日興リサーチセンター
SMBC日興証券グループ

テキストマイニングによる有価証券報告書の開示変更への対応に関する分析

Research Report
2018 年 9 月

投資工学研究所
主任研究員
田中　良典

要　約

金融審議会「ディスクロージャーワーキング・グループ」（平成 27 年度）において、会社法、金融商品取引法、証券取引所規則に基づく各開示書類を通じた企業情報開示のあり方が審議され、記載の整理や合理化等の見直しの方向性が報告された。これを踏まえ平成 29 年 10 月 24 日に「企業内容等の開示に関する内閣府令」等の改正案が発表され、有価証券報告書の「財政状態、経営成績及びキャッシュ・フローの状況の分析」に「業績等の概要」、「生産、受注及び販売の状況」が「経営者による財政状態、経営成績及びキャッシュ・フローの状況の分析」として統合され、この開示の変更は平成 30 年 3 月 31 日以後に終了する事業年度から適用された。ところで、昨今の AI 技術の発達に伴い決算短信やニュース等に対して、テキストマイニング技術を用いた非財務情報の研究が活発に行われている。有価証券報告書に関しても研究がなされているものの、先般の内閣府令の改正に伴う記載内容や記載箇所の変更は、これまでの枠組みでのテキスト分析にその企業の業績や方針等の変化とは無関係な変化をもたらすことになる。そこで本稿では、開示項目の平仄を合わせることを目的として、開示変更前後の有価証券報告書を用いて、「経営者による財政状態、経営成績及びキャッシュ・フローの状況の分析」の文章が、変更前のどの項目に属するのかを 5 つの方法を用いて分析し、その精度を比較した。また、各項目の平仄を合わせた上で、開示変更前後の各項目の記述量の変化を見ると、全体として減少していることが分かった。このことは、記載の整理・合理化が進んだことを示唆しているのかもしれない。

目次

1. はじめに
2. 方法
 2.1　対象データ
 2.2　分析方法
 2.3　分析単位
 2.4　評価方法
3. 結果
4. 考察
5. まとめ

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

1. はじめに

「日本再興戦略改定 2015」において、政府は廻りだした経済の循環を維持していくステージをアベノミクス第二ステージと位置付け、「未来投資による生産性革命の実現」と「ローカル・アベノミクスの推進」を打ち出した[1]。「未来投資による生産性革命の実現」の具体的な施策の 1 つとして、「株主への情報開示促進」が掲げられた。これは、会社法、金融商品取引法、証券取引所規則に基づく開示を検証し、統合的な企業情報開示のあり方を検討する[2]というもので、麻生金融担当大臣の諮問[3]のもと、金融庁、法務省等をはじめ東証、日証協、経団連、公認会計士、弁護士、学識経験者、事業会社、金融機関など 20 余名からなる金融審議会「ディスクロージャーワーキング・グループ」（平成 27 年度）が設置された。平成 27 年 11 月 10 日から計 5 回の会合の中で企業情報開示のあり方等が検討及び審議され、翌年 4 月 18 日に報告書が公表された。報告書[4]では、各法律・規則に基づく開示書類のあり方について、それぞれの目的・役割を整理し、記載の整理や合理化等の見直しの方向性が示された。中でも金融商品取引法に基づく有価証券報告書については、内容が重複している部分の合理化を図りつつも、本来意図されていた開示を充実させる目的で、「財政状態、経営成績及びキャッシュ・フローの状況の分析」、「業績等の概要」、「生産、受注及び販売の状況」を統合することが提案された。これを踏まえ平成 29 年 10 月 24 日に金融庁から「企業内容等の開示に関する内閣府令」等の改正案が発表[5]され、平成 30 年 3 月 31 日以後に終了する事業年度に係る有価証券報告書等から適用された[6]。既にこの変更が反映された 2018 年 3 月期の有価証券報告書が公表されており、前述 3 項目は、「経営者による財政状態、経営成績及びキャッシュ・フローの状況の分析」として記載されている。

有価証券報告書は、先の報告書によれば「投資者の投資判断に必要かつ重要な情報を提供することで、金融商品取引等の公正を確保し、有価証券の流通を円滑にするほか、資本市場の機能の十全な発揮による金融商品等の公正な価格形成等を図り、もって投資者保護に資するもの」で、投資判断の根拠として必要不可欠な企業開示情報の 1 つである。また、有価証券報告書の「事業等の状況」には、業績等の数字情報以外にも、その企業の考えるリスクや課題等に関する重要な情報がテキスト情報として記載されている。近年の機械学習やテキストマイニング技術の発展、それらの金融市場への応用（例えば、和泉他

