Vol.2022-GN-115 No.40
Vol.2022-CDS-33 No.40
Vol.2022-DCC-30 No.40
2022/1/21

Web データに基づく質問応答システムを用いた企業活動分析・評価システム WISDOM-DX の改善

久寿居 大[†1]　石川　開[†2]　市瀬　規善[†2]
奥村 明俊[†2]　鳥澤 健太郎[†3]　大竹 清敬[†3]

概要：
近年，デジタルトランスフォーメーション（DX）推進のために，アンケートなど各種調査に基づく分析・評価が実施されている．大規模調査では，アンケートの設計者と回答者，アンケートを分析・評価する専門家のコスト低減が課題となる．我々は，大規模かつタイムリーな調査を実現するため，専門家による評価項目を 5W1H の質問タイプとしてモデル化し， Web データに基づく質問応答システムを用いて企業の DX 活動を評価する WISDOM-DX を開発した．WISDOM-DX は，5W1H に関する 6 つの質問応答結果について，Web 上の掲載情報量，情報の信頼度，優良事例との類似度などを評価観点とする 8 つのスコア関数によって値を算出し，マルチクエリアンサンブルによって重みづけした値を結合スコアとした．DX 銘柄 2021 選定の調査に回答した企業 464 社から専門家が選定した企業 48 社を識別する実験を 8 つの結合スコアによって行った結果，各々の AUPR は 0.503 から 0.543 となり，AUPR が 0.543 の場合，上位 48 社の精度は 56.3%であった．しかし，特定のスコア関数が常に最高精度を有する保証はなく，スコア関数を精度よく一意に結合する必要があった．本論文では，過去の選定結果を学習データとして，掲載情報量，類似度，信頼度などの評価観点がどの程度重視されたかを推定し，8 つの結合スコア関数のスコアをさらに結合して計算するマルチクエリスコアアンサンブルを提案した．本手法を DX 銘柄 2021 の選定企業 48 社で実験した結果，AUPR が 0.541，上位 48 社の精度は 56.3%となり，8 つのスコア関数の最高精度と同等の精度を得られることを確認した．本手法により，複数の評価スコアを結合する基準が明白でない場合も，過去の評価結果から精度よく結合することが可能となり，今後の DX 活動や他のテーマに関する評価への適用可能性を向上させた．

キーワード：デジタルトランスフォーメーション，DX, テキストマイニング，質問応答

Improving Automatic Enterprise Evaluation System WISDOM-DX using QA System based on Web Information

DAI KUSUI[†1]　KAI ISHIKAWA [†2]　NORIYOSHI ICHINOSE[†2]
AKITOSHI OKUMURA[†2]　KENTARO TORISAWA[†3]　KIYONORI OHTAKE[†3]

Abstract:
In recent years, analysis and evaluation through various surveys, such as questionnaires, have been conducted in order to promote digital transformation (DX). It is important to reduce the cost of questionnaire designers, questionnaire respondents, and experts who analyze and evaluate the questionnaires in large-scale surveys. We developed WISDOM-DX, a system that automatically analyzes and evaluates an organization's DX initiatives instead of questionnaires using a question-and-answer (QA) system filled from Web information. By modeling the evaluation items of experts in the form of six question types (When, Who, Where, What, Why, How) , WISDOM-DX evaluates responses of the QA system for the six question types, by using eight score functions based on response volume, similarity to DX good practices, and confidence level. When the system was evaluated with 464 companies that replied to the DX Brand 2021 questionnaires by comparing the 48 companies selected by DX experts, the eight score functions all provided AUPR values between 0.503 and 0.543. When AUPR was 0.543, the precision was 56.3% for the 48 companies. However, there was no guarantee that a particular score function would always have the highest accuracy, and it was necessary to uniquely combine eight score functions with high accuracy. This paper proposes a multi-query score ensemble method, which makes it possible to combine the eight score functions using the past selection results as training data. The method can appropriately learn coupling coefficients regarding response volume, similarity to DX good practices, and confidence level. At the result of testing on the 48 companies selected for DX Brand 2021, we confirmed that the AUPR was 0.541 and the precision was 56.3%, which was equivalent to the highest accuracy of the eight score functions. The results show the method could combine multiple evaluation scores with high accuracy based on past evaluation results, even when the criteria for combining multiple scores are not obvious. The method improved the applicability of WISDOM-DX to future DX activities and evaluations on other topics.

Keywords:　Digital Transformation, DX, Text Mining, Question and Answering

1. はじめに

行政機関は, Society5.0 の実現に向けてデータ利活用とデジタル・ガバメントの観点から社会全体のデジタル化に取り組んでいる[1,2]．このような取組みでは，人的リソースや財源を最大限有効活用するためエビデンスに基づいた施策の立案と迅速な評価が求められる[3]．デジタル化の重要な施策として，デジタルトランスフォーメーション（DX）が推進されており[4]，各企業の DX に対する取り組みを調査・分析・評価する活動が行われている．評価活動は，アンケートやヒアリングなどの調査に基づいて専門家が分析して行うことが多い[5]．大規模調査では，アンケート設計者とアンケート回答者，分析の専門家の手間や費用などコスト低減が課題となる．まず，アンケートを設計すること

†1 1NEC ソリューションイノベータ株式会社
NEC Solution Innovators, Ltd.
†2 独立行政法人 情報処理推進機構 （IPA）
Information-technology Promotion Agency, Japan
†3 国立研究開発法人　情報通信研究機構（NICT）
National Institute of Information and Communications Technology

