象に依存しないテキスト処理について、Kashiwade プラットホームを利用して前処理を行う。その流れを図 8 に示す。

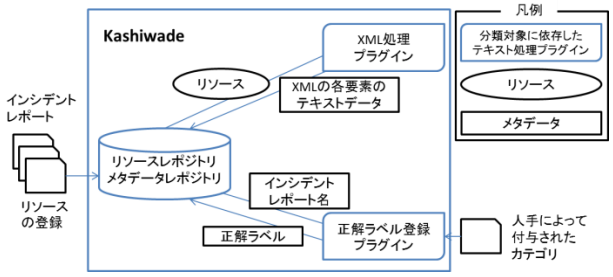

図 8: Kahiwade を利用した前処理.

まずインシデントレポートが記述された XML ファイルを訓練データとして Kashiwade に登録することで、バイナリデータがリソースリポジトリに、レポート名がメタデータリポジトリに保存される。次に XML から各要素を抽出する XML 処理プログラムをプラグインとして実行することで、2.1 章で説明した各項目がメタデータフィールドとして、そこに記述されたテキストがメタデータバリューとしてメタデータリポジトリに保存される。さらに人手によって付与されたカテゴリ（以下、正解ラベル）をメタデータとして登録するプラグインを実行し保存する。この前処理によって次節以降で必要となるテキストデータや正解ラベルが Kashiwade に保存され、これらを入力データとして用いることで分類対象に依存したテキスト処理プラグインを実行できる。

4.3 特徴ベクトルの作成

インシデントレポートの自動付与を行うには、まずインシデントレポートに記述されたテキストから特徴ベクトルを作成する必要がある。ここで特徴ベクトルとはテキストについて形態素解析を行い、分割された形態素を特徴量として持つベクトルのことである。なお、形態素解析エンジンには MeCab [5] を使用した。

特徴量として使用する形態素を選定する際、平ら[6]の研究を参考とし、「一般名詞」「固有名詞」「未定義語」「サ変接続名詞」を品詞としての持つ形態素を抽出した。また那須川[4]の研究を参考とし、「ない」「ん」のような否定語も特徴量として抽出した。これは「A が停止する」「A が停止しない」のように否定語の有無によってインシデントの症状の意味内容が反転するためである。

さらに専門用語、表記揺れを含む語の扱いに関して工夫した。インシデントレポートのような業務に特化した文書には専門分野に特化した語や製品特有の語（以下、専門用語という表現に統一）が存在するが、通常の形態素解析でこれらを抽出することはできない。例えば「管理コンソール」という用語は「管理」と「コンソール」という二つの形態素に分割され、分析対象とするインシデントレポートに出現する特徴量として正確に捉えることができない。この問題に対して中川ら[7]は出現頻度と連接頻度に基づいて専門用語を自動抽出する研究を行っている。本研究でも複合名詞を専門用語の候補として抽出し、その中から専門用語として使用する語を形態素解析器の辞書に登録することで専門用語を特徴量として抽出した。誤字・脱字・省略によって同義語であるのに対して異なる特徴量として抽出される形態素に関しては、編集距離という概念を用いて表記揺れの吸収を行った。編集距離とは二つの文字列がどの程度異なっているかを示す数値であり、文字の挿入や削除、置換によって、一つの文字列を別の文字列に変形するのに必要な手順の最小回数として与えられる。これが閾値以下のものについては表記揺れとして吸収した。例えば「message」とその複数形である「messages」に関しては「s」の挿入による編集距離 1 、「message」とその誤字である「messege」は「a」の「e」への置換による編集距離 1 として計算される。

これらの処理によって作成した特徴ベクトルについて、さらに各インシデントレポートに出現する全特徴量数で除すことで正規化し、TF-IDF 法を用いて特徴量の重み付けを行う。

4.4 SVM を用いたカテゴリ付与

4.3 章で作成した特徴ベクトルに対して、Support Vector Machines (以下、SVM)を用いてカテゴリの付与を行う。SVM はテキスト分類において非常に高い分類能力を持つことが既存研究[8][9]で証明されており、また本研究で対象とする各インシデントレポートが複数のカテゴリを持ち得ることから二値分類器である SVM を利用する。訓練データに付与された各カテゴリについて、そのカテゴリを正解ラベルとして持つ特徴ベクトルと、それ以外のカテゴリを正解ラベルとして持つ特徴ベクトルによってニ値分類器を生成し、テストデータがそのカテゴリに含まれるか否かを判別する。

