千葉商大論叢　第 55 巻　第 2 号（2018 年 3 月）pp. 113-133

〔論　説〕

事業等のリスクの分析
―記載内容の類似度にもとづくアプローチ―

土　屋　和　之

1. はじめに

近年の非財務情報に対する関心の高まり(1)を受け，法定開示書類においても非財務情報の開示の拡大が進んでいる(2)。例えば，最近では，2017 年 2 月に企業内容等の開示に関する内閣府令の一部改正が公布され，それまで決算短信の記載内容であった「経営方針」が，有価証券報告書の「事業の状況」に「経営方針，経営環境及び対処すべき課題」として記載されることとなった(3)。

この内閣府令の改正の背景にある，金融庁金融審議会による「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―」では，「非財務情報には，経営方針・経営戦略や MD&A などのほか，ガバナンスや社会・環境に関する情報など様々な情報が含まれる。近年，企業のガバナンス強化に向けた取組みの進展や社会・環境問題への関心の高まりなどにより，これらの非財務情報に対する関心が更に高まっている」とし，すでにさまざまな情報が任意開示の形で開示されているが，内容によっては，制度上開示を義務づけるものも出てくると述べている。こうしたことから今後さらに非財務情報が法定開示書類の中で開示されることも予想される。

定量的情報である財務諸表の分析と比べ，定性的情報である非財務情報の分析は非常に難しいことが指摘されている。非財務情報は自然言語による記述的情報であるため，利用者がそれを読んで分析することが必要になるからである。ところが，EDINET と XBRL の導入は，定性的情報である非財務情報を電子データとして入手することを可能にした。こうした非財務情報について，自然言語処理や機械学習の手法を用いることで，定性的情報を定量的情報として把握し，分析すること可能になる。

本研究では，定性的情報の中でも，比較的多くの研究が行われてきた有価証券報告書の記載事項である事業等のリスクを取り上げる。定性的情報を定量的情報に変換する手法には，さまざまものが考えられるが，今回は，クラスタリングと呼ばれる手法により記載内

(1)　例えば，日本経済新聞「環境・社会・統治「見えない価値」着目　ESG 投資市場の 3 割に」（2017 年 10 月 18 日，朝刊）では，長期的な企業価値を測定するために財務以外の情報が不可欠であるという指摘が紹介されている。

(2)　非財務情報の範囲などについては，さまざまな議論があるが，今回は取り上げない。ここでは自然言語処理と機械学習の手法を用いることのできる，自然言語による文章として記述された情報を指しており，広く使われている非財務情報の範囲とおおよそ一致していると思われる。非財務情報の範囲などについては，古庄［2010］を参照されたい。

(3)　2017 年 3 月 31 日に終了する事業年度から適用されている。

容の類似度を測定することで事業等のリスクの分析を行う。

本研究は，次のような構成となっている。まず，事業等のリスクの開示制度を整理した上で，事業等のリスクを調査した先行研究とその成果を確認する。特に，先行研究ではリスクをどのように分類しているか，その分類の結果から何が明らかになったかに焦点を当てる。

次に，本研究の分析の対象と分析の手法を明らかにする。2016 年度の 3668 社の事業等のリスクを分析の対象とし，TF-IDF 法によって特徴量を把握し，k 平均法でクラスタリングを行う。クラスタリングでは，先行研究の成果を踏まえ，クラスタリングによって分けられたクラスタと業種の関係を見ることで，業種ごとの事業等のリスクの記載内容が類似しているかどうかを明らかにしたい。

2. 事業等のリスクの開示制度

2002 年 8 月 6 日，金融庁は「証券市場の改革プログラム」を公表した。これは証券市場を幅広い投資家の参加する真に厚みのあるものとし，市場機能を中核とした我が国金融システムの中心を担うものとしていくため，(1) 誰もが投資しやすい市場の整備，(2) 投資家の信頼が得られる市場の確立，(3)効率的で競争力のある市場の構築の 3 つの柱に沿って，具体的な施策を提示するものであった。

これを受けて，2002 年 12 月 16 日，金融審議会は，第一部会報告として「証券市場の改革促進」を公表し，制度改正を伴う事項について具体的な内容を示した。そこでは，投資家の信頼が得られる市場の確立のため，会計・監査の充実・強化と並んで，ディスクロージャーの充実・合理化が挙げられている。その中の 1 つとして，開示内容の充実のため，有価証券報告書について，「リスク情報」，「経営者による財務・経営成績の分析（MD&A）」の開示の充実等を，翌年 3 月までに措置することが表明された(4)(5)。

その結果，2003 年 3 月 31 日に「企業内容等の開示に関する内閣府令等の一部を改正する内閣府令」（「開示府令」）の一部改正が行われ，有価証券届出書及び有価証券報告書において，経営者による財務・経営成績の分析（MD&A），コーポレート・ガバナンスに関する事項とならんで，リスクに関する事項についての情報開示が求められることになった。具体的には，「開示府令」の第三号様式の有価証券報告書では，2003 年 4 月 1 日以降に開始する事業年度から，「第一部【企業情報】」の「第 2【事業の状況】」に「4【事業等のリスク】」が新たに設けられた(6)。

現在，事業等のリスクの記載内容については，「企業内容等の開示に関する留意事項について（企業内容等開示ガイドライン）」（2016 年 8 月）によれば，以下が記載事例とし

(4)　リスク情報の開示の開示が求められたのは，これが初めてではない。1983 年に当時の店頭登録企業に対して，有価証券届出書に「事業の概況等に関する特別記載事項」を開示することが義務付けられ，その適用範囲が広げられた（和久［2003］，p. 51）。

(5)　「証券市場の改革促進」では，コーポレート・ガバナンスの強化の 1 つとして，ガバナンスに係るディスクロージャーの充実を挙げ，その中で同じく翌年 3 月までに有価証券報告書について「ガバナンス関連情報」の開示を充実することが表明されている。

(6)　詳しくは，小西［2008］，pp. 111-112。

て挙げられている。

(1) 会社グループがとっている特異な経営方針に係るもの
(2) 財政状態，経営成績及びキャッシュ・フローの状況の異常な変動に係るもの
(3) 特定の取引先等で取引の継続性が不安定であるものへの高い依存度に係るもの
(4) 特定の製品，技術等で将来性が不明確であるものへの高い依存度に係るもの
(5) 特有の取引慣行に基づく取引に関する損害に係るもの
(6) 新製品及び新技術に係る長い企業化及び商品化期間に係るもの
(7) 特有の法的規制等に係るもの
(8) 重要な訴訟事件等の発生に係るもの
(9) 役員，従業員，大株主，関係会社等に関する重要事項に係るもの
(10) 会社と役員又は議決権の過半数を実質的に所有している株主との間の重要な取引関係等に係るもの
(11) 将来に関する事項について