Vol.2022-GN-115 No.40
Vol.2022-CDS-33 No.40
Vol.2022-DCC-30 No.40
2022/1/21

が簡単ではなく，回答者にとっては回答そのものが手間であり，詳細な調査のため項目や記述内容が増えると負担感が増大し回答率が低下する．また，回答者が自由に記述した内容の分析（質的分析）は，専門家にとっても負担となる．そのため大規模調査をタイムリーに実施するのは容易ではない．質的分析は，数値化されたデータに対する量的分析とは異なり，分析法が標準化ないし規格化されているとは言い難く[6]，解釈が恣意的であり解釈に至る過程が不明確になることもある[7]．専門家一人で把握できるデータ量に限界があり，人手による質的分析には見落としや主観による偏りの問題が指摘されている[7]．このため，専門家による属人性を排除し評価の一貫性を維持するために膨大な手間や費用がかかることが課題であった．

我々は，大規模かつタイムリーな調査を可能とするため，専門家による評価項目を5W1Hの質問タイプとしてモデル化し， Web データに基づく質問応答システムを用いて企業の DX 活動を評価する WISDOM-DX システムを開発した[8]．WISDOM-DX は，5W1H の 6 つの質問タイプに対する質問応答システムの応答結果をスコア付けし，6 つのスコアから結合スコアを算出するマルチクエリアンサンブル（MQE）を開発した．8 つのスコア関数を準備して比較したところ，8 つの結合スコアはいずれも 50%以上の精度を有することを確認した[8]．しかし特定のスコア関数が常に最高精度を有する保証はなく 8 種類のスコア関数のいずれを選択するかの基準は明確ではなかった．今後の DX 活動や他のテーマに関する評価を自動的に行うために MQE を改良する必要がある．本論文では，様々な観点からの企業活動のスコアを統合する手法として，MQE を改良したマルチクエリスコアアンサンブル（MQSE）を提案する．過去の評価結果を学習データとして評価モデルに反映可能とすることにより統合基準を明確にして様々な調査に活用できるように汎用性向上を目指す．2 章では，関連調査研究として，DX に関する調査，DX 調査事例として DX 銘柄 2021 選定のアンケート調査や選定プロセス，質問応答システムについて概説する．3 章では，WISDOM-DX の内容とシステム構成，MQE について説明する．4 章では，MQE の課題とそれを解決する MQSE について説明する．5 章では，MQSE を用いて DX に関する企業のランキングを行い，ベースラインとして google の検索結果と，DX 銘柄 2021 選定の調査と比較した結果について考察する．

2. 関連調査と研究

2.1 DX に関する調査

近年，DX に関して進展状況の分析や助成事業の採択審査のため，民間調査会社[9,10,11]，業界団体[12,13,14]，地方自治体[15,16]，行政機関[17,18,19]など多数の組織によって調査が実施されている．調査の対象は，民間企業，自治体や官公庁など公的機関，各種団体など様々であるが，アンケートやヒアリングの結果や提案書などをもとに有識者など専門家が分析・評価することが一般的である．アンケートの回答は，選択式と自由記述式が用いられる．選択式はあらかじめ作成された選択肢を回答者が選び，自由記述式は回答者が自分の言葉で表現する．選択式は大量のデータの確保や回答者の分類など量的分析が容易であり，自由記述式は回答者の主張や意図などの質的分析に適する．

2.2 DX 銘柄 2021 の選定

DX 銘柄 2021 は，経済産業省と東京証券取引所とが共同で行う優れた DX 活動を行う企業を選定する取組である．国内企業を対象に選択式と自由記述式のアンケート調査を行い，評価委員が優れた企業を選定している[17,18]．DX 銘柄とは，東京証券取引所の上場企業の中から，企業価値の向上につながる DX を推進するための仕組みを社内に構築し，優れたデジタル活用の実績が表れていると選定された企業である．DX 銘柄 2021 の選定にあたって，DX 調査事務局が，2020 年 11 月から，東京証券取引所の国内上場会社約 3,700 社に対して，「経営ビジョン・ビジネスモデル」「戦略」「戦略実現のための組織・制度等」「戦略実現のためのデジタル技術の活用・情報システム」「成果と重要な成果指標の共有」「ガバナンス」の 6 項目に関するアンケート調査を実施した．アンケート調査は選択式と自由記述式の 2 種類である．銘柄選定は，一次評価と二次評価の 2 つのプロセスで行われた．一次評価として，DX 銘柄 2021 の調査に回答した東証株価指数 33 業種の企業 464 社からのアンケートの選択式回答 35 項目と各企業の 3 年平均 ROE（自己資本利益率）によるスコアリングが行われた．次に，評価者 9 名から構成される評価委員会においてアンケートの記述式回答 38 項目などから取り組みについて評価され，DX 銘柄 2021（グランプリを含む 28 社）と DX 注目企業 2021（20 社）などが選定され 7 か月後に発表された[17]．民間調査会社や業界団体においても調査開始から発表まで数か月かかることが多く[14]，調査に要するコストは小さくない．DX 銘柄 2021 と同様の調査が 2015 年から 2019 年までは「攻めの IT 経営銘柄」として，2020 年から「DX 銘柄」として実施されてきた．これらの調査のアンケート回答率は 6%～15%と高くない [18]．回答する企業の負担も大きく，回答および調査・分析のコストを低減しつつ，一貫性のある評価の大規模かつタイムリーな実現が今後の課題である．DX 銘柄に選定された企業は，単に優れた情報システムの導入，データの利活用にとどまらず，デジタル技術を前提としたビジネスモデルと経営の変革にチャレンジし続けていると評価された企業である．選定企業の優良な取組が他の企業における DX 戦略立案や DX 推進施策の参考となることが期待されている．

