利益伸び率に着目した有価証券報告書のテキストマイニング

吉田, 慎一郎
九州大学大学院システム情報科学府情報知能工学専攻

中藤, 哲也
九州大学情報基盤研究開発センター

御手洗, 秀一
株式会社Lafla

廣川, 佐千男
九州大学情報基盤研究開発センター

出版情報：情報処理学会研究報告. 2013 (A-5-2), pp.1-5, 2013-03-14. 情報処理学会九州支部
バージョン：

利益伸び率に着目した有価証券報告書のテキストマイニング

吉田 慎一郎[1,a)]　中藤 哲也[2,b)]　御手洗 秀一[3,c)]　廣川 佐千男[2,d)]

概要：将来有望な企業への投資のために，投資家は様々な企業の活動状況を分析している．企業の活動状況を知る方法の一つに有価証券報告書がある．本研究では企業の利益伸び率に着目し，有価証券報告書のテキストデータとの関係性を分析する．

Text Mining of Securities Reports by Profit-Growth Rate

Shin-ichiro Yoshida[1,a)]　Tetsuya Nakatoh[2,b)]　Shuichi Mitarai[3,c)]　Sachio Hirokawa[2,d)]

Abstract: A stock market is a base of the economic activity in present-day free economy society. In order to become a listed company, there is a severe examination. Furthermore, yearly duty is attached by law for the listed company to file the fixed form report about the financial condition. In this paper, correspondence of the numeric data about the activity of a company and text data is analyzed for the financial report.

1. はじめに

将来有望な企業へ投資するため，投資家は様々な企業の活動状況を分析している．株価変化や時事ニュースなどの分析による業界動向や株価の予測に多くの感心が集まっている．例えば，Lafla 社[*1]による UFOReader[*2] and UFOLenz[*3]は，その様な分析のための Web サービスによるツールの実例だ．

企業の分析のためのデータとしては，ニュース，更にツイッターやブログなど非常に多くの種類のデータがある．一方，有価証券報告書は，外部の風評や評価ではなく，各企業の実績そのものである．株式市場は現代の資本主義社会における経済活動の基幹であり，上場には厳しい資格審査がある．また，上場企業は毎年自社の経営状況を有価証券報告書という定まった形式で報告する義務が課せられている．このため有価証券報告書は，最も基礎的な企業の活動報告と言える．企業が出す情報としては，Web 上の IR 情報があるが，有価証券報告書は記述すべき項目が法律により規定され，期ごとの提出が義務付けられている．すなわち，より信頼性の高い公的文書である．

本稿では，この有価証券報告書のテキストデータを分析し，企業の利益伸び率との関連性を分析する．

2. 関連研究

Twitter，Blog，新聞記事などを用いて，株価予測を行う研究が盛んに行われている．例えば，企業の経営状況に関する Twitter の内容を適切に分析できれば，よりリアルタイムな株価予測が可能となる．Twitter のツイートをポジティブ，ネガティブに分類することで感情の評価値を算出し，株価の予測を行う研究がある [2], [10]．同様に，新聞記事を対象としてポジティブ／ネガティブ推定の研究もある [7]．また，一般の Blog 記事について，コメントの

1 九州大学大学院システム情報科学府 情報知能工学専攻
Graduate School of Information & Research Institute for Science and Electrical Engineering, & Information Technology,
Kyushu University, 6-10-1, Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan
2 九州大学 情報基盤研究開発センター
Research Institute for Information, Kyushu University, 6-10-1, Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan
3 株式会社 Lafla
Lafla Inc., Fukuoka Institute of System LSI Design Industry 507, 3-8-33 Momochihama, Sawara-ku, Fukuoka 814-0001, Japan
a) 2IE11096G@s.kyushu-u.ac.jp
b) nakatoh@cc.kyushu-u.ac.jp
c) mitarai@lafla.co.jp
d) hirokawa@cc.kyushu-u.ac.jp
*1 http://www.lafla.co.jp/
*2 http://www.uforeader.com/v1/
*3 http://www.yano.co.jp/ufolenz/index.php

数，コメントの長さなどと，株価変化との関連を調べる研究 [3], [4] がある．[12] らは，新聞記事に付与されるテーマに着目して，新聞記事の株価変動への影響と株価変動の外部要因を分析することによって株価データと新聞記事の関連性の分析している．[8] は，医薬品関連企業を対象として，株価と Web 上の IR 情報との関係を分析している．また，[6] は，米国企業を対象として好調な企業の security report に現れる特徴語の関連を可視化して分析を行っている．そこで分析に使われた株価は，投資家がその企業をどのように見ているかを評価する指標である．

