株主招集通知における議案別の開始ページの推定

高野海斗 [1*]　酒井浩之 [1]　坂地泰紀 [1]　和泉　潔 [2]
岡田奈奈 [3]　水内利和 [3]
Kaito Takano [1]　Hiroyuki Sakai [1]　Hiroki Sakaji [1]　Kiyoshi Izumi [2]
Nana Okada [3]　Toshikazu Mizuuchi [3]

[1] 成蹊大学 理工学部 情報科学科
[1] Department of Computer and Information Science, Faculty of Science and Technology, Seikei University
[2] 東京大学大学院 工学系研究科
[2] School of Engineering, The University of Tokyo
[3] 株式会社日経リサーチ
[3] NIKKEI RESEARCH INC.

Abstract: In this research, we aim to predict start pages of proposals stated in notice of the meeting of shareholders and classify which proposal the page is. We propose two methods that classification method of proposals. The first method heuristically predicts the page on which the proposal is described. Moreover our method extracts specialized terms of each proposal and assigns weights to them. After that, our method classifies proposals by specialized terms. The second method classifies proposals using deep learning. Each methods were evaluated, and the effectiveness of each methods was verified.

1　はじめに

人工知能分野の手法や技術を，金融市場における様々な場面に応用することが期待されており，例えば，膨大な金融情報を分析して投資判断の支援を行う技術が注目されている．その一例として，日本銀行が毎月発行している「金融経済月報」や企業の決算短信，経済新聞記事をテキストマイニングの技術を用いて，経済市場を分析する研究などが盛んに行われている [1][2][3][4][6].

日経リサーチでは，有価証券報告書などの公開資料を収集し，必要な箇所のデータを抽出する作業を行っている．データ作成にあたっては，例えば XBRL 形式のように値に付与されたタグ様の付加情報を利用し，作成しているものもあるが，データ分類用付加情報が付与されているデータはまだ少数で，人手による作業が大半を占めている．

手作業で必要な情報を抽出するには、専門的知識や経験が必要となる．人工知能分野の手法や技術を用いることで，正確性を担保しつつ，データを作成時間の短縮化をめざしている．そのための一環として，まずは株主招集通知に掲載されている議案の開始ページの推定を行う．従来は抽出したい議案（「取締役選任」「剰余金処分」などの十数区分の項目）が報告書のどのページに記載されているか人手により確認し，データを作成していたが，各社で報告書のページ数や議案数が異なるため，確認に時間を要していた．抽出したい議案がその報告書にあるのか，どのページに記載されているのかが自動で推定できれば，業務の効率化につながる．具体的には，株主招集通知の各ページが議案の開始ページであるかそうでないかを判別し，さらに，開始ページであると判断されたページに記述されている議案が，どのような内容の議案であるかを分類する．

関連研究としては白田らが，日本銀行政策委員会金融政策決定会合議事要旨のテキストデータから，トピック抽出の研究を行っている [7]．また酒井らは，決算短信 PDF から業績要因の抽出の研究を行っている [5]．それらの研究を踏まえ本研究では，株主招集通知のデータを扱う点や，トピックや業績要因の抽出ではなく，議案の開始ページの推定とその議案分類をする点が異なっている．

*連絡先：成蹊大学 理工学部 情報科学科
〒 180-8633 東京都武蔵野市吉祥寺北町 3-3-1
E-mail: dm166208@cc.seikei.ac.jp

2 提案手法 1 特徴語による議案分類

本章では，特徴語による議案分類について説明する．この提案手法は以下の 3 つのステップで議案の分類を行う．

Step 1: 議案ごとの特徴語の獲得

Step 2: 議案がある開始ページの推定

Step 3: 議案の分類

2.1 議案ごとの特徴語の獲得

2.1.1 特徴語候補の抽出

株主招集通知に出現する議案を分類するために，議案ごとの特徴語の抽出をする．例えば，「取締役選任」の特徴語として，「現任取締役」のような語を抽出する．議案ごとの特徴語の獲得をするために，2016 年 4 月から 8 月までの株主招集通知における議案別の開始ページとその議案の分類が記述されたデータ（4,729 件）を学習データとして使用する．特徴語の抽出は上記の学習データを形態素解析し，それから以下の条件のもと，各分類議案の開始ページに 2 回以上出現する語を特徴語の候補とする．

