E-021

TOPIX100 の有価証券報告書に含まれる表形式データの分析
Analysis of Tabular Data Contained in the TOPIX100 Annual Securities Report

佐藤 栄作[1)]　梶 縁[1)]　木村 泰知[1)]
Eisaku Sato　Yosuga Kaji　Yasutomo Kimura

1　はじめに

有価証券報告書は，金融商品取引所で株式を公開している会社が，事業年度ごとに外部へ公開する資料である．有価証券報告書は，原則として，EDINET (Electronic Disclosure for Investors' NETwork) への電子提出が義務付けられている[1)]．EDINET で公開されている有価証券報告書等は，企業の財務情報を記述する標準言語である XBRL（eXtensible Business Reporting Language）形式で記述されている．XBRL は，XML の技術を利用しており，複数企業の比較分析のようにデータの再利用を目的としており[2)]，特に，iXBRL(inline XBRL) は，関連する XBRL タグを元のテキストと一緒に表示できるという特徴がある．

しかしながら，実際に公開されている有価証券報告書を見ると，表形式データと本文の説明が対応付けられておらず，テキストが活用されていない．また，XBRL 形式でウェブ上に公開されているものの，会社ごとに表の構造が異なり，機械判読しづらいため，会社間の比較も困難である．

そこで，我々は，有価証券報告書の表形式データとテキストを結びつけることで，会社間の比較や会社の概要を容易に把握できるようにすることを目標としている．有価証券報告書に記載されているテキストは，図や表を参照しつつ報告している文章であり，非構造化データとみなすことができる．これらのテキストは，業績やリスクなどについて記載されていることから，業績要因や将来業績を抽出したり，リスクへの対応を分析するために研究が行われている [1] [2]．一方，有価証券報告書の表形式データは，各列に金額や割合表現などの数値表現が格納されていることから，構造化データとみなすことができる．構造化データと非構造化データを結びつけることは，自然言語処理分野でも期待されている．しかしながら，有価証券報告書に含まれる表形式データは，1 行目が項目名のものあれば，1 列目が項目名のものもあり，さまざまな形式が存在する．

そこで，本研究では，TOPIX100 の有価証券報告書に含まれる表形式データを対象として，どのような表のパターンが存在し，どのような項目名が存在するのかを明らかににすることを目的とする．また，表の項目に含まれる金額表現と本文で説明されている箇所を人手で対応づけ，金額表現がどのよう文脈で使われているのかを分析することにより，構造化データと非構造化データを自動で対応付ける際の問題点を明らかにする．本稿では，上記の目的の達成のために，有価証券報告書に含まれる表形式データおよび金額表現について分析した結果について述べる．

本研究の貢献は，下記の 2 つである．

- TOPIX100 の有価証券報告書に含まれる表の総数，表のパターン，項目名を調査し，表形式データを扱う問題点を述べている．
- 表形式データと説明しているテキストを手作業で結びつけ，分類することで，自動化の問題点を明らかにしている．

2　関連研究

有価証券報告書に関する研究は数多く存在しており，自然言語処理技術を活用する例がみられる．高野らは，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・業績結果文を抽出する研究を行っている [1]．安らは，有価証券報告書の訂正報告書と不適切会計処理に関する分析を行っている [3]．太田は，有価証券報告書の設備投資情報に関する実証分析をしており，設備投資情報，企業の設備投資の実態を明らかにしようとしている [4]．加藤らは，有価証券報告書を対象として 経営者による将来見通しと将来業績についてののテキスト分析をしている [2]．これらの研究は，テキストを対象としており，表形式のデータを考慮していない．

テキストデータと金額表現を結びつける研究としては，NTCIR16 QA Lab-PoliInfo-3 のサブタスクである Budget Argument Mining があり，議会会議録に含まれる金額表現と予算項目を結びつけている [5]．しかしながら，有価証券報告書のように同一ファイルに含まれる表形式データとテキストを結びつける研究ではない．

