3B2-01

文脈情報を利用した文書集合間の差異分析方法と決算報告書分析への応用

Text Mining Method Using Context Information and Its Application to Annual Report Analysis

*1竹内 広宜 Hironori Takeuchi　*1荻野 紫穂 Shiho Ogino　*1渡辺 日出雄 Hideo Watanabe　*2白田 佳子 Yoshiko Shirata

*1日本アイ・ビー・エム株式会社 東京基礎研究所
IBM Research, Tokyo Research Laboratory

*2筑波大学ビジネス科学研究科
Graduate School of Business Sciences, University of Tsukuba

In this paper, we consider the large documents in length and try to find differences between document collections. In the analysis of document collections such as project status reports or annual reports where each document and each sentence tend to be relatively long. Therefore, it is sometimes difficult to derive insights by looking only for representative concepts in the selected document collection based on a divergence metric. In this paper, we propose a analysis approach based on the context information. By extracting pairs of a topic word and a keyword and assessing their representativeness in the selected document collection, we are developing a method to extract insights from large documents in length. Applying the proposed method to the annual reports of bankrupt companies and going concern companies that are sound companies, we were able to derive insights that could not extracted from the conventional methods.

1. はじめに

近年、蓄積されたテキストデータを活用するテキストマイニング技術が研究・開発されている。特に CRM(Customer Relationship Management) の分野では蓄積された顧客の声をテキストマイニング技術を通して分析し、ビジネスに活用することが行われている [11]。

テキストマイニング分析の目的の一つは、テキスト内に出現する概念（キーワードや表現）について、出現頻度の時間的分布や特定の文書集合内における共起情報の中から特徴的なものを見つけ出し、知見につなげることである。その中でも、2つ以上の文書集合があった場合、集合間の差に相当する概念を抽出する差異分析は非常に有効な分析の一つである。例えば企業のコールセンターに蓄積されるデータ (コールメモ) にはエージェントが書くテキストデータの他に、問い合わせ対象製品やエージェント名等が定型データとして付与されている。このようなデータの分析の中で、製品ごと、またはエージェントごとの文書集合に注目し、それらの差を見つける分析は、特定の製品に固有の問題の同定や、優良エージェントが持つ knowhow の発見・共有などにつながることから、テキストマイニングの魅力的なアプリケーションである。

本研究では、文書サイズが大きいデータを対象にしたテキストマイニング分析手法を考える。今までテキストマイニング技術が主に適用されていたコールセンターにおけるコールメモデータは、エージェントによる会話の要約であり、1 文書のサイズは非常に短く、顧客の声を端的にまとめたものであった。それに対して、本研究ではプロジェクトの報告書、企業の決算報告書といった 1 文書のサイズが従来の対象データに比べて大きな文書データを対象にする。各文書のサイズが大きい文書は企業内または広く社会全体で読まれることを想定されており、文書中の各文も長く、複雑な構造を持っている。このような文書集合について文書集合間の差を求める分析手法を本研究では考える。

連絡先: 竹内 広宜, hironori@jp.ibm.com

文書集合間の差を求める手法の一つとして、分析観点 (カテゴリ) を事前に定義し、各観点に属する概念を辞書として準備する方法がある。各文書集合が辞書にあるキーワード・表現の中でどういったものを含んでいるかを分析することで、文書集合の差を分析することが可能となる。しかしながら、分析に有効な観点の設定や辞書の準備が分析の結果に影響する。従って有用な分析をするためには分析者の経験や勘が必要となる課題がある。

一方、特定の文書集合に対して、特徴的な表現を自動的に抽出する方法が行われている。これはある特定の文書集合において出現する表現に注目し、その文書集合内での出現頻度と他の文書集合内での出現頻度との差に注目し、特定の文書集合でのみ頻繁に出現するものを特徴的な表現として抽出するものである。このアプローチでは自動的に特徴的な表現が抽出されるが、上位に抽出される特徴的な表現は当たり前のものであることが多い。このような場合、既知の事実ばかりが得られ、今まで気がつかなかった有用知見が見つかることは少ない。また、各文書のサイズが大きい場合、抽出されたキーワード・表現だけでは知見を導き出せず、該当する元文書を精査する必要があり、分析は非効率となってしまう。