条件 1 名詞を対象

条件 2 分割は N-gram 単位

条件 3 ２５文字以上の長すぎる複合名詞は除外

2.1.2 特徴語候補への重み付け

特徴語の候補 n_i に対して分類ごとに重み付けを行い，特徴語を選択する．重み付けの式には式 1 を用いる．

$$W(n_i, C(t)) = (0.5 + 0.5\frac{TF(n_i, C(t))}{max_{j=1,\ldots,m} TF(n_j, C(t))}) \times H(n_i, C(t)) \log_2 \frac{N}{df(n_i)} \quad (1)$$

ここで，学習データにおいて，

$C(t)$**:** 議案分類 t の開始ページの文書集合．

$TF(n, C(t))$**:** $C(t)$ において，名詞 n が出現する頻度．

$H(n, C(t))$**:** $C(t)$ の各文書である d に名詞 n が出現する確率 $P(n, d)$ に基づくエントロピー．以下の式 2 によって求める．

$$H(n, C(t)) = -\sum_{d \in C(t)} P(n, d) \log_2 P(n, d) \quad (2)$$

ここで，$P(n, d)$ は d に名詞 n が出現する確率である．

$df(n)$**:** 名詞 n を含む文書の数．

N**:** 学習データにおける文書の総数．

エントロピーを用いた理由は，各議案分類の文書集合中で多くの文書に分散して出現している語の方が，少数の文書に集中して出現している語と比較して，よりその議案分類の特徴を表し，特徴語としても有効であるという仮定に基づく．

2.1.3 特徴語の選択

各議案分類ごとの特徴語候補の重み付けの平均値を算出し，平均値よりも重みの高いものを特徴語として選択する．すなわち, 以下の条件が成り立つ語 n_i を特徴語として選択する．

$$W(n_i, C(t)) > \frac{1}{m} \sum_{j=1}^{m} W(n_j, C(t)) \quad (3)$$

m**:** 議案分類 t の特徴語候補の総数．

例えば，取締役選任の特徴語の一部を表 1 に示す．

表 1: 取締役選任の特徴語

特徴語	重み
リーダーシップ	18.13
当社代表取締役社長就任	18.00
現任取締役	17.85
在任取締役	17.49

2.2 議案がある開始ページの推定

議案がある開始ページは，以下の条件のもと推定した．

条件 1 議案がある開始ページには，「議案」または「第 X 号議案」が先頭に含まれるページを対象とする

条件 2 目次ページが存在するため，目次ページは議案がある開始ページから除外する

条件 3 参考ページ [1] 以降の開始ページ推定の対象とする

[1] 株主招集通知において，議題について書かれている最初のページを参考ページと呼ぶことにする．多くの場合，第 1 号議案の開始ページが参考ページとなる．

「決議事項」または「目的事項」という表現が含まれているページを目次ページし,「参考書類」,「議題及び参考事項」または「議題および参考事項」という表現が含まれているページを参考ページとする．

「第 X 号議案」という語が含まれている場合，その株主招集通知に含まれる議案数は「X」の最大値である．また「第 X 号議案」という語が含まれていない場合，議案という語が含まれていれば，議案数は 1 であると推定できる．これらの情報を基にページの推定を以下のように行う．