3　有価証券報告書の表形式データの調査

3.1　目的

本調査の目的は，TOPIX100 の有価証券報告書に含まれる表形式データを対象として，表の総数，表のパターン，項目名の異なり数，項目名の表記揺れ，および，曖昧性の有無について明らかにすることである．具体的には，次の 7 つの疑問に答える．

Q1-1　表 (TABLE タグ) はいくつあるのか
Q1-2　TABLE タグは全て表なのか
Q1-3　どのような表が存在するのか
Q1-4　分析対象の表の条件とは
Q1-5　どのような項目名が存在するのか
Q1-6　項目名に表記揺れがあるのか
Q1-7　項目名に曖昧性があるのか

1)　国立大学法人 北海道国立大学機構 小樽商科大学

1)　EDINET とは，金融商品取引法に基づく有価証券報告書等の開示書類に関する電子開示システムである．

2)　http://itdoc.hitachi.co.jp/manuals/3020/30203N3920/N390006.HTM

表 1　TOPIX 100 の有価証券報告書に含まれる TABLE タグ数 (一部)

	企業名 \ 開始年	平成 28 年 2016 年	平成 29 年 2017 年	平成 30 年 2018 年	平成 31 年 2019 年	令和元年 2019 年	令和 2 年 2020 年	令和 3 年 2021 年	合計
1	ANA ホールディングス	175	182	179	180	0	178	0	894
2	HOYA	230	212	216	220	0	217	0	1,095
3	住友金属鉱山	204	215	291	255	0	249	0	1,214
4	JXTG ホールディングス	252	250	204	198	0	196	0	1,100
…	…	…	…	…	…	…	…	…	…
95	資生堂	0	231	218	202	0	207	224	1,082
96	野村ホールディングス	346	257	267	264	0	273	0	1,407
97	Z ホールディングス	227	208	218	218	0	223	0	1,094
合計		18,556	20,881	21,286	21,407	343	21,850	2,711	107,034

3.2　対象データの選定方法

分析対象のデータは TOPIX100 である．有価証券報告書は，受理した日から 5 年を経過する日まで保存することが義務づけられている．EDINET は，直近 5 年分の有価証券報告書を公開しており，WEB API を用いてダウンロードすることができる．そのため，本研究では有価証券報告書の開始年が 2016 年から 2021 年まで (平成 28 年から令和 3 年) を対象範囲とした．

- 対象会社 : TOPIX100 ※ 2021 年 10 月時点
- 対象年（開始年）: 2018-2021 年

有価証券報告書の「開始年」と記述している理由は，会社によって決算月が異なるためである．報告書の対象期間は，終了年が開始年の翌年となり，開始年と終了年が異なる場合も多いことから，開始年を基準に分けている．例えば，2016 年が開始年の場合，2016 年から 2017 年までの 1 年間の報告書のことである．

3.3　結果

Q1-1 表 (TABLE タグ) はいくつあるのか

表 1 は，TOPIX 100 の有価証券報告書に含まれる TABLE タグ数をカウントした結果の一部である．対象の会社は，TOPIX100 の 100 社であるが，3 社（三菱重工業，三菱商事，三菱ケミカルホールディングス）は通常と異なるフォーマットであったため対象外として，97 社とした．対象となる有価証券報告書は，97 社の 5 年間の **482 の報告書**であり，そのなかには，69 の訂正報告書が含まれる[3]．5 年間の有価証券報告書における**表 (TABLE タグ) の総数は 107,034** であった．また，一つの報告書に含まれる表の**最大値は 526，最小値は 103，平均値は 221.9** であることを確認した．

Q1-2 TABLE タグは全て表なのか

本研究で想定している表形式データは，項目名があり，その項目の値として数値表現（金額，増減比，パーセンテージなど）が含まれるデータである．TABLE タグには，長い文章，あるいは，空欄のみの表などが存在し，表の項目名や金額表現が含まれていない場合があることを確認した．

