Vol.2010-IOT-11 No.4
2010/10/15

自然言語処理を用いた企業相関関係の取得

斎藤祐一郎† 西森丈俊††

企業分析を行うにあたり、企業間の関係を確認したい場合がある。ある企業の動きが、関連する企業や業界の業績や株価に影響するためである。現在、この関係を求める場合は有価証券報告書や報道といった文章ベースの情報に依るところが大きい。しかし、これらの情報は定性的なものが多く、分析にあたってはアナリストや投資家の経験や勘に頼らざるを得ない部分があった。

そこで、筆者らは、この企業間の関係性をより定量的に分析するために、文書ベースの情報を形態素解析や頻度解析等を用いて定量化し、類似性の高い企業同士で相関図を作成するという手法の設計と評価を行った。データには東証一部上場企業の 2009 年度秋に発表された中間決算短信を用いた。

その結果、財務諸表や株価等の従来の定量的データからは得られなかった、企業の関係性を導き出すことに成功した。さらに、企業名ばかりでなく関連する製品名等からも関係性を得ることができ、業界構造を知るための新しい手法として有用であると評価できた。

1. はじめに

株式市場に上場している企業の分析を行うにあたり、代表的な定量的分析方法は大きく分けて「財務諸表」と「株価」の 2 種類が存在している。これらには、バリュエーションやテクニカル分析といった分析手法が普及しており、実際に金融機関をはじめとした機関投資家・アナリストや深く学習を重ねている個人投資家によって活用されている。

これらは過去の状況を分析する上でなくてはならない情報であるが、未来を予測するには不十分である。そこで、多くの投資家・アナリスト（以下、分析者）は分析対象企業に関する定性的情報を元に、未来の業績を予測している。その情報は、企業から定期的に発表される「有価証券報告書」や「決算短信」、報道機関から発表される「ニュース」である。

しかし、定性的情報を元に分析するとなると、データの性格上、分析者の経験と勘に頼らざるを得ない部分が存在している。そのため、分析者の知識や能力に左右されない客観的な評価を難しくしている現状がある。

本研究の目的は、定性的情報の分析精度に一貫性を持たせ、より多くの投資家・アナリストがより高い精度の分析を行えるよう、企業から発表された定性的情報を定量化し、経験と勘に頼る部分を削減してゆく手法の開発を行うことである。具体的には、「決算短信」を用い上場企業間のつながりを導き出すシステム(以下、本システム)を開発し、評価(以下、本実験)を行った。決算短信を選定した理由として、企業から発表されるため信頼性があり、かつ短信という性格から速報性が高いことによるものであるためである。

2. 先行研究

自然言語処理を用いた文書間の関係性を分析するシステムは、WISDOM[1]などの Web 上に散在する情報を集約し意思決定の支援を行うものや、専門情報の場合は医学用語関連の研究[2]が行われている。

また、経営に関する研究においては、社外取締役と企業業績の関係に関する研究[3]が行われている。

3. 決算短信情報について

決算短信は各企業が投資家に対しその活動を説明するために、決算期に発表される。その内容は、大きく分けて「業績」「経営成績」「経営方針」「財務諸表」の 4 つの章にわたって記述されている。この中で、定性的な情報として文章の記述があるのは、主に前期の「経営成績」、今期以降の「経営方針」の 2 箇所が該当している。これらの情報には、短信を発表している企業が関与している商品・サービスやそれに関連する単語が記述されている。この情報を元に、「特徴的な単語を用いている企業同士には関連性がある」という仮説に基づいて、実験を開始した。

本実験にて活用している文章は、先の 2 つのほかに、「財務諸表」作成時に記載されている「財務諸表作成のための基本となる重要な事項」も利用している。

解析期は 2009 年秋に発表されたデータとする。日本では 3 月末決算の上場企業は約 2,600 社（2010 年 4 月現在）と多いこと、それにあわせて 9 月末に中間決算期を迎える企業も数多く存在することから、短期間に豊富なデータ量を取得可能と判断した。

