4K2-2

新聞記事上の分散表現の時系列変化と企業業績の連動性

The correlation between time series change of distributed representation learned from newspaper and corporate performance

金子大輝 *1 大知正直 *2 森純一郎 *3 坂田一郎 *4
Daiki Kaneko Masanao Oochi Junichiro Mori Ichiro Sakata

*1東京大学大学院工学系研究科
School of Engineering, The University of Tokyo

Getting the idea from the study about distributed representation and time series meaning change of words, we propose the model acquired from news data which represents companies in time series on the assumption that distributed representations of companies change depending on the circumstances which surround them.

1. はじめに

近年, 企業の不祥事が数多く起こっている. 日本だけで見ても 2015 年に起きた旭化成建材のくい打ちデータ偽装問題や東芝の不正会計問題, 2016 年に起きた三菱自動車の燃費データの不正問題は記憶に新しい. 一度不祥事が明るみに出るとその企業のブランドイメージは大きく傷つくことになる. 不祥事が企業の業績や株価にどのような影響を及ぼすかということは, 経営者や投資家などにとって大きな関心事となっている.

それらの企業不祥事の影響を分析するため, テキストマイニングやウェブマイニングの分野では, 新聞記事情報 [和泉 11, 中山 14] やウェブ上の情報 [上野山 13] から企業に関する情報を抽出することで, 業績や株価を予測する研究 [藤本 13, 石黒 14] が数多く行われている. これらの研究は, 企業間の関係や企業と語の関係を共起情報に基づき抽出した上で, 企業とそれに関する情報がどのように記事やウェブに記述されているかを定量的に表現するモデルを構築している. それらのモデルは, 企業活動指標との相関 [上野山 13], 市場平均株価の予測, 株価の変動予測, など様々な応用に有効であることが示されている.

これらの従来研究の多くは, 不祥事の発生のような, 任意の時点において対象企業をよく表すようなモデルを, 企業と共起する語の情報を元に構築している. 一方, 不祥事を始めとした様々な要因による急速な業績の変化, 企業改革, 業態の変化など, 昨今の企業を取り巻く環境の変化が企業へ与える影響は大きいものであり, そのような変化を中長期に渡って逐次把握した上で的確な意思決定を行うことは, 経営者や投資家などにおいて重要課題となっている.

本研究では, 不祥事のような企業の経営に関わる事象が企業経営に与える影響を中長期的に分析するため, 大規模な時系列記事データから企業の情報を抽出しその変化を定量化する手法を提案する. 特に, 企業に関する情報とその変化は, その企業が記述される文脈の変化に表れているという仮定の元で, 企業に関する様々な情報が含まれる経済新聞データを対象として, 記事中に出現する企業エンティティの分散意味表現とその時系列変化を学習する. その上で, 学習した企業エンティティの表現とその変化について, ケース分析や業績などの外部指標との関連を分析することで, 不祥事などの環境変化によって企業の評判や世の中のイメージがどのように変化したのか評価を行う.

連絡先: 金子大輝, 東京大学工学系研究科, daikikaneko83@gmail.com

2. 手法

2.1 提案手法

本研究では, 企業に関する情報とその変化は, その企業が記述される文脈の変化に表れているという仮定の元で, 企業に関する様々な情報が含まれる経済新聞データを対象として, 記事中に出現する企業エンティティの分散意味表現とその時系列変化を学習する. その上で, 学習した企業エンティティの表現とその変化について, ケース分析や業績などの外部指標との関連を分析することで, 不祥事などの環境変化によって企業の評判や世の中のイメージがどのように変化したのか評価を行う. 以下, 本研究の提案手法の手順を説明する.

まず, 対象の企業群の基本となる分散意味表現を学習するため, 対象期間すべての記事データを用いて企業エンティティの表現ベクトルを獲得する. 本研究においては, 2012 年から 2016 年までの 5 年間を対象期間とし, その期間内の日本経済新聞電子版のすべて記事データを用いて学習を行う. このようにして学習された企業エンティティの表現ベクトル群をオリジナルモデルと呼ぶことにする.

次に, 対象期間の各年次ごとに企業エンティティの表現をその年の記事データのみを用いてオリジナルモデルに対して再学習を行うことで当該年のモデルを学習する. このように, オリジナルモデルを元に各年次ごとに再学習を行うことで, 異なる年の企業エンティティの表現ベクトルを分散意味表現の共通の特徴空間において比較することが可能になる.