本研究では SVM について詳細な説明は省略するが、SVM のタイプとして式(1)を最大化するような係数α_iの集合を求める最適化を行い、マージンを最大

にする超平面を求めるソフトマージン SVM を使用する。ここでC_iは誤分類によるペナルティとマージンの大きさの間のトレードオフを制御するパラメータである。またカーネル関数として線形カーネルを使用した。なお、実装には LIBSVM(A Library for Support Vector Machines)[10]を用いた。

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i} \cdot \boldsymbol{x_j}$$

$$\text{s.t.} \sum_{i=1}^{l} \alpha_i y_i = 0, \quad \forall i\colon 0 \le \alpha_i \le C_i \qquad \text{式(1)}$$

5 実験

5.1 実験概要

本研究では情報システム企業 A 社のインシデントレポートについて、人手によって正解ラベルが付与されたレポート全 100 件を対象として実験を行った。なお、ここでは 2.3 章で挙げた 11 カテゴリの中から 100 件のインシデントレポートに付与された 10 カテゴリを付与対象とする。

評価は交叉検定による Precision と Recall によって行う。Precision とは提案手法によって付与されたカテゴリの中に正解ラベルが含まれる割合、Recall は正解ラベルの中に付与されたカテゴリが含まれる割合を示す指標である。訓練データを一件のテストデータと残りの訓練データに分割し、テストデータの正解ラベルと提案手法によって付与されたカテゴリを比較することを全訓練データに対して行い、それらの Precision と Recall の平均値を結果として出力する。なお、ソフトマージン SVM におけるパラメータC_iに関しては、$C_i = 1$を使用した。

5.2 実験結果

全インシデントレポートに関する Precision と Recall を表 1 の上部に示す。Recall は 6 割の精度を示すが、Precision は 4 割の精度を示し、各インシデントレポートに平均 2.8 個のカテゴリが付与される結果となった。

また各カテゴリに関する Recall を表 1 の下部に示す。なお、左端の列に示す番号は各カテゴリに与えられた識別番号である。カテゴリによって Recall にばらつきがあることがわかる。Recall の高い「404: 環境設定ができない」というカテゴリには、「証明書の環境設定ができない」という症状が記述されたレポートが多く含まれ、「証明書」という形態素が本カテゴリを特徴づけていたため、高い Recall を示している。一方「370:表示がおかしい」というカテゴリに関しては「表示」という表現は本カテゴリ以外のインシデントレポートにも多く含まれ、かつ「おかしい」という症状に多様な表現が含まれているため、本カテゴリを特徴づける形態素が存在せず、低い Recall を示す結果になっていると考えられる。

表 1: 実験結果.

Precision	Recall	カテゴリが付与されなかったレポート数
0.41	0.59	2

	付与されたカテゴリ数	正解ラベル数	Recall
370	2	7	28.57
375	37	58	63.79
383	8	16	50.00
387	21	27	77.78
393	5	9	55.56
400	6	13	46.15
404	5	5	100
405	0	1	0
406	1	3	33.33
707	25	31	80.65

6 考察

6.1 カテゴリ自動付与手法に関する考察

本研究ではインシデントレポートの症状に関するカテゴリの自動付与を行ったが、各インシデントレポートに出現する形態素の出現頻度を特徴量として扱う bag of words 手法では 6 割の Recall を示す結果となった。この分類精度の向上に向けて、今後はインシデントの発生箇所と症状の係り受け構造を抽出する構文解析の利用を考える。これは「エラーメッセージが表示され、インストールできない。」、「インストール後、エラーメッセージが表示された。」という二つの例文からわかるように、bag of words 手法ではインシデントレポートに記述された内容を正確に捉えることに限界があるためである。