2.3 自然言語処理による質問応答システム

質問に対する回答の生成は，質問応答システムとして研究されている[20]．質問応答能力は，準備できるデータの質

Vol.2022-GN-115 No.40
Vol.2022-CDS-33 No.40
Vol.2022-DCC-30 No.40
2022/1/21

や量に依存しデータをいかに獲得・更新するかが課題となる．情報源として Web データやウィキペディアを活用する手法も提案されている[21,22,23]．すべての質問に完全に応答できるものではなく，実用として活用するためにはタスク設計が重要な課題となる．

WISDOM X [23]は，Web データに基づく質問応答システムである．Web 60 億ページから抽出した情報を基にして，ファクトイド型(例えば,「地球温暖化は何をもたらす？」)，なぜ型 （例えば,「なぜ地球温暖化が起きる？」）[24]，どうやって型（例えば,「どうやって地球温暖化を防ぐ？」）どうなる型 （例えば,「地球温暖化が進むとどうなる？」）[25]，定義型（例えば,「地球温暖化とは何？」）といったタイプの質問に回答することができる．WISDOM X は 350GB のテキストで事前学習した BERT や，BERT と敵対的学習と呼ばれる深層学習技術[25,26,27]を組み合わせることで，入力された質問中の語句やキーワード，さらにはそれとほぼ同義な表現等を，システムが Web ページを対象に検索し，質問に高精度に回答できる．なお，WISDOM X は，一般にその出力結果を WISDOM X 以外のシステム開発やデータベース構築に利用することを許諾していない．WISDOM-DX は，NICT の了解のもと WISDOM X を活用している．

3. 企業活動分析・評価システム WISDOM-DX

3.1 WISDOM-DX の概要

我々が開発した WISDOM-DX のシステム構成を図 1 に示す．WISDOM-DX では，アンケート設計者の労力を削減するため，評価項目を 5W1H（いつ，誰が，どこで，何を，なぜ，どうやって）を用いたマルチクエリの質問タイプでモデル化する．DX に活発に取り組んでいる企業ほど，報道，IR，プロモーションなどを通じて Web 上に多くの活動情報が掲載される，という仮説を立てた．調査目的（デジタルトランスフォーメーション）に合わせて，Web 上に掲載された企業活動の情報を，質問応答システムを用いて収集する．Web 上の「掲載情報量」と情報の「信頼度」，優良事例との「類似度」の 3 観点に基づいて各企業の活動を評価する．

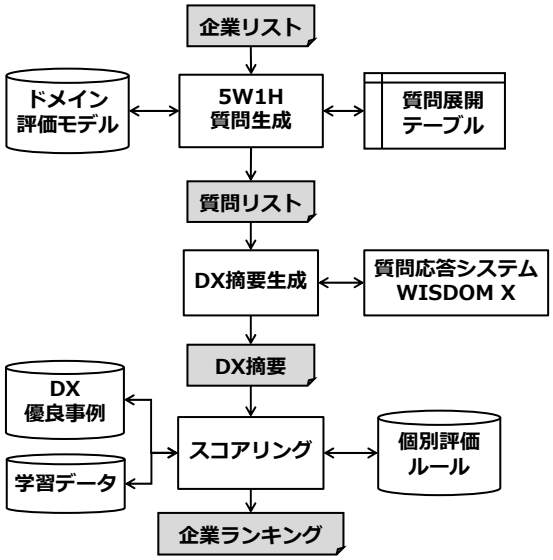

図 1 WISDOM-DX システム構成
Figure 1 WISDOM-DX System Configuration

5W1H 質問生成は，入力された企業リストの各企業に対し，ドメイン評価モデルのキーワードと質問展開テーブルを用いて，企業の DX 活動の 5W1H を問う多様な質問文を質問リストとして生成する．

次に，調査回答者の労力削減のため，Web データに基づく質問応答システム WISDOM X を利用し，調査回答者がアンケートに回答することなく Web から回答を取得する．DX 摘要生成が質問リストを入力として質問応答システム WISDOM X に与え，質問文に対する応答結果を得る．WISDOM X の応答結果から，同じ 5W1H の応答ごとにマージし，これを DX 摘要と呼ぶこととする．

最後に，分析専門家の労力削減のため，スコアリングでは DX 摘要を入力として企業のランキングを出力する．5W1H の質問タイプごとに,「掲載情報量」「信頼度」「類似度」の観点から複数のスコアを計算し，スコアをアンサンブルする手法によりランキングを決定する．最終的なランキングは個別評価ルールを用いて修正される．個別評価ルールは例えばランキング上位が特定業種に偏らないようにする等の特定企業のスコアを修正する仕組みである．DX 優良事例は，具体的には過去年度の DX 銘柄[18]や注目企業[18]として公開された取り組みの紹介記事からテキストデータを抽出したもので，類似した取り組みをしているほどスコアを高くする．学習データは，各企業の過去年度の選定結果の二値データであり，スコアを統合するパラメータの学習に用いる．

3.2 5W1H 質問生成

提案システムでは，様々な質問タイプへの応答が可能な質問応答システム WISDOM X の性能を十分活用するために，調査目的（デジタルトランスフォーメーション）に合わせて汎用的かつ網羅的な，5W1H に関する質問文の生成を行う．例えば，企業 A の情報を収集するため「企業 A はどうやって DX をしたか？」「企業 A はどこで DX をしたか？」「企業 A は誰が DX をしたか？」「企業 A は DX で何をしたか？」「企業 A はなぜ DX をしたか？」「企業 A はいつから DX をしたか？」の 6 タイプの質問を生成する．