*† （株）インタラクティブブレインズ
†† 筑波大学大学院企業科学専攻システムズマネジメントコース

4. 開発システム解説

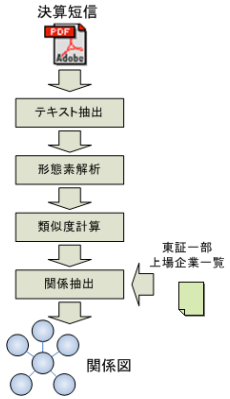
4.1 事前準備

本システムで用いる東証一部上場企業の決算短信は、企業が Web サイトを通じて決算期に発信した PDF ファイルを取得している。そのため、実験前にダウンロードを行っている。ファイル名は、企業が株式市場で区別される際に用いられる銘柄コードに書き換えている。これは、関係抽出時に企業名を照合しやすくするためである。

なお、修正などで同じ期に複数回発表されている場合は、最後に発表されている決算短信を優先して用いることとする。

4.2 システム構成図

本実験で構築ししたシステムの構成を図 1 に示す。



(図 1)

4.3 テキスト抽出

まず、決算短信データは PDF で配布されているため、形態素解析が行いやすいようテキストファイルへ変換を行う。この処理に xpdf[a]に含まれている”pdftotext”コマンドを用いている。

```
@list = grub ( "*.pdf" );
while $file ( @list ) {
 system( "pdftotext -f 3 -enc UTF-8 $file $file¥.txt" );
}
```

4.4 形態素解析

変換処理後は形態素解析器として mecab[b]を用い、PDF ファイルから抽出しファイルとして保存したテキスト情報を形態素解析する。辞書は mecab とともに配布されている ipadic[c]を利用した。

このとき、テキストのサイズが 8KB 未満の決算短信については次の形態素解析の対象から外している。理由として、多くの決算短信の「経営成績」「経営方針」の文量は 8KB を越えていること、そして後の「今後の課題」の章で説明するが、PDF ファイルに書き込まれている情報が原本をスキャンし画像データとして保存されているものが存在し、テキストとして抽出不可能なファイルが存在するためである。

英単語については、すべて表層(surface)を取得している。

形態素解析を行った後は、分かち書きされた情報から文章の特徴を示す可能性が低い情報を削除する。具体的には、品詞が「名詞」「動詞」「形容詞」に該当しない単語を一覧から削除する。

最後に、プログラムのメモリ内で、企業毎に区分されたハッシュテーブルに配列を紐付け、その配列中に文章毎の単語と頻度(TF)の一覧を記録する。

4.5 類似度計算

形態素解析が完了したすべての企業の組み合わせについて、類似度を TF/IDF モデルを用いて計算[4]し、関係抽出のための元となるデータを作成する。

まず、先の形態素解析で抽出した単語と頻度の一覧を元に、IDF を計算する。

続いて、形態素解析の対象となったすべての企業の決算短信同士の類似度をすべて計算し、その結果を「比較元銘柄コード, 比較先銘柄コード, 類似度(cos)」の順にファイルに出力する。

このとき、同時に残差 IDF[5]を計算し、決算短信内の特徴語上位 5 単語を抽出している。理由として、相関図を作成する際に影響を及ぼしたと考えられる単語を手早く調査可能にし、システムの検証を迅速に行うためである。また、特徴語抽出に IDF を用いず残差 IDF を用いたのは、全決算短信文章内の一般語の影響を最小限にしつつ特徴語を抽出したいという考えに基づいて行っている。

a) http://www.foolabs.com/xpdf/

b) http://mecab.sourceforge.net/
c) mecab-ipadic-2.7.0-20070801.tar.gz

4.6 関係抽出

分析対象としたい企業の銘柄コードを指定し、TF/IDF を用いて計算されたなす角の値が一定以上の企業同士を抽出し、図に表す。

先の類似度計算時に算出したファイルを元に、抽出された企業に対して同様になす角が一定以上の企業同士を抽出し、これを繰り返す。企業名は、銘柄コードをキーにし東証一部上場企業一覧から取得する。

指定した抽出階層分の計算が完了し次第、 graphbiz[d]を用いて相関関係図を PNG ファイルとして作成する。このとき、計算プログラムから graphviz のパーサーが読み込むための dot ファイル形式に情報を変換し、ファイルに保存した上で”dot”コマンドを用い相関図を作成している。なお、なす角の値がわかりやすいよう、計算結果を接続図に描画している。

