銘柄テキスト情報と銘柄数値情報をハイブリッド活用した企業間類似度の獲得

平松 賢士 [1]　伊藤 友貴 [2]
[1] 株式会社アイフィスジャパン　[2] 三井物産株式会社
kenji.hiramatsu@ifis.co.jp　Tomok.Ito@mitsui.com

概要

出資判断や自社 IR 活動，テーマ型投資信託の組成，事業推進時のパートナー選定といった金融実務の場において，類似企業間や競合企業間での比較分析を行うことが多い．ここで，「企業間類似度の算出」に関する有用なのがテキスト情報に対し BERT 等を活用することで獲得できる企業埋込表現である．このテキストデータをベースとする埋込表現は有効である一方，経済・金融分野においては株価や売上の推移やセグメント別売上等の数値データや株のや特定投資株式を始めとする株式保有情報等，企業間類似度を測る上で有用と考えられる数値データも多く存在し，これらの金融数値データとテキストデータを組み合わせることでより有用な「企業間類似度」を算出することが期待される．

そこで，本研究では，銘柄テキスト情報だけでなく，セグメント別売上や株価時系列データ，株式保有情報も活用した「テキスト情報」と「数値情報」をハイブリッドに活用した企業間類似度を提案する．

1　はじめに

出資判断や自社 IR 活動，テーマ型投資信託の組成，事業推進時のパートナー選定といった金融実務の場において，類似企業間や競合企業間での比較分析を行うことが多い．

例えば，自社 IR 活動においては自社と競合他社の比較を行う．また，テーマ型投資信託の組成においては，「生成 AI」や「半導体」等の特定のテーマに関連する、または影響のある企業を洗い出した上，比較・分析を行う．特定の企業への出資や事業推進時のパートナー選定時にも，類似企業間での比較を行うことが多い．以上を踏まえると，金融実務の場において，「類似企業を自動検索できること」は非常に有用であると考えられる．このような背景のもと，本研究では類似企業の検索に有効な「企業間類似度の算出手法」の開発を目指す．ここで，「企業間類似度の算出」に関する有用なのが BERT[1] や GCN[2] 等を活用した企業埋込表現である．例えば，[3] では決算短信や Form-10K 等の銘柄テキスト情報を活用した企業埋込表現を構築し，[4] では銘柄間の情報を元に埋め込みを獲得することで企業埋込表現を構築する．また，[5] では因果チェーンを用いて構築される銘柄間ネットワークと銘柄テキスト情報を組み合わせることで個別銘柄情報と銘柄間情報を同時に考慮した企業埋込表現を獲得している．

これらのテキストデータをベースとする埋込表現は有効である一方，経済・金融分野においては株価や売上の推移やセグメント別売上等の数値データや株のや特定投資株式を始めとする株式保有情報等，企業間類似度を測る上で有用と考えられる数値データも多く存在し，これらの金融数値データとテキストデータを組み合わせることでより有用な「企業間類似度」を算出することが期待される．一方，これらの数値データが既存手法において十分に活用されているとは言い難い．

そこで，本研究では，銘柄テキスト情報だけでなく，セグメント別売上や株価時系列データ，株式保有情報も活用した「テキスト情報」と「数値情報」をハイブリッドに活用した企業間類似度を提案する．また，提案手法を他の手法と比較し，その特徴やその実タスクにおける有効性を観察する．

2　関連研究

企業やファンドの埋め込み表現獲得手法か過去にいくつか行われている．例えば，[3] や [6] では決算短信や Form-10K 等の銘柄テキスト情報を活用した企業埋込表現を構築し，[4] では銘柄間の情報を元に埋め込みを獲得することで企業埋込表現を構築する．また，[5] では因果チェーンを用いて構築される

銘柄間ネットワークと銘柄テキスト情報を組み合わせることで個別銘柄情報と銘柄間情報を同時に考慮した企業埋込表現を獲得している．更には，経済情報のデータを学習に利用し、企業エンティティに特化したエンコーダーモデル[1)]等も開発されている．

埋込表現獲得手法についても近年、大きな進捗が見られる．例えば，[7] では大規模言語モデル (LLM) により生成される合成データを活用することで高品質な埋め込み表現獲得手法を提案している．また，近年、基盤モデルによる時系列データの埋め込み表現獲得手法もいくつか [8, 9] 提案されている．

また，数値情報とテキストデータを組み合わせた利用した企業埋め込み表現獲得手法として，[3] で提案されている手法が挙げられる．[3] では株価時系列に関する類似度を企業間類似度の教師として学習する手法を提案している．一方，数値情報，特に株価データ以外の金融数値情報とテキストデータをハイブリッドに利用した埋め込み表現獲得手法や企業間類似度の獲得はあまり行われておらず，金融数値情報の活用方法については多くの可能性があると考えられる．

3 提案手法

本節において，提案手法である，「テキスト情報」と「数値情報」をハイブリッドに活用した企業間類似度の算出手法を説明する．提案手法では「企業 A」及び「企業 B」の類似度 $Sim(A,B)$ を A 及び B それぞれの埋込表現 $\boldsymbol{h}_A$ 及び $\boldsymbol{h}_B$ を用いて以下のように計算する．

$$\begin{aligned} Sim(A,B) &= \alpha cos(\boldsymbol{h}_A^{text}, \boldsymbol{h}_B^{text}) \\ &+ \beta cos(\boldsymbol{h}_A^{seg}, \boldsymbol{h}_B^{seg}) \\ &+ \gamma cos(\boldsymbol{h}_A^{price}, \boldsymbol{h}_B^{price}) \\ &+ \epsilon \sum_{a \in Stock(A), b \in Stock(B)} cos(\boldsymbol{h}_a^{text}, \boldsymbol{h}_b^{text}) \end{aligned}$$