本研究では、このような問題に対して、文書間の差を見つけるために文脈を考慮した分析手法を提案する。以降ではまず 2 節において従来手法とその課題点について述べ、3 節で文脈情報を利用した分析手法を提案する。4 節では提案する分析手法の適用例としての企業の有価証券報告書の分析について述べ、5 節で分析結果を示す。そして最後に考察・まとめを行う。

2. 文書集合からの特徴量抽出

文書集合間の差を求める手法の一つとして、分析観点 (カテゴリ) を事前に定義し、各観点に属する概念を辞書として準備する方法がある [10]。例えば製造業のコールセンターでは部品名、苦情、要望、質問といった観点とそれぞれに該当する表現を事前に集め辞書に登録することがテキストマイニング分析が行われている。こうすることで製品ごとに分けられた文書集合

を分析する際、製品名を縦軸、設定した観点に属する表現を横軸にとり、両属性を満たす文書数を表す2次元表を作成することで、文書集合間の関係を概観することができる。しかしながら、分析観点の定義や辞書の準備は手作業であり、サンプル文書中における頻出語を元に分析者の勘を頼りに作成されることが多い。文書サイズが大きいデータの場合、文書中のどの箇所のどういった表現が分析に有効かということを事前に判断することは難しい。また、設定できたとしても重要な差を見つけられるとは限らないという課題がある。

一方、特定の文書集合に特徴的な表現を自動的に抽出する方法があり [5]、注目している文書集合の特徴を分析するために用いられている。特徴的な表現の抽出には文書集合内での出現頻度と他の文書集合内での出現頻度に注目し、相互情報量、χ^2 統計量、Kullback-Leibler 情報量といった指標が用いられている。分析だけでなく、文書分類においてもこういった指標を利用し、文書分類に有効な素性を選択することが行われており [8]、文書クラスタリングにおいても各クラスタリングのラベリングにこういった指標が用いられている [3]。しかしながら、こういった手法で抽出されるキーワードリストの上位に来る語は、注目している文書集合で非常に多く出るという観点で特徴的であると判定されるだけであり、必ずしも有用な知見につながる特徴語ではないことが多い。例えば、ハードディスクに関する文書集合において、ハードディスクや HDD といった語が上位に来ると推察されるが、これらは既知の概念であり知見には直接結びつかない。

また、ある文書集合に特徴的な語が見つかっても、知見を導き出すためには、なぜ特徴的な出現するのかを検証するために原文を参照する必要がある。分析の対象が数文から構成される文書集合である場合、該当文書を概観し、知見を導き出すことは比較的容易である。一方、1 文書辺りのサイズが大きくなると、特徴語からのみから知見を導き出すのは非常に難しく、該当文書で検証するためには結局本文全体を読むという非効率な状況になることが多くなるという問題点がある。

3. 文脈情報を利用した特徴量抽出と差異分析

従来の文書集合内における表現の頻度のみに注目した分析の課題点を踏まえ、本研究では文脈情報を利用した分析手法を提案する。

まず、分析手法の概要を述べる。文脈情報として、文脈を指し示す語（topic word）とそれが影響する範囲が与えられた時、分析手法は以下の 3 ステップから構成される。

- 各文脈において、出現するキーワードの注目している文書集合における特徴度を距離尺度を用いて計算する
- 注目している文書集合において特徴的な語を多く含む topic word（文脈語）を同定する
- 文脈によって注目している文書集合における特徴度が変化する語を抽出する