---

[1] 「日本再興戦略」改訂 2015 ー未来への投資・生産性革命ー　平成 27 年 6 月 30 日
http://www.kantei.go.jp/jp/singi/keizaisaisei/pdf/dai1jp.pdf

[2] 日本再興戦略改訂 2015 これまでの成果と新たな改革　2016 年 4 月
http://www.kantei.go.jp/jp/singi/keizaisaisei/pdf/new_seika_torikumi.pdf

[3] 金融審議会「ディスクロージャーワーキング・グループ」（第 1 回）資料 2「諮問事項」
https://www.fsa.go.jp/singi/singi_kinyu/disclose_wg/siryou/20151110/02.pdf

[4] 金融審議会　ディスクロージャーワーキング・グループ報告　－建設的な対話の促進に向けて－　平成 28 年 4 月 18 日
https://www.fsa.go.jp/singi/singi_kinyu/tosin/20160418-1/01.pdf

[5] 「企業内容等の開示に関する内閣府令」等の改正案の公表について　金融庁　平成 29 年 10 月 24 日
https://www.fsa.go.jp/news/29/sonota/20171024.html

[6] その他、それまで決算短信の記載内容とされていた「経営方針」が、有価証券報告書の「対処すべき課題」の記載内容に追加され「経営方針、経営環境及び対処すべき課題等」に変更されている（詳細は「企業内容等の開示に関する内閣府令」等の改正案に対するパブリックコメントの結果等について　https://www.fsa.go.jp/news/28/syouken/20170214-1.html 等を参照）。この変更に関する分析は別紙に譲りたい。

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

2011、坂地他 2015）に伴い、これらの非財務情報に関しても、因果関係等の特徴を含む文の選択的な抽出や、テキスト情報の数値化と外部変数との関係抽出など様々な分析が試みられている（佐藤他 2017、小寺他 2018、佐久間他 2018）。しかしながら、先般の内閣府令の改正に伴う記載内容や記載箇所の変更は、これまでの枠組みでの特徴文の数やテキスト情報の数値化にその企業の業績や方針等の変化とは無関係な変化をもたらすことになる。つまり、過去に遡って非財務情報から企業の状況変化を捉えるには、法令の変更などの外部要因によるテキストデータの構成変化を除去する必要がある。

そこで本研究では、2017 年 3 月期と 2018 年 3 月期の有価証券報告書を対象に、内閣府令対応前後の「財政状態、経営成績及びキャッシュ・フローの状況の分析」、「業績等の概要」、「生産、受注及び販売の状況」、「経営者による財政状態、経営成績及びキャッシュ・フローの状況の分析」に注目し、変更前を基準として変更後の各文が、変更前のどの項目に属するのかを同定することを試みた。

2. 方法

2.1 対象データ

EDINET から 2017 年 3 月期と 2018 年 3 月期の XBRL ベースの有価証券報告書をダウンロードし以後の分析に用いた。2018 年 3 月期の有価証券報告書に関しては、同年 4 月 1 日から 6 月 30 日までに情報が取得できるものを使用した。その後、2017 年 3 月期の各企業の XBRL ファイルからは、「業績等の概要」、「生産、受注及び販売の状況」、「財政状態、経営成績及びキャッシュ・フローの状況の分析」を抽出し、2018 年 3 月期の各企業の XBRL ファイルからは、「経営者による財政状態、経営成績及びキャッシュ・フローの状況の分析」を抽出した。

2.2 分析方法

本稿では文章の類似度を測る方法として cos 距離を用いた。cos 距離は、 2 つの文章について n 次元のベクトル$\vec{a}$、 $\vec{b}$とし、この 2 つのベクトルの距離を n 次元空間における 2 つのベクトルの挟む角度として考える方法である（山内 2017）。ベクトルの挟む角度は内積を使って計算出来る。

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\|\vec{b}\|} \qquad \vec{a} = (a_1, a_2, \dots a_n)、\ \vec{b} = (b_1, b_2, \dots b_n)$$