有価証券報告書を分析対象とする研究としては [9] がある．分析範囲を有価証券報告書の配当政策の部分に限定し，倒産企業と存続企業の違いをテキストマイニングにより抽出している．分析を配当政策に限定したのは彼らの経験に基くものであるが，我々は，より広い範囲で特徴語学抽出を行った．

[1], [5] は分析文書として倒産情報の記事を選び，倒産理由を表す文を限定し，それらの文に現れる特徴語の共起関係を分析している．文書群から特徴語を抽出し，あるいはさらに特徴語の相互関係を分析するという点では，本稿と同じ目的の研究である．しかし，数値データに基づいた特徴語等のテキスト分析が本稿の特色である．

[11] では，有価証券報告書の内容を報告書単位で分析し，特徴を求めている．本稿では，企業単位で特徴を求め，更に特徴語から条件にあった企業を抽出する事を試みている．

3. 解析対象データ

3.1 有価証券報告書

有価証券報告書とは，有価証券の発行者である会社が，経理状況その他事業の内容に関する重要な事項その他の公益または投資者保護のため必要かつ適当なものとして内閣府令で定める事項を記載した報告書である．有価証券報告書は毎事業年度 3 ヶ月以内に内閣総理大臣に提出することが義務付けられており，虚偽記載をした場合は罰せられるため，企業の情報として信頼のできる文書といえる．本研究において分析対象とした有価証券報告書の電子化テキストデータおよび数値データは，株式会社 Lafla から提供して頂いた．

有価証券報告書の書類様式は，大きく第一部と第二部に分かれている．第一部は企業情報，第二部は提出会社の保証会社等の情報が記載される．第一部の企業情報には，事業の内容や事業等のリスクなどが記載されており，その企業の経営状況を知るための重要な情報を得ることができる．本稿ではこれらの項目のうち，解析に足る量のテキスト部分を持った項目を取り上げて解析対象とした．解析対象項目については，後述する．

3.2 データの前処理

本稿では，株式会社 Lafla より提供いただいた電子化テキストデータ・数値データを分析に使用した．このデータには，西暦 2006 年から 2012 年までの 7 年間にかかる 76 業種，全 4493 社の有価証券報告書が含まれている．

これらの有価証券報告書は，企業等の統廃合やその他の理由により一部のデータが不完全であった．それらの部分は解析に向かないため，前処理で削除する事とした．具体的には，同一会計年度の有価証券報告書が 2 つ存在する場合，当期売上，もしくは純利益が空欄または 0 のデータを削除対象とした．その結果，解析対象の全企業数は 4493 社，その有価証券報告書数は 20339 件となった．

分析の第一段階として，分析対象の業種を限定した．具体的には，医薬品製造業（業種コード S0300）に分類される 66 社を対象とした．66 社の企業名については割愛する．

3.3 数値による対象企業の選択

企業の利益と有価証券報告書の関係を導き出す分析の第一ステップとして，業績が伸びている企業を目視で抽出した．図 1 は医薬品製造業 66 社の利益のデータを企業ごとの折れ線グラフにしたものであり，表記のスケールは，企業ごとに正規化してある．

このように一覧にすることで，相互に比較したい企業（ここでは医薬品製造業）の業績の変化を把握することが可能である．本報告ではこの中から，順調に業績を伸ばしている企業を目視により抽出した．抽出された企業は，表 1 にまとめる．本稿では，以降この 10 社を**好業績 10 社**と称する．

表 1 Growth Corporation

Code	Name of company
E00923	塩野義製薬
E00924	田辺三菱製薬
E00935	科研製薬
E00942	ロート製薬
E00944	久光製薬
E00947	持田製薬
E00960	ビオフェルミン製薬
E00974	東和薬品
E00975	富士製薬工業
E00976	沢井製薬