まず、文書集合が A における kwd の特徴度を計算について述べる。語の特徴度の計算については、$P(A|kwd)$ と $P(A)$ の間の距離を元に求める。距離尺度については前節で述べたような様々なものが提案されているが、その距離尺度は $P(A)$ に依存し、尺度の値から kwd が A に特徴的かどうかを適切に判定するのは難しい。例えば、距離尺度として Kullbeck-Leibler 距離を用いた際の $P(A)$ と $P(A|kwd)$ の関係を図 1 に示す。図 1 において KL 距離が a であるときの、各 $P(A|kwd)$ と baseline(KL 情報量が 0 である時、すなわち $P(A|kwd) = P(A)$) で囲まれる領域の面積を $S(a)$ と定義する。a を決定すると、各 $P(A)$ に対する $P(A|kwd)$ が求まり、$S(a)$ を計算することができる。

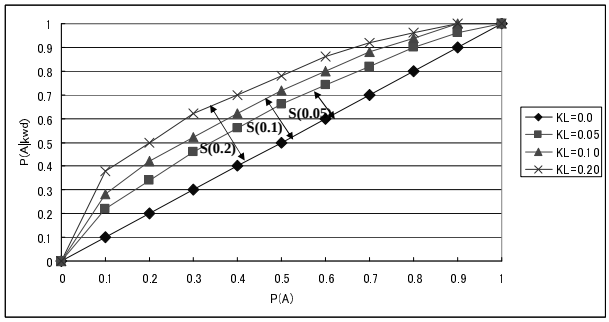

図 1: $P(A)$ と $P(A|kwd)$ との関係

ここで、$S(a)$ の値に注目し、τ_1 および τ_2 $(0 \leq \tau_1 < \tau_2 \leq 0.5)$ を特徴度を判定する閾値として設定する。τ_1 および τ_2 を用いて各 kwd の特徴度を以下のように判定する。

- $S(a) < \tau_1 \longrightarrow kwd$ は A において特徴的ではない
- $\tau_1 \leq S(a) \leq \tau_2 \longrightarrow kwd$ は A において特徴的である
- $\tau_2 < S(a) \longrightarrow kwd$ は A において非常に特徴的である

τ_i が決定されると $\tau_i = S(a)$ を満たす特徴度（距離尺度）スコア a を求めることができる。そして a が決定されれば、$P(A|kwd)$ と $P(A)$ との距離がこの値を越えるかどうかを判定することで kwd が特徴的かどうかを判定することができる。表 1 において距離尺度として KL 距離を用いた際の各 τ および $P(A)$ における $P(A|kwd)$ を示す。この表から、τ_1 および τ_2 を決定すると、kwd の特徴度を $P(A)$ に応じて判定することができる。

表 1: 各 τ および $P(A)$ における $P(A|kwd)$

		$P(A)$			
τ	KL	0.2	0.4	0.6	0.8
0.05	0.010	0.26	0.47	0.67	0.85
0.1	0.030	0.31	0.53	0.72	0.89
0.2	0.15	0.46	0.67	0.84	0.96
0.3	0.40	0.64	0.80	0.92	1.0

文脈語およびその影響範囲が与えられると、その文脈範囲内に出現する各キーワードや表現の特徴度を求めることができる。その結果、文脈語と文脈内のキーワードの組に対して特徴度が求まる。これらの結果から、文脈を利用した傾向抽出が可能となる。例えば、各文脈語ごとに、高い特徴度を示す語の数を求めることで、どの文脈が現在注目している文書集合を特徴付けているのか調べることができる。また、逆にキーワード・表現に注目し、文脈ごとの特徴度の推移を見ることで、文脈の違いで特徴度が異なるものを抽出できることができる。このような文脈によって特徴度が異なるキーワード・表現は文脈によって意味・役割が変わるものであると考えられ、気がつかなかった知見につながると考えることができる。さらに、特徴ありとして抽出される情報は文脈語とその範囲内に出現する表現との組であるため、原文を参照せずとも仮説を立てることができる可能性が高いという利点がある。