3. 事業等のリスクを調査した先行研究

事業等のリスクの開示が始まって，比較的早い時期から開示に関する調査・分析が行われている。中野［2010］によれば，こうした研究は，開示実態・開示行動に焦点を当てた研究と，開示情報の有用性に焦点を当てた研究の 2 つに分けられるが(7)，本研究では，前者の開示実態・開示行動に焦点を当てた研究の中で，事業等のリスクがどのように分類されているか，その分類から何が明らかなったかを整理する。

まず，財務会計基準機構［2005］は，対象企業を，日本を代表する大企業（グループ A，220 社），中堅企業（グループ B，100 社），小企業・店頭企業・新興企業（グループ C，100 社），継続企業の前提に重要な疑義が生じている企業（グループ D，29 社）の 4 つのグループに分け，2004 年 3 月期の有価証券報告書を事業等のリスクをはじめ，財政状態及び経営成績の分析，コーポレート・ガバナンスの状況を調査したものである(8)。

事業等のリスクについては，有価証券報告書の様式における記載上の注意に例示された項目などから，次の 17 に分類し（表 1），記載の有無を確認している。

財務会計基準機構［2005］は，明らかになったこととして，グループ A の企業は，他のグループに比べ開示した項目が多いこと，リスク項目の上位 3 項目（「将来に関する事項の記載」，「財政状態，経営成績及びキャッシュ・フローの状況の異常な変動」，「特定の取引先・製品・技術等への依存」）はグループ間で差異がほとんどないこと，グループ C では人材の確保を「将来に関する事項の記載」として開示していること企業が目立つことなどを挙げている。

また，参考として，対象企業の中で企業数の多い，電気機器，情報・通信業，化学の 3 業種について，分類ごとの記載事例が紹介されており，さらに，開示項目数の詳細も調査

(7) 中野［2010］，p. 135。
(8) 調査の要点は，小林［2005］にまとめられている。

表 1　財務会計基準機構［2005］による分類

1. 財政状態，経営成績及びキャッシュ・フローの状況の異常な変動
2. 特定の取引先・製品・技術等への依存
3. 特定の製品，技術等で将来性が不明確であるものへの高い依存度について
4. 特定の取引先等で取引の継続性が不安定であるものへの高い依存度について
5. 新製品及び新技術に係る企業化及び商品化期間にかかわるもの
6. 特有の法的規制・取引慣行・経営方針
7. 会社がとっている特異な経営方針に係るもの
8. 特有の法的規制等に係るもの
9. 特有の取引慣行に基づく取引に関する損害に係るもの
10. 法的規制等について
11. 重要な訴訟事件の発生に係るもの
12. 重要な訴訟について
13. 役員・従業員・大株主・関係会社等に関する重要事項に係るもの
14. 会社と役員又は議決権の過半数を実質的に所有している株主との間の重要な取引関係等に係るもの
15. 将来に関する事項の記載
16. 投資者の判断に重要な影響を及ぼす可能性のある事項
17. その他

されている。

張替［2008］は，2003 年度から 2006 年度までの 4 期分の有価証券報告書を開示している上場企業 3005 社を対象に，企業ごとのリスク項目数や文字数などを調査した上で，リスク情報の関連キーワード，172 個を抽出するテキストマイニングを行なっている。リスクの分類については，5 つの大分類を設け，さらに全部で 27 の細分類を行い（表 2），細分類項目に対してキーワードを設定している 。そのキーワードが事業等のリスクに 1 回でも出現すれば，そのリスクの該当企業としている。

張替［2008］は，以上のようなリスクの分類の結果として次の点を挙げている。キーワードが出現した企業数の全企業に対する比率を見ると，市場・経済リスクの市場，業界・競合リスクの販売，価格，環境リスクの規制が高く，7 割以上の企業がリスクとして開示している。また，鉄鋼や海運業では市況の影響が大きいため，市場・経済リスクが高いなど，業種によるリスクの開示の違いも指摘されている。

次に，中野［2010］は，日本企業は，リスク情報および MD&A 等の財務諸表外情報に関して，どのような種類の情報をどの程度開示しているのか，また，企業規模および事業リスク等の企業特性に応じて，当該開示行動は異なっているのかどうかについて分析を行ったものである 。

事業等のリスクの開示内容の分析では，2007 年の規模に応じた 200 社を選び，事業等のリスクを分析している。そこでは，事業等のリスクの表題からリスクを 7 つの要因に分類している（表 3）。

表 2　張替［2008］による分類

大分類	細分類
市場・経済リスク	市場（金利，為替等） 景気 原材料 人口
業界・競合リスク	販売，価格 競合 生産，コスト 新規，海外事業 R&D M&A，提携 顧客 流通 業界慣行
信用リスク	格下げ 特定取引先依存 与信
環境（自然，社会）	規制 自然災害 カントリーリスク 天候 疫病
オペレーショナルリスク	ブランド，安全危機管理 リーガル 品質管理 知財 人事・雇用 IT

表 3　中野［2010］による分類

分類	具体例
1. 市場・一般景気動向	・金利，為替レート，金融商品市場 ・一般景気動向
2. 個別事業の特質	・業界構造 ・研究開発，調達（売上原価変動要因）および販売リスク（収益変動要因） ・法的規制環境，知的財産権
3. 全社的戦略，特質	・全社的経営戦略 ・グループ経営，事業提携および M&A ・特定事業集中
4. オペレーション	・生産，調達および情報管理 ・コンプライアンス，ガバナンス
5. 自然環境，災害	・自然環境，災害リスク
6. 海外事業	・海外事業展開に伴うリスク
7. その他	・その他リスク

分析結果のうち，分類に関わるものとして明らかになったことして，もっとも多いのが個別事業の特質に関するものであること，市場・一般景気動向，オペレーションに関するものも比較的多いこと，また，市場・一般景気動向に関しては個別事業の特質とも関係していることが指摘されている。

