P3-20

有価証券報告書を対象とした機械判読が困難な表構造の分析

奥山 和樹(小樽商科大)　木村 泰知(小樽商科大)

有価証券報告書とは

企業の事業概況や経理の状況など、投資判断に有用な情報を記した文書
本研究で対象とした有価証券報告書には 1文書あたり平均**217**件　合計**8,673**件 の表が含まれていた

多種多様で複雑な表を多く含む文書から情報を機械的に抽出することは**困難**

機械判読が困難な表とは

機械判読が容易な表

	銘柄数(銘柄)	貸借対照表計上額の合計額(百万円)
非上場株式	43	672
非上場株式以外の株式		

表のセルを自然言語文へ変換

Attribute　Header　Data
非上場株式 の 銘柄数 は 43銘柄 である

NTCIR-17 UFOのサブタスクである
TDE(Table Data Extraction) データセットで対象とした表

機械判読が困難な表

	純営業収益							
	2020年3月期	2021年3月期	対前年同期増減率	構成比率				
リテール部門	166,430	169,505	1.8%	36.3%				
ホールセール部門	172,289	215,860	25.3%	46.3%				
グローバル・マーケッツ	121,301	161,730	33.3%	34.7%	28,191	62,777	122.7%	48.8%
グローバル・インベストメント・バンキング								
アセット・マネジメント部門								

①属性を示すセルが結合されていると機械的に処理が難しい

②Dataセルの属性を示すセルが三つ組より多くなってしまう

Attribute　Header　Data
2020年3月期 の ホールセール部門 の グローバル・マーケッツ の 純営業収益 は 121,301百万円 である

NTCIR-17 UFOのサブタスクである
TDE(Table Data Extraction) データセットで**非対象**とした表

表のセルを自然言語の文に変換することが困難な表について有価証券報告書内の
詳細で正確な実態が明らかになっていない

有価証券報告書に含まれる表のセルを自然言語の文に変換することが困難な表について、

機械判読が困難な表の特徴　**機械判読が困難な表が含まれる割合**　を明らかにする

有価証券報告書の分析・機械判読が困難な表の分類・集計

3名のアノテーターによって、表の各セルをAttribute, Header, Data のいずれかのクラスに分類

3種類のクラスを過不足なく抽出することが不可能だったケースの特徴を分類し、集計

門脇らによる研究

- 2人のアノテーターによる分析
- アノテーターの判断の正確性を検証していない(のべ数による集計)

本研究

- 3人のアノテーターによる分析
- 過半数の判断を採用し、より正確に集計
- 3人の判断がどの程度一致しているか Kappa値を算出して検証

分析の結果、有価証券報告書に含まれる機械判読が困難な表は、
5種類に分類することが適当であるとわかった

表のタイプ	Kappa
小見出し行を含む表	0.26
複数の Header や Attribute を持つ表	0.60
5 つ組があれば表現できる表	0.43
結合されたセルを含む表	0.55
空白セルを含む表	0.28
非スカラ値のセルを含む表	0.63
特殊な形の表	0.21
その他	0.14

それぞれのタイプにおけるKappa値
(評価者3人の判定一致度)

機械判読が困難な表の特徴・それぞれの割合

小見出し行を含む表

その行にはDataセルを持たないが、下のHeaderなどを修飾するための行がある

2019		40
当期発生額		
税効果調整前	8,324	26
税効果額	△2,338	△5
税効果調整後		21

1,666件 / 19.2%

複数のHeaderやAttributeをもつセルを含む表

	純営業収益			
	2020年3月期	2021年3月期	対前年同期増減率	構成比率
リテール部門	166,430	169,505	1.8%	36.3%
ホールセール部門	172,2			%
グローバル・マーケッツ	121,3			%
グローバル・インベストメント・バンキング	50,9			%

結合されたセルの存在で、3種類のクラスを過不足なく抽出することができない

銘柄	当事業年度 株式数(株)	当事業年度 貸借対照表計上額(百万円)
ナブテスコ株式会社	3,760,000	19,026
		2,684,200

Headerが2段組になっており
「AttributeのHeader①はValueである」
「AttributeのHeader②はValueである」
と5つ組のセル抽出が必要

2,020件 / 23.3%

不要な空白セルを含む表

	前事業年度(2020年3月31日)
	30.5%
	0.0
算入されない項目	△
	7.6

表の体裁を整える場合などで、Dataをもたない空白のセルが表中に混入している場合がある

1,503件 / 17.3%

非スカラ値の(単一の値ではない)セルを含む表

主な想定シナリオ
① リーマンショック級の世界金融危機が発生し、当社グループ保有資産の価値が大幅に下落する。
② 地政学リスクの顕在化等
① 政府の信用力低下により日本国債が暴落し、当社グループ保有資産の価値が
① 首都直 発生する。また、当社グループの事業継続に重大な影響が生じるほか、当社グループ

Dataセルに複数の情報を含む文章が記されている。

186件 / 2.1%

特殊な形の表

事業区分及び主要製品		
ゲーム&ネットワークサービス		
	家庭用ゲーム機	㈱ソニー・イン
	ソフトウェア	Sony Interactive E
	ネットワークサービス事業	Sony Interactive E
音楽		
音楽制作	販売	㈱ソニー・ミュ
	の収入	Sony Music Enter
音楽出版	楽曲の詞、曲の管理及びライセンス	Sony Music Publi

227件 / 2.6%