4. 有価証券報告書を用いた企業分析

本研究では企業の決算報告書の一つである有価証券報告書の分析を通して、提案手法の有効性を検証する。

4.1 分析の目的

投資などをはじめとした様々な目的のため、企業の財政状態や経営成績を評価する技法が試みられている。企業の財政状態や経営成績を客観的に示すものとして財務数値データがある。これを用いて企業を評価する試みは多く行われている。これに対し本研究では、非財務データを用いて企業評価を行うことを試みた。非財務データのうちテキストを対象とした先行研究としてはWｅｂ上のメッセージ [1]、株主への手紙 [4] を解析したものなどがある。これに対し、本研究では企業が提出する有価証券報告書内のテキストを解析対象とした。倒産企業と継続企業の差になる有価証券報告書中の記述を見つけるという差異分析が分析目的である。何らかの差が見つかれば、その知見を利用して企業評価を行なうことが期待できる。

4.2 データ

分析用データとして、倒産企業については 1999 年から 2005 年の間に倒産した企業 90 社の有価証券報告書を用いた。一方、非倒産企業については 2005 年の有価証券報告書を用いた。非倒産企業 (継続企業) は、全上場企業について SAF(Simple Analysis of Failure) 値 [9] を求め、得られた SAF 値を元に倒産企業のサンプルと同数 (90 社) となるよう、系統抽出を行った。SAF 値は財務情報を用いた倒産危険度を予測するモデルであり、企業格付けと相関が高い。したがって、選定した継続企業 90 社には、様々な財務状況の企業が含まれ、上場企業の代表値を表していると仮定することができる。有価証券報告書には多くの情報が記載されている。そこで本研究では、どの報告書にも記載がされている配当政策部分に注目して解析を行うこととした。

4.3 文脈情報の抽出と分析システム

企業の決算報告書やプロジェクト報告書といった日本語の報告文書では、各文の長さは長いものの、”～につきましては”、”～は”といった表現が文頭にあり、その文で何について報告するのかを述べる傾向がある。本実験では、この特徴を用いて、文頭の表現でパターンマッチングを行うことで文脈語を抽出する。また、文脈語の直後から、文の終わりまでを文脈範囲とする。抽出の例を図 2 に示す。

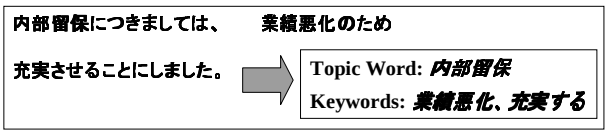

図 2: 文脈情報を用いた情報抽出の例

分析のために、報告書データに対して自然言語処理を適用し、動詞・名詞や特定の表現パターンを抽出し、文脈範囲や文書全体における出現頻度を計算できる分析システムを構築した。

5. 分析結果

文書集合全体の中で 3 文書以上に出てくる文脈語を分析の対象とした。以下は対象となる文脈語の例である。

配当金　資金　利益配分　内部留保　配当性向　中間配当

各文脈においてキーワード（および表現）を抽出し、継続企業に特徴的な表現を抽出した。閾値として $\tau_1 = 0.1$ および $\tau_2 = 0.2$ を設定し、各文脈ごとにキーワードの特徴度を求めた。表 2 は各文脈ごとの特徴語の分布を示している。この表において”Features”は各文脈において 10 文書以上で出現するキーワードおよび表現の異なり数を示す。各語に対して 3 節で述べた特徴度判定を行った (None, Slight, High)。”Other”は継続企業群ではなく、倒産企業群で特徴的である語の数を示している。倒産企業群の文書データについても同様の手順を適用した。