Q1-3 どのような表が存在するのか

TABLE タグは，前述の通り，一般的な表として利用されていない場合もある．そこで，どのような表のパターンが存在するのか，項目名に着目して分類してみる．下記の通り，項目名の有無によって分けることができる．「項目名あり」は，項目名が 1 行目の横並びに記述されている場合，項目名が 1 列目の縦並びに記述されている場合，1 行目にも 1 列目に記述されていないが項目名がある場合に分けられる．

1. 項目名あり
   - 1 行目に項目名がある
   - 1 列目に項目名がある
   - 上記のどちらでもないが項目名がある
2. 項目名なし

図 1 には，有価証券報告書に含まれるさまざまな形式の表の例を示す．

図 1　有価証券報告書に含まれるさまざまな表の例

Q1-4 分析対象の表の条件とは

分析対象の表は「1 行目に項目名が含まれる表」とした．具体的には，下記の条件に当てはまる場合には**分析対象から除外する**．

- 1 行目に数値のみの表現が含まれる
- 1 行目に空欄（欠損値）が 2 つ以上含まれる
- 2 行以下，あるいは，2 列以下の表
- 2 行 2 列以降の要素で，数値の正規表現にマッチしない

その結果，**分析対象の表は 35,263** となった．

Q1-5 どのような項目名が存在するのか

分析対象の表に含まれる**項目の総数は 110,754** であり，**項目の異なり数は 7,086** であった．表 2 に有価証券

3)　有価証券報告書に誤りがあった場合には，訂正報告書を提出する必要がある．

報告書に含まれる表の項目名と出現回数の一部を示す．表 2 に示した通り，「区分」「種類」「銘柄」などの出現回数が多いことがわかる．

表 2　有価証券報告書に含まれる表の項目名と出現回数

項目名	出現回数
区分	3,543
種類	2,460
銘柄	1,941
合計	1,675
前事業年度	1,164
当事業年度	1,164
...	...

Q1-6 項目名に表記揺れがあるのか

表の項目の表記揺れは，項目名に単位が付与されたり，注記が付与されることにより，数多く存在する．例えば，「時価」と「時価（百万円）」のように単位が付与されることによる表記揺れがあった．また，包含関係による難しさもある．例えば，「土地，建物 及び構築物」「土地」「建物及び構築物」などの項目名がそれぞれ存在する．

Q1-7 項目名に曖昧性があるのか

曖昧性とは表の項目名が同じ字面（じづら）でも，異なる意味になるものである．例えば，「区分」「種類」「その他」という項目名は，表によって，意味が異なる．

スタイルシートにより機械判読が困難な例

他にも，有価証券報告書の表の特有の問題として，CSS により階層を識別して記述している問題がある．例えば，勘定科目を記述している表では「(資産の部)」の下の階層に「流動資産」が記述され，その下の階層に「現金及び現金同等物」を記述するために，margin の値を変えてインデントすることで階層を表現している．このような記述は，人間にとって読みやすいが，機械で識別するのが困難である．

4　有価証券報告書に含まれる金額表現の分析

4.1　目的

本分析では，有価証券報告書に含まれる表形式データに含まれる金額表現に対して，本文で説明されている箇所を人手で対応づけることにより，有価証券報告書における構造化データと非構造化データを自動で対応付ける場合の問題点を明らかにする．具体的には，下記の 2 つについて明らかにする．

Q2-1　表形式データに含まれる金額表現に対して，本文で説明されている箇所は，どのぐらい結びつけられるのか．

Q2-2　本文中に記述されている金額表現はどのように使われているのか．

4.2　分析方法

分析対象の有価証券報告書は TOPIX100 から 2 社を選択し，大和ハウス工業の第 78 期 (平成 28 年 4 月 1 日から平成 29 年 3 月 31 日まで)，バンダイナムコホールディングの第 12 期 (平成 28 年 4 月 1 日-平成 29 年 3 月 31 日) である．

