健康経営度調査テキストに対する定量評価および レコメンドアルゴリズムの提案

林和希 [1]　参木裕之 [1]
[1] 株式会社大和総研 データドリブンサイエンス部
{kazuki.hayashi,hiroyuki.mitsugi}@dir.co.jp

概要

近年、従業員等への健康投資を行うことで、従業員の活力や企業の生産性を向上させる「健康経営」に対する関心が高まっている。本研究では、健康経営度調査票内の「健康経営課題、それに対する施策実施の結果、効果検証結果」の文章を定量的に評価する手法を提案した。また、この評価スコアを利用して各企業の健康経営課題に対して適切かつ多様、なおかつ各企業の健康経営評価の改善に繋がるような施策を提案するレコメンドシステムを開発した。

1　はじめに

1.1　研究背景

健康経営とは、従業員等への健康投資を行うことで、従業員の活力や企業の生産性を向上させる取り組みのことである。健康経営を実施している企業は、当然ながら従業員の健康状態も良好な傾向がある。例として、健康経営度調査結果の中央値で高スコア群と低スコア群の 2 群に企業を分けて分析したところ、年間医療費平均、メタボ該当率など複数の指標において、高スコア群が低スコア群をいずれも下回る結果が得られた [1]。加えて、追跡期間 14 年間 (2000 年から 2014 年) において、健康経営に関する顕彰を受けた企業の株価は市場平均 (S & P 500 株価指数) を大きく上回る上昇を示した [2]。従って、健康経営を行うモチベーションとしては従業員の健康状態を改善できるのみならず、企業の労働生産性を向上させ結果的に企業価値の上昇に繋がるといった点も考えられる。

各法人の健康経営の情報は健康経営度調査 [3] に記載されており、これらのうち数値データについては [1] など複数の研究において分析されているものの、テキストデータについては着目されて来なかった。ただし、具体的な施策やその実施結果などの有用性の極めて高い情報がテキストデータとして記載されているため、これらの評価を行うための手法を確立することが重要と考えられる。

1.2　データセット

本研究では、令和 5 年度の健康経営度調査の評価結果（フィードバックシート）のデータを使用した [3]。健康経営度調査は、評価結果の開示に同意した健康経営度調査回答法人に対する評価結果をまとめたものである。

このデータは多数の列を含むが、今回は総合評価 (数値データ) および「課題内容、施策実施結果、効果検証結果」(テキストデータ) の列を使用した。表 1 に例を示す。

1.3　関連研究

表 1 のような健康経営施策に関するテキストを評価する手法として、 LLM-as-a-Judge が考えられる。LLM-as-a-Judge とは、LLM を利用してテキストの有用性、無害性、信頼性、関連性などを評価する手法であり、人間の感覚や常識に基づく評価を LLM で自動化することが可能である [4]。LLM-as-a-Judge においては絶対評価を行う手法も存在する。ただし、今回のケースにおいては相対評価を複数回実施した上で比較観点を抽出し、それらの観点に基づく評価を行う方が一貫性および説明性の面から望まし

表 1　使用したデータの例 (架空の企業のデータ)

総合評価	620.4
課題内容	冬に社内でインフルエンザが流行し、業務に多大な影響が出る。
施策実施結果	全従業員のうち希望者にワクチン接種会社費用負担で実施した。
効果検証結果	インフルエンザ感染者が減少し、業務への影響を低減できた。

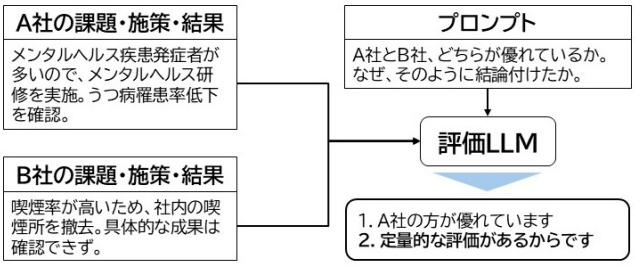

図 1　ペアワイズ評価による観点抽出の例

いと考えられる。そのための手法の 1 つとして、2 つの内容を LLM に比較させることでランク付けを行うペアワイズ評価が存在する [5]。ペアワイズ評価は LLM の出力評価に使用されることが多いが、人手で記載されたテキストに対しても有用と考えられる。そのため、本研究においては各企業の健康経営度調査票のテキストに対してペアワイズ評価を行った。

