テキストマイニングを用いた株主招集通知の重要ページ抽出

高野 海斗[1]　　酒井 浩之[2]　　中川 慧[3]

[1] 成蹊大学大学院理工学研究科理工学専攻

[2] 成蹊大学理工学部情報科学科

[3] 野村アセットマネジメント株式会社

[1]dd186201@cc.seikei.ac.jp

[2]h-sakai@st.seikei.ac.jp

[3]k-nakagawa@nomura-am.co.jp

1　はじめに

近年，機械学習などの手法が注目を集め，様々な分野への応用研究が活発に行われている．特に，データが紙媒体から電子媒体に移行したことにより，これまで必要だった多くの業務の自動化や半自動化が，今後より進むと予測される．金融業界でも，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，膨大な金融情報を分析して投資判断を支援する技術にも注目が集まっている．さらに，最近では証券市場における個人投資家の比重が増大しており，個人投資家に対して投資判断の支援を行う技術の必要性が高まっている [1, 2, 3]．野村アセットマネジメントでは，様々な部署で文書に目を通して重要な部分を確認することや，レポートを作成するなどのテキストに関わる業務を多く行っている．これらの業務を効率化することにより，時間を大幅に削減する事ができれば，より良い働き方ができるようになることが期待される．

2　問題定義

本研究では，責任投資調査部が行っている業務の一部である，株主招集通知の内容確認業務の効率化を目的とする．業務では，数十ページにわたる資料から，株価などに与える影響が大きいと思われる部分のテキストを人手で探し確認する必要がある．また，確認する項目が多く存在するため，多くの時間を要している．現状は，株主招集通知を印刷し，必要な情報を数人で人手にて確認しているが，必要のないページ（重要な項目が載っていないページ）も多く存在する．そのため，必要な情報が記載されているページを探す時間や，印刷の待ち時間などに無駄な時間がかかってしまっている．

そこで，そのような無駄な時間を削減するために，テキストマイニング技術を活用し，必要な情報が記載されているページだけを自動で抽出する手法を提案する．具体的には，株主招集通知から，表紙や目次となるページ，大株主に関するページ，会社役員に関するページ，決議事項について記載されたページ，独立役員に関して記載されたページを印刷するページとして自動で抽出する．また，株主総会で扱われる決議事項はタイトルによって重要であるかどうかを予測したりすることができることや，大株主の情報は重要度が高いため，それらの情報も自動抽出できるような方法を提案する．

目標は，図 1 のような情報の抽出である．

```
証券コード:6145
表紙・目次ページ:1, 3,
議案について記載されたページ: 5-9 # 終了ページは候補です
大株主に関するページ:18, # 10%以上の保有者:1名
会社役員に関するページ: 19-22 # 終了ページは候補です
独立役員というワードが含まれるページ:8, 9, 19,
印刷ページ:1目次1, 3目次2, 5議案3, 6議案4, 7議案5, 8議案(独立)6, 9議案(独立)7,
18大株主8, 19役員(独立)9, 20役員10, 21役員11, 22役員12,
株主提案: なし
--------------------議案一覧start--------------------
第1号議案 剰余金の配当の件 p.5-6 該当ワードなし
第2号議案 定款一部変更の件 p.6-7 該当ワードなし
第3号議案 取締役2名選任の件 p.7-8 該当ワードなし
第4号議案 監査役1名選任の件 p.8-9 該当ワードなし
第5号議案 補欠監査役1名選任の件 p.9-9 該当ワードなし
--------------------議案一覧end--------------------
--------------------大株主情報start--------------------
株%日本トラスティ・サービス信託銀行株式会社4,071,10022.5 #条件を満たす 22.5%
SMC株式会社1,285,5007.1 #条件を満たさない 7.1%
--------------------大株主情報end--------------------
印刷:1, 3, 5-9, 18-22
```