本実験では、抽出条件の既定値として、なす角のしきい値を 0.5、企業の抽出階層を 3 階層と設定して実施している。

5. 実験結果

5.1 正常例

図 2.a は東京電力を基準として、決算短信の内容が類似している企業を抽出した結果である。すべての電力会社が抽出され、2 ステップ以内で相関関係が表現されていることを導き出すことができている。

図 2.b は特徴語上位 5 位のデータである。ここからも、電力会社の特徴的な単語が列挙され手いることが確認でき、これらが軸となって図が作成できたと考えられる。

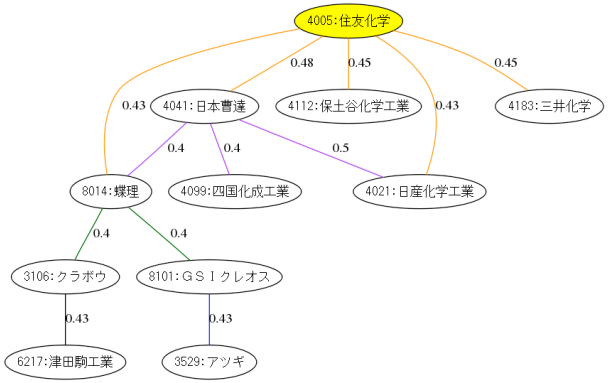
(図 2.a)

順位	単語
1	発電
2	燃料
3	電気
4	程度
5	号機

(図 2.b)

次の図 2.c は住友化学のデータである。こちらは、TF/IDF を用いて計算したなす角を 0.5 から 0.4 に引き下げて計算を行っている。これは、0.5 のままでは関係性を全く抽出できなかったため、0.1 引き下げた上で再試行したためである。

図 2.d は特徴語上位 5 位のデータである。農業が 4 番目に来ているのは、製品として農薬を取り扱っているからである。



(図 2.c)

順位	単語
1	化学
2	住友
3	利
4	農業
5	石油

(図 2.d)

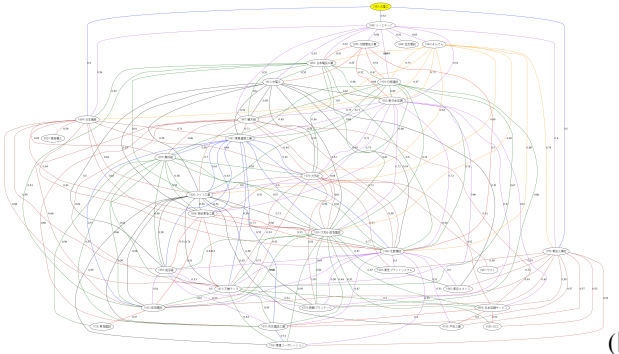
5.2 失敗例

図 3.a は、九電工のデータである。電設関連企業をはじめとした類似の企業が抽出

d) http://www.graphviz.org/

されているのだが、これらの企業同士が密に接続しているため、図示に無理があるデータとなってしまった。

図 3.b は、特徴語上位 5 位のデータである。ここで抽出されている単語は土木業などとも共通する単語である。そのため、電設工事関連企業ばかりでなく図 3.a のように多くの企業が接続する形になったと考えられる。



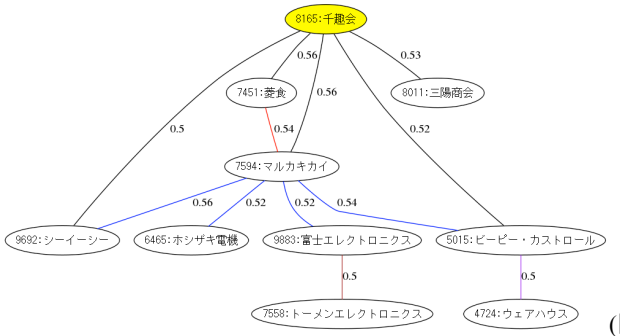
(図 3.a)

順位	単語
1	九電工
2	工事
3	個別
4	完成
5	線

(図 3.b)

図 3.c は千趣会のデータである。こちらは、通販型の小売業という企業特性を活かした結果を得ることができず、関係性があまりないと考えられる企業型が多数抽出されてしまった。