2　提案スコアの作成

2.1　提案スコアの計算フロー

本研究では図 1 に示すように、上記の課題内容、施策実施結果、効果検証結果に対して、下記の手順でペアワイズ評価を適用して比較観点の抽出を行った。

1. 課題内容、施策実施結果、効果検証結果の文章を 1 つに結合し、結合済みテキストを作成
2. 各社の結合済みテキストから、ランダムに 2 つ (A 社と B 社) を選び、LLM に比較させる
3. LLM に「A 社と B 社、どちらが優れているか。なぜ、そのように結論付けたか。」を問いかける

上記の 2,3 を繰り返すことで、「LLM がどのような観点から比較を行い、優劣をつけたのか」という情報を得ることが可能である。このようにして得られた観点をドメイン知識と照らし合わせて絞り込んだ結果、A.1 に示す 9 つの比較観点が抽出された。

加えて、各法人の結合済みテキストが上記の 9 つの基準を満たすか否かを LLM に判定させたところ、上記 9 つの観点を多く満たすほど、健康経営度調査における総合評価も向上する傾向が見られた。ただし、各観点ごとに充足難易度は異なると考えられる。具体的には、多くの企業において健康経営施策の実施結果に対する定量的な評価が行われている一方、継続的にこれらの施策を実施している企業は少ない。そのため、TF-IDF により達成難易度が高い

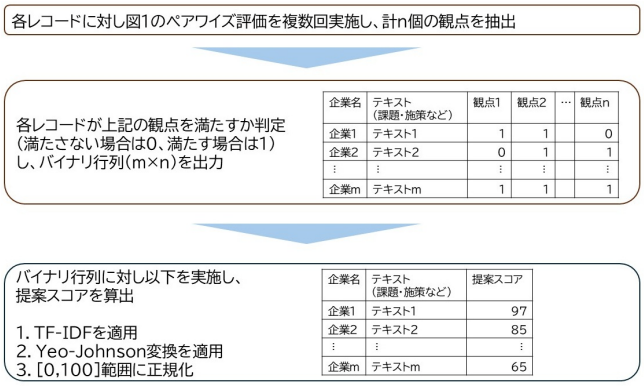

図 2　提案スコアの計算フロー

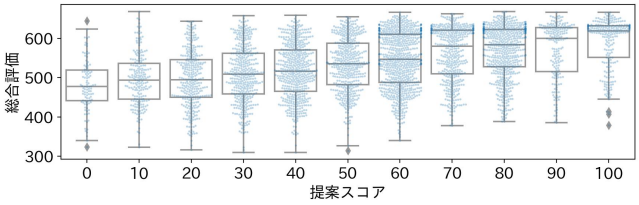

図 3　提案スコアと総合評価の関係

観点ほど高く得点が付けられるように重み付けを行い、更に Yeo-Johnson 変換および [0,100] 範囲への正規化を行った。これを「提案スコア」と呼称する。一連の処理の手順を図 2 に示す。

2.2　提案スコアの検証

図 3 に、10 刻みにビニングした提案スコアと健康経営度調査の総合評価の関係を示す。中央値などの統計値で確認すると、提案スコアが増えるほど総合評価も上昇する傾向があることが分かる。ゆえに健康施策を検討する際には、A.1 に示した 9 つの観点を意識して立案を行うことで、健康経営度の総合評価の上昇を目指すことが可能となると考えられる。

3　提案スコアの活用

3.1　応用例

提案スコアの応用例は複数考えられるが、1 つの例としては自社にスコアが近しい企業を探し出し、その企業の施策を参考にすることが考えられる。今回は考えられる応用例の 1 つとして、「自社と類似した課題を持つ企業」の施策をレコメンドすることで、自社の施策立案に活かすためのレコメンドシステムの提案を行った。

3.2 レコメンドシステムの構築

3.2.1 データの前処理

テキスト間の類似性の比較を行うため、各法人の課題内容、施策実施結果の文章を Azure OpenAI Embeddings API text-embedding-ada-002 version2 によりベクトル化した。更に、後のレコメンドに使用するために上述の手法で得られた施策実施結果ベクトルを BERTopic [6] でクラスタリングし、各法人の施策に施策クラスタ番号を付与した。また、A.2 に示すように各施策に施策タグを付与した。

