

---

一般口演

## 一般口演26

## 機械学習・HISデータ解析

2017年11月23日(木) 09:15 ～ 10:45 F会場 (10F 会議室1004-1005)

---

**[4-F-1-OP26-2] データマイニング技術を用いた電子カルテからのインシデント検出の試み**

五家 花梨<sup>1</sup>, 岡本 和也<sup>1,2</sup>, 廣瀬 昌博<sup>3</sup>, 平木 秀輔<sup>1,2</sup>, 山本 豪志朗<sup>2</sup>, 杉山 治<sup>4</sup>, 南部 雅幸<sup>4</sup>, 黒田 知宏<sup>1,2</sup> (1.京都大学 大学院情報学研究科, 2.京都大学医学部附属病院 医療情報企画部, 3.島根大学 医学部, 4.京都大学医学部附属病院 先制医療・生活習慣病研究センター)

一般的にインシデントは医療事故に繋がる前段階と定義されており、インシデントを把握して対応策を取ることで医療事故を未然に防ぐことができると考えられている。そのため、各病院ではインシデントの報告を義務化しているが、発生したインシデント全てが報告されるわけではないという問題があり、報告漏れのインシデントを発見する技術が求められている。そこで、本研究では、データマイニング技術を応用し、電子カルテからインシデントに関する記録を検出する手法を提案する。提案手法では、TF/IDF法を応用して電子カルテとインシデントレポートからインシデントの特徴を抽出し、その特徴を含む電子カルテを見つけることでインシデントの検出を行う。インシデントレポートは分析を目的として「転倒」「注射」などのカテゴリに分類されており、本提案手法ではインシデントの特徴を抽出するためにこのカテゴリを利用する。まず、カテゴリをTF/IDF法の処理単位とみなすことでインシデントと関連の強い単語とインシデントの各カテゴリに関連の強い単語を抽出する。続いて、電子カルテにおいてインシデントと関連の強い単語とインシデントの各カテゴリに関連の強い単語が共起するものを、各カテゴリのインシデント候補として検出する。本提案手法の妥当性を評価するために京都大学医学部附属病院の電子カルテとインシデントレポートを用いて実験を行った。評価実験では、電子カルテから抽出した各カテゴリのインシデント候補とインシデントレポートとを比較し、それぞれに紐づく患者とインシデントのカテゴリが一致する場合に同じインシデントを扱っているものとして評価した。結果として、「転倒」カテゴリのようにカテゴリに関連の強い単語がインシデントとの関連も強い傾向があるカテゴリでは、比較的電子カルテからインシデントの抽出が行いやすいことが明らかになった。

# データマイニング技術を用いた電子カルテからのインシデント検出の試み

五家花梨<sup>\*1</sup>、岡本和也<sup>\*1,2</sup>、廣瀬昌博<sup>\*3</sup>、平木秀輔<sup>\*1,2</sup>、山本豪志朗<sup>\*2</sup>、杉山治<sup>\*4</sup>、南部雅幸<sup>\*4</sup>、黒田知宏<sup>\*1,2</sup>

<sup>\*1</sup> 京都大学 大学院情報学研究科、<sup>\*2</sup> 京都大学医学部附属病院 医療情報企画部、

<sup>\*3</sup> 島根大学 医学部、<sup>\*4</sup> 京都大学医学部附属病院 先制医療・生活習慣病研究センター

## Attempt to Detect Incidents from Electronic Medical Records Using Data Mining Technology

K. Goka<sup>\*1</sup>, K. Okamoto<sup>\*1,2</sup>, M. Hirose<sup>\*3</sup>, S. Hiragi<sup>\*1,2</sup>, G. Yamamoto<sup>\*2</sup>, O. Sugiyama<sup>\*4</sup>, M. Nambu<sup>\*4</sup>, T. Kuroda<sup>\*1,2</sup>

<sup>\*1</sup> Graduate School of Informatics, Kyoto University,

<sup>\*2</sup> Division of Medical Information Technology and Administration Planning, Kyoto University Hospital,

<sup>\*3</sup> Faculty of Medicine, Shimane University,

<sup>\*4</sup> Preemptive Medicine & Lifestyle-Related Disease Research Center, Kyoto University Hospital