それを裏付ける物として、図 3.d の特徴語上位 5 位のデータを見ると、1 位の「頒布」以外はどの企業の決算短信でも記載されるような事項であった。そのため、本企業の決算短信内に特徴的な表現が余りなかったことが背景であると考えられる。



(図 3.c)

順位	単語
1	頒布
2	1 月
3	連結
4	四半期
5	円

(図 3.d)

6. 今後の課題

本実験結果から、より精度を高めるために次の課題が導き出された。

6.1 データ量の問題

図 3.d では、決算短信内に企業の特性を表す特徴的な単語が存在せず、実際の相関性が低い結果を導き出す例が出てしまっている。

そのため、更なる精度向上には、より長文の情報を得ることができる決算月に発表された決算短信、並びに有価証券報告書を用いることが必要である。

特徴語上位 5 単語を比較すると、失敗例である千趣会は「頒布」「1 月」「連結」「四半期」「円」であり、成功例である図 2.a の東京電力の特徴語は「発電」「燃料」「電気」「程度」「号機」[e]であった。失敗例は企業の特徴を表す単語が「頒布」のみだったのに対し、成功例は電源供給を行う企業の特徴を良く表した単語が上位にもれなく並

e) この時期は柏崎刈羽原子力発電所の事故が問題になった時期でもあったため、決算短信内に原発に関する記事が多く記述されていたことが影響している。

んでいる。

6.2 絞り込み方法のチューニング

本実験における失敗例の九電工の場合、検索結果が大きくふくれあがってしまう事例も確認した。これは、すべての企業において同じ検索条件を用いることが難しいことを示している。そのため、企業毎に動的に検索条件を変化させられる仕組みを導入することを予定している。

方法として、階層が深くなるたびになす角のしきい値をあげていくことや、一定以上の企業が抽出された場合に企業の上限数を設けることを計画している。

6.3 データ自体の問題

これはシステムの問題ではないが、トヨタ自動車をはじめとして決算短信が記述されている PDF ファイルの文章情報が画像で記録されておりテキストが抽出できなかったものがあった。OCR を用い認識させる仕組みを構築するか、またはテキストで抽出できるフォーマットに統一されることを待つ必要がある。

7. 終わりに

本実験を通じ、企業に関する定性的情報をもとに、自然言語処理を用いることで定量的に分析することができた。

現在広く用いられている定量的分析の手法と組み合わせて、より多面的、かつ精度の高い企業分析を行う手法を提案できたと考えている。

また、自然言語処理の違った側面での活用方法として提案できたことも、本実験の成果であった。

しかし、データの量や鮮度の問題が確認されたため、今後はよりデータ量を増やして検証を進めて実験を重ねて行き、企業分析における本システムの価値を高めて行きたいと考えている。加えて、情報の鮮度を高めるために新聞記事や QUICK[f]をはじめとした機関投資家が主に利用する市場情報配信サービスから受信するニュースなどの速報性の高い情報を活用することも、重要になってくると考える。

平行して、類似度計算のチューニングも進めて行きたい。

8. 謝辞

本論文の制作にあたり、XBRL 勉強会[g]のメンバーの方の多大なるご支援と貴重なアドバイスをいただきました。ここに記して、御礼申し上げます。

9. 参考文献

[1] 赤峯享,宮森恒,加藤義清,中川哲治,乾健太郎,黒橋禎夫,木俵豊: Web 情報の信頼性検証のための情報分析システム WISDOM, 言語処理学会 第 14 次大会 (2008)
[2] 田中昌昭,竹内孔一: 医学用語辞書で学習した分類器による放射線読影レポート用語の分類, 言語処理学会 第 14 次大会 (2008)
[3] 三輪晋也: 日本企業の社外取締役と企業業績の関係に関する実証分析, 日本経営学会誌 VOL.25, pp.15-27 (2010)
[4] Christopher D. Manning and Prabhakar Raghavan and Hinrich Schuetze: Introduction to Information Retrieval, Cambridge University Press. (2008)
[5] 北研二, 津田和彦, 獅々堀正幹: 情報検索アルゴリズム, 共立出版 (2002)

f) 株式会社 QUICK が運営する金融市場情報配信サービス

g) 企業財務情報の統一 XML フォーマット XBRL に関する勉強会 http://xbrl-study.pbworks.com/