続いて, 対象期間において各企業の表現ベクトルがどれほど変化したのかを評価する尺度として, cos 類似度とユークリッド距離の 2 通りの方法を試している. これらを用いて表現ベクトルの変化が大きい企業を割り出している.

次に, 企業の対象期間での表現ベクトルの変化を可視化している. これは対象企業について各年での表現ベクトルの類似語を取り出し, t-SNE により次元圧縮することで行っている.

最後に対象企業について, 得られた分散意味表現と業績とに相関性があるかどうかの検証を行う. ここでは再学習の際に用いる一年分の記事データを四半期ごとにずらして再学習することで合計 17 個のモデルを作成している. 各企業について前年同期のモデルとの cos 類似度と前年同期比の売上高増加率の絶対値との相関について検証した.

以上が本研究の主な手順である. 以下で詳細を述べていく.

2.2 各期間での分散意味表現の再学習

本研究では分散的意味表現の学習に gensim の skipgram モデルを用いている. 全てのモデルを次元数 200 で学習してい

る．各年の記事について別々に学習を行うと，それぞれの特徴空間が異なるために比較することができない．この問題を解決するために本研究ではモデルの再学習を行っている．図 1 に示すように，まず 2012 年から 2016 年の 5 年分の記事データについてまとめて学習を行いオリジナルモデルを作成する．その後，各年の記事を使ってオリジナルモデルに対して再学習を行い各年のモデルを作成した．

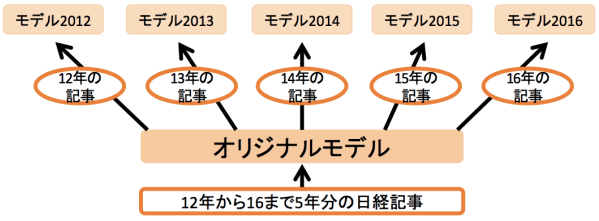

図 1: モデルの再学習のイメージ図

2.3 cos 類似度とユークリッド距離

本研究では，対象企業の分散表現が対象期間においてどれだけ変化したのかを評価する尺度として cos 類似度とユークリッド距離を用いた．cos 類似度は分散表現の局所的変化を検出するため，ユークリッド距離は対象期間に渡って特徴空間上をどれだけ移動したかという大域的な変化を検出するために使用した．この節では，説明のため 2012 年から 2016 年までの各年のモデルが存在しているとし，企業 $\mathbf{w}$ の t 年でのモデルの分散表現を $\mathbf{w}_t$ と表すとする．

ベクトル $\mathbf{x}$ と $\mathbf{y}$ の cos 類似度は $\cos(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{|\mathbf{x}||\mathbf{y}|}$ と求められる．cos 類似度については，評価対象の各企業について再学習された各年のモデルの分散表現を取り出して各年同士の cos 類似度を計算し，その最小値をその企業の cos 類似度としている．企業 $\mathbf{w}$ の cos 類似度の計算の組み合わせは $\cos(\mathbf{w}_{2012},\mathbf{w}_{2013}), \cos(\mathbf{w}_{2012},\mathbf{w}_{2014}), \cos(\mathbf{w}_{2012},\mathbf{w}_{2015}), \ldots, \cos(\mathbf{w}_{2015},\mathbf{w}_{2016})$ の 10 通りである．この 10 個の値の最小値を企業 $\mathbf{w}$ の cos 類似度としている．

ベクトル $\mathbf{x} = (\mathbf{x_1},\mathbf{x_2},\ldots,\mathbf{x_n}), \mathbf{y} = (\mathbf{y_1},\mathbf{y_2},\ldots,\mathbf{y_n})$ のユークリッド距離 $d(\mathbf{x},\mathbf{y})$ は $d(\mathbf{x},\mathbf{y}) = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\ldots+(x_n-y_n)^2}$ と求められる．ユークリッド距離に関しては，前年度のモデルの分散表現とのユークリッド距離を対象期間で合計したものをその企業のユークリッド距離としている．企業 $\mathbf{w}$ のユークリッド距離 $d_{\mathbf{w}}$ は，$d_{\mathbf{w}} = d(\mathbf{w}_{2012},\mathbf{w}_{2013}) + d(\mathbf{w}_{2013},\mathbf{w}_{2014}) + d(\mathbf{w}_{2014},\mathbf{w}_{2015}) + d(\mathbf{w}_{2015},\mathbf{w}_{2016})$ と表すことができる．

