企業リスク分析のための重要単語抽出と因果関係ネットワークの構築

五十嵐 光秋[1]　坂地 泰紀[1]　和泉 潔[1]　島田 尚[1]　松島 裕康[1]　須田 真太郎[2]
[1] 東京大学 大学院工学系研究科　[2] 株式会社 三菱 UFJ トラスト投資工学研究所
m2018higarashi@socsim.org
{sakaji, izumi, shimada, matsushima}@sys.t.u-tokyo.ac.jp
suda@mtec-institute.co.jp

1　研究背景と先行研究

近年，企業経営や投資活動においては，リスクマネジメントが強く意識されるようになった．これの最初のステップにあたるのが，潜在的なリスクを事前に発見するリスクマイニングであり，例えば Leidner ら [3] は，Web 上のテキストデータを対象にパターンマッチングをベースとしたブートストラップ手法によってリスク事象を抽出することを試みた．こうした研究で取り組まれたタスクは，既存のリスクとして記述されている事象を獲得するものであったが，より進んだリスクマイニングを目指す上では，それが発生する上流での事象の因果に注目することが有効である．

こうした背景に対して，社会事象の流れを捉えることを目的として，因果関係ネットワークを構築する研究が報告されている．Ishii ら [2] は，ニュース記事を対象として，SVO 構造に注目した語義の一致による因果関係の接続について検討した．これに類似した手法として，Radinsky ら [5] は，事象を Action，Actor，Object，Instrument，Location の 5 項目に整理して因果関係の連鎖を表現した．彼らの手法は事象の表現に対して厳しい制約をかけており，本来的には接続されるべきであった因果関係が抜け落ちてしまう可能性をはらむ．

語彙や構文などの表現方法による制限をかけずに因果関係を接続するという考え方から，Nishimura ら [4] は，抽出した因果関係表現に対して FastText モデル [1] を用いた因果関係の接続を試みた．本研究においても同様の目的から，単語分散表現を用いた因果関係ネットワーク構築を目指す．

因果関係の連鎖に現れるそれぞれの因果関係は，それ単体で意味が特定できる（特定可能性）ものである必要性がある．例えば,「地球温暖化が進行する→海水面が上昇する」という因果関係は，誰でも同一の事象を想起することができる．しかし，例えば「狂牛病が流行する→売上が減少する」という文は，実際に売上が減少する会社・業種という必要な情報が欠落しているため，事象として捉えるには特定可能性の要件を満たさない．Nishimura らの手法では，このような特定可能性の観点を考慮していないために意味の齟齬が発生することが課題であった．そこで本研究では，事象を表現するキーワードの一致機構を導入することにより，因果関係ネットワークにおける特定可能性の課題を解消し，企業の潜在的なリスク事象を獲得するモデルの開発を目指す．

2　因果関係ネットワーク構築手法

本研究では，決算短信と日経新聞記事から抽出された因果関係を接続することで因果関係ネットワークを構築する．入力時に分析対象の業種や銘柄を指定し，該当する企業の決算短信から最初の因果関係接続を行う．これが第 1 階層である．これによって，該当する企業・企業群においての原因に焦点を当てて検索する．獲得された因果関係を起点として，次の第 2 階層以降では，日経新聞記事を対象として因果関係接続（2.2 節にて後述）を繰り返す．この流れを，図 1 に示す．

2.1　因果関係の抽出

決算短信と新聞記事からの因果関係の抽出には，青野ら [6]，Ishii ら [2] も適用している坂地ら [7] の手法を用いた．これにより，表現に制約をかけることなく，重文や複文にまたがって記述されるような因果関係を抽出することができる．詳しいアルゴリズムについて