野田［2016］は，2003 年度から 2012 年度までの 10 年間について，金融を除く東証一部上場の約 1200 社を対象にした研究である。文字数を記述量として業種別 の平均記述量が，事業等のリスクだけでなく，経営成績及びキャッシュ・フローの状況の分析，コーポレート・ガバナンスに関する状況，対処すべき課題とともに示されている。

野田［2016］では，事業等のリスクを 12 のリスクカテゴリーに分け，各リスクカテゴリーで 5 から 19 のキーワードを指定し（表 4），事業等のリスクにキーワードが記載されている会社数を調査している。

表 4　野田［2016］による分類

リスクカテゴリー	キーワード
取引及び法的問題	談合，カルテル，優先的地位，ダンピング，不正，倒産，遅延，未遵守，法令，国際法，慣習，規制
社会・経済	売上，販売，収入，費用，コスト，支出，景気，景況，市況，不買運動，供給途絶，金利，為替，地価，株価，原材料価格
自然現象	地震，津波，台風，噴火，落雷，雪害，豪雨，洪水，高潮，竜巻，火災，地盤沈下，液状化，渇水，天候不順
政治	戦争，紛争，革命，テロ，暴動，反社会的
技術	ライフライン，電力，水道，情報通信，ITC，IT，技術革新，イノベーション，陳腐，流出
経営及び内部統制	インサイダー，重要事実，非開示，操作的，虚偽，脱税，過少申告，合併，買収，M&A，知財，知的財産権，侵害，リスク情報，隠ぺい，隠蔽，改ざん，株主代表訴訟，オペレーションリスク，詐欺
財務	横領，粉飾，運用の失敗，資金不足，貸し渋り，貸し剝がし
製品・サービス	欠陥，瑕疵，ミス，事故，過誤，失敗，使用禁止物，クレーム
情報セキュリティ	システムダウン，不正使用，不正アクセス，機密情報，ウイルス，ウィルス，ハッカー，ハッキング
環境問題	土壌，大気，水質，放射能，汚染，騒音，異臭，グリーン，不法投棄
労働安全衛生	労働災害，労災，伝染病，感染症，健康，食中毒，交通事故，過労，ストレス，職業病，メンタルヘルス
雇用	人材，モラール，モラル，人権，差別，不法就労，セクハラ，セクシャルハラスメント，パワハラ，パワーハラスメント，スキャンダル，プライバシー，個人情報，就業規則違反，ストライキ

こうした分類の結果，明らかになったこととして，以下の点が挙げられている。(1) 増加比率が高いカテゴリーは自然現象，情報セキュリティ，環境問題で，社会・経済といった事業活動に伴う経済リスク以外の分野の開示が増えている。(2) リスクのカテゴリーを CSR 関連リスクと非 CSR 関連リスク，システムリスクと個別リスクに分類した結果，CSR 関連リスクと個別リスクの開示の比率が増加している。

4. 分析の対象

4.1 対象となる会社の範囲と期間

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EDNIET（Electronic Disclosure for Investors' NETwork）は，2001年に運用を開始した。2008年にはEDINETで提出される書類のうち財務諸表について，XBRL（eXtensible Business Reporting Language）形式での提出が義務付けた。さらに，2013年からは次世代EDINETとして，XBRLの対象範囲が拡大されている。具体的には，XBRLの対象書類の拡大と対象項目の拡大である。対象書類の拡大では，それまでの有価証券報告書，半期報告書，四半期報告書，有価証券届出書の4書類から発行登録書等の発行開示書類，臨時報告書等の継続開示書類，公開買付届出書等の公開買付関連書類，大量保有報告書等の大量保有関連書類の44様式に拡大されている。また，対象項目の拡大では，各報告書の財務諸表本体のみとなっているXBRLによる記述を報告書全体に拡大している。

XBRLの対象範囲が各報告書の全体に拡大されたことで，非財務情報がデジタルデータとして従来よりも容易に入手できるようになった。有価証券報告書については，2013年12月31日を決算日とする事業年度にかかる有価証券報告書から全体がXBRLの対象となっている。

このうち対象となる有価証券報告書を，企業内容等の開示に関する内閣府令の第三号様式の有価証券報告書する。また，対象となる期間は，会社間の事業等のリスクの類似度を分析することから，一事業年度を対象とする。したがって，本研究の対象となるのは，2016年4月1日から2017年3月31日の間の決算日にかかる3668社の有価証券報告書である。なお，業種ごとの分析を行うため，業種を特定する必要がある。今回はEDINETで配布されているEDINETコード一覧で，EDINETコードに付与された業種によっている(9)。これは証券コード協議会の業種分類の中分類，いわゆる東証33業種である。

4.2 対象となる事業等のリスク

本研究の分析の対象となる事業等のリスクは，対象会社の有価証券報告書のXBRLインスタンスの中から，事業等のリスクを指定する要素であるBusinessRisksTextBlockというタグに囲まれたテキストを抽出したものである(10)。ただし，このテキストには事業等のリスクの見出しである「【事業等のリスク】」が含まれるので，これは除去している。

先行研究と同様に分析対象会社の事業等のリスク全体の様子を把握するため，ここでは事業等のリスクの文字数を数えている。 文字数を数えるにあたっては，表などを用いて金額や数字を記述している場合もすべて1文字として数えている。

2016年度の3668社の事業等のリスクの文字数に関する記述統計量は，平均値が2958文字，最大値が39294文字，最小値が47文字，中央値が2186文字となっている。参考ま

(9) EDINETコード一覧では書類提出日の業種によっているので，証券コード協議会が定期的に行なっている業種の変更は反映されていない。

(10) 2017年度版EDINETのタグの詳細については，次のファイルで開示府令の様式ごとに確認できる。
https://disclosure.edinet-fsa.go.jp/download/ESE140114.xls

でに現行の EDINET になってから比較のできる，2014，2015 年度の事業等のリスクの文字数も挙げておく（表 5）。

先行研究のうち事業等のリスクの文字数を調査した中野［2010］，野田［2016］の調査と合わせると，当初は記述内容が少なかった事業等のリスクは次第に文字数が増加し，ここ数年は 2000 字から 3000 字で推移していると考えることができるようである。

表 5　会社数と文字数の記述統計量

	2014	2015	2016
会社数	3570	3622	3668
平均値	2778	2879	2958
最大値	42685	37656	39294
最小値	11	11	47
中央値	2022	2106	2186

5. 分析の手法

5.1 分析の手法とその意義

事業等のリスクの開示実態・開示行動に焦点を当てた先行研究は，事業等のリスクを，内容を見るなど，なんらかの方法でラベルを決めておいて，各会社の事業等のリスクにラベルを割り当てる，分類による分析であった。