角度θが小さくなると、2 つのベクトルが近くなり(距離が短くなり)、角度θが 0 になるとベクトルは重なる。この時 *cosθ*は 1 であり、つまり、*cosθ*が 1 に近いほど文章の類似性が高いことになる。文章をベクトル化するにあたっては、対象とした文章の中で、BoW（Bag of Words。いくつかの注目した単語がそれぞれ何回出てきたか。単語の出現頻度）を用いた。以下に簡単な例を示す。

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

文章１：リンゴとミカンと犬と猫が好きです。でも１番はリンゴです。
文章２：リンゴとミカンとメロンと犬が好きです。やはり１番はリンゴです。
文章３：リンゴとメロンが好きです。メロンが１番です。

文章のベクトル化は、単に注目した７つの単語（ここでは名詞）の頻度を求めるだけである。各文章における単語の頻度を求め、各文章を７次元のベクトルに変換する。

［リンゴ、ミカン、メロン、犬、猫、好き、１番］
文章１：[2、1、0、1、1、1、1]
文章２：[2、1、1、1、0、1、1]
文章３：[1、0、2、0、0、1、1]

あとは３つの文章から２つを選びベクトル化された文章の cos 距離をそれぞれ求めればよい。

文章１と文章２の cos 距離[7]：0.88
文章２と文章３の cos 距離：0.75
文章３と文章１の cos 距離：0.50

この場合、文章１と文章２の類似度が最も高いことになる。

上記では、文章内にある単語が何回出現するかのみを考慮し、その単語の用いられやすさや単語の語順などを考慮していない。そこでこの点も考慮するために TF-IDF を用いた文章のベクトル化と形態素 N-gram を用いた文章のベクトル化を行った。TF-IDF は TF (Term Frequency)と IDF (Inverse Document Frequency)の積を取る文章のベクトル化で、それぞれ以下のように表す。

TF: その文章 d 中に出てきた単語 i の頻度 fi / その文章 d 中に出現した単語の総数 F
IDF: 単語 i が全文章 D のなかでいくつの文章に出てきたかの割合の逆数の log を取った値

TF は文章 d 中にどれくらいその単語 i が出てきているかを表している。しかしながら、多く出現する単語を単に重要だとすると、他の文章でも万遍なく出現する単語（上記事例では、”1 番”など）が重要な単語になってしまう。これを調整するのが IDF であり、どの文章でも万遍なく出てくる単語の IDF

[7] 例えば、$\vec{a}$ を文章１、$\vec{b}$ を文章２とすると、$\|\vec{a}\| = \sqrt{2^2+1^2+0^2+1^2+1^2+1^2+1^2} = 3$ となる。同様に $\|\vec{b}\| = \sqrt{2^2+1^2+1^2+1^2+0^2+1^2+1^2} = 3$となる。また、$\vec{a}\cdot\vec{b} = 2\cdot2+1\cdot1+0\cdot1+1\cdot1+1\cdot0+1\cdot1+1\cdot1 = 8$となる。よって文章１と文章２の cos 距離は、$8/3\cdot3 \fallingdotseq 0.88$となる。

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

は 0 に近い値を取るようになっている。TF-IDF は TF と IDF の積を取ることで、文章内のある単語の頻度が高く TF が高いとしても、その単語が他の文章でも見られる場合 IDF が低くなることでその単語の重要度を下げることになる。逆を言えば、ある文章には頻繁に出現するが、他の文章にはあまり見られないような単語について TF-IDF の値は高くなる。つまり、TF-IDF は各文章の特徴語を数値化していると言える。TF-IDF による文章のベクトル化を先の例で行うと以下のようになる。

［リンゴ、ミカン、メロン、犬、猫、好き、1 番］
文章 1[8]：［0、0.025、0、0.025、0.068、0、0］
文章 2 ：［0、0.025、0.025、0.025、0、0、0］
文章 3：［0、0、0.070、0、0、0、0］

これをもとに TF-IDF ベースの cos 距離を計算してみると、

文章 1 と文章 2 の cos 距離[9]：0.37
文章 2 と文章 3 の cos 距離：0.57
文章 3 と文章 1 の cos 距離：0

となり、BoW ベースの cos 距離と異なり、文章 2 と文章 3 の類似度が最も高くなる。”リンゴ”や”好き”、”1 番”といった単語はどの文章中にも出てきており、IDF の値は 0 となっている（＝log(3/3)）。また、”メロン”は文章 3 に 2 回出ているものの、文章 1 では出てきておらず、文章 3 における特徴語となっている。文章 1 における”猫”も値が高くなっており、TF-IDF はその文章における特徴語をうまく数値化している様子が分かる。つまり、TF-IDF の cos 距離は、特徴語ベースでの文章の類似度を表していることになり、先の BoW ベースの cos 距離とは異なる結果となることがある。