次節でこれらの企業に共通する特徴を求める．

4. 特徴語

4.1 特徴語抽出方法

有価証券報告書には法律に定められた様式があり，複数の項目に分かれている．このうち，重要でかつ多くの有価

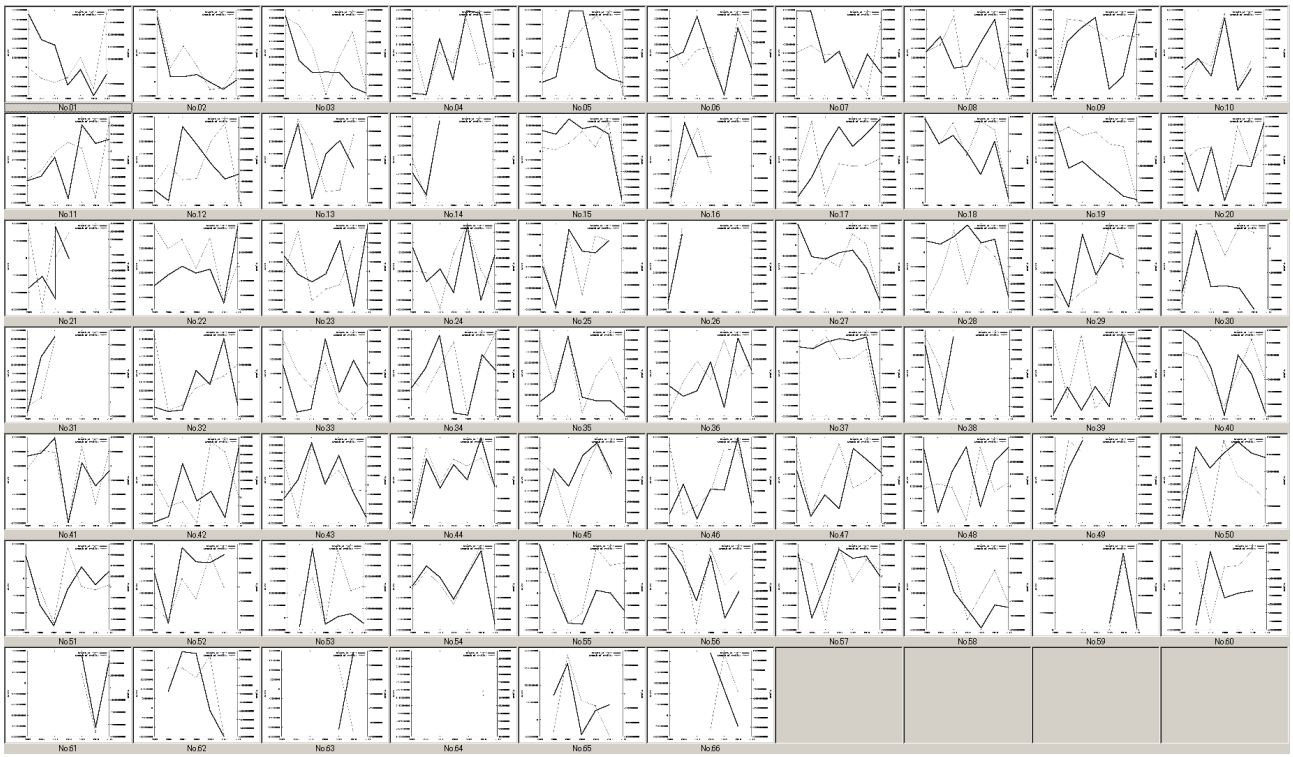

図 1 企業の利益の変動グラフ

証券報告書でデータを含んでいる項目を選択し，特徴語を抽出する対象データとした．本稿で特徴語の抽出に用いた有価証券報告書における項目とその項目を持つ有価証券報告書データ数及び企業数を表 2 に示す．

単語の特徴量の計算，及び特徴語の抽出には，汎用連想計算エンジン GETA[*4]を用い，weight の計算アルゴリズムには GETA の持つ SMART を用いた．特徴計算の対象語句は，名詞及び動詞とした．また，各有価証券報告書に出現した語がどの項目に現れているのかを区別するために，元の語句に加え，語句の前に項目名を “abst:” （表 2 参照）の形式で追加した語句も特徴計算の対象とした．これにより，特定の項目に出現する語句が大きな特徴を持っている場合に，他の項目に出現する同じ語句と区別して抽出することが可能となる．

4.2 特徴語抽出

医薬品製造業全 66 社の有価証券報告書を企業ごとにまとめ，3.3 節で抽出した企業の特徴語を求めた．得られた特徴語の weight 値上位 20 件を表 3 に示す．

得られたこれらの特徴語が継続して利益が増える企業を特定できるかを判断するために，これらの個々の単語を用いて文書検索を行い，精度，再現率，及び F 値を求めた．結果を表 4 に示す．