DX 銘柄 2021 選定の調査では大きく 6 項目，A）経営ビジョン・ビジネスモデル，B）戦略，C）戦略実現のためのデジタル技術の活用・情報システム，D）戦略実現のための組織・制度等，E）ガバナンス，F）成果と重要な成果指標の共有,に関するアンケート調査を実施している．HOW（どうやって）と WHY（なぜ）の質問文で，企業 A の活動に関する A，B，C の項目に対応する表現を Web ページから検索して抽出する．WHERE（どこで）と WHO（誰が）の

Vol.2022-GN-115 No.40
Vol.2022-CDS-33 No.40
Vol.2022-DCC-30 No.40
2022/1/21

質問文で D の項目，WHO（誰が）の質問文で E の項目，WHAT（何を）の質問文で F の項目，に対応する表現を抽出する．WHEN（いつから）の質問文は DX 銘柄 2021 選定の調査では必ずしも対応するアンケート項目はないが，より汎用的に調査を行えるように質問タイプに加えている．

質問展開テーブルには，質問タイプと対応付けて，次のようなスロット付きの質問文のテンプレートが保持される．

1) 質問タイプ 1:「<sub>はどうやって<obj>を<pred>か？」
2) 質問タイプ 2:「<sub>はどこで<obj>を<pred>か？」
3) 質問タイプ 3:「<sub>は誰が<obj>を<pred>か？」
4) 質問タイプ 4:「<sub>は<obj>で何を<pred>か？」
5) 質問タイプ 5:「<sub>はなぜ<obj>を<pred>か？」
6) 質問タイプ 6:「<sub>はいつから<obj>を<pred>か？」

これらのテンプレートのスロットにあてはまる表現部分は，ドメイン評価モデルに保持される．具体的には，スロット<obj>にあてはまる目的語の表現として「デジタルトランスフォーメーション」が，<pred>に対する表現としては，「した」，「成し遂げた」，「達成した」といった同義語が登録されている．企業リストに含まれる企業名の異表記も，ドメイン評価モデルに登録され，企業名リストが入力された際に，企業名の異表記についても，スロット<sub>に組み合わされ，質問文が生成される．質問応答システム WISDOM X は文の構造と語彙の一致によって Web から応答情報を見つけ出すので，日本語としてやや不自然であっても問題はない．

3.3 DX 摘要の生成

5W1H 質問生成では，企業リストに記載の各企業に対して複数の質問文が質問リストとして生成される．質問リストを入力として質問応答システムから質問文ごとに複数の応答が得られる．ドメイン評価モデルに登録された異表記や同義語に由来する複数の応答は各企業の質問タイプごとにまとめられる．各応答から(1)回答の情報源として用いられた Web ページのスニペット，(2)URL，(3)質問応答システムが推定する信頼度，の 3 つ組が抽出される．各企業の質問タイプごとに，重複している応答の 3 つ組は一つにまとめられ，DX 摘要となる．

3.4 スコアリング

3.4.1 スコア関数

提案システムでは，企業の DX に対する取り組みの優良さや活動の活発さなどを，質問応答システムから得られる情報の掲載情報量，DX 優良事例との類似度，情報の信頼度の 3 観点を軸に 8 つのスコア関数を定義している．掲載情報量から計算（**cnt**），掲載情報量と語彙的類似度を類似度として計算（**sim**），掲載情報量と語彙的類似度を単語の重みで補正した類似度から計算（**sim_idf**），掲載情報量と語彙的類似度を単語の重みと単語頻度で補正した類似度から計算（**sim_tf_idf**）した 4 つのスコア関数と，それぞれのスコア関数に信頼度を組み合わせて計算した 4 つのスコア関数（**cnt_conf, sim_conf, sim_idf_conf, sim_tf_idf_conf**）である．

上記 8 つのスコア関数の定式化において，全ての質問タイプに関する DX 摘要の集合を$\boldsymbol{D}$, 質問タイプtに対する DX 摘要を$\boldsymbol{D}_t$，さらに$\boldsymbol{D}_t$ の要素をd_t で表す．DX 摘要の要素d_tは，応答要素由来のスニペット，URL，信頼度の組を持つ．このうち要素d_t の信頼度を$conf(d_t)$，要素d_t のスニペットに含まれる単語集合の要素を$w_t \in d_t$ で表す．また，DX 優良事例のテキストをd_h，同テキストに含まれる単語集合を$\{w_h\}$ で表す．

スコアリングを構成する 3 観点のうち，「信頼度」については要素d_t の信頼度$conf(d_t)$，「掲載情報量」については DX 摘要$\boldsymbol{D}_t$ に含まれる要素d_t の要素数を用いることができる．一方，「類似度」については， DX 摘要$\boldsymbol{D}_t$ に含まれる要素d_t と，DX 優良事例のテキストd_h の間の類似度を用いることができる．ここでは，DX 摘要 $\boldsymbol{D}_t$ に含まれる単語 w_t と DX 優良事例のテキストd_h に含まれる単語集合$\{w_h\}$ の間に定義される次式のような語彙的類似度$sim(w_t, \{w_h\})$ を「類似度」の具体的な実装として用いる．

$$sim(w_t, \{w_h\}) = \max_{\{w_h\}} \frac{\boldsymbol{v}(w_t) \cdot \boldsymbol{v}(w_h)}{\|\boldsymbol{v}(w_t)\| \|\boldsymbol{v}(w_h)\|}$$