ところで、文章 1 、 2 において”1 番”は最後から 2 番目に出てくるが、文章 3 では最後に出現している。上記の例では、この順番に関する情報を無視している。そこでこの点を考慮するために形態素 N-gram による文章のベクトル化を行った。日本語は英語と異なり、「リンゴが 1 番好きです。」のように語と語の間に空白を置かない。そのため、先のように頻度を数える際には、その前処理として語と語を切り分ける必要があり、これを形態素解析と呼ぶ。また、切り分けられた 1 つ 1 つを形態素と呼ぶ。テキストマイニングでは、隣り合う文字や語を一纏まりにし、それを新たな「単位」として分析を行うことがある。この N 個の要素の連なりのことを N-gram と呼ぶ（山内 2017）。本稿では、形態素解析

[8] 例えば、文章 1 の tf は、文章 1 内の総単語数 7(=2+1+0+1+1+1+1)で各要素を割ればよい(［2/7、1/7、0/7、1/7、1/7、1/7、1/7］)。一方で、idf は各単語の全文章に対する出現割合を求め(［リンゴ、ミカン、メロン、犬、猫、好き、1 番］= [3/3, 2/3, 2/3, 2/3, 1/3, 3/3, 3/3])、逆数の log をとればよい。最後に両者の積をとることで文章 1 の TF-IDF が計算できる。
[9] 注釈 7 と同様に計算。

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

に MeCab[10]を用い、3 つの形態素の連なりを単位とした形態素トライグラムにより文章をベクトル化した。以下に MeCab による形態素解析とそれに基づいた形態素トライグラムの例を示す。

例文：
リンゴとミカンと犬と猫が好きです。

MeCab を用いた形態素解析の例：

リンゴ	名詞,一般,*,*,*,*,リンゴ,リンゴ,リンゴ
と	助詞,並立助詞,*,*,*,*,と,ト,ト
ミカン	名詞,一般,*,*,*,*,ミカン,ミカン,ミカン
と	助詞,並立助詞,*,*,*,*,と,ト,ト
犬	名詞,一般,*,*,*,*,犬,イヌ,イヌ
と	助詞,並立助詞,*,*,*,*,と,ト,ト
猫	名詞,一般,*,*,*,*,猫,ネコ,ネコ
が	助詞,格助詞,一般,*,*,*,が,ガ,ガ
好き	名詞,形容動詞語幹,*,*,*,*,好き,スキ,スキ
です	助動詞,*,*,*,特殊・デス,基本形,です,デス,デス
。	記号,句点,*,*,*,*,。,。,。

形態素とともに品詞、原形、読み方等の情報が得られる。

形態素トライグラムの例：
リンゴミカン犬、ミカン犬猫、犬猫好き
1 語ずつずらしながら、連続する 3 つの形態素（この場合は名詞のみ）を各々一つの単位としている。これにより例文は語順を考慮した要素を持つ[1, 1, 1]のベクトルにすることが出来る。

本稿では、cos 距離を計算する際、2.1 で抽出した有価証券報告書に対して、1.BoW、2.TF-IDF、3.形態素 N-gram、4.BoW+形態素 N-gram、5.TF-IDF+形態素 N-gram の 5 つのベクトル化を行い、それぞれ cos 距離を算出した。

2.3　分析単位

2017 年 3 月期の有価証券報告書では、「業績等の概要」等はそれぞれ項目立てて述べられており、これらが統合された 2018 年 3 月期の有価証券報告書においても、ある程度これらの項目に沿って述べら

[10] MeCab は京都大学情報学研究科-日本電信電話株式会社コミュニケーション科学基礎研究所　共同研究ユニットプロジェクトを通じて開発されたオープンソースの形態素解析エンジン。http://taku910.github.io/mecab/