何れの単語も好業績 10 社の一部を抽出しているものの，抽出の精度，再現率は共に充分とは言えない．一つの単語では情報不足の面が否めない為，これらの単語の組み合わせによる AND 検索を試みた．ランキング上位から単語を選び，AND 検索の検索語として追加して言った場合の，精度，再現率，F 値を求めたものを，表 5 に示す．この表から，単純な AND 検索では目的とする文書を特定できない事が分かる．一方で，2 単語から 3 単語に増やした場合のように，単語を増やすことで再現率が下がらずに精度が上がる場合もある．これは，適切な組み合わせが必要であることを示していると言える．

5. まとめと今後の課題

本稿では，有価証券報告書を対象として，企業の活動についての数値データとテキストデータの対応を分析した．具体的には，医薬品関連の上場企業 66 会社から，業績が順調に伸びている企業を選びその理由を表す特徴語を抽出した．この結果は，数値データの変化に関する理由を，有価証券報告書をつぶさに読み解くことなく自動的に抽出できる可能性を示している．

しかし一方で，特徴語から対象文書を求めることは容易では無いことが明らかになった．今後は，適切な特徴語の組み合わせなどを発見し，あるいは数値データに対する文書群を自動分類するなど，分析に適した手法を明らかにする必要がある．

[*4] http://geta.ex.nii.ac.jp/geta.html

表 2 有価証券報告書の項目

項目名	有価証券報告書における項目名
abst	第 2【事業の状況】の 1【業績等の概要】
busi	第 1【企業の概況】の 3【事業の内容】
conn	第 1【企業の概況】の 4【関係会社の状況】
cont	第 2【事業の状況】の 5【経営上の重要な契約等】
empl	第 1【企業の概況】の 5【従業員の状況】
fin	第 2【事業の状況】の 7【財政状態及び経営成績の分析】
hist	第 1【企業の概況】の 2【沿革】
ind	第 1【企業の概況】の 1【主要な経営指標等の推移】
issue	第 2【事業の状況】の 3【対処すべき課題】
prod	第 2【事業の状況】の 2【生産，受注及び販売の状況】
rd	第 2【事業の状況】の 6【研究開発活動】
risk	第 2【事業の状況】の 3【事業等のリスク】
seg	第 5【経理の状況】の 1【連結財務諸表等】中の【事業の種類別セグメント情報】

表 3 好業績 10 社の特徴語

順位	weight	単語	順位	weight	単語
1	2.350	issue:続ける	11	1.972	issue:後発
2	2.129	prod:変更	12	1.968	fin:人税
3	2.097	hist:営業	13	1.968	fin:払法
4	2.081	issue:品質	14	1.963	abst:加算
5	2.075	hist:福岡	15	1.958	hist:支店
6	2.053	福岡	16	1.951	人税
7	2.009	リーディング	17	1.951	払法
8	1.998	処方せん	18	1.946	物価
9	1.991	続ける	19	1.936	issue:競争
10	1.973	適正	20	1.934	hist:大阪

参考文献

[1] Takahiro Baba, Tetsuya Nakatoh and Sachio Hirokawa, “Text Mining of Bankruptcy Information using Formal Concept Analysis,” Proc. of *3rd International Conference on Awareness Science and Technology (iCAST2011)*, pp.527–532, 2011.

[2] Johan Bollen, Huina Mao and Xiao-Jun Zeng, “Twitter mood predicts the stock market.” *Journal of Computational Sci- ence*, Volume 2, Issue 1, March 2011, pp. 1–8, 2011.

[3] De Choudhury, M., Sundaram, H.,John, A.Seligmann, D.D. “Can blog communication dynamics be correlated with stock market activity?”, *Proceedings of the 19th ACM Conference on Hypertext and Hypermedia, HT ’ 08 with Creating ’ 08 and WebScience ’ 08*, pp. 55–59, 2008.

[4] Eric Gilbert and Karrie Karahalios, “Widespread Worry and Stock Market”, *4th International AAAI Conference on Weblogs and Social Media(ICWSM)*, 2010.

[5] Sachio Hirokawa, Takahiro Baba and Tetsuya Nakatoh, “Search and Analysis of Bankruptcy Cause by Classification Network,” *Model and Data Engineering, Lecture Notes in Computer Science*, Volume 6918/2011, pp.152–161, 2011.