Step 1: 保存されている参考ページの中で最も後ろのページから最後のページまでに,「第 1 号議案」から「第 N 号議案」の順に出現するかどうかを調べる．

Step 2: 出現しない場合は，保存されている参考ページの一つ若いページを用いて，同様の探索を行う．

Step 3: 参考ページが目次ページ以下になった場合，目次ページの次のページから同様の探索を行い，それでも見つからない場合は議案はないものとする．

2.3 議案の分類

2.1 節で得られた重み付けと 2.2 節で推定した議案の開始ページを用いて，開始ページごとの議案の分類を行う．議案分類 t の開始ページ j に対するスコア付与は式 4 を用いる．

$$score(j,t) = \frac{\boldsymbol{V(t)} \cdot \boldsymbol{V(j)}}{|\boldsymbol{V(t)}||\boldsymbol{V(j)}|} \tag{4}$$

ここで,

$\boldsymbol{V(t)}$**:** 議案分類 t の特徴語を要素，特徴語の重みを要素値とするベクトル

$\boldsymbol{V(j)}$**:** 開始ページ j の名詞 N-gram を要素，出現数を要素値とするベクトル

複数の議案が同ページに存在する場合，スコアが上位のものから順に選ばれるものとする．

3 提案手法 2 深層学習による議案分類

提案手法 1 では，学習データから各議案の特徴語を抽出し，それに基づいて議案を分類している．この学習データを使用すれば，機械学習手法に基づく手法でも議案分類が可能である．そこで，本研究では深層学習を用いた議案分類も試みた．

3.1 素性選択

株主招集通知に記載されている議案の開始ページの議案分類を，深層学習により行う．すなわち，議案の開始ページが，ある議案分類であるかそうでないかを判別する分類器を議案分類の数だけ生成し，テストデータとなる議案の開始ページがどの議案分類に属するかを判定する．従って，例えば議案分類「取締役選任」を判別するための学習データは,「取締役選任」の開始ページが正例，それ以外の議案分類の開始ページが負例となる．また，テストデータは，学習データとして使用した株主招集通知を除き，株主招集通知を 1 ページごとに分割したうえで,「号議案」の文字列が含まれているページとした．

まず，入力層の要素となる語（素性）を選択する．具体的には，学習データにおいて正例に含まれる内容語（名詞，動詞，形容詞）に対して，以下の式 5 にて重みを計算する．

$$W_p(t,S_p) = TF(t,S_p)H(t,S_p) \tag{5}$$

ただし,

S_p**:** 学習データにおいて正例に属する文の集合

$TF(t,S_p)$**:** 文集合 S_p において，語 t が出現する頻度

$H(t,S_p)$**:** 文集合 S_p における各文に含まれる語 t の出現確率に基づくエントロピー

$H(t,S_p)$ が高い語ほど，正例の文集合に均一に分布している語であることが分かる．$H(t,S_p)$ は次の式 6 で求める．

$$H(t,S_p) = -\sum_{d \in S_p} P(t,s)\log_2 P(t,s) \tag{6}$$

$$P(t,s) = \frac{tf(t,s)}{\sum_{s \in S_p} tf(t,s)} \tag{7}$$

ここで，$P(t,s)$ は文 s における語 t の出現確率を表し，$tf(t,s)$ は文 s において語 t が出現する頻度を表す．次に，負例に含まれる内容語（名詞，動詞，形容詞）に対しても，同様に重みを計算する．

$$W_n(t,S_n) = TF(t,S_n)H(t,S_n) \tag{8}$$

ただし，S_n は学習データにおいて負例に属する文の集合である．

ここで，ある語 t の正例における重み $W_p(t,S_p)$ が負例における重み $W_n(t,S_n)$ より大きければ，その語 t を素性として選択する．もしくは，語 t の負例における重み $W_n(t,S_n)$ が正例における重み $W_p(t,S_p)$ の 2 倍より大きければ，その語 t を素性として選択する．すな

わち，以下の条件のどちらかが成り立つ語 t を素性として選択する．

$$W_p(t, S_p) > W_n(t, S_n) \quad (9)$$

$$W_n(t, S_n) > 2W_p(t, S_p) \quad (10)$$

上記の条件を課すことで，正例，負例における特徴的な語のみを素性として選択し，ともによく出現するような一般的な語を素性から除去する．以下に議案分類「取締役選任」を判別するための学習データから選択された素性の一部を例示する．