本研究では，事業等のリスクの記載内容の類似度に応じて，グループ（クラスタ）に分ける，クラスタリングによる分析を行う。事業等のリスクの分類による分析では，その事業等のリスクに，あらかじめ用意されたラベルを必ずつけなければならないため，どのようなラベルを用意するか，どのラベルをつけるかは，非常に難しい問題となる場合もある。

一方，クラスタリングは，そもそもどのようなラベルをつけるかは，事前に決める必要はない。内容の特徴を把握し，その特徴が類似する事業等のリスクをグループに分けるだけなので，厳密で機械的な分類の窮屈さから解放される(11)。

事業等のリスクについてクラスタリングを行うことで，事業等のリスクの記載内容が類似している会社を明らかにすることが可能になる。これによって，1 社 1 社の事業等のリスクを読んでいるだけでは把握できない，対象会社全体の事業等のリスクの記載内容を把握することができる。これは 1 社 1 社の事業等のリスクを読んで分析する場合にも，有用な情報となるはずである。

あるいは，事業等のリスクについては，いわゆるボイラープレートであるという批判がある。つまり，開示が決まり文句化して，横並びになっているという批判である。もしほとんどの会社の事業等のリスクが同じ内容であるということになれば，そうした批判を支持することになるかもしれない。このように事業等のリスクをクラスタリングすることで，

(11) 吉田［1993］，p. 105。

従来の研究をより充実したものにすることも期待できる。

また，クラスタリングはあらかじめラベルを用意しないため，機械学習のうちの教師なし学習の１つとされる。Bao and Datta[2014] によれば，米国のリスク情報[(12)]を対象にした研究は，自動化されたテキスト分析とリスク情報の開示効果に関する研究に分けられるという。前者はさらに，辞書による方法，教師あり学習による方法，教師なし学習による方法に分類される[(13)]。事業等のリスクに関する先行研究をこの分類に当てはめれば，辞書による方法ということになるだろう。これに対して本研究は教師なし学習による方法に分類される。

クラスタリングは，(1) 事業等のリスクの特徴量を把握する，(2) 把握された特徴にもとづいて事業等のリスクをクラスタに分ける，という２つのステップで行われる。本研究では，(1) については，自然言語処理のベクトル空間モデルを，(2) については，教師なし学習のアルゴリズムである k 平均法を用いることにしたい。

5.2 特徴量の把握

事業等のリスクを類似するクラスタに分ける場合，分ける基準となる特徴を決めなければならない。機械学習ではこの特徴のことを特徴（attribute）と呼び，その値を特徴量と呼んでいる。

まず，特徴量を把握する前に，日本語では文章である事業等のリスクを単語に分かち書きしなければならない[(14)]。先に示したように，先行研究では事業等のリスクに現れるキーワード（名詞）をもとに分類を行なっている。そこで，分かち書きされた事業等のリスクのうち，今回は名詞以外の単語をストップワード（不要語）として除去し，名詞のみを使用することとする。ある会社の事業等のリスクを分かち書きして，名詞のみを取り出すと次の通りとなる。

事業 等 リスク 当社 グループ 経営成績 株価 連結財務諸表 等 影響 可能性 リスク 当期末 現在 主要 もの 以下 通り これら リスク 発生 可能性 認識 発生 回避 発生 場合 影響 最小化 国内 経済状況 変動 個人消費 動向 影響 当社 グループ 有利子負債 額 金利 変動 金利 負担 増加 季節 要因 販売 状況 左右 商品 取扱い 売行き 不振 季節 経過 商品 価値 下落 発生 可能性 存在 当社 グループ 有価証券 保有 市況 悪化 投資 先 業績 不安 評価 損 計上 可能性 および 株価 変動 資金調達 額 制約 可能 性 存在 当社 グループ 保有資産 実質的 価値 低下 等 減損 処理 必要 場合 当社 グループ 業績 影響 可能 性 存在 海

(12) SEC は 2005 年に登録会社に対し，年次報告書 Form 10-K に Item 1A. Risk Factors を設け，Regulation S-K の第 503（c）項（§ 229. 503（c））に記載されているリスク要因を開示ように求めている（近藤［2014］）。

(13) Bao and Datta[2014] によれば，辞書による方法としては，リスク情報を調査した研究からキーワードを抽出して，キーワードのリストを作って分類している Campbell et. al[2014] が，教師あり学習による方法としては，１つのリスク情報に複数のラベルを割り当てた，Huang and Li[2011] が，そして Bao and Datta[2014] が教師なし学習による方法として挙げられている。

(14) 本研究では，日本語の分かち書きを MeCab（http://taku910.github.io/mecab/）によって行なっている。また，MeCab が使う辞書として，mecab-ipadic-NEologd（https://github.com/neologd/mecab-ipadic-neologd/）を使用している。

外 生産 海外 調達 活動 為替レート 変動 現地 通貨 価値 変動 経済状況 変化 生産 調達 コストアップ 生産管理 上 トラブル 製品事故 等 予期 事象 発生 不動産 賃貸 競争激化 ため 賃貸 条件 悪化 影響 可能 性 存在 従業員 年齢 構成 バランス 悪さ 後継者 養成 制約 可能 性 存在 和装 事業 洋装 事業 成熟 産業 成長産業 進出 シフト 遅れ 当社 グループ 財政状態 および 経営成績 影響 可能性 存在

こうして分かち書きされた事業等のリスクについて，自然言語処理の文書検索で用いられるベクトル空間モデルを適用する。文書検索では，質問に対してもっとも関連性の高い文書を検索しなければならない。ベクトル空間モデルでは，質問と検索対象となる文書をベクトルで表現し，質問のベクトルともっとも近いベクトルの文書を，もっとも関連性の高い文書と考えるのである。

文書をベクトルとして表現したものには，文書内の単語の出現頻度をベクトルの要素の値とする頻度ベクトルや，単語の出現の有無をベクトルの要素の値とする二値ベクトルがある(15)。本研究では，単語の出現頻度をもとに単語の重みづけをベクトルとする TF-IDF 法によるベクトルを用いることにする。

TF-IDF 法は，文書検索の分野で広く用いられており，出現する文書内での重要性と文書集合全体から見た重要性の 2 つの観点から単語について重み付けを行うものである(16)。文書に何度も出現する単語は，その文書を特徴づける単語であると考えられるが，どの文書にも高い頻度で出現する単語は重要ではないと考える。また，低い頻度でも少数の文書に出現する単語は，その文書を特徴づける単語であると考えるのである(17)。