In this research, we proposed two methods to detect incidents from electronic medical records by applying data mining technology. One was a method based on the TF-IDF method and the other was based on the SVM method. In the TF-IDF based method, we exploited incident categories. First, by considering categories as processing units of the TF-IDF method, the important words relevant to the incidents of each category were extracted. Subsequently, electronic medical records including the words strongly related to each incident category were detected as incident candidates of each category. In the SVM based method, we learned medical reports related to incidents and medical reports having no relation to incidents by using a SVM with a linear kernel. In order to evaluate the validity of the proposed methods, we conducted experiments using incident reports and electronic medical records of Kyoto University Hospital. As a result, it was shown that it was relatively easy to extract incidents from electronic medical records in categories that tend to have strong association with incidents by using the TF/IDF based method. On the contrary, the SVM based method could extract incident candidates of category ‘injection’, which has little association with incidents directly, though the method needed the data to learn.

**Keywords:** incident, incident report, electronic medical record, TF-IDF, SVM.

### 1. はじめに

近年話題に上ることが多い医療事故に関連して、病院にはインシデントと呼ばれるものが存在している。一般的にインシデントは医療事故につながる前段階と定義されており、インシデントを把握して対応策をとることで医療事故を未然に防ぐことができると考えられている。そのため、一部の重要性が高いと考えられるインシデントは国が定める医療事故調査・支援センターへの報告が義務化されており<sup>1)</sup>、インシデントレポートを作成するためのシステムが多くの病院に導入されている。インシデントレポートは患者 ID、発生日、発生状況、インシデントを転倒や注射薬といった種類ごとに分類するカテゴリ、報告者の職種などの項目からなっている。

医療事故を防ぐ手段の一つであるインシデントレポートだが、医師の倫理観の欠如や安全に対する意識の低さに起因する報告の遅れが近年の研究で指摘されている<sup>2) 3)</sup>。インシデントの報告が遅れば、インシデントへの対応も遅れることとなり、医療事故が発生する可能性が高まる。さらに、必ずしも起こったすべてのインシデントが報告されるわけではないとも考えられており、インシデントの報告が無ければ、インシデントへの対応を取ることも困難になる。したがって、インシデントを検出する仕組みを構築することは有用である。

Li らは新生児集中治療室におけるインシデントの自動検出手法を提案した<sup>4)</sup>。しかし、Li らの検出手法は新生児集中治療室に関連する項目を用いたルールベースの手法であり、特定の場所や患者の属性に依っているため、他の場所や異なる患者属性に関するインシデントを発見することは難しいという問題がある。

本研究では、報告が漏れているインシデントを検出することを目指し、患者属性に依らない幅広いインシデントを抽出す

る手法の確立を目的とする。

### 2 提案手法

本研究では、2つの手法を用いてインシデントの検出を試みる。

1つ目は、TF-IDF 法を用いてインシデントレポートからインシデントを特徴づけるキーワードを抽出し、抽出したキーワードを利用して電子カルテからインシデントの検出を試みる手法である。本研究では、抽出するインシデントを特徴付けるキーワードをインシデント・キーワードと定義する。

2つ目は、機械学習を用いて電子カルテを学習し、インシデントを含む電子カルテを分類することによりインシデントを検出する手法である。

#### 2.1 TF-IDF 法を用いたインシデント検出手法

TF-IDF 法を用いた手法では、まず、インシデントレポートの文章を形態素解析により単語に分割する。その後、TF-IDF 法を利用して単語の重要度を計算し、インシデント・キーワードとして抽出する。続いて、インシデント・キーワードを用いて電子カルテを検索することで、インシデントの検出を試みる。

##### 2.1.1 インシデント・キーワードの抽出

形態素解析によりインシデントレポートの1項目である発生状況を単語に分割する。形態素解析とは単語の区切りを判定する作業である<sup>5)</sup>。なお、本提案手法で使用する品詞は、名詞、動詞、形容詞、副詞である。

続いて、TF-IDF 法によりインシデント・キーワードを抽出する。TF-IDF 法は、1つの文書に対する単語の頻度に応じた重みを表す TF (Term Frequency) 値と、文書間での出現の偏りに応じて重みをつける IDF (Inverse Document Frequency) 値を算出し、それらの値を掛け合わせることで単語の重要度を

表す TF-IDF 値を得る手法である<sup>6)</sup>。TF-IDF 値は以下の式で求めることができる。

$$tf_t^d = \frac{tf(t, d)}{\sum_{s \in d} tf(s, d)}$$
$$idf_t = \log_e \frac{N}{df(t)} + 1 \quad (1)$$
$$w_t^d = tf_t^d \cdot idf_t$$

