言語処理技術による企業報告書分析への応用

Application to the Company Integrated Report Analysis by the Language Processing Technology

黄 海湘[*1]，大坪 史治[*2]
HaiXiang Huang, Fumiharu Otsubo

Email: huang@dokkyo.ac.jp

日本企業では，環境報告書等の非財務報告書が作成・公表されるようになり，およそ 20 年が経過する.公開される内容は企業の環境保全活動をはじめ,社会活動や社会的責任に関する情報に拡張,結合してきた．現在，KPI（Key Performance Indicators）を用いた選別，そして国際的に統合へと歩み始めようとしている．しかし現状では，企業側が何を重視しているかによって，統合報告に開示される情報が異なる．本研究は，言語処理技術を用いて，統合を意識する企業報告書に使用されている単語を抽出し，属性を与え，経年的動向を観察することによって，企業側の意図や重要課題の明確化を試みる.

In Japanese companies, non-financial report will be created and published, about 20 years have passed. But in the present circumstances, information to be disclosed to the integration report by the companies is different. In this paper, we use the morphological analysis, analyzing the part of speech information and words contained in the text of the report. In Addition, we gave an attribute to the words, and observed these changes. By observing the words that are frequently used in the report, we aimed to clarify the key issues and intention of the author side of the report.

*1: 獨協大学経済学部
*2: 和光大学

1. はじめに

日本では，環境報告書，レスポンシブル・ケア報告書，環境社会報告書，サステナビリティ報告書，CSR 報告書などの非財務報告書を作成・公表するようになり，およそ 20 年が経過する．

報告書の内容は環境情報，あるいは化学物質の安全管理に関わる情報からはじまり，拡張と結合していく中で多種多様になってきた．

現在，KPI（Key Performance Indicators）を用いた選別，そして統合（integrate）へと国際的に歩みはじめようとしている．これを牽引する国際統合報告委員会（IIRC：International Integrated Reporting Committee）は，2020 年までに，ビジネスモデルを中核に財務情報と非財務情報を統合した報告書へ移行するビジョンを提唱している[3],[4],[14]．

統合報告が注目を集める背景には，CSR と収益性のリンケージ，サステナビリティとビジネスモデル，ESG（Environmental, Social, and Governance）情報に対する株主・投資家の関心の高まりなど，企業実態を判定するうえで財務情報と非財務情報を関連視する必要性があるとの認識が定着しつつある．アメリカで最初の統合報告書と称される UTC 社のアニュアルレポートでは，「企業の責任（responsibility）と利益（profitability）は不可分である」と統合報告の意義について示唆している[5]．

しかし現状では，報告書の作成が任意であるため，作者側の都合によって統合報告に開示する情報が異 なって，報告書の内容が様々である．

従って，統合報告書で開示すべき財務情報と非財務情報の画定，アニュアルレポート等との一元化，非財務情報開示の制度化，非財務情報の財務情報化などといった情報開示のあり方を議論するために，既存の統合報告を意識した報告書の性質を明らかにして，非財務報告書に掲載されていた情報が統合報告の進行により，いかなる影響を受けているのかについて考察すべきである．

そこで，本研究では，統合報告を意識する 34 社[1]の報告書（総冊数：135 冊程度）のテキスト化を行い，言語処理技術を応用し，テキストマイニングを活用して分析する．各報告書に使用されている単語とそれらの出現回数を観測し，さらに出現した単語に属性を与え，経年的動向を観察する．報告書で頻繁に使用される単語とその単語の経年的な変化を属性別に観察することで，作成者である企業側の意図，報告書の性質及び重要課題を明確化することができると考えられる．

報告書の中から情報を探し出すには，文書を読み，要点をまとめる必要がある．情報要約の手段として，文書内容を特徴付ける重要な単語を見つけ出す作業がある．人手によりこの作業を行うには膨大な時間と手間がかかる．さらに，単語の選定には個人の主観が入りやすく，抽出漏れのようなミスも考えられる．