取締役，監査，議案，配当，株主，社外，変更，事業，代表，現任，責任，部長，社長

上記の学習データでは，2,845 語が素性として選択された．

3.2 モデル

入力は，学習データから抽出された語（素性）を要素，語 t における $\log(W_p(t, S_p))$，もしくは，$\log(W_n(t, S_n))$ の大きいほうを要素値としたベクトルとする．モデルの入力層のノード数を入力ベクトルの次元数（すなわち素性の数）と同じとし，隠れ層は，ノード数 1000 が 3 層，ノード数 500 が 3 層，ノード数 200 が 3 層，ノード数 100 が 3 層の計 12 層とする．出力層は 1 要素である．

4 評価

本手法を実装した．学習データとして，2016 年 4 月から 8 月までの株主招集通知から，人手にて議案開始ページとその分類を作成し使用した（学習データ数は 4,729 件）．実装にあたり，形態素解析器として MeCab[2] を使用した．

評価において，正解データとして，2016 年 9 月から 10 月までの株主招集通知から，人手にて議案開始ページとその分類を作成した（正解データ数は 345 件）．次に，正解データと同じ 9 月から 10 月までの株主招集通知から，各手法を用いて議案開始ページとその議案分類を推定した．表 2 はある企業の株主招集通知における正解データと提案手法 1 による議案開始ページと議案分類の推定結果を示す．そして，各手法の推定結果と正解データが一致すれば正解とし，議案ごとの適合率，再現率，F 値を算出した．評価結果を表 3 に示す．表 3 の手法 1 は提案手法 1 の特徴語による議案分類による手法であり，手法 2 は提案手法 2 の深層学習による手法であることを示す．

[2]http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

表 2: 提案手法 1 による議案開始ページと議案分類の推定結果とその正解データ

正解データ		分類結果	
開始ページ	議案分類	開始ページ	議案分類
34	剰余金処分	34	剰余金処分
35	定款変更	35	定款変更
37	取締役選任	37	取締役選任
38	監査役選任	38	監査役選任
39	退職慰労金	39	退職慰労金

表 3: 評価結果

議案分類	手法	適合率 (%)	再現率 (%)	F 値
会計監査人選任	手法 1	100.00	100.00	1.000
	手法 2	100.00	100.00	1.000
監査役選任	手法 1	72.73	96.00	0.828
	手法 2	85.19	90.20	0.876
企業再編	手法 1	100.00	50.00	0.667
	手法 2	50.00	25.00	0.333
ストック	手法 1	81.25	100.00	0.897
オプション	手法 2	61.11	84.62	0.710
退職慰労金	手法 1	52.63	100.00	0.69
	手法 2	90.91	100.00	0.952
定款変更	手法 1	94.74	94.74	0.947
	手法 2	94.44	89.47	0.919
取締役選任	手法 1	95.79	87.27	0.910
	手法 2	95.79	82.73	0.888
役員賞与	手法 1	83.33	83.33	0.833
	手法 2	100.00	100.00	1.000
役員報酬	手法 1	84.21	94.12	0.889
	手法 2	100.00	66.67	0.800
剰余金処分	手法 1	93.06	93.06	0.931
	手法 2	96.97	88.89	0.928