3. 実験

3.1 使用したデータ

本研究では 2012 年から 2016 年までの 5 年間の日本経済新聞の日経電子版の全ての記事を使用した．各年の記事数を表 1 に示している．

3.2 対象とした企業

本研究では研究対象の企業として，日経平均株価の構成銘柄及びその主要子会社の計 279 社を用いた．ただし学習が十分に行われていることを保証するために，全ての実験において評価対象の企業をこれらの企業の中でも 5 年間の記事での出現回数が 500 回以上のもの計 129 社に絞っている．

表 1: 本研究で使用した各年の記事数

年	2012	2013	2014	2015	2016
記事数	103363	74584	97532	96594	89627

3.3 結果と考察

3.3.1 分散表現の変化が大きい企業

表 2: cos 類似度及びユークリッド距離から求めた分散表現の変化が大きい企業

年	cos 類似度	企業	距離	企業
12, 15	0.7926	東芝	31.04	東京電力
12, 15	0.8603	日立製作所	30.77	東芝
12, 16	0.8843	東京電力	28.42	ソニー
13, 16	0.8903	日産自動車	26.99	日産自動車
12, 15	0.8911	パナソニック	26.91	日立製作所
12, 15	0.8944	ソニー	26.86	全日本空輸
12, 15	0.8959	全日本空輸	26.63	ソフトバンク
12, 16	0.9008	三菱自動車	26.54	トヨタ自動車
12, 15	0.9015	本田技研工業	25.83	イオン
12, 14	0.9026	関西電力	25.60	パナソニック

表 2 は対象企業において cos 類似度が小さい企業及びユークリッド距離が大きい企業を上位 10 社を示したものである．cos 類似度の年の列は最小値の cos 類似度に対応する年の組み合わせを示している．どちらの指標も東芝，日立製作所，パナソニックなどの大手総合電気メーカー，日産自動車などの自動車会社，東京電力などの電力会社が並んでいることが分かる．どちらの指標でも不正会計問題が多く報じられた東芝が上位に位置している．特に cos 類似度の方には東芝以外にも東京電力や三菱自動車などの不祥事を起こした企業が並んでおり，不祥事を起こした企業を特定する指標として有効であると言える．

3.3.2 ケーススタディ（東芝）

次にベクトルの変化度合いが大きかった企業のうち，上位に位置している東芝について詳しく見ていく．

表 3 は 2012 年及び 2015 年のモデルでの東芝の類似語を 10 個示したものである．どちらの年も東芝の競合企業や関連企業が並んでいることでは共通している．違いとして 2015 年の類似語には，2011 年に粉飾決算事件を起こしたオリンパスや経営危機が記事に取り上げられたシャープ，ルネサスといったネガティブな点を記事に書かれた企業が含まれていることが挙げられる．これは東芝が不正会計問題を起こしたことでネガティブな内容の記事を書かれてそれらの記事を再学習したことで，上で挙げた企業との類似度が上がったためと考えられる．

次に東芝の類似語の時系列での変化を視覚化した．具体的には，各年のモデルの東芝の類似語を cos 類似度が大きいものから順に 10 個取り出し，東芝及びその類似語のベクトルを t-SNE を用いて二次元に次元圧縮し視覚化した．視覚化の際に各年のモデルの東芝を区別するために，2012 年のモデルは東芝 2 のように，語尾にその年の一の位の数字を付け足している．類似語については複数年で重複して出現するものがあるため，ベクトルは再学習前のオリジナルモデルのものを用いてい

表 3: 2012 年及び 2015 年のモデルの東芝の類似語

2012 年		2015 年	
cos 類似度	企業	cos 類似度	企業
0.8137	パナソニック	0.7384	オリンパス
0.8074	三菱電機	0.6717	パナソニック
0.8023	日立製作所	0.6325	シャープ
0.7590	富士通	0.6298	ソニー
0.7469	ＮＥＣ	0.6250	三菱電機
0.7370	三菱重工業	0.6142	富士通
0.7332	デンソー	0.6017	キヤノン
0.7085	富士電機	0.5896	ルネサス
0.7003	キヤノン	0.5850	三菱重工業
0.6995	ソニー	0.5820	京セラ

表 4: 2012 年及び 2015 年のモデルの東芝の固有名詞以外の類似語

2012 年		2015 年	
cos 類似度	企業	cos 類似度	企業
0.5854	半導体大手	0.5294	粉飾決算
0.5492	プラズマテレビ	0.5258	損失隠し
0.5273	リチウムイオン電池	0.5042	不正会計
0.5220	マイコン	0.5033	半導体事業
0.5209	フラッシュメモリー	0.4846	テレビ事業
0.5177	画像センサー	0.4746	プラズマテレビ
0.5145	半導体事業	0.4500	半導体大手
0.5114	パワー半導体	0.4485	会計処理
0.5106	中小型液晶パネル	0.4478	会計
0.4998	航空機エンジン	0.4389	不祥事