またオントロジーを用いて用語の概念関係を把握することも分類精度の向上につながると考えられる。例えば各製品を「ハードウェア」と「ソフトウェア」に分類し、「ハードウェア」「ソフトウェア」それぞれと共起率の高い症状をオントロジーとして関連付けることによって、より正確にインシデントレポートに記述された内容を捉えることができると考えられる。また本研究ではインシデントレポートの全項目を一様に抽出して分析を行ったが、インシデントの症状が記述された「質問概要」とその原因と解決策が記述された「回答概要」の関係に着目し、インシデントの症状、原因、解決策の関係をオントロジーを用いて体系化、およびクラスタリングすることによって、分類精度の向上だけでなく、カテゴリを用いた検索を行うユーザの検索効率の向上に寄与で

きると考えられる。

6.2 分析プラットホームに関する考察

本研究では分類対象に依存したテキスト処理以外の共通部分をプラットホーム化した「Kashiwade」を提案した。本研究から得られた本プラットホームの利点を以下にまとめる。

① 単純なリソース・メタデータバリューの管理を行うプログラムは Kashiwade の機能を利用することができる。そのため研究テーマ毎に個別にリソースを管理する環境を開発する必要がなく、アルゴリズムの実装に集中することができる。

② 研究テーマの担当者が実行環境の管理を行う必要がないため、安定した実行環境を提供することができる。

③ リソースの登録、プラグインの実行などの操作はプラットホーム上で実装するため、異なるテーマに対するプログラムの場合も操作方法は統一される。

④ プラットホーム上に多数のプラグインを同時に実行することができるため、ある研究テーマで実装したプログラムが過去の研究テーマで作成したメタデータを再利用することができる。

実際に他の製品に関するレポートを登録した例を図 9 に示す。このように Kashiwade に登録されたレポートに関しては、本研究で提案したカテゴリ自動付与プログラムなどをプラグインとして適用することが可能で、研究者の開発の労力低減や既存プログラムの再利用の促進などの利点が挙げられる。

図 9: 異なる製品のレポートを登録した例.

7 結論

本研究では、インシデントレポートに対するカテゴリの自動付与を行った。専門用語や表記揺れ等の専門分野に特化したテキストに基づいた特徴量を用いて特徴ベクトルを作成し、SVM を用いたカテゴリの自動付与を行った結果、Precision が 4 割、Recall が 6 割という結果を示した。今後、構文解析やオントロジーを利用することでインシデントレポートから特徴量をより正確に抽出し、分類精度の向上を目指す。また本研究では分類対象に依存したテキスト処理以外の共通部分をプラットホーム化した「Kashiwade」を開発・提案し、分析に用いたプログラムをプラグインとして実行することで、分析における労力の低減などの有用性を示した。

謝辞

本研究を行うにあたり、多大なご指導をいただいた富士通株式会社ミドルウェア事業本部の方々に感謝いたします。

参考文献

[1] Hahn U. et al.: Deep Knowledge Discovery from Natural Language Texts., Proceedings of KDD-97, pp.175-178 (1997)

[2] Knight M.: Mining Online Text., Communications of the ACM, Vol.42, No.11, pp.58-61 (1999)

[3] Mladenic D.: Text-Learning and Related Intelligent Agent: A Survey, IEEE Intelligent Systems, Vol.14, No.4, pp.44-54 (1999)

[4] 那須川哲哉: コールセンターにおけるテキストマイニング, 人工知能学会誌, Vol. 16, No. 2, pp. 219-225, (2001)

[5] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, Availabled at <http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html> Accessed on: Feb 18th 2013

[6] 平博順, 春野雅彦: Support Vector Machine によるテキスト分類における属性選択, 情報処理学会論文誌, Vol. 41, No. 4, pp. 1113-1123, (2005)

[7] 中川裕志, 湯本紘彰, 森辰則: 出現頻度と連接頻度に基づく専門用語抽出, 自然言語処理, Vol. 10, No. 1, pp. 27-46, (2003)

[8] Dumas, S., Platt, J., Heckerman, D. and Sahami, M.: Inductive Learning Algorithms and Representations for Text Categorization, Proc., 7th International Conference for Information and Knowledge Management, (1998)

[9] Joachims, T.: Text Categorization with Support Vector Machines, Proc., 10th European Conference on Machine Learning (ECML), (1998)

[10] LIBSVM -- A Library for Support Vector Machines, Availabled at <http://www.csie.ntu.edu.tw/~cjlin/libsvm/> Accessed on: Feb 18th 2013