ここで， $\boldsymbol{v}(w_t)$ および $\boldsymbol{v}(w_h)$ はそれぞれ，単語 w_t および w_h の単語埋め込みベクトルを表す．提案システムでは，自然言語処理ライブラリ spaCy [28]と， UD Japanese GSD [29] 由来の spaCy 用日本語言語モデル“ja_core_news_lg”を用いて形態素解析を行い，解析結果から単語とその単語埋め込みベクトル（48 万語，300 次元）を取得する．また「類似度」評価のバリエーションとして，上記の語彙的類似度$sim(w_t, \{w_h\})$に情報検索で一般的な単語重み$idf(w_t)$や$tf(w_t)$ による補正を加えた「類似度」も導入する．

ここまでの議論で導入された「掲載情報量」，「類似度」，「信頼度」の 3 観点を軸とするスコアリングの基本設計に沿って，DX 摘要$\boldsymbol{D}_t$ に対するスコア関数$Score(\boldsymbol{D}_t)$ の定式化を行う．まず，「掲載情報量」の 1 軸のみを用いて，次式のようにスコア関数を定式化する．

$$Score_{cnt}(\boldsymbol{D}_t) = \sum_{d_t \in \boldsymbol{D}_t} 1$$

「掲載情報量」と「信頼度」の 2 軸から，次式の定式化が可能である．

$$Score_{cnt\_conf}(\boldsymbol{D}_t) = \sum_{d_t \in \boldsymbol{D}_t} conf(d_t)$$

「掲載情報量」と「類似度」の 2 軸からは，まず語彙的類似度$sim(w_t, \{w_h\})$ のみを「類似度」の評価値として用いれば，次式の定式化が可能である．

$$Score_{sim}(\boldsymbol{D}_t, \{w_h\}) = \sum_{d_t \in \boldsymbol{D}_t} \sum_{w_t \in d_t} sim(w_t, \{w_h\})$$

「類似度」の評価に，語彙的類似度$sim(w_t, \{w_h\})$ を単語の重み$idf(w_t)$ で補正した定式化を用いれば，次式のスコア

Vol.2022-GN-115 No.40
Vol.2022-CDS-33 No.40
Vol.2022-DCC-30 No.40
2022/1/21

関数が得られる．

$$Score_{sim\_idf}(\boldsymbol{D}_t,\ \{w_h\}) = \sum_{d_t \in \boldsymbol{D}_t} \sum_{w_t \in d_t} sim(w_t, \{w_h\}) \cdot idf(w_t)$$

単語頻度 $tf(w)$ も併用すれば，次式のスコア関数となる．

$$Score_{sim\_tf\_idf}(\boldsymbol{D}_t,\ \{w_h\}) = \sum_{d_t \in \boldsymbol{D}_t} \sum_{w_t \in d_t} sim(w_t, \{w_h\}) \cdot tf(w_t) \cdot idf(w_t)$$

「掲載情報量」,「類似度」,「信頼度」の 3 軸を用いる場合，語彙的類似度$sim(w_t, \{w_h\})$ を「類似度」の評価に用いれば，次式のスコア関数が得られる．

$$Score_{sim\_conf}(\boldsymbol{D}_t,\ \{w_h\}) = \sum_{d_t \in \boldsymbol{D}_t} \sum_{w_t \in d_t} sim(w_t, \{w_h\}) \cdot conf(d_t)$$

「類似度」の評価値に単語の重み$idf(w_t)$ で補正した語彙的類似度を用いた場合，次のスコア関数を得る．

$$Score_{sim\_idf\_conf}(\boldsymbol{D}_t,\ \{w_h\}) = \sum_{d_t \in \boldsymbol{D}_t} \sum_{w_t \in d_t} sim(w_t, \{w_h\}) \cdot idf(w_t) \cdot conf(d_t)$$

「類似度」に，さらに単語頻度 $tf(w)$ の補正を加えると次式となる．

$$Score_{sim\_tf\_idf\_conf}(\boldsymbol{D}_t,\ \{w_h\}) = \sum_{d_t \in \boldsymbol{D}_t} \sum_{w_t \in d_t} sim(w_t, \{w_h\}) \cdot tf(w_t) \cdot idf(w_t) \cdot conf(d_t)$$

提案システムでは，以上によって導入された計 8 つのスコア関数を DX 摘要のスコアリングに用いる．

3.4.2 マルチクエリアンサンブル

提案システムでは，一つの企業に対して 6 種類の質問タイプごとに，6 つの DX 摘要が得られる．また，それぞれの DX 摘要に対して 8 つのスコア関数が適用されるので，一つの企業に対して 48 のスコアが得られる．

複数のスコアを統合する手法として，教師無しの統合手法 Reciprocal Rank Fusion (RRF)が提案されている. RRF は，順位の逆数（Reciprocal Rank，ただし定数補正項付き）を重み無しで足し合わせた単純な定式化であるが，NIST TREC における複数の関連文書ランキングの統合では，標準的な統合手法である Condorcet や 他の学習ベースの統合手法と比較しても良い性能が得られることが報告されている[30]. そこでスコアの統合に RRF を適用した．ただし，48 のスコアの中でどのスコアが有効かを事前に知ることはできないため，アンサンブル法の一種である重みづけ平均化[31]の考え方を RRF に応用し, 結合パラメータを組み込んだ拡張版 RRF をマルチクエリアンサンブル（MQE）として提案した[8]．MQE では過去の調査結果（選定結果）を学習データとして，その調査ではどの質問タイプがどれくらい重視されていたかを推定し，6 つの質問タイプのスコアを結合した結合スコアを計算した．8 つのスコア関数から「掲載情報量」,「類似度」,「信頼度」のいずれを重視するかによって 8 つの結合スコアが計算される．

3.4.3 個別評価ルール

DX 銘柄 2021 選定の調査において, 過去年度の DX 銘柄，注目企業，攻めの IT 経営銘柄を分析すると，専門家による企業の選定において，特定業種に選定企業が集中するのを避けるため，業種毎の選定数が一定数を超えないような調整が行われている[17]．このような個別の調査に特有の評価指標に対応するため，条件にマッチした特定企業のスコアを修正する仕組みを用意する．