一方，情報検索システムでは大量のテキスト文書を対象にして，特定な語句を探索するために言語処理技術を利用している[8]．計算機を用いて人手にかかる負担の軽減を計り，単語抽出に個人的な主観が入ることも防げることができるため，文書中の単語に着目する今回の報告書分析も言語処理技術の応用が可能であると考えられる．

言語処理技術を情報検索システムに応用する研究は古くから存在している[6]．また，テキストマイニングの手法は様々な分野の研究で使用されている[9],[10],[12]．企業報告書に関する研究においては，財務報告書から企業行動を把握する試みや会計史研究への応用がある[1],[2],[7],[11]．しかし，我々が調べたところでは，本研究のように企業を横断して，経年分析に応用する試みはまだ存在しない．

以下 2 章では，分析手法について説明する．次に 3 章では実際の結果を考察し，言語処理技術を利用する分析手法の有効性を考察する．

2. 提案する分析手法

2.1 概要

言語処理技術を活用してテキストマイニングを行うためには，まず，①分析対象である報告書を PDF 形式からテキストデータに変換する．次に，②テキストデータに対して，言語処理技術の基礎である形態素解析を行う．最後に，③形態素解析で得られた単語と単語の品詞情報を利用して名詞を抽出し，分析用の単語リストが出力する．

図 1 は提案する分析手法の概要を示している．以下，手順に沿って分析手法の詳細を説明する．

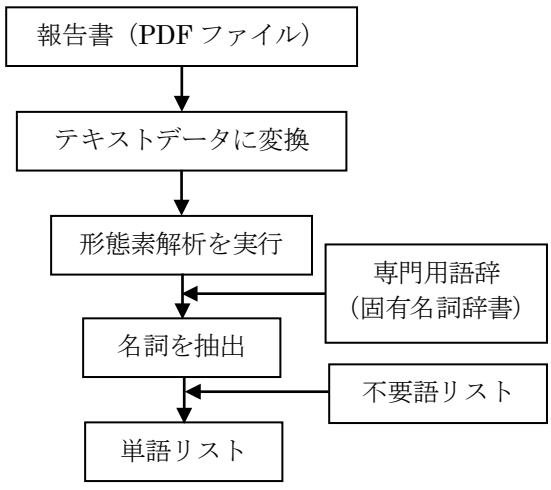

図 1 分析手法の概要

[1] 本論文では，過去に公表していた環境報告書等の作成を取りやめ，アニュアルレポート等に一元化する企業を「統合報告を意識する企業」と呼ぶ．この条件で，我々が収集していた 1994 年～2011 年までの 1,331 組織の各種報告書から「統合報告を意識する企業」34 社の報告書を分析対象に選定した．

2.2 テキストデータ変換

報告書から重要な単語を抽出するためには，対象となる報告書を分析可能な形式であるテキストデータに変換する必要がある．

現在，極稀な状況を除いて，公開されている報告書は主に PDF 形式となっている．これはコンピューター環境を意識することなく報告書を閲覧するためである．

フリーソフト Adobe Reader の「高度検索機能」を利用して，特定の単語の出現回数と所在を検索できる．しかし，この方法では先に検索用単語を予測する必要があり，人手で操作しなければならないので，大量文書の分析に不向きである．従って，各報告書を文字レベルのテキストデータに変換し，分析可能な状態に加工する必要がある．

PDF ファイル内のテキストデータを抽出するために Adobe Acrobat をはじめ，様々な商用とフリーなソフトウェアがある．しかし，これらのソフトは一つ，あるいは少量の PDF ファイルを対象に作られているため，今回のように大量の PDF ファイルを一括してテキストデータに変換するには不向きである．

従って，本研究では，Linux の環境で pdftotext コマンドを利用して変換用プログラムを自作し，報告書の PDF ファイルに含まれているテキストデータを抽出する．ただし，PDF ファイル内の画像データに関して変換できない部分がある．

2.3 形態素解析

変換したテキストファイルに対して，言語分析技術の基礎技術である形態素解析を使用して，各報告書中の文章に含まれている単語とそれぞれの品詞情報を解析する．

形態素解析（morphological analysis）とは，コンピューターで文中の意味を担う最小の言語要素（形態素）を自動的に制定する処理である[13]．例えば，「環境にやさしい技術の開発と普及を促進する．」という文には，図２のように単語ごとに判別され，さらに品詞情報などが出力される．