ここで、 $tf_t^d$ は単語  $t$  の文書  $d$  に対する TF 値、 $idf_t$ は単語  $t$  の IDF 値、 $w_t^d$ は単語  $t$  の文書  $d$  に対する TF-IDF 値である。そして、 $tf_t^d$ に関する式で用いられている  $tf(t, d)$ は単語  $t$  の文書  $d$  内での出現回数を、 $\sum_{s \in d} tf(s, d)$ は文書  $d$  内でのすべての単語の出現回数の和を示している。また、 $idf_t$ に関する式で用いられている  $N$ は全文書数を、 $df(t)$ はある単語  $t$  が出現する文書の数を示している。

本研究では各カテゴリを 1 つの文書として取り扱うことで、各カテゴリに対する各単語の TF-IDF 値を算出し、TF-IDF 値の大きい単語をインシデント・キーワードとする。

しかし、各カテゴリを1つの文書として取り扱った場合、1文書あたりの単語量が多くなることで IDF の影響が弱くなり、求めるべき結果が得られないことが考えられる。したがって、(1)式の IDF 値を求める式の一部を以下のように変更し、一般化する。

$$idf_t = \log_y \frac{N}{df(t)} + x \quad (2)$$
$$x = \begin{cases} 1 \\ -\log_y 2 \end{cases}, y = \begin{cases} e \\ 2 \\ 1.5 \end{cases}$$

(2)式への変更は算出される TF-IDF 値に対する IDF 値の影響を大きくすることが目的である。 $x$  として  $-\log_y 2$ を用いたのは半分以上のカテゴリで出現する単語をわかりやすく差別化しようと考えたためであり、 $y$  で底の調整を行っているのは、底が小さくなると単語の IDF 値の差が大きくなるためである。

## 2.1.2 インシデント・キーワードによる電子カルテの検索

前項で抽出されたインシデント・キーワードを用いて、電子カルテからインシデント・キーワードを含む診療記録を検索し、検索された診療記録をインシデントの内容を含むものとして検出する。

## 2.2 機械学習を用いたインシデント検出手法

機械学習を用いた手法では、機械学習手法としてサポートベクターマシンを利用する。サポートベクターマシンは過学習を防ぐための理論的に正当化できるアプローチである点で、他の機械学習手法にない重要な優位性を持っている<sup>7)</sup>。

本提案手法では、電子カルテにおいてインシデントに関する記載を含む診療記録を正例とし、含まないものを負例とし、教師データを作成する。

続いて、教師データの各文書を形態素解析によって単語に分割し、単語を要素とした文書ベクトルを作成する。その文書ベクトルをサポートベクターマシンによって学習して学習モデルを作成する。

最後に、作成した学習モデルによって、形態素解析を用いて文書ベクトルとした診療記録がインシデントに関する記載を含むか判定し、電子カルテからインシデントの検出を試みる。

## 3. 評価実験

提案手法を用いることで電子カルテからインシデントを検

出することができるか検証するために、実データを用いて評価実験を行う。

### 3.1 TF-IDF 法にもとづく手法の評価

本評価実験では、形態素解析器としてオープンソースの形態素解析エンジン MeCab<sup>8)</sup>を利用し、単語への分割と品詞の判定を行う。

以下では、評価実験において使用するデータ、および、抽出された結果に対してインシデントであると判断するための基準と評価方法について記述する。

#### 3.1.1 利用データ

本評価実験で使用するデータは京都大学医学部附属病院のインシデントレポートと電子カルテに蓄積された診療記録である。

インシデント・キーワードの抽出を行うために 2015 年 1 月 1 日から 2015 年 12 月 31 日を発生日として報告されたインシデントレポートを用いる。また、診療記録を検索し、検索結果の検証を行うために、2017 年 2 月 1 日から 2017 年 2 月 7 日を発生日として報告されたインシデントレポートと 2017 年 2 月 1 日から 2017 年 2 月 7 日に記載された診療記録を用いる。

また、診療記録の検索・検索結果の検証を行う際には、インシデントレポートの 21 項目存在するカテゴリのうち「転倒」のカテゴリのインシデント・キーワードを用い、カテゴリが「転倒」であるインシデントの検出を試みる。

