有価証券報告書からのリスク情報の抽出の試み

平田勝大[1) 2)]　梅村恭司[1) 2)]　関根聡[3)]
豊橋技術科学大学 情報工学系[1)]
インテリジェントセンシングシステム リサーチセンター[2)]
ニューヨーク大学[3)]

Extracting Risk Information from Securities Report

Katsuhiro Hirata[1) 2)]　and Kyoji Umemura[1) 2)]　and Satoshi Sekine[3)]
Toyohashi University of Technology, Information & Computer Scienses[1)]
Intelligent Sensing System Research Center[2)]
New York University[3)]

概要

有価証券報告書は、上場企業等の事業に関する情報が詳細に記載されている。しかしながら、その情報は文書で記載されている情報が多く、文書量も膨大である。本研究では、ルールとSVMを組み合わせて、この文書から事業のリスクに関する情報を抽出したことを報告する。

1 はじめに

有価証券報告書は、上場企業等が年度ごとに事業の状況、財務状態や経営成績等を記載して公表されている書類である。この書類に記載されている情報は、投資家が投資をする際には重要な判断材料となっている。また、これらの情報は、投資家のみならず、一般的にも有益な情報である。

有価証券報告書には、100 ページ以上にもわたって企業情報が詳細に記載されており、上場企業のみでも4500社以上あるため、これら書類をすべて調べることは非常に困難である。このため、これらの書類から文書すべてを見ることなく情報を得ることができれば、企業について調べる上で有用である。

文書すべてを見ることなく文書の内容を見るには、必要な文書のみを抽出することが有効ある。今回は、”事業等のリスク”の項目からリスク情報を抽出する。具体的には、リスクについて述べている文を抽出し、リスク内容ごとに分割したことを報告する。この問題は、文書分類のタスクであり、実際に需要のある処理であるが、分類するための文書の区切りが不明であり、標準的な分類手法[4,5,6]で分類できるかもわからない。

”事業等のリスク”の文書には、箇条書きで記載された文書と箇条書き無しで記載された文書があり、それぞれに別の手法で区切る。分類は標準の方法を使用し、機械学習法であるサポートベクターマシン(SVM)を利用する。初めに、句点で区切られる文書を文、各文を形態素解析した単語

の出現頻度を特徴ベクトルとして、各文がリスクについて述べている文であるかを分類する。箇条書きを含む文書では、ルールベースで内容ごとに分割し、初めの処理で分類した文を含むかどうかでリスクについて述べている文書であるかを判断する。箇条書きを含まない文書では、初めに分類した文を順番に結合し、結合部分における前後の文の単語出現頻度を特徴ベクトルとして、内容が分割されるかを判定する。以上の処理を行い、有価証券報告書からリスク情報を内容ごとに分割して抽出したことを報告する。

2 有価証券報告書

有価証券報告書とは、有価証券である株券や債券を使って 1 億円以上の資金調達をする企業や株式を証券取引所などに上場や公開している企業が年度ごとに提出を義務付けられている書類である。この書類には、年度ごとの企業の事業内容や一年間の業績、設備投資の状況等が記載されており、本研究では、特に“事業等のリスク”の項目からリスク情報を抽出する。

3 サポートベクターマシン

サポートベクターマシンとは、2 つのクラスを識別する識別器を構成するための学習法である。学習データを用いて、2 つのクラスを線形分離し、分離超平面と最も近い学習データとの距離が最大になるようなモデルを学習する。この手法により、線形分離できるデータにおいては、未学習のデータにおいても高い識別性能を持つ学習手法の一つである。また、線形分離不可能なデータにおいても、非線形写像によって、線形分離可能なより高次元の空間に変換することができれば、同様に 2 つのクラスを識別することができる。さらに、非線形写像前のデータを利用して、写像後のデータにおける学習と等価な計算をすることで写像の計算を簡単化することができ、これをカーネルトリックという。カーネルトリックに利用可能なカーネルには以下のようなものがある。

・多項式カーネル

$$K(X_i, X_j) = (X_i, X_j + 1)^h$$

・ガウスカーネル

$$K(X_i, X_j) = e^{-\|X_i \bullet X_j\|^2 / 2\sigma^2}$$

本研究では、SVM ツールである SVMlight[2]を利用する。

4 形態素解析

日本語は文書中の単語が区切られていないため、どの部分文字列が単語であるか判定する必要がある。形態素解析は、文字列の単語候補を辞書から調べ、単語の品詞の接続情報を用いて単語候補から単語を判別する解析手法である。この手法を用いることで、意味を持つ最小単位である単語に分割することができる。本研究では、形態素解析ツールである茶筌[3]を使用して、文書を単語に分割する。

5 リスク情報の抽出

リスク情報の抽出には、4 つの手順を用いて抽出する。それぞれの手順は、データを調査して決定した。手順にはルールが含まれている。手順 1 で、文書を句点および箇条書き表現によって分割する。手順 2 では、SVM を使用して手順 1 で分割した文書がリスクに関する文書であるかを判定する。手順 3 では、元の文書から、箇条書き表現を利用したルールによってひとつのリスク