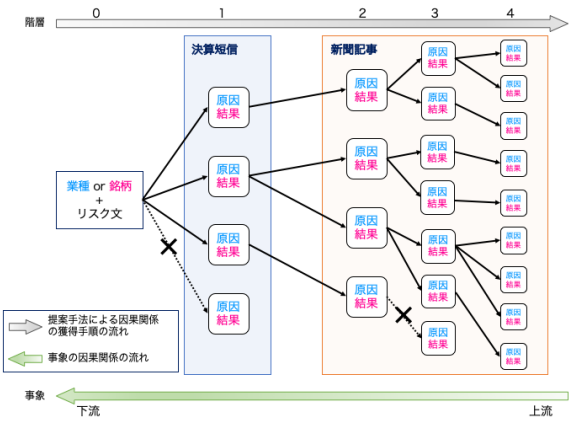

図 1: 因果関係接続によりネットワークが構築される手順

は引用論文を参照されたい．本研究では，2014 年 1 月 1 日から 2015 年 12 月 31 日までに発行された決算短信，日経新聞記事から因果関係を抽出した．決算短信の食料品業種からは 14413 件が，日経新聞記事からは 80433 件が抽出された．

2.2 因果関係の接続

因果関係 $X \rightarrow Y$ に対して，また別の因果関係 $Y' \rightarrow Z$ が接続されることで，$X \rightarrow Y \rightarrow Z$ という連鎖を獲得することができ，これを繰り返すことでネットワークを構築する．2 つの因果関係 $X \rightarrow Y$ と $Y' \rightarrow Z$ が接続されるか否かについては Y と Y' の 1) 文間類似度判定，2) キーワード判定によって決定される．具体的には，文間類似度が閾値 α 以上であり，かつ類似キーワードが存在すると判定された因果関係のみ，接続される．

文間類似度 S_{wei} は idf 値と FastText[1] による単語分散表現を用いて，以下のように計算した．

$$v_Y = \sum_{y \in Y} idf(y) \cdot wv_y \quad (1)$$

$$S_{wei}(Y, Y') = \phi(v_Y, v_{Y'}) \quad (2)$$

ここで，y は Y に所属する内容語で，$idf(y)$ と wv_y は y の idf 値と分散表現，ϕ は 2 ベクトルのコサイン類似度を表す．idf 値は日経新聞記事から計算したものを，単語分散表現は日経新聞記事，決算短信，Wikipedia コーパスを用いて学習したものを用いた．

キーワードに関しては，Y から獲得したキーワード群 $\{k_i\}$ と，Y' から獲得したキーワード群 $\{k'_j\}$（獲得方法に関しては 3 章にて詳述する）について，類似したものが存在するかを判定する．それぞれの集合内の要素について全組み合わせを検証し，単語間類似度 $\phi(k_i, k'_j)$ が閾値 δ 以上の単語が存在するか否かを判定する．単語間類似度には，上述の FastText モデルを用いた．

2.3 因果関係ノードのスコアリング

因果関係ネットワークにおいて，獲得された事象がどの程度「重要である／重要でない」の具合を提示するために，獲得された因果関係に対してスコアリングを行う．スコアリング対象の因果関係ノード A，A から波及先の因果関係ノード $B_i (i = 1, 2, \ldots, n)$，因果関係 A と B_i の類似度を w_{AB_i} を定める．この時，A の重要度スコア NI_A は以下のように求められる．

$$NI_A = \sum_{i=1}^{n} w_{AB_i} \cdot NI_{B_i} \quad (3)$$

3 キーワード獲得

単語が因果関係接続におけるキーワードとなるか否かは，4 つの指標によって定められる．業種キーワードに関わる指標（3.1 節参照）として，1) 業種 PMI による重要度 I_S，2) 業種 PMI による信頼度 R_S，そして，時系列キーワードに関わる指標（3.2 節参照）として，3) 時系列 PMI による重要度 I_T，4) 出現パターンの時系列的特性である．