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

れていると想定される。また、XBRL ファイルにおいて、一纏まりの文章を区切る空白行を示す部分が散見されたため、解析する際の文章の単位を空白行で区切られる文章（以後、段落とする）とした。機械的に区切ったため、タイトルのみの段落からタイトルを含む 10 行程度の文で構成される段落まで、その粒度はまちまちであった。また、数字は毎年値が変わるため削除し、「生産、受注及び販売の状況」では表の占める割合が高いがこれらを全て削除し、タイトルや注釈のみを残した。分析に用いた形態素は、“名詞”, “動詞”, “形容詞”, “形容動詞”, “副詞”, “連体詞”, “接続詞“である。

2.4　評価方法

まず、時価総額の大きい企業の中から業種が重複しないように有価証券報告書を選び、2018 年 3 月期の「経営者による財政状態、経営成績及びキャッシュ・フローの状況の分析」の新項目に属する各段落が、2017 年 3 月期の「業績等の概要」、「生産、受注及び販売の状況」、「財政状態、経営成績及びキャッシュ・フローの状況の分析」の旧 3 項目のうちどれに該当するかの割当てを人手により行った。このデータは、6 社 246 段落からなり、これを正解データとした。なお、2018 年 3 月期に新出した段落は対象外とした。次に、この 6 社について個別に 2018 年 3 月期の「経営者による財政状態、経営成績及びキャッシュ・フローの状況の分析」に属する各々の段落に対し、2017 年 3 月期の「業績等の概要」、「生産、受注及び販売の状況」、「財政状態、経営成績及びキャッシュ・フローの状況の分析」の全ての段落との cos 距離を算出し、最も cos 距離の高かった組合せからどの項目に属するかを決定した。これを前述の 5 つの方法で行い、正解データと照らし合せることで、それぞれの方法の正解率を算出した。ただし、選ばれた組合せの段落に含まれる文章同士が一致しているかを検証した際、あまり一致していないと判断できる場合や、旧項目の中によりよい文章の候補がある場合が確認された。その時は、割り当てられた項目が偶然正解していたとしても、その根拠となる新旧の段落内の文章が一致していないと判断し、不正解とした。また、一つの段落に旧項目の 2 つの内容が含まれている場合、文字数の多い項目をその段落の正解とした。図表 1 に正解、不正解についての具体的な文例を示す。

図表 1　正解、不正解の文例

	割当の対象となる段落（2018 年 3 月期）	割当の根拠となる段落（2017 年 3 月期）
正解例	触媒・ファイン事業触媒事業では、国内シェアの回復、輸出案件の拡販および顧客との協業体制の維持・拡大等を重点施策として取り組んだ結果、FCC 触媒が伸長したほか、ケミカル触媒および環境保全触媒も好調に推移いたしました。ファイン事業においては、・・・	触媒・ファイン事業触媒事業では、国内シェアの回復、輸出案件の拡販および既存顧客の深耕等を重点施策として取り組んだ結果、FCC 触媒の新規開拓案件やインドネシア向け大口案件を受注したほか、脱硝触媒原料の欧米向け輸出も好調に推移いたしました。ファイン事業においては、・・・ ：（「業績等の概要」）

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

不正解 例 1	(d)営業利益(b)と(c)の結果、営業利益は、前期比 277,802 百万円（27.1%）増の 1,303,801 百万円となりました。	(m)純利益(k)と(l)の結果、純利益は、前期比 916,189 百万円（164.1%）増の 1,474,430 百万円となりました。 ：（「業績等の概要」）
不正解 例 2	④住宅ストック事業住宅ストック部門では、当社施工の戸建・賃貸住宅を所有されているオーナー様に対し、インスペクション（点検・診断）を通じたリレーションの強化や保証期間延長のためのリフォーム提案を強化するなど、受注拡大を図ってきました。さらに、･･･⑤商業施設事業商業施設部門では、テナント企業様の事業戦略に対応した適切な出店計画の提案や、エリアの特性を活かしたバリエーション豊富な企画提案を行ってきました。また、市街地開発や･･･	⑤商業施設事業商業施設部門では、テナント企業様の事業戦略に対応した適切な出店計画の提案や、エリア特性を活かしたバリエーション豊富な企画提案を行うとともに、市街地開発・大型物件への取り組みを強化することで、業容の拡大を図ってきました。特に･･･ ：（「業績等の概要」） ④住宅ストック事業住宅ストック部門では、戸建・賃貸住宅を所有されているオーナー様に向け、インスペクション（点検・診断）を通じたリレーションの強化や、キャンペーンを通じた様々なリフォーム提案を行い、事業を推進してきました。また、･･･ ：（「業績等の概要」）