4. マルチクエリスコアアンサンブルによる改良

4.1 マルチクエリアンサンブルの課題

大規模な調査において，評価指標は明確であっても各評価指標をどれだけ重視するかは明確でなく，複数の評価者同士での調整やすり合わせが必要になることが多い．「掲載情報量」,「類似度」,「信頼度」のいずれが重視されるかは明確でなく，8 つのスコア関数のうち，いずれが最も性能が良いかは自明ではなかった．

4.2 マルチクエリスコアアンサンブル

過去の選定結果を学習データとして「掲載情報量」「類似度」「信頼度」がどれくらい重視されていたかを推定し，8 つのスコア関数のスコアを結合したスコアを計算するマルチクエリスコアアンサンブル（MQSE）を提案する．

MQSE では，6 種類の質問タイプと 8 種類のスコア関数を二段階に分けて統合することとした．以下の手順によって結合係数の推定を行う．

Step 1: 各企業の 6 種類の質問タイプに対して得られた各 DX 摘要 $\boldsymbol{D}_t$ から，前節の 8 種類のスコア関数を用いてスコアを求める．

Step 2: Step 1 で求められた DX 摘要のスコアを，同じ質問タイプのグループに分けてソートし，質問タイプごとに企業集合内の各企業の順位 $rank\big(Score_s(\boldsymbol{D}_t)\big)$ を求める．

Step 3: それぞれの質問タイプ, スコア関数の組合せごとに求めた企業の順位から，全質問タイプに対する DX 摘要 $\boldsymbol{D}$ に対する総合スコア $Score_{ens}$ を次式で求める．

$$Score_{ens}(\boldsymbol{D}) = \sum_{\{s,t\}} \frac{\widehat{c_{s,t}}}{rank\big(Score_s(\boldsymbol{D}_t)\big)}$$

ここで，$\{s\}$ は 8 種のスコア関数，$\{t\}$ は 6 種の質問タイプ，$rank\big(Score_s(\boldsymbol{D}_t)\big)$ は質問タイプ t に関する DX 摘要にスコア関数 s を適用して得られる企業のスコア順の順位，$\widehat{c_{s,t}}$ は結合係数を表す．

Step4: 結合係数は，次式のように，ランキングの評価尺度 AUPR を目的関数としてこれを最大化する値を用いる．

$$\widehat{c_{s,t}} = \underset{c_{s,t}}{\operatorname{argmax}}\, AUPR(Score_{ens}(\boldsymbol{D}), y_{true})$$

Vol.2022-GN-115 No.40
Vol.2022-CDS-33 No.40
Vol.2022-DCC-30 No.40
2022/1/21

ここで, AUPR（Area Under the Precision-Recall Curve）は，引数のスコア関数$Score_{ens}(\boldsymbol{D}_t)$ によるランキング結果と，学習データ中の二値分類の教師ラベルy_{true} を用いて描かれる Precision-Recall 曲線下の面積を計算したものである．一般的に，分類システムの精度と再現率はトレードオフの関係にあるため，Precision-Recall 曲線は右下がりの曲線を描く．AUPR が高いほどシステムの予測精度が高いことを示す．AUPR を最大化するための学習データとしては，例えば調査レポート等で優れた DX 活動を行うと判定された（上位にランクされた）企業を正例，そうでない企業を負例として与える．

また，$\widehat{c_{s,t}}$ の推定には，グリッドサーチを適用する．その計算量を抑えるために, スコア関数$\{s\}$と質問タイプ$\{t\}$の最適化を交互に進めながら，漸近的に最適化を行う．具体的には次の漸化式の中で積の関係$c_{s,t} = \alpha_s\beta_t$ を近似的に仮定し，$l = 1,\cdots$ に対して交互反復的に結合係数$\widehat{\alpha_s}^{(l)}$ および $\widehat{\beta_t}^{(l)}$ を求める．

$$\widehat{\alpha_s}^{(l)} = \underset{\alpha_s}{\operatorname{argmax}}\, AUPR\left(\sum_{\{s,t\}} \frac{\alpha_s \widehat{\beta_t}^{(l-1)}}{rank\left(Score_s(\boldsymbol{D}_t)\right)}, y_{true}\right)$$

$$\widehat{\beta_t}^{(l)} = \underset{\beta_t}{\operatorname{argmax}}\, AUPR\left(\sum_{\{s,t\}} \frac{\widehat{\alpha_s}^{(l)} \beta_t}{rank\left(Score_s(\boldsymbol{D}_t)\right)}, y_{true}\right)$$

$\widehat{\alpha_s}^{(l)}$ と $\widehat{\beta_t}^{(l)}$ に対するスコア関数は次式で計算される．

$$Score_{ens}{}^{(l)}(\boldsymbol{D}) = \sum_{\{s,t\}} \frac{\widehat{\alpha_s}^{(l)} \widehat{\beta_t}^{(l)}}{rank\left(Score_s(\boldsymbol{D}_t)\right)}$$