TF（Term Frequency）は，ある単語のある文書における出現頻度である。DF（Document Frequency）は，ある単語が少なくとも 1 回出現する文書数である。IDF（Inverted Document Frequency）は，DF の逆数について対数をとったもので，単語が出現する文書数が少ないほど，IDF は大きくなる。TF-IDF 法は，この TF と IDF の積を単語のベクトルの値とする方法である。

本研究では，先に示したように，会社によって事業等のリスクの文字数に違いがあるため，文字数の違いが TF-IDF の計算に影響が出ないように正規化を行なっている。3668 社の TF-IDF を計算した結果，29001 個の単語が識別された(18)(19)。参考までに，TF-IDF が高い順に 10 単語を示すと次の通りである。

nidec，野村，富士電機，NEC グループ，当行，OKI，エプソン，東京海上グループ，ソニー，アサヒグループ

---

(15) 例に挙げた，分かち書きされた事業等のリスクのうち「グループ」という単語は，6 回出現しているので頻度ベクトルでは 6 が，出現しているので二値ベクトルでは 1 が与えられる。

(16) この場合，文書が 1 社の事業等のリスク，文書集合が今回の分析対象の 3668 社の事業等のリスクということになる。

(17) 秋葉［2013］，p. 136。

(18) 例に挙げた，分かち書きされた事業等のリスクの「グループ」という単語には，約 0.219 が与えられている。

(19) TF-IDF の計算，k 平均法によるクラスタリングには，python と機械学習のライブラリである scikit-learn（http://scikit-learn.org）を使用している。

会社名などの固有名詞が多いのは，その会社の事業等のリスクにしか現れないが，その会社の事業等のリスクには頻繁に現れるためと考えられる。

5.3 クラスタリングアルゴリズム

クラスタリングアルゴリズムは，文書集合をクラスタという部分に分けるアルゴリズムである。クラスタ内の文書はお互いにできる限り似通っているが，他のクラスタの文書とはできる限り異なっていることが望ましい(20)。この類似度（非類似度）の尺度として，距離が使われる。クラスタリングには，クラスタ間を関係付ける明示的な構造を持たないクラスタのフラットな集合を生成する非階層クラスタリングと，クラスタの階層を作る階層クラスタリングがある。本研究では，非階層クラスタリングの１つであり，代表的なクラスタリングの手法である k 平均法によるクラスタリングを行う。

k 平均法は，とりあえず適当に分けてしまって，それからよりうまく分かれるように調整していくことによってクラスタリングを行う方法である(21)。クラスタリングの手順は次の通りである。まず無作為に k 個の代表ベクトルを決める。そして，どの代表ベクトルに近いかという基準にしたがって，各ベクトルをどこかのクラスタに帰属させる。次に，各クラスタに含まれているベクトルの平均を計算し，これを新たな代表ベクトルとする。そして，この代表ベクトルにしたがって各ベクトルを再び k 個に分ける。これをクラスタに変化がなくなるまで繰り返すことによってクラスタリングが達成できる(22)。

各クラスタの平均は重心と呼ばれる。ベクトル x が属するクラスタ ω の平均 μ は次のように定義される。

$$\mu(\omega) = \frac{1}{|\omega|} \sum_{x \in \omega} x$$

k 平均法は，k 番目のクラスタω_kに属するベクトル x と，その平均 $\mu(\omega_k)$ までの距離の二乗のすべてを足した残差平方和（RSS）

$$\mathrm{RSS} = \sum_{x \in \omega_k} \left| x - \mu(\omega_k) \right|^2$$

が最小になるように平均を決めるアルゴリズムである(23)。

6. 分析の結果

2016 年度の 3668 社の事業等のリスクについて，TF-IDF 法により特徴量を把握くし，k 平均法によりクラスタリングを行なった結果は次の通りである。先行研究で指摘されているように，事業等のリスクの開示の内容は業種によって異なっているという。そこで，

(20) Manning, Raghavan, Schütze[2012], p. 310。
(21) 高村［2010］, p. 82。
(22) 高村［2010］, p. 83。
(23) Manning, Raghavan, Schütze[2012], p. 320。

本研究では，クラスタリングによって分けられたクラスタと業種の関係を見ることで，事業等のリスクの記載内容が業種ごとに類似しているのかどうかを明らかにしたい。

まず，クラスタ数である k の値を決めなければならない。今回は，RSS の最小値とクラスタ数の関係をグラフにしたときの屈曲点[24]から k=10 としてクラスタリングを行う[25]。

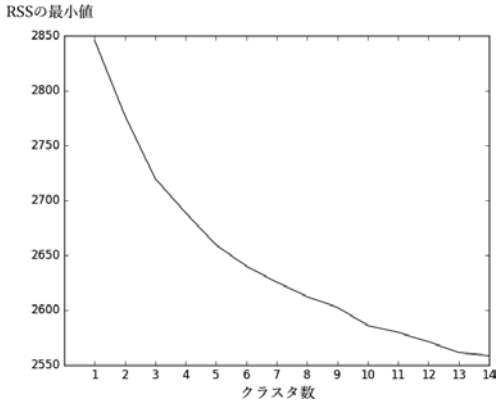

図 1　全業種を対象とした RSS の最小値とクラスタ数

クラスタリングの結果，クラスタとそのクラスに属する会社数は，次のようになっている（表 6）。

表 6　クラスタと会社数

クラスタ	会社数
0	1290
1	74
2	303
3	1251
4	750

クラスタと業種の関係をみるため，各クラスタに属するそこで，各クラスタに属する業種ごとの会社数を見てみよう（表 7）。

クラスタ 1 は銀行業 74 社である。銀行業の事業等のリスクは，他の事業会社の事業等のリスクと大きく異なっていることを表しているが，TF-IDF の上位 10 語に「当行」が

(24) 詳しくは Manning，Raghavan，Schütze［2012］，p. 325。

(25) クラスタリングは，似ているものをグループ化することが目的であるから，どの程度まで似ているものを一緒のグループとみなすか，あるいはクラスタ数をいくつにするか，について一般的な解はないといってよい（高村［2010］，p. 94）。したがって，k = 10 が正解ではなく，k = 10 のとき，どのようにクラスタに分けられるかを見ているにすぎない。