本研究では，フリーツールの MeCab[2]を使用して形態素解析を行っている．MeCab は他の形態素解析用ソフトより解析速度が速く，辞書の追加なども容易である．

```
環境　名詞,一般,*,*,*,*,環境,カンキョウ,カンキョー
に　助詞,格助詞,一般,*,*,*,に,ニ,ニ
やさしい　形容詞,自立,*,*,形容詞・イ段,基本形,やさしい,
技術　名詞,一般,*,*,*,*,技術,ギジュツ,ギジュツ
の　助詞,連体化,*,*,*,*,の,ノ,ノ
開発　名詞,サ変接続,*,*,*,*,開発,カイハツ,カイハツ
と　助詞,並立助詞,*,*,*,*,と,ト,ト
普及　名詞,サ変接続,*,*,*,*,普及,フキュウ,フキュー
を　助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
促進　名詞,サ変接続,*,*,*,*,促進,ソクシン,ソクシン
する　動詞,自立,*,*,サ変・スル,基本形,する,スル,スル
。　記号,句点,*,*,*,*,。,。,。
```

図２ 形態素解析の例

[2] http://code.google.com/p/mecab/

2.4 名詞抽出と単語リスト

形態素解析により報告書に含まれる単語とその単語の品詞情報が得られる．

しかし，形態素解析の過程で使用する辞書によって単語の過分割問題が存在する．この問題を解決するために，図１で示したように専門用語辞書（固有名詞辞書）を作成し，抽出結果にフィルタリングをかける．

例えば，文の中に「環境会計」という単語が含まれている場合，専門用語辞書を利用しなければ，「環境」と「会計」の二つの単語に分割され抽出されてしまい，「環境会計」は結果に表れないことになる．

また，抽出された単語から「これ」，「それ」，「物」などのような不要語（stop word）に対して，図１で示したように本研究で作成した「不要語リスト」でフィルタリングをかける．

さらに，全ての単語をデータマイニング分析に用いることはない．例えば，動詞，形容詞，副詞，助詞など，文の内容と関係ない単語を分析前に除く必要がある．本研究では，名詞以外の品詞を抽出対象から除外している．

最後に，分析対象となる単語リストを出力する．リスト中の各単語とその出現回数をカウントし，報告書の意図や重要事項について観察を行う．

3. 考察

3.1 データの分布

34 社の報告書から抽出した単語の数（延べ数）は 45,605 語である．図３は全単語の出現回数によるヒストグラムである．縦軸の左側は単語数，右側は単語の累積割合を示し，横軸は単語の出現回数に関するデータ区分を示している．

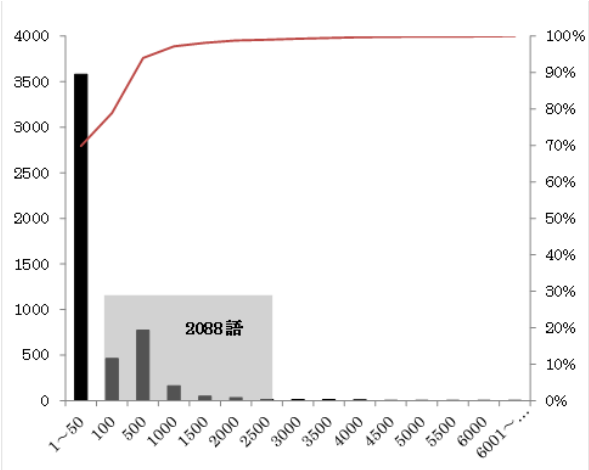

図３ 単語出現回数の度数分布図

図３の中で，一番左側は出現回数が 1 から 50 までの下位層の単語数（異なり語数）を示し，全部で 3,580 語がある．抽出した全単語数（異なり語数）の 69.87%を占める．このデータ区分に含まれた単語はほとんど意味を持たない単語である．