る．図 2 が視覚化した図である．東芝に注目すると 12 年, 13 年は図の中央付近にいるが 14 年以降は上下左右に大きく動いていることが分かる．また, 図の右下には業績低迷に苦しんでいたシャープ, ルネサス, ソニーや不祥事を起こしたオリンパスなどネガティブなイメージを持つ企業が集まっていることが分かる．東芝は 2015 年に不正会計問題を起こしそのことを記事で数多く取り上げられたため, この東芝 5 がネガティブな企業群の近くに寄っていると考えられる．さらに 2016 年になると, 東芝 6 はネガティブな企業群からは遠ざかっている．2016 年においても東芝の不正会計問題に関しては多くの記事が出ていたので, 本来なら東芝 6 もネガティブな企業群の近くに位置するはずであるが, その点に関しては期待通りの結果とはならなかった．

表 3 から分かるように, 東芝の類似語の上位 10 単語はどちらの年も全て企業名となっている．そこで企業名以外の類似語を見るために, 固有名詞以外の類似語上位 10 単語を抜き出した．表 4 に 2012 年と 2015 年のモデルの固有名詞以外の類似語を示している．2012 年の類似語を見ると, プラズマテレビやリチウムイオン電池など 2012 年当時の東芝の事業内容に関する単語が多く並んでいることが分かる．一方で 2015 年の類

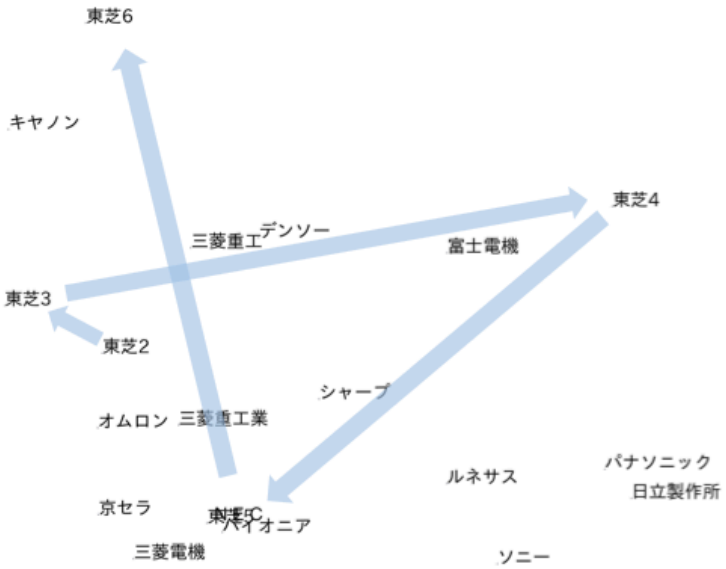

図 2: 東芝とその類似語の分散表現を視覚化した図

似語を見ると, 粉飾決算や損失隠し, 不正会計など不正会計問題に関係する単語が多く並んでいることが分かる．これは不正会計問題について多くの記事が書かれたためと考えられる．続いて, 先ほど同様に東芝と各年のモデルの固有名詞以外の東芝の類似語上位 10 単語を t-SNE により視覚化した．図 3 がその図である．図中央の少し下の部分に NAND 型フラッシュメモリーやテレビ事業, リチウムイオン電池, マイコンなど東芝が当時扱っていた商品に関係する単語のクラスタがある．このクラスタ上部には不正会計や会計処理, 会計など不正会計に関係する単語が集まっている．東芝の各年での移り変わりに注目すると, 東芝 3, 東芝 4 は商品のクラスタの近くにあるが東芝 5 は不正会計に関する単語群に移動している．そして東芝 6 はさらに図の上部に移動していることが分かる．不正会計問題が起こり不正会計に関する単語との類似度が上がったことで, 東芝 5 は事業に関する単語のクラスタから離れ不正に関する単語の方へ移動したと考えられる．一方で東芝 2 も東芝 5 とほぼ同じ所に位置している．不正会計事件が起こる前なので東芝 3 や東芝 4 の近くに位置するはずだが期待通りの結果にはならなかった．