図 1: 株主招集通知から抽出したい情報の例

3　株主招集通知

株主招集通知とは，株式総会を開催する際に，株主に発送することが義務付けられている，開催要項や企業情報が記載された通知である．企業の WEB サイト上でも閲覧が可能であり，PDF データとして，取得可能である．企業概要や，大株主情報，決議事項である議案など，多くの情報が記載されていることから，投資判断の基準に用いられることもある．ページ数は十数ページのものから百ページを超えるものも存在する．また，株主総会前に発行されるため，発行時期が集中する特徴を持つ．

4　関連研究

株主招集通知に関するテキストマイニングの研究としては，高野らの研究が存在する [4]．これらは業務中に蓄えられた大量の議案分類のタグが付いたデータを学習データとして用いることで，議案の分類をしているが，本研究とは，学習データが存在しない点，および，大株主に関する情報等の重要な情報の抽出を行っている点で大きく異なる．金融テキストから，重要な情報を抽出する研究としては，高野らの研究が存在するが [5]，これらは文章単位で抽出を行うが，本研究ではページ単位で重要な部分を抽出する点で異なる．

5　提案手法の詳細

手法のアウトラインは以下の通りである．

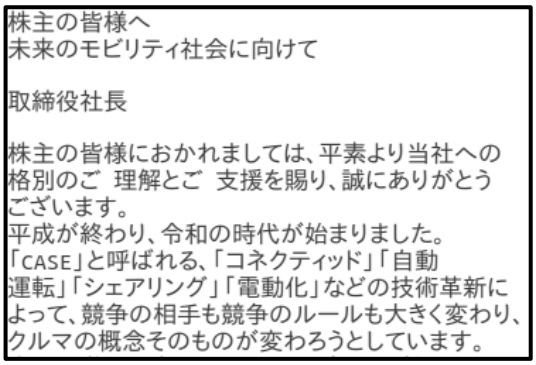
1 株主招集通知 PDF データをテキストデータに変換

2 テキストデータを用いて以下のような情報を抽出
- ・表紙・目次となる候補ページ
- ・大株主に関するページ
- ・会社役員に関する開始ページ
- ・各議案のタイトル
- ・各決議事項の開始ページ
- ・独立役員に関するページ

3 抽出情報を用いて印刷するページの決定

5.1　PDF データをテキストデータに変換

株主招集通知は PDF データとして，web ページ上で公開されているため，PDF データをテキストデータに変換する必要がある．PDF データをテキストデータに変換するためには，「k2pdfopt[1]」を用いた．実務としては印刷するページの抽出だけでなく，追加で表の中の情報などを抽出する必要があるため，多少時間はかかるが，形式が崩れにくい，「k2pdfopt」を用いることにした．「k2pdfopt」を用いることで，PDF データから図 2 のようなテキストデータを取得することが可能である．このテキストデータを用いて，重要ページの抽出などを行う．


株主の皆様へ
未来のモビリティ社会に向けて

取締役社長

株主の皆様におかれましては、平素より当社への格別のご 理解とご 支援を賜り、誠にありがとうございます。
平成が終わり、令和の時代が始まりました。
「CASE」と呼ばれる、「コネクティッド」「自動運転」「シェアリング」「電動化」などの技術革新によって、競争の相手も競争のルールも大きく変わり、クルマの概念そのものが変わろうとしています。

図 2: PDF データからテキストの抽出例

5.2　表紙・目次となる候補ページの抽出

表紙ページとは，株主招集通知の先頭にあるページで，日時，場所，目次，社名などが書いてあることが多い．目次ページとは，表紙ページの後に出てくるページで，証券コード，日時，場所，報告事項と決議事項の一覧などが記載されている．表紙や目次となるページには，話し合われる議案のタイトルなどが記載されているため，重要なページである．特徴として，表紙ページは，先頭のページにあることが多いが，目次ページは，企業情報などがそれよりも前に来ることもあるため，必ずしも表紙ページの次に来るとは限らない．