業種キーワードに関わる指標と時系列キーワードに関わる指標のどちらかを十分に満たす単語を，本研究におけるキーワードとして定める．

3.1 業種キーワード

統計学的な自然言語処理では共起の指標としてたびたび用いられる自己相互情報量を応用して，業種と単語の関連度合いの定量化を試み，この値を業種 PMI と呼ぶ．この業種 PMI が文中キーワードの抽出に寄与することに期待する．以下の式により業種 PMI を計算する．

$$PMI_{sector}(w, S) = log_2 \frac{P(w, S)}{P(w)P(S)} \quad (4)$$

ここで，$P(w)$ は単語 w が出現する確率，$P(S)$ は文書全体に対して業種 S の文が出現する確率である．

$P(w,S)$ は単語 w が業種 S の文において出現する確率である．

業種 PMI の計算には，業種（銘柄）と紐づいたテキストデータが必要である．そこで本研究では，決算短信と，各銘柄に関する Web テキストを用いてそれぞれ業種 PMI を計算した．決算短信で求めた業種に関する業種 PMI の最大値 I_{fr} と Web テキストで求めた業種に関する業種 PMI の最大値 I_{web} に対して，調和平均をとることで業種に関する単語の重要度 I_S を求める．

$$I_S = \frac{2}{\frac{1}{I_{fr}} + \frac{1}{I_{web}}} \tag{5}$$

また，決算短信由来の PMI と Web テキスト由来の PMI の相関を，スピアマンの順位相関から計測する．これを単語の信頼度 R_S として定める．

重要度 I_S と信頼度 R_S の閾値を決定するために，目視で 1000 件の単語を無作為抽出し，目視で 2 値評価を行った．閾値を変更したときの Accuracy による評価から，$I_S = 2.0$ と $R_S = 0.3$ と決定した．

3.2 時系列キーワード

業種を問わず幅広い影響を及ぼすマクロ事象は，業種 PMI による方法では獲得できない可能性が高い．大規模なマクロ事象が発生した時，その事象はどの業種においても記述されるからである．しかし一方で，こうした単語は観測時期においてのみ頻出するような出現頻度の偏りが観測されると考えられる．これを定量化するために，以下のように時系列 PMI を定める．

$$PMI_{time}(w,T) = log_2 \frac{P(w,T)}{P(w)P(T)} \tag{6}$$

ここで，$P(w)$ は単語 w が出現する確率，$P(T)$ は文書全体に対して時系列区間 T の文が出現する確率である．$P(w,T)$ は単語 w が時系列区間 T の文において出現する確率である．日経新聞記事を対象として，四半期ごとに時系列区間を定め，時系列 PMI と計算した．前節とは異なりこちらでは単一の文書を対象としているため，重要度 I_T は区間に関する時系列 PMI の最大値をそのまま適用する．

本研究において時系列キーワードの対象とするのは，ただ時系列 PMI が高いピークを有するだけではなく，継続的に話題性を有する単語である．継続的な話題性の根拠には，PMI が正となる期間が連続 4 区間（1 年間）以上存在することをもって判定する．

前節と同様に，重要度 I_T の閾値を決定するために，目視で 1000 件の単語を無作為抽出し，目視で 2 値評価を行った．閾値を変更したときの Accuracy による評価から，I_T の閾値を 1.7 と決定した．

4 因果関係ネットワーク構築実験の結果と考察

入力文：「売上が減少した」，対象業種：「食料品業種」の条件で因果関係ネットワーク構築実験を行った．評価に際しては，エッジを無作為に 100 個抽出してラベル付けを行うことにより算出した Precision と，ノードの重要度スコア（2.3 節参照）により算出した DCG@10 及び nDCG@10 を用い，各階層ごとに結果を整理した．DCG の評価のための基準は，特定可能性と入力事象との関連度に注目して以下のように定め，1 6 の目視による判定を行った．

表 1: 因果関係ノードに関する評価基準と配点

		関連度		
		✓✓	✓	✗
特定可能性	✓	6	4	2
	✗	5	3	1

比較手法として Ishii ら [2] による手法と Nishimura ら [4] による手法を実装した．表 2 と表 3 に実験結果を示す．なお，提案手法 Proposed は文間類似度による接続判定のみ用いた結果，提案手法 Proposed+KW は，文間類似度とキーワードの接続判定を併用した結果である．