3.3.3 分散表現と業績の相関

最後に対象企業の分散表現と業績との相関について検証を行った．ここでは再学習の際に用いる一年分の記事データを四半期ごとにずらして再学習することで合計 17 個のモデルを作成した．各企業について前年同期のモデルとの cos 類似度と前年同期比の売上高増加率の絶対値との相関係数を求めた．表 5 は, 対象企業のうち無相関検定を行い有意な相関が認められた企業を示している．東京ドーム, セコム, 住友化学の 3 社は負の相関であることから, 分散表現が変化に応じて業績も変動し

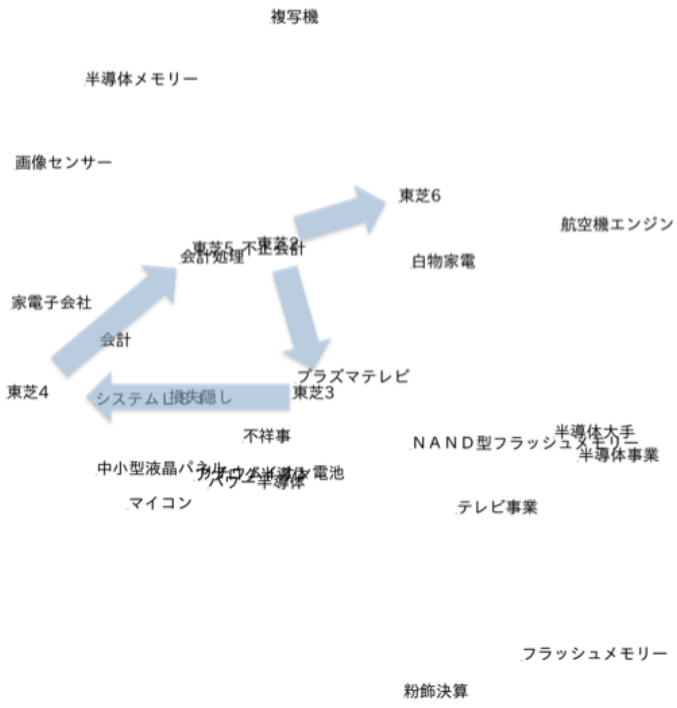

図 3: 東芝と固有名詞以外の類似語の分散表現を視覚化した図

たと言える. 一方, ヤフーと大林組は分散表現が変化しないほど業績が変動したと言える.

表 5: 有意な相関が確認できた企業とその相関係数

企業	相関係数
ヤフー	0.6759
東京ドーム	-0.5895
セコム	-0.6095
住友化学	-0.6320
大林組	0.7864

4. まとめ

本研究では, 時系列の大規模記事データを用いて, 不祥事の前後における企業情報の分散意味表現とその中長期的な変化を学習することで不祥事の影響の定量化を行った. 本研究では期間内のすべての記事データを用いて学習を行いモデルを作成した後にそのモデルに対し各期間ごと別々に再学習を行うことで, 異なる期間の単語の分散表現を比較するできる新しい手法を提案している. また, 表現ベクトルの変化を評価する手段として cos 類似度とユークリッド距離, 次元圧縮による圧縮による可視化の尺度として t-SNE を用いている. 再学習の手法を企業群に適用して cos 類似度及びユークリッド距離により表現ベクトルの変化度合いが大きい企業を抽出したところ, 不祥事を起こした企業を抽出することができた. また, これらの企業の類似語には不祥事に関連する単語が含まれており, 手法の有用性を確認した. また, 一部の企業については分散表現と業績とに相関があることも確認した.

これからやるべきこととしては再学習の手法を改良すること, 業績との相関の取り方を見直すことが挙げられる.

参考文献

[和泉 11] 和泉潔, 後藤卓, 松井藤五郎: 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 (2011).

[中山 14] 中山大, 坂地泰紀, 勝田研一郎, 酒井浩之: 株価に影響を与える重要な出来事が記載された記事の自動抽出, 成蹊大学理工学研究報告, Vol. 51, No. 2, pp. 53-60 (2014).

[上野山 13] 上野山勝也, 松尾豊: Web を用いた企業認知状況の把握と企業 PR への活用, 情報処理学会論文誌, Vol. 54, No. 11, pp. 2392-2401 (2013).

[藤本 13] 藤本貴久, 和泉潔, 吉村忍, 石田智也, 中嶋啓浩, 松井藤五郎 : 新聞記事のテキストマイニングによる長期市場動向の分析, 人工知能学会論文誌, Vol. 28, No. 3, pp. 291-296 (2013).

[石黒 14] 石黒祐輔: 為替ニュース記事を用いた SVM による株価動向予測, 東京大学大学院 情報理工学研究科 (2014).