表 2: 各文脈における継続企業群の特徴語の分布

		Representativeness			
Topic Word (Context)	Features	No	Slight	High	Other
配当金	10	1	3	6	0
資金	13	4	1	6	2
利益配分	10	6	1	2	1
内部留保	41	10	8	11	12
配当性向	3	0	2	0	1
中間配当	26	4	19	3	0

この表から内部留保の文脈ではいくつかのキーワードは継続企業群で特徴的である一方、倒産企業群に特徴的な語も存在する。以下は内部留保の文脈における、継続企業群、倒産企業群それぞれにおける特徴語の上位リストである。

- 継続企業群：充当する、成長、設備投資、研究開発、競争、合理化、企業価値、生産設備、新規事業
- 倒産企業群：基本、充実する、応じる、安定、利益還元、状況

資金の文脈も同様の特徴語の分布を示している。

次に表 3 に従来通り文書全体に注目して抽出した特徴語を示す。この表から、”研究開発”や”企業価値”といったキーワードが継続企業群において特徴的であることがわかる。しかし、”内部留保”は両方の文書群において同程度に出現していた。また、”設備投資”、”生産設備”といったキーワードも表 3 では継続企業に特徴的であるとは判定されていない。提案手法では、従来のキーワードのみに注目する方法では特徴的であると判定されなかったこのようなキーワードが”内部留保”の文脈において非常に特徴的であると抽出されている。

表 3: 各文書群における特徴語

継続企業群	年間　研究開発　充当する　連結業績　中間　企業価値　取得　自己株式　開催　中間配当
倒産企業群	遺憾ながら　引き続き　損失　回復　大幅な　全力で　早く　見送る　至る　従って

この結果は、どの企業も内部留保について有価証券報告書において言及しているが、その文脈に出てくるキーワードは継続企業群と倒産企業群では異なるということを示している。表の 3 の結果に比べ、文脈語とその文脈で特徴的な語を抽出する本手法の結果から、”継続企業は企業価値を高めるために研究開発を行っている”という知見を対象文書を参照しないでも導出することができる。この結果、**”企業は、研究開発投資や新規事業を行うことにより収益を獲得できるのではなく、収益があり内部留保が確保できて初めて研究開発投資や新規事業への投資が可能になる”**ということがテキストマイニング分析では結論付けることができる。

次に抽出されたキーワードに注目する。表 3 から”中間配当”が継続企業群に、また”遺憾ながら”といった謝罪に関する表現が特徴的であるということがわかる。財務分析では、継続企業は配当は行えるが倒産企業は財務状態が非常に悪いため配当を行うことができない、という知見は広く知られた事実である。つまり、本分析において文書全体に注目して特徴語を抽出した場合には良く知られた知見しか得られないことを示している。テキストからの情報抽出において、”数字 + 円”といった表現パターンを金額表現として抽出した。この金額表

現は文書全体に注目した場合、特徴語としては抽出されなかったが、配当金の文脈において継続企業群における特徴語として抽出された。一方、業績の文脈において倒産企業群における特徴語として抽出された。この結果から、文脈によって意味・役割が変わるキーワード・表現があり、キーワードのみに注目して分析すると誤った知見を導出する可能性があるということがわかる。逆にこのような文脈によって意味・役割が変わるキーワード・表現の抽出は分析者が気がつかなかった知見につながる可能性があり、重要であると考えられる。

6. 考察

本研究では、特定の文書集合に特徴的な文脈語およびその文脈におけるキーワード・表現の組を抽出し、差異分析を行う手法を提案した。テキストマイニング分析では、抽出された語のみに注目するだけでは、知見を得るのには不十分であることが従来から指摘されている。この問題を解決するため、語と語の間の係り受け情報が頻繁に用いられている [10]。

例えば、製造業のコールセンターなどでは、”製品 AAA... 壊れた”、”修正パッチ ... 欲しい”といった名詞と動詞の係り受け情報は、製品の問題点や顧客の要望を同定する上で有効である。以下は有価証券報告書の分析において配当金、内部留保を含む名詞と動詞の頻出係り受け情報である。