ここで，$cos(\boldsymbol{h}_A, \boldsymbol{h}_B)$ は $\boldsymbol{h}_A$ 及び $\boldsymbol{h}_B$ のコサイン類似度を表し，$\boldsymbol{h}_A^{text}, \boldsymbol{h}_A^{seg}, cos(\boldsymbol{h}_A^{price}, \boldsymbol{h}_A^{stock}$ はそれぞれ後述の通り「銘柄テキスト情報」「セグメント別売上を考慮したテキスト情報」「株価時系列データの埋め込み」及び「株式保有情報」の情報に関する埋め込み表現であり，$\alpha, \beta, \gamma, \epsilon$ はそれぞれ各情報の重み（スカラー量）を表す．

また，$Stock(A)$ は有価証券報告書から得られる企業 A が株式保有する企業の一覧となる．

1) https://huggingface.co/uzabase/UBKE-LUKE

3.1 銘柄テキスト情報の埋め込み

「企業 A」に関する銘柄テキスト情報の埋め込み表現は BERT[1] や LUKE[10] 等の言語モデルを利用した埋め込み表現獲得により，以下で表す．

$$\boldsymbol{h}_A^{text} := EMB_{emb}^{text}(T_A) \tag{1}$$

ここで，T_A は企業 A に関するテキスト情報の記載文である．例えば，決算短信や有価証券報告書内の「事業内容」等から取得する情報が該当する．

3.2 セグメント別売上を考慮したテキスト情報の埋め込み

「企業 A」に関するグメント別売上を考慮したテキスト情報の埋め込みは以下で表す．

$$\boldsymbol{h}_A^{seg} := \sum_{s \in S_A} \frac{V(s)}{\sum_{s \in S_A} V(s)} EMB_{emb}^{seg}(T_A^s) \tag{2}$$

ここで，S_A は事業セグメントの集合，$V(s)$ はセグメント s の売上を表す．また，T_A^s はセグメント s に関する事業内容の記述を指す．

3.3 株価時系列データの埋め込み

本埋め込み表現は株価時系列のデータ $[p_m, p_{m-1}, \cdots, p_0]$ 及びその時系列に関する基盤モデル LLM_{emb}^{price}[8] による埋め込み表現への変換により獲得する．

$$\boldsymbol{h}_A^{price} := LLM_{emb}^{price}([p_m, p_{m-1}, \cdots, p_0]) \tag{3}$$

p_m は現在時点から m 日前の株価を表す．

4 評価実験

本節では実データを用いて本手法の有効性を検証すると共に性質を観察する．本評価のため，[3] にて公開されている類似銘柄検索に関するデータセット[2)]を用いた検証を行う．

また，性能評価のため，提案手法を以下の二つの手法との比較を行う．

- ベースライン [3]: $\boldsymbol{h}_A^{text}$ のみを用いた手法
- UBKE-LUKE: UBKE-LUKE[3)] によるエンティティのの埋め込み表現を用いた手法

4.1 結果・考察

評価実験の結果，及び，各手法の比較・考察については当日発表する．

2) https://github.com/itomoki430/Company2Vec
3) https://huggingface.co/uzabase/UBKE-LUKE

5 結論

本研究では，銘柄テキスト情報だけでなく，セグメント別売上や株価時系列データ，株式保有情報も活用した「テキスト情報」と「数値情報」をハイブリッドに活用した企業間類似度を提案した．

今後の方向性としては，PBR の推移等を含む複数の数値データの活用や取引情報や顧客情報を考慮した埋込表現・類似度算出等が考えられる．

参考文献

[1] Ken ton Lee Jacob Devlin, Ming-Wei Chang and Kristina Toutanova. Bert: pre-training of deep bidirectional transformers for language understanding. In **NAACL-HLT 2019**, p. 4171–4186, 2019.

[2] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In **International Conference on Learning Representations**, 2017.

[3] Hiroki Sakaji Steven Schockaert Tomoki Ito, Jose Camacho Collados. Learning company embeddings from annual reports for fine-grained industry characterization. In **The Second Workshop on Financial Technology and Natural Language Processing In conjunction with the 29th International Joint Conference on Artificial Intelligence**, 2021.

[4] Z. Wei Y. Chen and X. Huang. Incorporating corporation relationship via graph convolutional neural networks for stock price prediction. In **2Proceedings of the 27th ACM International Conference on Information and Knowledge Management**, pp. 1655–1658, 2018.

[5] Takehiro Takayanagi, Hiroki Sakaji, and Kiyoshi Izumi. Setn: Stock embedding enhanced with textual and network information. In **2022 IEEE International Conference on Big Data (Big Data)**, pp. 2377–2382, 2022.

[6] Dimitrios Vamvourellis, Máté Tóth, Snigdha Bhagat, Dhruv Desai, Dhagash Mehta, and Stefano Pasquali. Company similarity using large language models. In **2024 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr)**, pp. 1–9, 2024.

[7] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 11897–11916, 2024.

[8] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. In **Advances in Neural Information Processing Systems**, Vol. 36, 2024.

[9] Che Liu, Zhongwei Wan, Sibo Cheng, Mi Zhang, and Rossella Arcucci. Etp: Learning transferable ecg representations via ecg-text pre-training. In **ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 8230–8234, 2024.

[10] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6442–6454, Online, 2020. Association for Computational Linguistics.