含まれていることから，「当行」が含まれている事業等のリスクを同じクラスタに分けていることが考えられる。一方で，他の事業会社の事業等のリスクは，クラスタ 2 に小売業が集中している他は業種とクラスタの関連はないように見える。

そこで，次に「当行」と「当社」を分かち書きされた単語の中から除去し，改めて TF-IDF を計算し，クラスタリングを行うことにする。再度 RSS の最小値とクラスタ k の関

表 7　全業種を対象としたクラスタと業種ごとの会社数

	0	1	2	3	4
水産・農林業	3	0	0	7	0
鉱業	1	0	0	7	0
建設業	35	0	0	135	7
食料品	58	0	6	64	3
繊維製品	34	0	2	18	0
パルプ・紙	20	0	0	6	0
化学	159	0	1	53	3
医薬品	15	0	0	32	21
石油・石炭製品	8	0	0	5	0
ゴム製品	15	0	0	4	0
ガラス・土石製品	30	0	0	30	0
鉄鋼	28	0	0	19	0
非鉄金属	25	0	0	11	1
金属製品	54	0	0	37	0
機械	140	0	0	86	5
電気機器	205	0	1	50	9
輸送用機器	69	0	0	26	0
精密機器	36	0	0	12	3
その他製品	62	0	3	40	4
電気・ガス業	3	0	0	21	0
陸運業	27	0	0	34	4
海運業	4	0	0	10	0
空運業	1	0	0	3	1
倉庫・運輸関連業	15	0	0	20	4
情報・通信業	9	0	0	67	325
卸売業	145	0	12	145	36
小売業	29	0	258	38	26
銀行業	17	74	0	4	0
証券，商品先物取引業	3	0	0	16	21
保険業	3	0	0	4	5
その他金融業	3	0	0	18	15
不動産業	11	0	2	72	35
サービス業	23	0	18	157	222

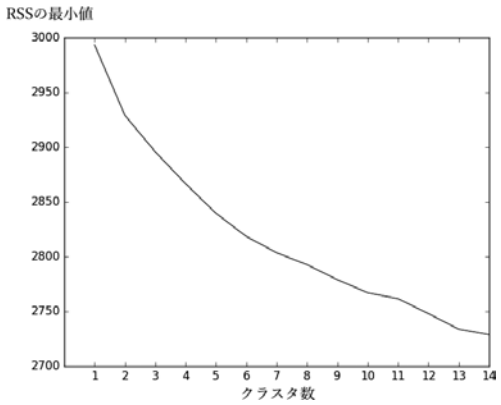

図 2 「当行」等を除去した全業種を対象にした RSS の最小値とクラスタ数

係をグラフにして，その屈曲点から，今回は k=6 としてクラスタリングを行う。

クラスタリングの結果，クラスタとそのクラスに属する会社数は，次のようになっている（表 8）。

表 8　クラスタと会社数

クラスタ	会社数
0	288
1	96
2	1122
3	302
4	1240
5	620

クラスタと業種の関係をみるため，各クラスタに属するそこで，各クラスタに属する業種ごとの会社数を見てみよう[(26)]（表 9）。

この結果からは，建設業と不動産業（クラスタ 0），銀行業（クラスタ 1），小売業（クラスタ 3），情報・通信業とサービス業（クラスタ 5）に分けられていると思われるが，「当行」等を除去しても，はやり金融業の事業等のリスクは他の業種の会社の事業等のリスクと比べ大きく異なっているようである。

そこで，東証 33 業種のうち，銀行業，証券，商品先物取引業，保険業，その他金融業の 4 業種を除いた 29 業種，3485 社について，改めて，TF-IDF 法で特徴量を求め，クラ

(26) クラスタリングの結果をどう評価するかは非常に難しい問題である。本研究では，業種との関連でクラスタリングを評価している。ただし，これは自分の期待する分け方をクラスタリングがたまたま再現しているかを見ているのであって，側面的な評価である。つまり，この指標が悪くても，別の側面からみれば非常によいクラスタを作っているかもしれない（高村［2010］，p. 95）。

表９ 「当行」等を除いた全業種を対象としたクラスタと業種ごとの会社数

	0	1	2	3	4	5
水産・農林業	0	0	7	0	3	0
鉱業	0	0	7	0	1	0
建設業	148	0	17	0	8	4
食料品	0	0	58	6	65	2
繊維製品	0	0	17	2	35	0
パルプ・紙	0	0	5	0	21	0
化学	0	0	50	1	162	3
医薬品	0	0	40	0	18	10
石油・石炭製品	0	0	5	0	8	0
ゴム製品	0	0	2	0	17	0
ガラス・土石製品	2	0	26	0	32	0
鉄鋼	0	0	16	0	31	0
非鉄金属	1	0	12	0	24	0
金属製品	4	0	28	0	59	0
機械	3	0	84	0	141	3
電気機器	0	0	50	1	207	7
輸送用機器	0	0	23	0	72	0
精密機器	0	0	12	0	38	1
その他製品	0	0	39	3	64	3
電気・ガス業	0	0	21	0	3	0
陸運業	5	0	32	0	18	10
海運業	0	0	10	0	4	0
空運業	0	0	3	0	1	1
倉庫・運輸関連業	2	0	19	0	0	4
情報・通信業	2	0	98	0	7	294
卸売業	13	0	147	14	135	29
小売業	1	0	46	255	25	24
銀行業	0	92	3	0	0	0
証券，商品先物取引業	0	3	22	0	3	12
保険業	0	0	6	0	3	3
その他金融業	2	1	18	0	1	14
不動産業	91	0	23	1	2	3
サービス業	14	0	176	19	18	193

スタリングを行なってみよう。全業種のクラスタリングの場合と同様に RSS の最小値とクラスタ数の関係をグラフにして，その屈曲点から，今回は k=7 としてクラスタリングを行う。