図３の中で，灰色の枠に囲まれている中位層の

単語（出現回数 51～2500 回，単語数 2,088 語，全体の 29.41%）では，重要な単語が多く含まれている．特に，専門用語はこの中位層に埋没している．今後は，この中位層に注視して観察を行い，分析の充実を図りたい．

残り出現回数の多い上位層の単語（出現回数 2,501 回以上，単語数 40 語，全体の 0.72%）は，一般的に共通して多く使用される単語の群であり，あらかじめ想定される単語が多く存在する．

図 4 は抽出した単語の中で，出現回数上位 60 位までの単語リストである．

図 4 を見ると，上位 60 位までの多くは汎用性の高い単語であり，想定範囲内の結果である．

例えば，「事業」，「環境」，「技術」，「情報」，「社会」，「企業」，「体制」などは，単体で使用されるよりも他の単語と結びついて複合語として一つの意味を形成する．例えば，「環境」は，「企業環境」，「事業環境」，「市場環境」，「職場環境」，「環境問題」，「環境破壊」と使用されることが多い．

前述した通り，このような複合語の過分割を想定して事前に専門語辞書と固有名詞辞書を作成している．しかし，全てをカバーするには限界がある．現段階では，事象を眺めながら辞書を精緻化していく作業を繰り返し行うしかない．

順位	語句	回数	順位	語句	回数	順位	語句	回数
1	事業	17449	21	市場	3571	41	拠点	2498
2	環境	15356	22	売上高	3566	42	コンプライアンス	2468
3	グループ	11943	23	目標	3529	43	リスク	2425
4	製品	7681	24	株主	3300	44	お客	2413
5	技術	5724	25	子会社	3288	45	中国	2203
6	情報	5692	26	お客様	3285	46	当期純利益	2109
7	社会	5637	27	エネルギー	3266	47	戦略	2097
8	企業	5431	28	世界	3175	48	課題	2076
9	工場	4774	29	自動車	3102	49	社長	2062
10	体制	4551	30	部門	3080	50	米国	2018
11	日本	4409	31	従業員	3067	51	欧州	1955
12	地域	4254	32	状況	3055	52	有価証券	1921
13	社員	4205	33	グローバル	2966	53	効率	1920
14	取締役	4156	34	方針	2954	54	財務	1910
15	商品	4008	35	制度	2933	55	資源	1894
16	海外	3730	36	基準	2840	56	実績	1848
17	CO2	3710	37	部品	2818	57	効果	1786
18	品質	3696	38	廃棄物	2667	58	責任	1772
19	国内	3668	39	役員	2652	59	費用	1712
20	CSR	3605	40	資産	2628	60	米ドル	1681

図 4 出現回数上位 60 位までの単語リスト

3.2　属性と経年分析

繰り返し使用される単語は報告書の内容を表す重要なキーワードとしての側面があり，作成者が何を重要項目に挙げているのか，あるいは強調しているのかについて，概括的ではあるものの把握することができる．

しかし，単語の出現回数だけでは，作成者側が挙げる重要事項（単語）の変化については明らかにすることはできない．

図 4 で示した出現回数上位 60 位までの単語リストを概観すると，事業関連，財務関連，製品関連，環境関連など多様な領域の単語が混在している．これは，環境報告書から持続可能性報告書，CSR 報告書，さらにはアニュアルレポートへと開示する情報範囲の拡大と情報量の集約を繰り返す過程で異なる領域の情報が錯綜していることから生じている．

作成者側が挙げる重要事項（単語）の変化を経年的に分析するために，抽出した単語に領域区分（属性）を与え，カテゴリごとに整理することによって，重要事項の変化をより鮮明に把握でき，膨大な結果情報をより効率的に要約できると考えられる．

本研究では，8 種類の属性を設定した．

- ステークホルダー属性
  例：株主，投資家，経営者，従業員，顧客，消費者など
- 地域属性
  例：東京，関西，米国，中国，アジア，EU，中東など
- テクニカル属性
  例：環境管理会計，コンプライアンス，エコ効率性，MFCA，JEPIX，LIME，CSR 会計など
- プロダクト属性
  例：エコプロダクト，環境技術，トップランナー，省エネ設計，DFE など
- 環境負荷インベントリ属性
  例：CO2，NOx，SOx，COD，VOC，煤塵，PRTR 対象物質，排水，排熱，廃棄物など
- 財務属性
  例：売上高，当期純利益，経常利益，営業利益，資産，負債，内部留保，ROA，ROE，キャッシュフロー，包括利益など
- 一般用語とその他属性
  例：グローバル，不況，景気，需要，供給，デフレ，インフレなど