注）正解例について：割当対象段落（左）が、根拠段落（右）に内容がマッチしていると判断し、「業績等の概要」への割り当てを正解とした。不正解例 1 について：割当対象段落（左）が、根拠段落（右）に内容がマッチしていないと判断し、不正解とした。不正解例 2 について：割当対象段落（左）の後半部分に cos 距離が最も高かった根拠段落（右上）とマッチする部分があり「業績等の概要」に割り当てられた。しかしながら、この他にもマッチすると判断できる根拠段落が存在（右下）し、後者の方が僅かに文字数が多かった。双方の根拠段落とも「業績等の概要」の段落であり、結果としては「業績等の概要」に割り当てられるものの、根拠段落がマッチしていないとし不正解とした。
出所）EDINET データより日興リサーチセンター作成

3. 結果

どの方法も 80%以上の高い正解率を示した(図表 2)。中でも BoW のみ、もしくは BoW と形態素 N-gram の組合せの正解率が最も高く 90%を超えた。一方で、TF-IDF と形態素 N-gram の組合せは TF-IDF のみよりも正解率が低かった。また、2018 年 3 月期に初めて出てきた段落が全段落に対して 10%程度存在した。

図表 2　各手法による同定の正解率と cos 距離の平均値

	正解	不正解	新出段落	合計	正解率	cos距離
BoW	202	18	30	250	0.918	0.81
TF-IDF	200	20	30	250	0.909	0.72
形態素N-gram	179	41	30	250	0.814	0.60
BoW+形態素N-gram	203	17	30	250	0.923	0.75
TF-IDF+形態素N-gram	188	32	30	250	0.855	0.59

注）正解率は、合計値から新出段落を除いた値に対する割合。cos 距離は正解、不正解を合わせた平均。
出所）EDINET データより日興リサーチセンター作成

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

4. 考察

本稿では、いくつかの手法で文章をベクトル化し cos 距離を求めることで、有価証券報告書の開示変更前後の文章の類似度を測り、統合された項目の各文章が旧 3 項目のどれに属するのかを同定した。

cos 距離というシンプルな方法であったが、どれも比較的高い正解率を示した（図表 2）。同様の研究には、企業 web ページを用いた関連企業の抽出(本間他 2015)がある。この研究では、企業の web ページにおける名詞についてエントロピーを考慮した TF-IDF を用い、特徴語に関しての cos 距離によって企業の類似度を測定し、関連企業の抽出について高い精度を達成している。一方で、本稿では BoW の方が TF-IDF よりも高い正解率を示した。これは開示変更の前後で段落ベースの表現自体にはそれほど違いがなく、特徴的な表現が少なかったために、特徴語で距離を測るよりも文章にある言葉を等しく扱って（つまりは単純に頻度を用いて）距離を測った方が適していたと考えられる。また、経済環境など毎年書きぶりが変わるものの、BoW、TF-IDF では頻度のみで語順を考慮していないために、寧ろ表現のゆらぎに柔軟に対応したのかもしれない。そのため形態素 N-gram は、単独でも他との組合せにおいても相対的に正解率が低かったと考えられる。連続する 3 つの語の順番を固定せず、組合せにするなど少し条件を緩和するとよいだろう。ある程度マッチングに近い文章の類似度を測る際には、あまり条件を厳しくしない方法の方が良いのかもしれない。

一方で、平均して 10%程度は不正解であったが、それらの段落にはいくつか特徴があった。業績に関する記述で、「売上高」や「収益」など主語が異なり、それ以降述語部分の「減収」や「増収」の表現がある程度共通する場合や、セグメント別の業績の説明に関して、内容に引っ張られセグメントのタイトルが新旧で一致しない場合があった(図表 1)。また、タイトルに変更が無くても、内容を記述する代わりに別の項目の内容を参照するように促す記述に変更されていると不正解になりやすかった。タイトルや主語、書き出し部分について重み付けをすると改善できると思われる。また、「減収」や「増収」など主語が変わりやすい語の辞書を作成し、係り受け解析によってこれらに係る語を検出し、重みを付与することも考えられる。