に関する文書候補を抽出し、手順 2 でリスクに関すると判断された文書を含む文書を抽出結果とする。手順 4 では、箇条書き表現を含まない文書に対して、手順 1 の結果であるリスクに関する文書を合わせて、どの文書までがひとつのリスクに関する文書であるかを SVM で判断して抽出結果とする。

○手順 1 リスク文書の分割

有価証券報告書の“事業等のリスク”の文書を以下のルールにより分割する。

－ルール

・句点で終わる文書を分割する

・①,②,…,⑳で始まる文書を分割する

○手順 2 リスクに関する文書の判定

手順 1 の結果である文書がリスクに関する文書であるか SVM を利用して判定する。学習データには、500 企業に対して手作業で判定したデータを使用する。SVM の特徴ベクトルには、学習データを形態素解析で分割した単語のうち、ひらがなを含む単語は 4 文字以上の単語、含まない単語は 2 文字以上の単語を合わせたものから出現頻度順上位 3000 語と以下の箇条書き表現の出現頻度を使用する。

－箇条書き表現

・①,②,…,⑳

・(1), (2),…, (20)

・1.,2.,…,20.

○手順 3 ルールベースによる抽出

有価証券報告書の“事業等のリスク”の文書を以下の箇条書き表現を用いて、ひとつのリスクに関する文書候補を抽出する。

－箇条書き表現

・〔リスク名〕,[リスク名]

・＜リスク名＞,(リスク名)

・①,②,…,⑳

・(1), (2),…, (20)

・1.,2.,…,20.

抽出した文書候補が手順 2 のリスクに関する文書を含んでいるかを調べ、含んでいればリスクに関する文書とする。

○手順 4 箇条書き表現のない文書の処理

手順 1 の文書を合わせ、どの文書がひとつのリスクに関する文書の境目であるかを SVM によって判定し、ひとつのリスクに関する文書を抽出する。SVM の学習データには、手順 2 で使用したデータもの同じものを使用し、特徴ベクトルには、手順 2 と同様の単語の境目となる 2 つの文書における別々の出現頻度を利用する。

6 抽出結果

以上の処理によって、上場企業 4122 企業の有価証券報告書からリスク情報を抽出した。しかしながら、手順 2 のリスクに関する文書の判定において、短い文書はすべてリスクに関する文書でないと判定されてしまい、手順 3 でリスクを短く箇条書きにした文書が抽出されなかった。そのため、手順 3 と手順 4 に使用する手順 2 の結果を別にした。手順 3 に使用する結果には、SVM の学習にガウスカーネルを使用して学習し、短い文書についても抽出できるように変更した。抽出結果と学習データとは異なるデータを実際に見て作成した 50 企業の正解データと比較した。この結果を表 1 に示し、

抽出例を図 1 に示す。

	再現率	適合率
ルールのみ	84.6	91.3
ルールと SVM	85.6	90.1
ガウスカーネルを使用	85.3	90.7

表 1.リスク情報の抽出

(1)経済状況による影響当社グループが販売している製造装置の需要は、その製造装置で製造される液晶・半導体などのエレクトロニクス部品の需要に影響を受け、特にエレクトロニクス部品が消費されている国の経済状況の影響を受ける。従って北米、欧州、アジア、日本などの国の景気後退と需要の縮小により、当社グループの業績に悪影響を及ぼす可能性がある。

原材料等の価格の上昇足元好調に推移するグローバル経済に支えられ、日本経済も民間需要中心に引き続き順調に成長し、とりわけ企業部門の好調さが家計部門に波及してゆくことが期待されます。しかしながら、米国・中国の金融引き締めや中東情勢の悪化など先行き不透明要因も数多く、予断を許さぬ状況が続くものと思われます。とりわけ前期より顕著となっている原油・穀物などの国際商品市況の高騰はエネルギー、原材料価格の上昇を通して、生産コストの大幅な圧迫要因となり、当社の業績に影響を与える可能性があります。

図 1.抽出例

7 考察

ルールに加えて、SVM でリスク抽出をすることにより、ルールでは抽出できないリスク情報も抽出できるようになった。手順３の判定では、ルールの処理で箇条書きではあるがリスクには関係のない文書を判定できるが、同時にリスクに関する短い文書がリスクに関係しないと判定されてしまう場合があった。そのため、ガウスカーネルを使用して、短い文書を抽出できるようにすると、適合率が増加し、再現率が減少した。これは、リスクに関する長い文書が影響を受けてしまったためと考えられる。

8 謝辞

本研は文部科学省 21 世紀 COE プログラム「インテリジェント ヒューマンセンシング」の援助により行われた。

9 参考文献

[1]栗田多喜夫:サポートベクターマシン入門, 産業技術総合研究所 脳神経情報研究部門 (2002)

[2]Thorsten Joachims : Support Vector Machine, Cornell University (2004)

[3]松本裕治:日本語形態素解析システム『茶筌』 version 2.2.1 使用説明書 (2000)

[4]Christopher D. Manning and Hinrich Schuetze. : Foundations of Statistical Natural Language Processing,The MIT Press (1999)

[5]Ian H. Witten and Eibe Frank : Data Mining,Morgan Kaufmann Publishers (2000)

[6]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(2001)