なお、検索対象期間となる 2017 年 2 月 1 日から 2017 年 2 月 7 日を発生日として報告されているカテゴリが「転倒」であるインシデントレポートは 13 件である。

#### 3.1.2 評価方法

特定のカテゴリのインシデント・キーワードを用いて検索された診療記録がインシデントの内容を含むと判断する基準について述べる。検索された診療記録の患者 ID・発生日と検索に用いたインシデント・キーワードのカテゴリが、報告されているインシデントレポートの患者 ID・発生日・カテゴリと一致した場合、提案手法によるインシデントの検出ができたと考え、一致しなかった場合、誤検出として取り扱う。

前節で一般化を行った TF-IDF 法の評価を行うために  $x$  と  $y$  それぞれの組み合わせにおいて一致数・抽出数・再現率・精度・F 値を求める。なお、再現率は該当インシデントのうち抽出されたインシデントの割合を示すものであり、精度はインシデント・キーワードにおいて抽出された結果のうちインシデントレポートのインシデントと一致した割合を示すものであり、F 値とは精度と再現率の調和平均のことである<sup>9)</sup>。

### 3.2 機械学習にもとづく手法の評価

本評価実験では、形態素解析器として MeCab<sup>8)</sup>を、サポートベクターマシンとしてオープンソースである機械学習ライブラリ LIBSVM<sup>9)</sup>を利用し、学習モデルの作成とインシデントの判定を行う。なお、サポートベクターマシンで用いるカーネルとして、自然文の分類を行う際に高い性能を示す線形カーネルを用いる。

以下、利用データと評価方法について述べる。

#### 3.2.1 利用データ

本評価実験で使用するデータは京都大学医学部附属病院の電子カルテに蓄積された診療記録である。2016 年 1 月 1 日から 2016 年 2 月 29 日に記載された診療記録のうち、「注射薬」のカテゴリに分類されたインシデントレポートの患者 ID

と発生日が一致する 2611 件の診療記録を用いる。この診療記録のうち、注射薬に関するインシデントの内容を含む 92 件の診療記録を正例とし、含まない 2519 件を負例とする。

### 3.2.2 評価方法

前述したデータを用いて、10 分割交差検定を行う。10 分割交差検定では、データセットを 10 分割し、10 個のデータセットを作る。そして、1 個目のデータセットを評価データ、残りの 9 個を教師データとして用いて、学習と評価を行う。その後、2 個目のデータセットを評価データ、残り 9 個を教師データとして用いて、学習と評価を行う。この動作を 10 回目まで繰り返し、結果を統合して評価する評価方法である<sup>5)</sup>。

## 4 結果

TF-IDF 法を用いたインシデント・キーワード抽出結果と、インシデント・キーワードを用いた診療記録からのインシデント検索結果、および、サポートベクターマシンを用いたインシデントに関連する記述を含む診療記録の判定結果について記述する。

### 4.1 TF-IDF 法にもとづく手法の結果

#### 4.1.1 インシデント・キーワードの抽出

一般化を行った TF-IDF 法を適用させて抽出した「転倒」カテゴリのインシデント・キーワードを表 1 に示す。

表1 カテゴリ「転倒」の TF-IDF 値による  
インシデント・キーワードランキング上位 10 語

| x    | 1   |     |     | -log <sub>y</sub> 2 |     |     |
|------|-----|-----|-----|---------------------|-----|-----|
|      | e   | 2   | 1.5 | e                   | 2   | 1.5 |
| 1 位  | する  | する  | する  | 尻餅                  | 尻餅  | 尻餅  |
| 2 位  | いる  | いる  | 転倒  | 崩す                  | 崩す  | 崩す  |
| 3 位  | 転倒  | 転倒  | いる  | 転ぶ                  | 転ぶ  | 転ぶ  |
| 4 位  | 患者  | 患者  | 打撲  | 転ける                 | 転ける | 転ける |
| 5 位  | 時   | 時   | 患者  | 滑る                  | 滑る  | 滑る  |
| 6 位  | ない  | 打撲  | トイレ | 立案                  | 立案  | 立案  |
| 7 位  | 打撲  | ない  | 時   | 手すり                 | 手すり | 手すり |
| 8 位  | トイレ | トイレ | ない  | 強打                  | 強打  | 強打  |
| 9 位  | ため  | ため  | 歩行  | 物音                  | 物音  | 物音  |
| 10 位 | ある  | ある  | 尻餅  | スコア                 | スコア | スコア |