本研究では，各属性の中でステークホルダー属性にクローズアップし，抽出された単語の出現回数を経年的に考察する．また，財務属性に対して個別の事例を挙げて考察する．

ステークホルダー属性

表 1 は，数多く抽出したステークホルダー属性の単語のうち主要なステークホルダー及び関心の高いステークホルダーに限定して，2006 年から 20011 年までの経年変化を示している．中には，2010 年と 2011 年のデータはアニュアルレポートから得られている．表中に網掛けした数字は，それぞれの単語において出現回数が上位 2 位であることを示している．

表 1 ステークホルダー属性の経年変化

単語	2006	2007	2008	2009	2010	2011
取締役	18	16	10	8	27	26
株主	3	3	5	6	32	32
役員	15	26	17	12	33	31
親会社	0	0	9	0	14	14
子会社	0	1	7	2	41	32
支店	8	15	11	16	1	1
研究所	37	41	55	42	16	21
従業員	44	56	70	84	21	16
社員	10	25	22	48	13	11
社長	20	6	13	4	23	11
投資家	0	0	2	2	1	4
地域社会	10	6	9	14	3	3
銀行	0	0	0	0	5	5
顧客	7	15	18	32	16	13
お客様	11	2	3	2	0	0
家族	9	10	15	10	7	1
企業市民	4	3	4	6	5	0

表 1 をみると, 2010 年から,「取締役」,「株主」,「親会社」,「子会社」,「役員」といった内部のトップ・マネジメント層や外部の財務的持分関係者に関する単語が増え，報告書における出現回数も急増している．また，出現回数は少ないものの,「銀行」,「投資家」など新たなステークホルダーの出現も見られる.

一方で,「支店」,「研究所」,「従業員」,「社員」,「地域社会」,「顧客」,「お客様」,「家族」,「企業市民」の出現回数が減少する傾向にある.

アニュアルレポートは，主に株主・投資家に向けてディスクローズされる報告書であり，事業内容，財務状況，経営戦略，経営ビジョン等の概要を報告するものである．従って，一般報告書からアニュアルレポートに変更したことにより，株主やトップ・マネジメント層のステークホルダーの出現回数が多くなるのは当然の帰結である. 2010 年以降では，明らかに財務的持分関係者にウエイトを置いている意図が推察される.

財務属性

財務属性は，企業の経営や会計に関連するプライベートエリアに関わる単語に付与する属性である.

事例に挙げる企業は，2005 年の合併により商号変更した企業である．合併前から両社共に非財務報告書の作成経験がある．2010 年より，CSR 報告書を中止し，アニュアルレポートの一元化を図っている.

表 2 は 2011 年のアニュアルレポートに含まれた財務属性の単語を示している.

表 2 事例企業の 2011 年報告書に含まれた財務属性

単語	出現回数
戦略	56
売上高	41
売上	40
アニュアルレポート	33
収益	29
当期純利益	21
財務	21
経営計画	18
包括利益	15
有価証券	15
現金	15
純資産	15
営業利益	14
資産	12
キャッシュフロー	11
借入金	9
利益	8

2006 年の社会・環境報告書において，抽出された財務属性の単語とそれぞれの出現回数は，収益（4 回），費用（4 回），資金（2 回）のみであった．表 2 の結果と比較すると，2011 年のアニュアルレポートでは財務属性の単語が格段に増えていることが分かる.

4. おわりに

本研究では，統合報告を意識する 34 社の報告書を対象に形態素解析を利用し，データマイニングを活用した分析を行った.

結果として，対象報告書 135 冊から約 50,000 語を抽出した．全単語を横断的に見ると，アニュアルレポート等の統合報告を意識した報告書において非財務情報の量は後退し，経年的に財務属性の単語が増加していることが分かった.