3.2.2 提案アルゴリズム

Algorithm 1 提案するレコメンドアルゴリズム

```
Require: k                          ▷ 必要な施策数
Require: criterion                  ▷ ソート基準 (提案スコアなど)
Require: data                       ▷ 健康経営度調査票データ
Require: own_arr                    ▷ 自社の課題内容ベクトル
Require: other_arr                  ▷ 他社の課題内容ベクトル
Ensure: selected_policies           ▷ k 個の施策リスト
  function GEN_CANDIDATES(own_arr, other_arr, k)
      similarities ← cosine_sim(own_arr, other_arr)
      sorted_sim ← sort(similarities)
      clusters ← get_clusters(sorted_sim)
      top_k_clust ← drop_duplicates(clusters)[: k]
      return top_k_clust
  end function
  function GET_BEST_POLICY(clust_n, criterion)
      filtered ← data.query([policy_num] == clust_n)
      best_data ← filtered.sort_by(criterion)[0]
      return best_data
  end function
  function RANKING(clusters, criterion)
      policies ← []
      for all i ∈ clusters do
          policies.append(get_best_policy(i, criterion))
      end for
      return policies
  end function
  clusters ← gen_candidates(own_arr, other_arr, k)
  selected_policies ← ranking(clusters, criterion)
```

一般的な EC サイトと異なり、健康経営施策においてはクリック数や購入回数などに基づく一般的なレコメンドアルゴリズムは使用できない。そのため、クラスタリングに基づく 2-stage レコメンドシステムを開発した。2-stage レコメンドシステムとは、全体集合から自社に関連する候補アイテムを選出し、候補として選ばれたアイテムの中から、自社にとって望ましいと思われる順番にソートすることで、自社に対してアイテムをレコメンドするシステムのことである。本研究で提案するレコメンドアルゴリズムは下記の通りである。

1.候補集合生成
   1.「自社の課題内容のベクトル」と「レコメンド候補となる複数他社の課題内容のベクトル」のコサイン類似度を計算
   2.類似度基準で降順にソート
   3.ソート結果の上位 k 個の施策クラスタ番号 (3.2.1 にて作成) を選定
2.ランク付け
   1.選択されたそれぞれの施策クラスタの中で総合評価や提案スコアの最も高い施策をレコメンド

3.2.3 比較手法

コサイン類似度を基準に自社課題に近い他社課題に対する施策を任意の数レコメンドするルールベースのアルゴリズムを比較手法とした。

3.3 レコメンドシステムの評価

一般的な EC サイトと異なり、健康経営施策においてはクリック数、購入回数などに基づく Precision@K やコンバージョン率といった一般的な評価指標を使用することはできない。そのため、以下で定義する 3 つの指標で評価した。ただし、これら 3 要素はトレードオフの関係にあり、全ての水準において高いスコアを両立することは不可能である。

また、本実験では自社が必要とする施策数を $k = 5$ とした。そのため、200 件のクエリ (自社の課題内容) に対して 1000 件の施策がレコメンドされた。提案アルゴリズム 3.2.2 においては提案スコアによるソートを実施した。

3.3.1 妥当性

課題に対して、適切な施策がレコメンドされるかを示す指標である。本研究では、自社の課題内容に

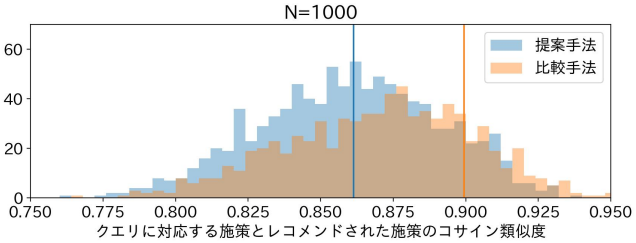

図 4　コサイン類似度による妥当性比較結果

表 2　改善性の比較

手法	改善性
提案手法	99.4 %
比較手法	72.0 %

対する施策とレコメンドされた施策のベクトル間のコサイン類似度として定義した。自社の課題内容に対する施策テキストのエンベディングを A、レコメンドされた施策テキストのエンベディングを B として、下記のように求められる。

$$\text{cosine\_similarity}(A, B) = \frac{A \cdot B}{\|A\|\|B\|} \tag{1}$$

図 4 に示すように、この基準において優れていたのは比較手法であった。これは、3.2.3 に示した比較手法のアルゴリズムを考慮すれば当然の結果である。

3.3.2　改善性

自社のスコアの向上に繋がるような施策がレコメンドされるかを示す指標である。本研究では、k 個のレコメンドされた施策のうち自社より提案スコアが高いものの割合として定義した。

$$\text{改善性} = \frac{\text{自社より評価が高い施策数}}{\text{レコメンドされる施策数}} \tag{2}$$

表 2 に示すように、提案手法は比較手法に比べて自社より高評価な施策を表示する傾向があった。これは提案手法内におけるランク付けによるものである。なお、提案手法にルールベースでのフィルタリングなども導入することで更に改善性は向上すると思われるが、他の指標に悪影響を及ぼすと考えられるため、注意が必要である。

3.3.3　多様性 (カバレッジ)