#### 4.1.2 インシデント・キーワードを用いた検索

表 2 に  $x = 1, y = 1.5$ 、表 3 に  $x = -\log_2 y, y = 1.5$  として求めたインシデント・キーワードを利用した診療記録からのインシデント検索結果に対して評価した結果を示す。

また、インシデント・キーワードの上位 10 語によって OR 条件で抽出された検索結果に対する評価結果を表 4 に示す。

### 4.2 機械学習にもとづく手法の結果

評価データを 10 分割交差検定により評価した結果を表 5 に示す。精度は 37.2%、再現率は 34.8%、F 値は 0.360 であった。

## 5 考察

初めに、全てのインシデントを検出できることが望ましいが、報告が漏れているインシデントを検出するという本研究の目標に照らし合わせて考えると、全てのインシデントを検出できなくとも検出手法として意味はある。また、最終的な判断は安全管理部門に任せるという運用も現実的であると考えられるこ

とから、本研究においては再現率よりも精度が重要である。

表 2  $x = 1, y = 1.5$  のときの評価結果

|     | 一致数 | 抽出数   | 再現率   | 精度    | F 値   |
|-----|-----|-------|-------|-------|-------|
| する  | 13  | 9298  | 1     | 0.01  | 0.03  |
| 転倒  | 13  | 1022  | 1     | 0.013 | 0.025 |
| いる  | 11  | 9155  | 0.846 | 0.001 | 0.002 |
| 打撲  | 12  | 111   | 0.923 | 0.108 | 0.194 |
| 患者  | 9   | 11333 | 0.692 | 0.001 | 0.002 |
| トイレ | 7   | 678   | 0.538 | 0.010 | 0.020 |
| 時   | 13  | 20490 | 1     | 0.001 | 0.001 |
| ない  | 13  | 10129 | 1     | 0.001 | 0.003 |
| 歩行  | 9   | 1231  | 0.692 | 0.007 | 0.014 |
| 尻餅  | 6   | 21    | 0.462 | 0.286 | 0.353 |

表 3  $x = -\log_2 y, y = 1.5$  のときの評価結果

|     | 一致数 | 抽出数 | 再現率   | 精度    | F 値   |
|-----|-----|-----|-------|-------|-------|
| 尻餅  | 6   | 21  | 0.462 | 0.286 | 0.353 |
| 崩す  | 0   | 6   | 0     | 0     | 0     |
| 転ぶ  | 0   | 8   | 0     | 0     | 0     |
| 転ける | 0   | 4   | 0     | 0     | 0     |
| 滑る  | 0   | 2   | 0     | 0     | 0     |
| 立案  | 0   | 324 | 0     | 0     | 0     |
| 手すり | 0   | 144 | 0     | 0     | 0     |
| 強打  | 0   | 6   | 0     | 0     | 0     |
| 物音  | 1   | 17  | 0.077 | 0.077 | 0.067 |
| スコア | 0   | 205 | 0     | 0     | 0     |

表 4 上位 10 語を用いた OR 条件での検索結果

| x                   | y   | 一致数 | 抽出数   | 再現率   | 精度    | F 値   |
|---------------------|-----|-----|-------|-------|-------|-------|
| 1                   | e   | 13  | 21421 | 1     | 0.001 | 0.001 |
|                     | 2   | 13  | 21421 | 1     | 0.001 | 0.001 |
|                     | 1.5 | 13  | 21421 | 1     | 0.001 | 0.001 |
| -log <sub>2</sub> 2 | e   | 7   | 686   | 0.538 | 0.01  | 0.02  |
|                     | 2   | 7   | 686   | 0.538 | 0.01  | 0.02  |
|                     | 1.5 | 7   | 686   | 0.538 | 0.01  | 0.02  |

表 5 機械学習にもとづく手法の結果

|       |    | 実際の判定 |      |
|-------|----|-------|------|
|       |    | あり    | なし   |
| テスト結果 | 陽性 | 32    | 54   |
|       | 陰性 | 60    | 2465 |

まず、一般化した TF-IDF 式である(2)式で抽出したインシデント・キーワードの評価、および、抽出したインシデント・キーワードを用いたインシデントの検出結果に対する考察を以下に記述する。