[6] Kun Qian, Sachio Hirokawa, Kenji Ejima and Xiaoping Du, “A fast associative mining system based on search engine and concept graph for large-scale financial report texts,” Proc. *2nd IEEE ICIFE (Information and Financial Engineering)*, pp.675–679, 2010.

[7] Hiroyuki Sakai, Shigeru Masuyama, “Estimation of Impact Contained in Articles about each Company in Financial Articles,” *Information Processing Society of Japan (IPSJ)*, vol 94, pp.43–50, 2006. (in Japanese)

[8] Toshihiko Sakai, Masashi Matsushita, Brendan Flanagan, Jun Zeng and Sachio Hirokawa, “Analysis of influence of Investor Relation Documents to stock price”, Proc. of *FSKD2012 : 9th International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 1280–1284, 2012.

[9] Cindy Y. Shirata and Manabu Sakagami, “An Analysis of the ”Going-Concern Assumption”: Text Mining from Japanese Financial Reports,” The Journal of *Emerg-*

表 4 特徴語から好業績 10 社を抽出した場合の精度と再現率

順位	単語	weight	文書頻度 (好業績 10 社)	文書頻度 (全 66 社)	Recall	Precision	F 値
1	issue:続ける	2.35	6	15	0.60	0.40	0.48
2	prod:変更	2.129	7	20	0.70	0.35	0.47
3	hist:営業	2.097	10	37	1.00	0.27	0.43
4	issue:品質	2.081	9	33	0.90	0.27	0.42
5	hist:福岡	2.075	6	20	0.60	0.30	0.40
6	福岡	2.053	7	23	0.70	0.30	0.42
7	リーディング	2.009	5	9	0.50	0.56	0.53
8	処方せん	1.998	5	8	0.50	0.63	0.56
9	続ける	1.991	7	24	0.70	0.29	0.41
10	適正	1.973	9	37	0.90	0.24	0.38
11	issue:後発	1.972	8	21	0.80	0.38	0.52
12	fin:人税	1.968	9	33	0.90	0.27	0.42
13	fin:払法	1.968	9	33	0.90	0.27	0.42
14	abst:加算	1.963	7	17	0.70	0.41	0.52
15	hist:支店	1.958	7	27	0.70	0.26	0.38
16	人税	1.951	9	34	0.90	0.26	0.41
17	払法	1.951	9	34	0.90	0.26	0.41
18	物価	1.946	4	6	0.40	0.67	0.50
19	issue:競争	1.936	10	43	1.00	0.23	0.38
20	hist:大阪	1.934	10	49	1.00	0.20	0.34

表 5 特徴語の AND 検索による企業選択の精度と再現率

単語 （AND 検索）	文書頻度		Recall	Precision	F 値
	好業績 10 社	全 66 社			
issue:続ける	6	15	0.6	0.4	0.48
issue:続ける prod:変更	4	8	0.4	0.5	0.44
issue:続ける prod:変更 hist:営業	4	5	0.4	0.8	0.53
issue:続ける prod:変更 hist:営業 issue:品質	3	4	0.3	0.75	0.43
issue:続ける prod:変更 hist:営業 issue:品質 hist:福岡	3	3	0.3	1	0.46
issue:続ける prod:変更 hist:営業 issue:品質 hist:福岡 福岡	3	3	0.3	1	0.46
issue:続ける prod:変更 hist:営業 issue:品質 hist:福岡 福岡 リーディング	1	1	0.1	1	0.18
issue:続ける prod:変更 hist:営業 issue:品質 hist:福岡 福岡 リーディング 処方せん	0	0	-	-	-

ing Technologies in Accounting, Strategic and Emerging Technologies Section of the American Accounting Association, pp.1–16, 2009.

[10] Mike Thelwall, evan Buckley and Georgios Paltoglou, “Sentiment in Twitter Events,” Journal of the *American Society for Information Science and Technology*, Volume 62, Issue 2, pp. 406–418, 2011.

[11] Shin-ichiro Yoshida, Tetsuya Nakatoh, Shuichi Mitarai and Sachio Hirokawa, “Text Mining of Securities Reports for Discovering Reason of Change”, Proc. of *CAINE-2012: ISCA 25th International Conference on Computer Applications in Industry and Engineering*, New Orleans, Louisiana, USA, 2012.11.14-16.

[12] He Zhang and Shigeki Matsubara, “Quantitative Analysis of Relevance between News Articles and Stock Price Change”, *Information Processing Society of Japan (IPSJ)*, pp.183–184, 2008. (in Japanese)