ただし，$\widehat{\beta_t}^{(0)}$ は結合係数全て 1 を初期パラメータとした．

4.3 個別評価ルールの MQSE への対応

個別評価ルールは，MQSE では統合後のスコア関数の値に対して調整を適用する．具体的には前節の漸化式中の結合係数 $\widehat{\alpha_s}^{(l)}$ と $\widehat{\beta_t}^{(l)}$ の最適化に用いられる統合スコア $\sum_{\{s,t\}} \frac{\alpha_s \widehat{\beta_t}^{(l-1)}}{rank\left(Score_s(\boldsymbol{D}_t)\right)}$，$\sum_{\{s,t\}} \frac{\widehat{\alpha_s}^{(l)} \beta_t}{rank\left(Score_s(\boldsymbol{D}_t)\right)}$，および最適化後の統合スコア$Score_{ens}{}^{(l)}(\boldsymbol{D})$ の 3 つが補正対象となる．例えば，ランキング上位に特定業種の企業が集中するのを避けるためのルールとして，補正対象のいずれも，各スコアによる順位 r_{total} とコスト $cost(r_{seg})$ の和の逆数 $\frac{1}{r_{total}+cost(r_{seg})}$ を調整後の値とする．コスト関数$cost(r_{seg})$ には以下のヒンジ型のコスト関数を用いる．

$$cost(r_{seg}) = \begin{cases} a \cdot N(r_{seg} - n_{max}) & (r_{seg} > n_{max}) \\ 0 & (r_{seg} \le n_{max}) \end{cases}$$

ここで，Nは 464 であり，n_{max}とaはコスト関数のパラメータである．同一業種内で三位以下となった企業は，さらに順位が下がるように，ただしあまり下がりすぎないように，例えば $n_{max} = 3$，$a = 0.5$ として設定する．

5. 評価実験

5.1 実験方法

提案手法である MQSE を評価するため，MQE による WISDOM-DX の評価と同様の実験[8]を行う. 実験データは DX 銘柄 2021 選定の調査にエントリした企業 464 社（33 業種）を対象とした．この調査では DX 活動が活発な DX 先進企業として DX 銘柄(28 社）と DX 注目企業（20 社）の合計 48 社が評価委員によって選定された．この 48 社を上位 48 社とみなし，正解企業とした．企業 464 社のリストを WISDOM-DX の入力とし，出力結果である企業ランキングの上位 48 社に正解企業がどれくらい含まれるかを調べた．学習データは, 2020 年の DX 銘柄, 注目企業, および, 2015-2019 年の攻めの IT 銘柄として選定された企業を正例，それ以外の企業を負例とした．DX 優良事例としては，2015 年から 2020 年の経産省発行の DX 関連レポートに記載された攻めの IT 経営銘柄，DX 銘柄，DX 注目企業に関するテキストデータを用いた．ランキング上位に特定業種の企業が集中するのを避けるためスコアを補正する個別評価ルールを追加した．

ベースラインとして，Google 検索を用いた結果と比較する．具体的には，464 社について，企業名と“デジタルトランスフォーメーション”の 2 つのキーワードの AND 検索を Google Custom Search で実行し，検索件数の大きい順に 464 社のランキングを生成し，これをベースラインとした．評価尺度は，正例が少ないインバランスデータであることを考慮し，ランキング上位 48 社の精度，および AUPR を用いた．以降，精度とは上位 48 社の精度を表す．

5.2 評価結果

提案手法の MQSE の精度は 56.3%（上位 48 件中 27 件が正解）となり，AUPR は 0.541 であった．ベースラインの精度は 22.9%（上位 48 件中 11 件が正解）であり，AUPR は 0.181 であった. 提案手法とベースラインの Precision-Recall 曲線を図 2 に示す．いずれも提案システムがベースラインを上回る結果となった．

表 1 に，6 つの質問タイプと 8 つのスコア関数のそれぞれの AUPR，および，過去データから学習した 8 つのスコア関数ごとに 6 つの質問タイプのスコアを結合した MQE の AUPR，8 つのスコア関数のスコアを結合した MQSE の AUPR を示す．個別の AUPR は 0.303 から 0.434 の間になっており，MQE の AUPR は 0.503 から 0.543 の間，MQSE の AUPR は 0.541 となっている．MQSE の AUPR は最高値ではないがほぼ同等の値となっており，過去データからそれぞれのスコアをどれくらい重視して結合すべきかをうまく学習できていると言える．

Vol.2022-GN-115 No.40
Vol.2022-CDS-33 No.40
Vol.2022-DCC-30 No.40
2022/1/21

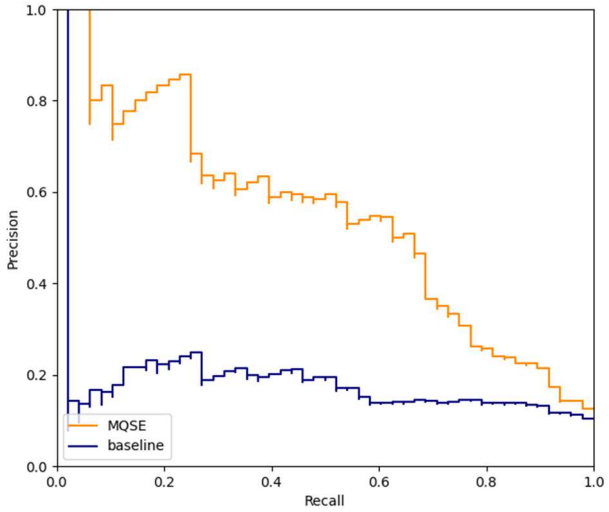

図 2　提案手法の Precision-Recall 曲線
Figure 2　Precision-Recall Curve of the proposed method