表 1 より「する」や「いる」などの一般的な動詞や名詞が多く出現していることから、 $x = 1$  である場合は TF の影響が大きく出ていることがわかる。一方、 $x = -\log_2 y$  の場合は、「する」や「いる」などの一般的な動詞や名詞が出現しておらず一般化の目的であった IDF の影響を大きくするという効果が現れていると考えられる。

次に、表 2 および表 3 に着目する。表 2 の結果は前述してきたように TF の影響が大きく「する」や「いる」などの一

般的な動詞や名詞が出現しており、対象期間ほとんど全ての診療記録を検索してしまっている。そのため、 $x = 1$  を利用してインシデントを検出するのは難しいと考えられる。一方、 $x = -\log_2 2$ は(2)式の変更の目的である IDF の影響を大きくすることに於いては上手く機能していることが伺える。さらに、抽出数が絞れているため、検索された診療記録から人手でインシデントを検出することも可能ではあると考えられる。

しかし、本手法は「転倒」のカテゴリのように「転倒」のカテゴリに特徴的な単語が直接的にインシデントに関連する場合には有用と考えられるが、「注射薬」のカテゴリのように「注射薬」のカテゴリに特徴的な単語がインシデントに直接関連すると考えにくい場合にはインシデントの検出が難しいと推測される。

続いて、2 つ目の手法であるサポートベクターマシンを用いた検出手法の評価結果に対する考察を以下に記述する。

サポートベクターマシンを用いた検出手法の評価ではカテゴリとして「注射薬」を用いた。先に述べたように「注射薬」のカテゴリに特徴的な単語が直接的にインシデントに関連するとは考えにくい、本手法では精度が 37.2%となっており、最終的な判断は人手で行うという運用を考えた場合には実用的な精度であると考えられる。機械学習では各単語の重要度だけではなく、単語間の関連を考慮することができるため、十分な精度が出たと推測される。ただし、本手法は単純に機械学習手法を適用しただけであり、パターン認識などを用いることで精度を高める余地があると考えられる。

## 6 おわりに

本研究では、電子カルテからインシデントを検出することを目的とし、2 つの手法を提案した。1 つ目は、変形した TF-IDF 法を用いたインシデントレポートからのインシデント・キーワードの抽出、および、抽出したインシデント・キーワードを用いた電子カルテからのインシデントの検出を行う手法である。2 つ目は、サポートベクターマシンによって教師データから作成した学習モデルを用いた、診療記録からのインシデント検出手法である。

1 つ目の手法において、インシデントの 1 カテゴリである「転倒」を対象として評価実験を行い、TF-IDF 法の変形を行うことで F 値が高くなることを示し、インシデントの検出のためには TF-IDF 法の変形が有用であるという示唆を得た。

また、2 つ目の手法においてはインシデントの 1 カテゴリである「注射薬」のカテゴリにおいて評価実験を行い、提案手法の有用性を確認した。

## 謝辞

本研究は JSPS 科研費 JP16K15392, JP15K00466 の助成を受けたものである。

## 参考文献

- 1) 厚生労働省. 医療法施行規則. 1948.  
[<http://law.e-gov.go.jp/htmldata/S23/S23F03601000050.html>]  
(cited 2017-Sep-5)]
- 2) 津田佳彦, 廣瀬昌博, 塚本 篤. 医療安全文化醸成度と lag time との関連性に関する検討 看護管理 2014; 24 : 172-174.
- 3) M.Hirose, S E Regenbogen, and A A Gawande. Lag time in an incident reporting system at university hospital in japan. Qual Saf Health Care 2007 ; 16 :101-104.
- 4) Li Q, Melton K, Lingren T et al. Phenotyping for patient safety: algorithm development for electronic health record based

automated adverse event and medical error detection in neonatal intensive care. JAMIA 2014; 21:776-84.

- 5) 高村大也. 言語処理のための機械学習入門(自然言語処理シリーズ). コロナ社. 2010
- 6) 徳永健伸. 情報検索と言語処理(言語と計算). 東京大学出版会.1991.
- 7) Ronen Feldman, James Sanger.テキストマイニングハンドブック(辻井潤一, IBM 東京基礎研究所). 東京電機大学出版局. 2010.
- 8) 京都大学情報学研究科-日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクト. MeCab: Yet another part-of-speech and morphological analyzer  
[<http://taku910.github.io/mecab/>].
- 9) C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1-27:27, 2011.