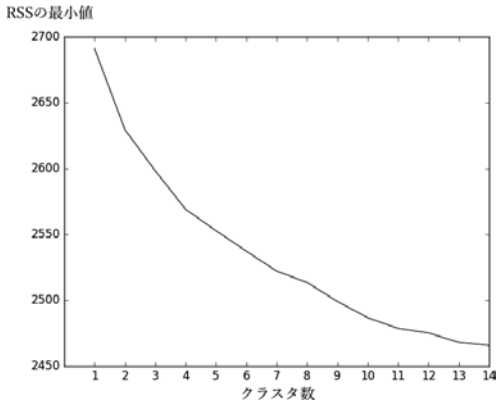

図 3　金融業を除く業種を対象にした RSS の最小値とクラスタ数

クラスタリングの結果，クラスタとそのクラスに属する会社数は，次のようになっている（表 10）。

表 10　クラスタと会社数

クラスタ	会社数
0	410
1	950
2	612
3	142
4	935
5	138
6	298

クラスタと業種の関係をみるため，各クラスタに属するそこで，各クラスタに属する業種ごとの会社数を見てみよう（表 11）。

表によると，クラスタ 2，3，5，6 には，特定の業種が集中している。つまり，クラスタ 2 では，情報・通信業（316 社），クラスタ 3 では，建設業（124 社），クラスタ 5 では，不動産業（98 社），クラスタ 6 では小売業（254 社）が，各クラスタに集中している。クラスタリングでは似たものをクラスタに分けるので，多くの会社が特定のクラスタに分けられている業種では，事業等のリスクの記載内容が類似していることを示している。したがって，本研究のクラスタリングによれば，以上の 4 業種，情報・通信業，建設業，不動産業，小売業の事業等のリスクの記載内容の類似度は高いと結論できる。一方，他の業種

表 11　金融業を除いたクラスタと業種の会社数

	0	1	2	3	4	5	6
水産・農林業	0	7	0	0	3	0	0
鉱業	0	7	0	0	1	0	0
建設業	2	17	3	124	16	15	0
食料品	18	54	3	0	51	0	5
繊維製品	9	17	0	0	26	0	2
パルプ・紙	4	5	0	0	17	0	0
化学	58	49	2	0	106	0	1
医薬品	2	30	19	0	17	0	0
石油・石炭製品	2	4	0	1	6	0	0
ゴム製品	5	3	0	0	11	0	0
ガラス・土石製品	10	25	0	1	23	1	0
鉄鋼	8	17	0	0	22	0	0
非鉄金属	9	11	1	0	15	1	0
金属製品	18	28	0	5	40	0	0
機械	43	81	1	3	103	0	0
電気機器	95	46	7	0	116	0	1
輸送用機器	37	23	0	0	35	0	0
精密機器	10	11	2	0	28	0	0
その他製品	17	39	1	0	49	0	3
電気・ガス業	0	18	0	0	6	0	0
陸運業	9	30	1	0	24	1	0
海運業	0	9	0	0	5	0	0
空運業	0	3	0	0	2	0	0
倉庫・運輸関連業	4	18	1	0	14	2	0
情報・通信業	3	70	316	0	10	2	0
卸売業	26	128	28	5	134	5	12
小売業	15	36	19	0	26	1	254
不動産業	1	14	2	0	4	98	1
サービス業	5	150	206	3	25	12	19

はそれぞれのクラスタに点在していることから，その業種に属する会社の事業等のリスクの記載内容の類似度は 4 業種に比べ低いと考えられる。

こうした結果と先行研究の成果を検討してみよう。先行研究では，開示されているリスクを分類し，どのようなリスクが開示されているかを分析してきた。それに対して，本研究では事業等のリスクの記載内容を直接見るのではなく，事業等のリスクの記載内容が類

似しているかどうか特に業種との関連を明らかにした。

一方，先行研究では業種ごとに多様なリスク項目が開示されていることが指摘されている。本研究の分析の結果によれば，その多様性は業種によって幅があると考えられる。つまり多様なリスク項目が開示される業種とそうでない業種があるということである。

また，事業等のリスクについて，ボイラープレートという批判がある。今回の分析の結果によれば，確かに記載内容が類似していると考えられる業種もある。今回の分析の結果によれば，確かに記載内容が類似している業種もある。一方で記載内容が類似していない業種もあることから，記載内容と業種という点から見ると，一律にボイラープレートとなっているという批判は当たらないのではないかと思われる。

7. おわりに

本研究では，EDINET と XBRL の導入によって，デジタルデータとして入手可能となった非財務情報について，自然言語処理と機械学習の手法を用いて，定量的情報として分析できることを明らかにするため，2016 年度の 3668 社の有価証券報告書の事業等のリスクについて，TF-IDF 法で特徴量を把握し，k 平均法によってクラスタリングを行なった。先行研究の成果を踏まえ，クラスタリングによって分けられたクラスタと業種の関係を見ることで，業種ごとの事業等のリスクの記載内容が類似しているかどうかを分析した。

本研究で明らかになったことは以下の通りである。東証 33 業種のうち，銀行業，情報・通信業，建設業，不動産業，小売業はそれぞれがあるクラスタに多くの会社が分けられていることから，これら業種に属する多くの会社の事業等のリスクの記載内容は類似していると見られる。一方，その他の業種に属する会社の事業等のリスクの記載内容は，業種内での類似度は低いと見られる。

こうした結果は，先行研究のような研究に対しても貢献できると考えられる。先行研究では，それぞれの立場から事業等のリスクを分類して，対象企業の事業等のリスクを分析しているが，前提となる事業等のリスクの分類において，クラスタリングの結果を踏まえた分類を行うことが考えられる。また，事業等のリスクの有用性を検討した研究に対しても，変数の決定にクラスタリングの結果を反映させることができるかもしれない。

さらに，本研究のように個別の企業の事業等のリスクを見ていくのではなく，開示企業全体の傾向を把握することで，これまでは言語の違いのため，比較が難しかった国ごとのリスク情報の開示内容を比較できるようになる。すでに EDGAR で開示が行われている米国をはじめ，EU でも 2020 年 1 月から欧州単一電子フォーマット（European Single Electronic Format）による開示が義務付けられている(27)。こうした国々のリスク情報を同じようにクラスタリングすることで，リスク情報の開示を国ごとに比較できるようになるだろう。

(27) European Securities and Markets Authority（ESMA）は，2017 年 12 月 18 日，欧州単一電子フォーマットにかかる Regulatory Technical Standards に関する最終報告書を公表した。最終報告書によれば，IFRS 連結財務諸表は，IFRS タクソノミによって XBRL で記述されること，年次財務報告書のその他の部分は，XHTML で記述されることが求められている。詳しくは ESMA［2017］を参照のこと。