新出の文章も 10%程度存在していた。この中には、これまで表で表記していた負債や資産に関して新たに記述したものや、短期的及び長期的資金調達についての記述、経営者の視点として項目を設け、業績に関しての総括や目標達成状況と今後についての記述等があった。本稿では、企業毎に開示変更前後の 2 期分の有価証券報告書のみを比較しているため、これら新出の記述に対応できなかった。また、段落に旧項目の 2 つの内容が含まれている場合、2 つの属性を振ることが出来なかった。分析単位を文単位にし、文ごとに属性を割り当てれば、段落内に複数の属性があっても対応できるだろう。さらに、全企業を対象として文ごとに属性を割り振れば、その企業においては新出の文であっても、他企業において既出の文である場合もあると考えられ、新出の文にもある程度対応できるであろう。

最後に、2814 社の 2018 年 3 月期の有価証券報告書に対して BoW+N-gram を用いて項目を同定し、2017 年 3 月期の 3 項目と比較することでどの程度、各項目の段落数や文数、文字数が変化したかの平均値を示す(図表 3)。これをみると開示変更に伴い、「業績等の概要」が僅かに増えたのに対し、「生産、

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

受注及び販売の状況」、「財政状態、経営成績及びキャッシュ・フローの状況の分析」は減少し全体としても減少した。このことは、開示変更の目的であった記載の整理やその合理化がしっかりと一歩進んだことを表しているのかもしれない。

図表 3　有価証券報告書あたりの 3 項目の構成内容の変化

		業績等の概要	生産、受注及び販売の状況	財政状態、経営成績及びキャッシュ・フローの状況の分析	合計
段落数	2018年3月期	7.8	3.1	6.4	17.2
	2017年3月期	7.3	3.6	6.7	17.5
	⊿	0.5	-0.5	-0.3	-0.3
文数	2018年3月期	32.1	7.1	27.7	66.9
	2017年3月期	31.0	7.9	28.6	67.4
	⊿	1.2	-0.9	-0.8	-0.5
文字数	2018年3月期	2,463	310	1,975	4,748
	2017年3月期	2,443	326	2,029	4,798
	⊿	20	-16	-54	-50

注）2018 年 4 月 1 日から 6 月 30 日で取得した有価証券報告書が対象（N=2814）。2018 年 3 月期、2017 年 3 月期は有価証券報告書あたりの平均値を示す。⊿は平均値の差分。文数は句点の数を示す。
出所）EDINET データより日興リサーチセンター作成

5. まとめ

本稿では、有価証券報告書の開示変更に対し、新項目における各段落が旧 3 項目のどれに該当するのかを、異なる 5 つの文章のベクトル化を行い、cos 距離を用いて分析し比較した。その結果、BoW+形態素 N-gram を用いた分析が最も正解率が高かった。また、この手法を用いて新項目を旧 3 項目に割振り、開示変更前後でのその構成内容を検証したところ、全体の記述量が僅かに減少していた。今後は、文を単位とした割振りを行うとともに、記載内容の変化について分析していきたい。

＊ 本稿の作成にあたり、東京大学大学院工学系研究科教授和泉潔先生と特任講師坂地泰紀先生に、多大なるご指導とご助言を頂いた。心から感謝の意を表したい。なお、本稿の内容・意見は全て当社に属する。

[END]

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

参考文献

和泉潔・後藤卓・松井藤五郎[2011], 「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, Dec. 2011.

小寺俊哉・佐藤史仁・田中良典[2018], 「テキストマイニングによる有価証券報告書の因果関係文以外の特徴文の抽出」, 日興リサーチレビュー, 2018 年, 1 月号.

坂地泰紀・酒井浩之・増山繁[2015],「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, pp811-822, May 2015.

佐藤史仁・佐久間洋明・小寺俊哉[2017], 「テキストマイニングによる有価証券報告書の因果関係文の抽出」, 日興リサーチレビュー, 2017 年, 10 月号.

佐久間洋明・田中良典[2018],「有価証券報告書に含まれるテキスト情報と企業業績の関係について」, 日興リサーチレビュー, 2018 年, 3 月号.

本間友実子, 酒井浩之, 坂地泰紀[2015], 「企業 Web ページを用いた関連企業の抽出」, ARG 第 7 回 Web インテリジェンスとインタラクション研究会, pp.13-14, 2015.

山内長承 [2017], 「Python によるテキストマイニング入門」, 株式会社オーム社.

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。