- 配当金： 配当金 ... つきましては、配当金 ... なる、配当金 ... 決定する
- 内部留保： 内部留保 ... 必要だ、内部留保 ... 確保する、内部留保 ... 努める、内部留保 ... 充実する

抽出された情報は一般的なもので知見につながるものはほとんどない。コールセンターにおけるコンタクトメモはエージェントによる会話の要約であるため、”PC-ABC の HDD が壊れた”や”FIX パックが欲しい”といったように重要な点のみが簡潔に書かれている。一方、報告文書のようなデータでは各文が長くなり、構造が複雑となっており、文脈語は重要な概念と直接の係り受け情報を持っていない。そのため、適切な係り受け情報を抽出するためには、各データごとに抽出手段が必要となる。従来手法に比較し、本手法は意味のある情報を複雑な構造を持つ長い文からも頑健に抽出できることがわかる。

本手法では、文脈語と特徴あるキーワードの組を抽出している。文書内の共起を元に有用なキーワードの組み合わせを抽出し、提示する手法がある [2][6]。これらの手法では、単なら語の組み合わせであり方向の情報がなく、結果の解釈が難しい。一方、提案手法は文脈語からその文脈での特徴語という方向情報があるので、解釈が容易に行える。

本研究では文脈語とその範囲の抽出について日本語の報告書特有の性質を用いた。実験結果からヒューリスティックな抽出手法でも有用な知見が得られることがわかった。しかしながら、他の種類の文書にも適用するためには、文脈語と文脈範囲を抽出する汎用な手法も必要であり、今後の課題である。また、財務データのマイニングは、数値データを中心に広く行われている [7]。テキスト分析と数値データの分析との組み合わせも今後の課題である。

7. まとめ

本研究では、各文書が長いデータにおいて、文脈情報を利用した特徴語の抽出し分析する手法を提案した。そして提案手法を有価証券報告書の分析に適用し、継続企業群と倒産企業群の差を見つけ出すことを試みた。実験の結果、提案手法を適用することによって、従来の文書レベルでの特徴キーワード抽出による分析では見つけ出せなかった有用な知見が得られることが確認できた。今後の課題としては、文脈情報の抽出について汎用的な手法の導出や財務分析におけるテキスト分析と数値データ分析との統合などがある。

参考文献

[1] W. Antweiler and M. Z. Frank. Is all thattalk just noise? the information of internet stock message boards. *The Journal of Finance,*, 59(3):1259–1294, 2004.

[2] Y. Aumann, R. Feldman, Y. Yehuda, D. Landau, O. Liphstat, and Y. Schler. Circle graphs: New visualization tools fo text-mining. In *Proceedings of the 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pp. 277–282, 1999.

[3] F. Beil, M. Ester, and X. Xu. Frequent term-based text clustering. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 436–442, 2002.

[4] M. Clatworthy and M. J. Jones. Financial reporting of good news and bad news: evidence from accounting narratives. *Accounting and Business Research,*, 33(3):171–185, 2003.

[5] T. Hisamitsu and Y. Niwa. A measure of term representativeness based on the number of co-occurring sailent words. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, pp. 1–7, 2002.

[6] Y. Ohsawa, N. E. Benson, and M. Yachida. Keygraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In *Proceedings of IEEE International Forum on Research and Technology Advances in Digital Libraries (ADL)*, pp. 12–18, 1998.

[7] E. Vityaev and B. Kovalerchuk. Data mining fo financial applications. In *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practioners and Researchers*, pp. 1203–1224, Springer, 2005.

[8] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning (ICML)*, pp. 412–420, 1997.

[9] 白田. 企業倒産の予知モデル. 中央経済者, 2003.

[10] 那須川. コールセンターにおけるテキストマイニング. 人工知能学会誌, 16(2):219–225, 2001.

[11] 那須川. 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.