表紙や目次には，特徴的な単語が多く存在するので，「第 n 号議案[2]」と下記の単語が対象とするページにすべて含まれているかどうかで判定した．

株主総会，日時，決議事項，通知，招集

また，目次ページは複数ページにまたがる可能性があるため，上記の単語がすべて含まれているページが存在しない場合には 2 ページごとにテキストをまとめて対象となるページの抽出を行った．

[1] https://www.willus.com/k2pdfopt/
[2] n は対象とする株主招集通知の中で一番大きい番号

5.3　大株主に関するページの抽出

大株主に関する事項が記載されているのは，株主招集通知の株式に関する事項の一部分である．大株主上位１０名の株主名，持株数，持株比率などが記載されており，株価への影響を与える可能性があるため重要な確認項目の１つである．

大株主に関するページには，特徴的な単語が多く存在するので，その単語がすべて含まれているかどうかで判定した．

株主, 株式, %, 持株, 発行, 総数

5.4　会社役員に関するページの抽出

会社役員に関する事項も重要である．会社役員に関するページには，以下の単語のいずれかが含まれているかどうかで判定した．

役員に関する事項，役員に関する状況，役員の状況

5.5　各議案タイトルの抽出

決議事項である各議案タイトルの抽出は，全ページを対象に文字列マッチングで以下のいずれかに当てはまるものを抽出した．

第 N 号議案～件，第 N 号議案～について

ある議案に対するタイトルが複数とれる場合もあるため，さらに細かいルールを用いて，各議案ごとにタイトルを決定した．

5.6　各決議事項の開始ページの抽出

抽出した各議案タイトルと議案番号を用いて，議案ごとの開始ページの抽出を行った．最初に抽出した表紙・目次となる候補ページよりも後ろのページを対象に，第 1 号議案から順に，議案タイトルと第 n 号議案が出現するページをその議案の開始ページとして抽出した．

5.7　独立役員に関するページの抽出

独立役員に関する情報も，株価に影響を与えることがあるため重要である．独立役員に関するページは，ページに「独立役員」という単語が出現するものをすべて抽出した．

5.8　抽出情報を用いた印刷ページの決定

ここまでに抽出した情報以外にも,「連結計算書類」,「貸借対照表」,「損益計算書」,「メモ」,「新株予約権」,「会計監査人」などに関する事項が書いてあるページの抽出も行った．これらの情報自体はあまり重要ではないが，抽出したい情報が記載されていることが終了するページを抽出することは現状難しくできないため，次の項目が始まる場所までを範囲とすることで印刷するページを決定した．例えば,「会社役員に関する事項」の次に来るのは，高い可能性で「会計監査人」に関するページであるため，そのページを「会社役員に関する事項」の終了ページとした．

特にうまくいかないのは，最後の議案がどこまで記載されているかを決めることであった．最後の議案の終了ページは，5.1 節にて説明したテキストデータにおいて,「以上」のみの行が出現したページとしているが，ルールにヒットしないものも多かった．終了ページが見つからない場合は，他の項目に関するページが始まるまでのすべてのページを印刷することで対応した．その結果，重要な情報が印刷できないファイルはほとんど存在しない状態にすることに成功した．

6　評価の方法とその結果

今回対象とした株主招集通知 PDF のデータは，2019 年 1 月から 2019 年 9 月までに企業が公表した，1,527 件のデータを用いた．1,527 件のデータから情報を抽出したところ，47,645 ページの圧縮に成功した（1 ファイルあたり 31 ページ短縮）．今回の課題は，印刷するページ数を少なくしつつ，必要な情報が記載されたページを漏れなく印刷することが求められるため，適合率はもちろんだが，再現率が高いことがとても重要である．人手で 1,527 件すべてのファイルを評価することは難しいため，ランダムに選んだ 10 件のファイルを人手で確認し，印刷ページ抽出の適合率，再現率と，議案タイトル抽出の精度を用いて評価を行った．結果を表 1 に示す．

7　考察

開始ページの抽出は，そもそも課題としてあまり難しくないが，終了ページの抽出はまだまだ改良の余地が存在する．今回扱った課題では，印刷されたページが必要な部分を含んでいないと，二度手間になってしまうため多めに範囲を指定することでフォローしてい