一方で，本研究に関連して，さらに検討を要する点として，次の３点を挙げることができる。まず１点目は，分析のためデータの前処理の問題である。１回目のクラスタリングで示したように，銀行業における「当行」のように，特定の業種で使用される単語によって，クラスタリングの結果が大きく左右される可能性がある。どのような単語を除去すべきか検討が必要だろう。また，文章を単語に分かち書きするためには辞書が必要になる。事業等のリスクなど，有価証券報告書等に記載される文章で使用される単語のための辞書を用意することも考えられる。XBRL にはタクソノミがあるので，ここから辞書を生成することも考えられる。

２点目は把握すべき特徴量の選択の問題である。 TF-IDF 法では，単語の出現頻度による重み付けを行なっており，単語の意味を考慮しているわけではない。近年では，単語の意味や関係に基づいて単語の重み付けを行う方法も提案されており，こうした方法に基づいて特徴量を把握することも検討する必要があるだろう。

３点目は，クラスタリングのアルゴリズムの選択の問題である。本研究で用いたｋ平均法は，各文書がどれか１つのクラスタに割り当てられるハードクラスタリングの１つである。しかし，事業等のリスクでは，さまざまな関連するリスクが記載されているため，１つのクラスタに割り当てるのは無理があることも考えられる。ハードクラスタリングではなく，ある事業等のリスクを確率によって複数のクラスタに割り当てるソフトクラスタリングによれば，事業等のリスクの開示の様子をより的確に把握できるかもしれない。

非財務情報の開示が拡大することで，今後自然言語による記述的な情報がますます増えることが予想される。本研究のように自然言語処理と機械学習の手法を用いることで，開示企業全体の非財務情報の開示の状況を把握することは，従来の，個別企業の非財務情報を読み込んで分析する場合に有用であるといえるだろう。

〔参考文献〕

Bao, Yang and Anindya Datta, Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures, *Management Science*, Vol.60 No. 6, June 2014, pp. 1371–1391.

Campbell, John L., Hsinchun Chen, Dan S. Dhaliwal, Hsin-min Lu, Logan B. Steele, The information content of mandatory risk factor disclosures in corporate filings, *Review of Accounting Studies*, March 2014, Vol.19 No.1, pp. 396–455.

European Securities and Market Authority, *Final report on the RTS on the European Single Electronic Format*, Final Report ESMA32–60–204, 18 December 2017, https://www.esma.europa.eu/sites/default/files/library/esma 32-60-204 _final_report_on_rts_on_esef.pdf

Ke-Wei Huang and Zhuolun Li, A multilabel text classification algorithm for labeling risk factors in SEC form 10 -K. *ACM Transactions on Management Information Systems*, Vol. 2 No.3, October 2008, pp. 1–19.

秋葉友良，「検索・質問応答システム」，中川聖一編，『音声言語処理と自然言語処理』，コロナ社，2013 年 3 月，pp. 120–150。

石川慎一郎，前田忠彦，山崎誠，『言語研究のための統計入門』，くろしお出版，2010 年 12 月。
金融庁，「証券市場の改革促進プログラム」，2002 年 8 月 6 日，http://www.fsa.go.jp/news/newsj/14/syouken/f-20020806-2b.pdf
金融庁金融審議会，「第一部会報告　証券市場の改革促進」，2002 年 12 月 16 日，http://www.fsa.go.jp/singi/singi_kinyu/siryou/kinyu/dai1/f-20021216_sir/01b.pdf
金融庁金融審議会，「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―」，2016 年 4 月 18 日，http://www.fsa.go.jp/singi/singi_kinyu/tosin/20160418-1/01.pdf
小林一男，「有価証券報告書における「事業等のリスク」等の開示実態調査（調査報告）について」，『JICPA ジャーナル』，第 594 号，2005 年 1 月，pp. 26–30。
近藤汐美，「米国におけるリスク情報開示の制度的展開」，『現代マネジメント学部紀要』，第 3 巻第 1 号，2014 年，pp. 15–23。
財務会計基準機構編，『有価証券報告書における『事業等のリスク』等の開示実態調査』，財務会計基準機構，2005 年 2 月。
高村大也，『言語処理のための機械学習入門』，コロナ社，2010 年 8 月。
中野貴之，「財務諸表外情報の開示実態」，山﨑秀彦編『財務諸表外情報の開示と保証―ナラティブ・リポーティングの保証―』，2010 年 10 月，pp. 133–150。
野田健太郎，『有価証券報告書における定性情報の分析と活用―リスク多様化にともなう望ましい対話のあり方―』，経済経営研究，第 37 巻第 1 号，日本政策投資銀行設備投資研究所，2016 年 5 月。
張替一彰，「有価証券報告書事業リスク情報を活用したリスク IR の定量評価」，『証券アナリストジャーナル』，第 46 巻第 4 号，2008 年 4 月，pp. 32–44。
古庄修，「財務諸表外情報の位置づけ」，山﨑秀彦編『財務諸表外情報の開示と保証―ナラティブ・リポーティングの保証―』，2010 年 10 月，pp. 21–44。
吉田類，『分類学からの出発―プラトンからコンピュータへ』，中公新書，中央公論社，1993 年 9 月。
Christopher D. Manning，Prabhakar Raghavan，Hinrich Schütze 著，岩野和生，黒川利明，濱田誠司，村上明子訳，『情報検索の基礎』，共立出版，2012 年 6 月。
和久友子，「どう変わる？ガバナンス関連情報，リスク情報，MD&A の開示」『旬刊経理情報』，第 1009 号，2003 年 3 月，pp. 50–54。

（2018.1.19 受稿，2018.2.21 受理）

〔抄　録〕

本研究では，EDINET と XBRL の導入によってデジタルデータとして入手が可能になった有価証券報告書の事業等のリスクを，TF-IDF 法によって単語の重みづけを行ったベクトルとして表現し，k 平均法によるクラスタリングを行った。先行研究は記載されたリスクにもとづいて事業等のリスクを分類する研究であった。一方，類似したものをグループに分けるクラスタリングによれば，記載内容が類似する事業等のリスクをグループに分けることができるので，対象会社全体の事業等のリスクの開示の傾向を把握することができる。分析の結果，銀行業，小売業，建設業，不動産業，情報・通信業では，事業等のリスクの記載内容の類似度が高いことが明らかになった。こうした結果は先行研究のような事業等のリスクの開示行動を分析する研究だけでなく，開示の有用性に関する研究に対しても分析モデルの構築に役立つと期待できる。また，本研究で用いた自然言語処理と機械学習の手法を非財務情報の分析に適用することで，定性的情報である非財務情報を定量的に分析することが可能になる。