4.2.1　分析の流れ

分析の流れは，次の 2 つのステップからなる．

1. 表に含まれる金額表現と本文中に含まれる金額表現を対応づける
2. 対応付けられた金額表現がどのように使われているのか人手で分類する

図 2 に本文に含まれる金額表現と表の金額表現を対応づけ，金額表現を分類する例を示す．本分析では，有価証券報告書から表と本文に分けて，表に含まれる金額に対応する本文をみつける作業を行う．最初に，表の一覧を作成する．表の一覧は図 2 の「（１）連結経営指標等」のようにスプレッドシートに貼り付けている．次に，本文から「円」が含まれる文の一覧を作成する．文の一覧は，「円」が含まれる一文が一行になるように作成しており，図 2 の本文に含まれる No.1 の文であれば，下記のように記述される．

有価証券報告書に含まれる金額表現

以上の結果、当連結会計年度における売上高は**3,512,909 百万円**（前連結会計年度比 10.0 % 増)、

分析では，本文に含まれる金額表現に対応する表の値をみつける作業をする．その際に，金額表現が本文においてどのように使われているのか下記の７つのカテゴリに人手で分類する．

1. 報告：実際の金額の報告をする表現
2. 理由：ある金額の経緯，理由を表す表現
3. 目標・指標：成長目標や今後の指標などを示す表現
4. 比較：比較元の表現
5. 増減：多年度と比較して増減，または，差額
6. 内訳：ある金額の内訳を表す表現
7. その他：単位など

4.3　結果

Q2-1 金額表現をどのぐらい結びつけられるのか

表 3 は，表形式データに含まれる金額表現に対して，本文で説明されている金額表現が対応付けられた数を示したものである．大和ハウス工業は，有価証券報告書のテキストに 116 の金額表現があり，そのうち，63 (=54.3%) の金額表現を表と対応付けることができた．バンダイナムコホールディングは，有価証券報告書のテキストに 181 の金額表現があり，そのうち，67 (=37.0%) の金額表現を表と対応付けることができた．この結果から，本文中に含まれる **4〜5 割程度の金額表現を表と対応付けることができる**ことを確認した．

金額表現のなかには，対応付ける際に，判断が難しい例が存在した．例えば，バンダイナムコホールディングに含まれる下記の表現は，「その他」に該当する箇所が数多く存在したため，対応付けるのが困難であった．