表 1: 評価結果

証券コード	印刷ページ適合率	印刷ページ再現率	議案タイトル抽出精度	全ページ数	圧縮率
3182	12 / 13 = 0.923	12 / 12 = 1.000	1 / 2 = 0.500	51	24%
3341	17 / 19 = 0.895	17 / 17 = 1.000	7 / 7 = 1.000	64	27%
3676	10 / 10 = 1.000	10 / 10 = 1.000	1 / 2 = 0.500	50	20%
3694	6 / 7 = 0.857	6 / 6 = 1.000	2 / 2 = 1.000	48	13%
4620	22 / 26 = 0.846	22 / 22 = 1.000	6 / 7 = 0.857	56	39%
6121	10 / 12 = 0.833	10 / 10 = 1.000	3 / 3 = 1.000	28	36%
6145	12 / 13 = 0.923	12 / 12 = 1.000	5 / 5 = 1.000	36	33%
6923	17 / 18 = 0.944	17 / 17 = 1.000	2 / 2 = 1.000	68	25%
7780	18 / 28 = 0.643	18 / 19 = 0.947	2 / 2 = 1.000	44	41%
8589	14 / 15 = 0.933	14 / 14 = 1.000	3 / 3 = 1.000	39	36%

るが，可能であるならば，不必要なページは印刷しないようにしたい．

今回の手法はルールベースを用いているが，良好な結果は得られている．しかし，人間が確認すればすぐにわかるような些細な違いでさえも，ルールベースに当てはまらないことで対応できなくなる．ランダムでファイルを選べば成功しているファイルばかりであるが，ルールベースにヒットしなかったものにのみ限定して出力し確認を行うと，やはりページの抽出に失敗し必要なページが印刷範囲に指定されていない例も存在した．また，株主招集通知の構成に依存するため，記載順番や見出しのタイトル変更などには，自動で対応することが現状できない．

8 発展的課題解決策の提案

今回の手法を用いることで，かなり高い精度で重要なページの抽出をすることができた．また，抽出がうまくいっていない可能性があるものと，うまくいっているもので区別することができるため，抽出がうまくいっていることが保証できるようなデータを学習データとして，機械学習を用いた分類をすることが可能であると考えられる．現状は考察で述べた通り，ルールから少しでも外れたりするものは，抽出がうまくできなくなるが，機械学習を用いてページ全体や前後情報を加味することで，多少の違いに影響されないモデルを作成することが期待できる．具体的には，1 ファイルごとにページごとのデータを，何らかの手法を用いて分散表現で表し，BiLSTM 層と CRF 層を使った深層学習モデルで，ページにタグ付けを行うことで，印刷するページを抽出できるモデルを考えている．

9 おわりに

本研究では，業務効率化を目的とした株主招集通知の重要ページ抽出を行った．実験を行った結果，本手法の有用性を確認できた．

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] 藏本貴久, 和泉潔, 吉村忍, 石田智也, 中嶋啓浩, 松井藤五郎, 吉田稔, 中川裕志. 新聞記事のテキストマイニングによる長期市場動向の分析. 人工知能学会論文誌, Vol. 28, No. 3, pp. 291–296, 2013.

[3] 酒井浩之, 増山繁. 企業の業績発表記事からの重要業績要因の抽出. 電子情報通信学会論文誌 D, Vol. J96-D, No. 11, pp. 2866–2870, 2013.

[4] 高野海斗, 酒井浩之, 坂地泰紀, 和泉潔, 岡田奈奈, 水内利和. 株主招集通知における議案タイトルとその分類及び開始ページの推定システム. 自然言語処理, Vol. 25, No. 1, pp. 3–32, 2018.

[5] 高野海斗, 酒井浩之, 北島良三. 有価証券報告書からの事業セグメント付与された業績要因文・業績結果文の抽出. 自然言語処理, Vol. 34, No. 5, pp. 1–22, 2019.