表 2 より，提案手法においては比較手法よりも多くのノードを獲得できていることが分かる．また，キーワードの制約を追加することで，Precision が低くなる傾向にある深い階層においても比較的高い値を獲得できている．さらに，表 3 における DCG@10 の結果からも，全ての階層において高い値を記録している．これは，入力事象との関連度の高い事象が多く獲得できていることを意味する．

5 まとめ

本研究では，企業にとってネガティブな事象が発生する要因となるリスク事象を，因果関係ネットワークの構築を通じて獲得する手法の開発に取り組んだ．因

表 2: 実験結果：ノード数と Precision

手法	ノード数				Precision			
	Layer1	**Layer2**	**Layer3**	**Layer4**	**Layer1**	**Layer2**	**Layer3**	**Layer4**
Ishii	**20**	4	1	0	0.950	**0.500**	0.000	-
Nishimura	**20**	21	48	186	0.917	0.300	0.360	0.245
Proposed	**20**	51	**133**	**417**	0.875	0.091	0.235	0.255
Proposed+KW	**20**	**52**	112	279	**1.000**	0.438	**0.500**	**0.475**

表 3: 実験結果：DCG@10 と nDCG@10

手法	DCG@10				nDCG@10			
	Layer1	**Layer2**	**Layer3**	**Layer4**	**Layer1**	**Layer2**	**Layer3**	**Layer4**
Ishii	27.23	7.00	2.00	-	0.872	**1.000**	**1.000**	-
Nishimura	27.89	**18.28**	14.48	11.36	**0.965**	0.785	0.771	0.799
Proposed	**28.25**	15.11	18.02	13.70	0.947	0.783	0.932	**0.893**
Proposed+KW	**28.25**	16.60	**24.47**	**14.41**	0.947	0.747	0.906	0.831

果関係ネットワークの構築には，決算短信と新聞記事から抽出した因果関係表現を接続するという方法を採用することで，企業が把握していなかった潜在的なリスクの獲得を目指した．接続においては，1) 単語分散表現を利用した意味的な類似度による判定，2) 文から抽出したキーワードの一致による判定の 2 点を組み合わせることで，語彙や構文の自由度を高つつも，特定可能性の低い事象の表現を排除することを目指した．食料品業種の売上減少リスクに関する因果関係ネットワーク構築実験により，キーワードの制約を導入する提案手法が最も高い精度で関連度の高い因果関係事象を多数獲得できることを確認できた．

留意事項

本稿の内容は筆者が所属する組織を代表するものではなく, すべて個人的な見解である．また，当然のことながら，本稿における誤りは全て筆者の責に帰するものである．

参考文献

[1] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T.: Enriching Word Vectors with Subword Information, *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146 (2017).

[2] Ishii, H., Ma, Q. and Yoshikawa, M.: Incremental Construction of Causal Network from News Articles, *Journal of information processing*, Vol. 20, No. 1, pp. 207–215 (2012).

[3] Leidner, J. L. and Schilder, F.: Hunting for the black swan: risk mining from text, *Proceedings of the ACL 2010 System Demonstrations*, Association for Computational Linguistics, pp. 54–59 (2010).

[4] NISHIMURA, K., SAKAJI, H. and IZUMI, K.: Creation of Causal Relation Network using Semantic Similarity, 人工知能学会全国大会論文集 2018 年度人工知能学会全国大会 (第 32 回) 論文集, 一般社団法人 人工知能学会, pp. 1P104–1P104 (2018).

[5] Radinsky, K., Davidovich, S. and Markovitch, S.: Learning causality for news events prediction, *Proceedings of the 21st international conference on World Wide Web*, ACM, pp. 909–918 (2012).

[6] 青野壮志, 太田学: 要因検索による因果関係ネットワークの構築と因果知識の獲得, Forum on Data Engineering and Information Management (2010).

[7] 坂地泰紀, 酒井浩之, 増山繁: 決算短信 PDF からの原因・結果表現の抽出, 電子情報通信学会論文誌 D, Vol. J98-D, No. 5, pp. 811–822 (2015).