5 考察

提案手法と深層学習を比較すると，会計監査人選任の項目の分類推定は，両手法ともに良好な結果を示している．ストックオプション，定款変更，取締役選任，役員報酬，剰余金処分の項目は提案手法 1 が良好な結果が得られ，監査役選任，退職慰労金，役員賞与の項目は提案手法 2 によって良好な結果が得られた．提案手法 2 の適合率は全体的に高い傾向にあるが，これは「号議案」の文字列が含まれているページに限定して分類を行っているため，議案数が 1 つしかない株主招集通知は分類ができていないためである．よって，全体的に再現率が低くなってしまっている．提案手法の分類推定が誤っていたものを確認したところ，誤分類は 46 件だった．その誤分類の詳細を確認したところ，取締役選任の項目が，監査人選任の項目に誤分類されている件数が 15 件あった．これはどちらの項目も選任の件であり，分類が難しいことと，議案としての出現確率が高いことに起因している．また，分類にはその他といった項目が存在するが，今回はその他への分類をしていないため，15 件が誤分類となった．そして，同一ページに対し同じ議案が出てこないことが前提で議案分類を行っているため，同じ議案が複数出てきた開始ページの推定に影響を与え，6 件の誤分類となった．

また，その場合の推定は退職慰労金に分類される傾向にあり，それに起因して退職慰労金の分類適合率が低くなってしまった．その際の正解データと出力結果を表 4 に示す．

表 4: 同じ議案が複数出てきた開始ページ

正解データ		分類結果	
開始ページ	議案分類	開始ページ	議案分類
40	剰余金処分	40	剰余金処分
40	定款変更	40	定款変更
48	取締役選任	48	取締役選任
50	取締役選任	50	取締役選任
52	役員賞与	52	役員賞与
52	退職慰労金	52	退職慰労金
53	役員報酬	53	役員報酬
53	役員報酬	53	退職慰労金

提案手法 1 の分類推定を向上させるためには，取締役選任と監査人選任の項目の特徴語の選択をヒューリスティックに調整することが考えられる．また，その他の項目も，同様の手法で分類できるようにすることも考えられる．同じ議案が存在するページに関しては，退職慰労金の項目に分類されることが多いため，退職慰労金への分類に制約を与えることで，解消されると考えられる．

6 まとめ

本研究では，株主招集通知における議案の開始ページを推定し，その議案を分類する手法を提案した．議案の開始ページ推定は，議案がある開始ページには「議案」または「第 X 号議案」が先頭に含まれるといった規則に基づく．議案分類の推定は，議案ごとの特徴語を抽出し，その特徴語のスコアに基づき分類した．また，深層学習を用いた議案分類も行った．評価の結果，特徴語による議案分類，深層学習による議案分類ともに良好な適合率，再現率を達成した．

今後の展望としては，各手法には得手不得手が存在するため，本研究での評価に用いた結果を開発データとし，ハイブリッド手法の実現を予定している．

参考文献

[1] 和泉潔, 後藤卓, 松井藤五郎: 経済テキスト情報を用いた長期的な市場動向推定, 情報処理学会論文誌, Vol. 52, No. 12, pp. 3309–3315 (2011).

[2] 北森詩織, 酒井浩之, 坂地泰紀: 決算短信 PDF からの業績予測文の抽出, 電子情報通信学会論文誌 D, Vol. J100-D, No. 2, p. 150161 (2017).

[3] Peramunetilleke, D. and Wong, R. K.: Currency exchange rate forecasting from news hcadlines, *Proceedings of the AAAI Spring Symposium on Exploring At titude and Affect in Text*, pp. 131–139 (2002).

[4] Sakai, H. and Masuyama, S.: Cause Information Extraction from Financial Articles Concerning Business Performance, *IEICE Trans. Information and Systems*, Vol. E91-D, No. 4, pp. 959–968 (2008).

[5] 酒井浩之, 西沢裕子, 松並祥吾, 坂地泰紀: 企業の決算短信 PDF からの業績要因の抽出, 人工知能学会論文誌, Vol. J98-D, No. 5, pp. 172–182 (2015).

[6] 坂地泰紀, 酒井浩之, 増山繁: 決算短信 PDF からの原因・結果表現の抽出, 電子情報通信学会論文誌 D, Vol. J98-D, No. 5, pp. 811–822 (2015).

[7] 白田由香利, 橋本隆子, 佐倉環: LDA 方式による金融政策トピック抽出, 第 159 回 DBS・第 115 回 IFAT 合同研究発表会 (2014).