レコメンドされた施策の多様性を示す指標である。本研究では、k 個のレコメンドされた施策に含まれる施策タグの数を全施策の数で割ったものとした。

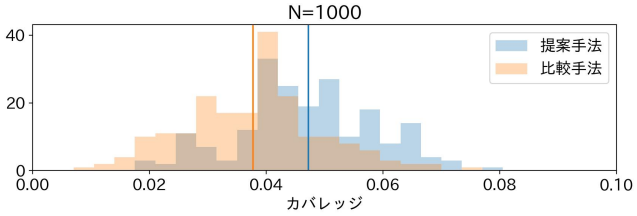

図 5　カバレッジ

$$\text{多様性} = \frac{\text{レコメンドされる施策タグ数}}{\text{全施策タグ数}} \tag{3}$$

各手法のカバレッジは、図 5 に示す通りである。比較手法と比べて、提案手法は多様な施策をレコメンドする傾向があることが明らかになった。

4　結論

本研究では、健康経営度調査票内のテキストをペアワイズ評価により比較することで観点を抽出した。さらに各レコードが観点を満たすか否かから計算可能なスコアを提案した。更に、上記のスコアを活用して自社の課題に対して施策をレコメンドする 2-stage のレコメンドシステムを提案した。提案手法は、課題に対して適切かつ、評価が高く多様な施策をレコメンドできることを示した。

5　今後の展望

現時点では比較対象となるレコードをランダムサンプリングにより選定して比較している。そのため、類似した傾向を持つ企業間の比較が多く行われてしまっているため、乖離した状態にある企業間の比較で着目するべき観点が見落とされてしまっている可能性がある。故に、k-means 法などによるクラスタリングを行い総合評価が離れた企業同士を比較するといった手法により、更に多くの観点を抽出することで健康経営度評価の上昇に繋がるような施策の立案および評価に役立てることができると考えられる。

また、提案手法を用いることで企業の健康経営状態（「課題内容、施策実施結果、効果検証結果」から算出した提案スコア）の類似度を計算することができる。この際、各企業がどの観点を満たすか否かといった情報も合わせて取得することができるため、自社の不足点を深く考慮した上で他社の施策を参考にした施策立案を行うといったことが可能となる。

参考文献

[1] 経済産業省. 健康経営の推進, 2016. https://kenko-keiei.org/document_dl/symposium0403.pdf.

[2] Ron Z Goetzel, Raymond Fabius, Dan Fabius, Enid C Roemer, Nicole Thornton, Rebecca K Kelly, and Kenneth R Pelletier. The stock performance of c. everett koop award winners compared with the standard & poor's 500 index. **Journal of Occupational and Environmental Medicine**, Vol. 58, No. 1, pp. 9–15, 2016.

[3] 健康経営優良法人認定事務局（日本経済新聞社）. 認定企業評価結果（フィードバックシート）の開示, （2024-12 閲覧）. https://kenko-keiei.jp/houjin_list/feedback/.

[4] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In **Proceedings of the 37th International Conference on Neural Information Processing Systems**, pp. 46595–46623, 2023.

[5] Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, et al. Large language models are effective text rankers with pairwise ranking prompting. In **Findings of the Association for Computational Linguistics: NAACL 2024**, pp. 1504–1518, 2024.

[6] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. **arXiv preprint arXiv:2203.05794**, 2022.

A　付録

A.1　抽出された 9 観点

1. 定量的な評価が行われているか
2. 前年度と比較して改善がなされたか
3. 施策による明確な改善があったか
4. 従業員からの肯定的なフィードバックがあったか
5. 従業員の健康リテラシー向上や意識・行動の変容、企業風土の変革などに結び付いたか
6. 全社的な取り組みか行われているか
7. 継続的な取り組みか行われているか
8. 明確な課題・目標設定に基づいた計画的な取り組みか行われているか
9. テーマに合致した取り組みが行われているか

A.2　施策タグの作成

可用性向上のため、各法人の施策実施結果テキストに施策タグを付与した。手順は下記の通りである。

- 施策実施結果のテキストから、LLM によって具体的な施策名を抽出する
- 抽出された施策名を Azure OpenAI Embeddings API text-embedding-ada-002 version2 によりベクトル化する
- 上記の施策名を BERTopic でクラスタリングする
- クラスタごとに名称を付与し、これを施策タグとする

1 つの施策実施結果テキストに対して、複数の施策タグが付与されることに注意が必要である。また、3.3.3 のように全施策タグのうち何種類がレコメンドされたのかを計算することで、レコメンドアルゴリズムにより提案された施策の多様性を評価することが可能となる。