判断が難しい例

「その他」29 百万円として組替えております。

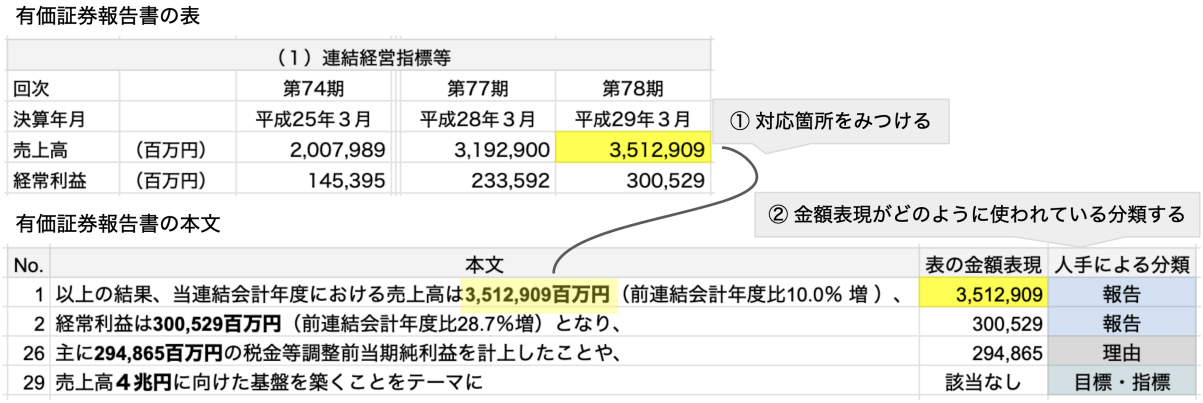

図 2　本文に含まれる金額表現と表の金額表現を対応づけ，どのように使われているか人手で分類する例

表 3　金額表現が対応付けられた数

	大和ハウス工業		バンダイナムコホールディングス	
対応ありの金額表現	63	54.3%	67	37.0%
対応なしの金額表現	53	45.7%	113	62.4%
判断が難しい表現	0	0.0%	1	0.6%
合計	116	100.0%	181	100.0%

Q2-2 金額表現はどのように使われているのか

表 4 は大和ハウス工業，および，バンダイナムコホールディングの有価証券報告書の文中に含まれる金額表現が，どのように使われているのかを人手で分類した結果である．金額表現に付与されたラベルは**「報告」のラベルが最も多く，3〜4 割程度**であった．「報告」や「目標・指標」に使われている金額表現は，表の内容をそのまま記述しており，それほど重要な情報とはいえない．一方，「**理由」「比較」「増減」「内訳」が付与されたラベルは，注目した方が良い情報，あるいは，表のみでは判断しづらい情報**であり，これらのラベルを対象に金額表現を抽出し，表と結びつけることが良いと考えられる．下記にこれらのラベルが付与された例文を示す．

表 4　金額表現に対して人手で分類した結果

	大和ハウス工業		バンダイナムコホールディングス	
報告	49	42.2%	61	33.7%
理由	2	1.7%	16	8.8%
目標・指標	5	4.3%	4	2.2%
比較	7	6.0%	0	0.0%
増減	15	12.9%	13	7.2%
内訳	32	27.6%	57	31.5%
その他	6	5.2%	30	16.6%
合計	116	100.0%	181	100.0%

理由

2,017 億円の親会社株主に帰属する当期純利益を**計上したことによるもの**です。

比較

前連結会計年度末の **3 兆 2,578 億円と比べ** 2,980 億円の増加となりました。

増減

前連結会計年度末の 3 兆 2,578 億円と比べ **2,980 億円の増加**となりました。

内訳

その内訳は、建物及び構築物 1,593 百万円、機械装置及び運搬具 50 百万円、... です。

5　まとめ

本研究では，TOPIX100 の有価証券報告書に含まれる表形式データを対象として，表の項目について調査した．また，表の項目に含まれる金額表現と本文で説明されている箇所を人手で対応づけることにより，有価証券報告書における構造化データと非構造化データを自動で対応付ける場合の問題点を明らかにした．

謝辞

本研究は JSPS 科研費 21H03769 の助成を受けたものである．

参考文献

[1] 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. 人工知能学会論文誌, Vol. 34, No. 5.

[2] 大輔加藤, 五島圭一. 有価証券報告書のテキスト分析 : 経営者による将来見通しの開示と将来業績. 金融研究, Vol. 40, No. 3, pp. 45–75, 07 2021.

[3] 安珠希, 金川一夫. 有価証券報告書の訂正報告書と不適切会計処理に関する予備的分析. 産業経営研究所報, 第 51 号, pp. 1–9, 2019.

[4] 太田裕貴. 有価証券報告書の設備投資情報に関する実証分析. 環境と経営 : 静岡産業大学論集 = Environment and management : journal of Shizuoka Sangyo University, Vol. 23, No. 1, pp. 105–119, 06 2017.

[5] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Madoka Ishioroshi, Masaharu Yoshioka, Tomoyoshi Akiba, Yasuhiro Ogawa, Minoru Sasaki, Kenichi Yokote, Kazuma Kadowaki, Tatsunori Mori, Kenji Araki, Teruko Mitamura, and Satoshi Sekine. Overview of the ntcir-16 qa lab-poliinfo-3 task. *Proceedings of The 16th NTCIR Conference*, 6 2022.