また, ステークホルダー属性では,「株主」や「投資家」といった財務的持分関係者にウエイトを置き，CSR 報告書では多く取り上げられていた「従業員」や「社員」,「地域社会」,「支店」,「研究所」,「顧客」,「お客様」,「家族」,「企業市民」などのステークホルダーは減少傾向にあった．これは，アニュアルレポートとしての機能や役割を保ちながら，CSR 報告書に掲載していた非財務情報の一部を組み込んだと解釈できよう.

統合報告を意識した報告書における非財務情報は，情報利用者に効果的かつ視覚的にインパクトを与える情報に絞られている．ただし，様々な企業や組織の中には，アニュアルレポートに非財務情報を積極的に組み込んでいく一方で，より詳細

な情報については，CSR 報告書等で提供する工夫をしている．統合報告書より詳細な情報を提供する補完的役割として，非財務報告書は存続すべきだと考えられる．例えば，環境パフォーマンスに関わるインベントリーデータは，あらゆる分析や検証の基礎情報として肝要なのである．

以上の分析結果から，言語処理技術を利用することによって各報告書の定量的な評価と比較に結びつけることで成果があると言える．また，報告書分析に新たな検証視点と手段を提供した．

今後の課題として，単語を抽出する際の過分割問題を解決するために専門辞書を充実させることである．単語抽出精度の向上には，辞書の整備が重要である．使用する形態素解析ツールで，辞書に登録されていない専門用語の辞書登録が不可欠である．

また，同じ意味の単語は同義語辞書を作成する必要がある．例えば，「MFCA」と「マテリアルフローコスト会計」のような同義語を統一させてまとめて集計することが必要である．

さらに，単純な単語の頻度分析だけではなく，その他のデータマイニング手法の適用である．例えば，係り受け分析による単語間の共起関係を明確して分析に活用する手法や，単語の出現頻度からクラスタへ分ける分析手法や，分析対象文書の内容から分類して傾向を分析する手法などが挙げられる．

参考文献

(1) Cho, H., R. Roberts, and D. Patten, “The language of US corporate environmental disclosure”, Accounting, Organizations and Society, 35, (2010)
(2) Don Tapscott, Robert G. Eccles, Michael P. Krzus, “One Report: Integrated Reporting for a Sustainable Strategy”, Wiley,（2010）
(3) International Integrated Reporting Committee, “Consultation draft of the international <IR> Framework, IIRC draft Paper”, (2013)
(4) International Integrated Reporting Committee, “Towards Integrated Reporting, IIRC discussion Paper”, (2011)
(5) UTC, “Annual Report: 2008 Financial and Corporate Responsibility Performance, titled “More with less””,（2009）
(6) 奥村学, “自然言語処理の基礎” コロナ社,（2010）
(7) 川上直哉，中條良美，朴恩芝，前田利之, “テキストマイニングによる環境コスト支出要因の時系列分析,” 2013 年経営情報学会秋季全国研究発表大会,（2013）
(8) 北研二, 津田和彦，獅々堀正幹, “情報検索アルゴリズム,” 共立出版,（2002）
(9) 金明哲, “テキストデータの統計科学入門,” 岩波書店,（2009）
(10) 金明哲，村上征勝，永田昌明，大津起夫，山西健司, “言語と心理の統計-ことばと行動の確率モデルによる分析,” 岩波書店,（2003）
(11) 中野常男，橋本武久，清水泰洋, “わが国における会計史研究の過去と現在：テキストマイニングによる一試論,” 國民經濟雜誌 200(4)，1-23,（2009）
(12) 那須川哲哉, “テキストマイニングを使う技術/作る技術-基礎技術と適用事例から導く本質と活用法,” 東京電機大学出版局,（2006）
(13) 田中穂積監修, “自然言語処理－基礎と応用－”電子情報通信学会,（1999）
(14) 湯田雅夫, “財務情報と非財務情報を統合する統合報告の動向”『松蔭論叢』(2013)

(2013 年 9 月 30 日受付)
(2013 年 12 月 18 日採録)