6.考察

6.1 DX 銘柄 2021 選定の調査での評価結果分析

DX 銘柄 2021 選定の調査で選定された 48 社に関する MQSE の出力結果を分析する．MQSE の出力したランキング上位 48 件の精度は 56.3%（27 件が正解）であった．不正解の原因について分析した結果，False Positive（上位 48 件中の不正解企業）の 21 件中 11 件は，1 次評価での機械的な選抜に用いられている ROE や選択式項目への企業の回答など，提案システムが利用していないデータに起因するものであり，残り 10 件は選定企業に準ずる一定以上の DX 推進活動を行っていることが分かった．専門家による実際の DX 先進企業の選定では，464 社が提出した選択式項目の回答および 3 年平均 ROE（自己資本利益率）が選考の中で考慮されているが，今回の実験では，あくまでも現状の提案システムの性能と課題の把握に主眼を置いているため，これらの情報は用いずに評価した．一方，False Negative（上位 48 件に含まれなかった正解企業）の 21 件中 13 件は，質問応答システムの応答数の不足に起因した，さらに，21 件中 2 件は，東証マザーズに上場して 3 年未満の若く先鋭的な企業であり，専門家による評価を介した事例のタイムリーな登録スキームが必要なことが分かった．

6.2 MQSE の多様な評価観点への適用性検討

提案システムにおいて，6 つの質問タイプと 8 つのスコア関数のそれぞれの AUPR は 0.303 から 0.434 の間に分布している（表 1）．Google 検索によるベースラインの AUPR は 0.181 であり，質問応答システムを用いる提案システムの方が，一般的な検索エンジンを用いるアプローチよりも，精度の良いランキングを生成できたと言える．これは，検索エンジンに比べて，質問応答システムが企業の DX 関連情報のみを精緻に収集できたためと考えられる．

MQE を用いた場合の AUPR は 0.503 から 0.543 の間に分布しており（表 1），用いなかった場合の AUPR の最高値 0.434 よりも高い．さらに，MQSE の AUPR は 0.541 となっていて，最高値 0.543 とほぼ同等の値である．過去データから様々な評価指標をどれくらい重視して評価すべきかを学習できていると言える．大規模調査においては評価者間でどの評価指標をどれくらい重視して評価するかのすり合わせが発生するが，その部分が MQSE によってカバーされていると考える．

東京証券取引所の国内上場会社約 3700 社を対象にした場合でも，アンケート未提出企業を含めて，従来の評価に準ずる形で DX 先進企業が自動的に推定できる可能性は高いと考えられる．今後，アンケート調査を行うことなく Web 情報を用いて大規模かつ客観的な調査・分析を実現するために，提案システムを改良していく．

7. おわりに

企業の DX 活動を自動的に分析・評価するシステム WISDOM-DX の開発において，汎用性を高めて様々な調査に活用できるように，様々な観点からの企業活動のスコアを統合する手法として，マルチクエリスコアアンサンブル（MQSE）を提案した．調査における評価基準が必ずしも明確になっていなくても，過去の調査結果から，様々な企業活動を Web での掲載情報量，優良事例との類似度，情報抽出の確からしさの観点からどれくらい重視すべきかを MQSE は学習できる．DX 銘柄 2021 選定の調査のデータを

スコア関数	質問タイプ						MQE	MQSE
	1	2	3	4	5	6		
cnt	0.376	0.414	0.363	0.387	0.378	0.420	0.517	
cnt_conf	0.400	0.405	0.319	0.433	0.410	0.404	0.527	
sim	0.396	0.403	0.366	0.397	0.388	0.423	0.527	
sim_conf	0.401	0.397	0.305	**0.434**	0.398	0.395	**0.543**	
sim_idf	0.395	0.404	0.369	0.395	0.386	0.425	0.529	
sim_idf_conf	0.398	0.404	0.306	**0.434**	0.399	0.398	0.520	
sim_tf_idf	0.384	0.402	0.353	0.383	0.380	0.411	**0.503**	
sim_tf_idf_conf	0.393	0.395	**0.303**	0.428	0.397	0.395	0.516	

表 1　464 社のランキングに対する AUPR
Table 1 AUPR values of ranked 464 companies

Vol.2022-GN-115 No.40
Vol.2022-CDS-33 No.40
Vol.2022-DCC-30 No.40
2022/1/21

用いて評価実験を行い, ベースラインとした Google 検索よりも性能が上回ることを確認した．また，MQSE は 464 社の上位 48 社に含まれる企業を 56.3%（27 件）正解した．

今後の課題として，質問応答システムの応答数不足解消のための Web ページ収集強化，DX 銘柄 2021 選定の調査のデータを用いた実験結果の企業ランキングに対して専門家による妥当性の検証が考えられる．また，DX 以外のドメインへの適用は重要な取り組みと考えている．組織のカーボンニュートラル，SDGs，セキュリティ，ダイバーシティなど多様なドメインにおいて，提案システムの有効性を検証していく．個別企業が競合他社や業界動向を調べるための評価・シミュレーションツールとしての活用についても検討を進める．

謝辞 経済産業省の平井裕秀氏と渡辺琢也氏，NICT の内元清貴氏と田仲正弘氏など関係者の皆様のご支援に厚く御礼申し上げます．

参考文献

[1] 奥村明俊:デジタルアーキテクチャデザイン特集 Society 5.0 の実現に向けた挑戦者へのエール，情報処理 Vol.62, No.6, pp. 284-287. (May 15, 2021)
[2] 首相官邸:デジタル社会の実現に向けた改革の基本方針, (Dec.25,2020) https://www.kantei.go.jp/jp/singi/it2/dgov/201225/siryou1.pdf
[3] 小林庸平:日本におけるエビデンスに基づく政策形成(EBPM)の現状と課題”，日本評価研究 Vol. 20, No.2, pp.33-48, (July 2020)
[4] 経済産業省:DX 推進ガイドライン. (Dec 12, 2018) https://www.meti.go.jp/press/2018/12/20181212004/20181212004.html
~~[5]~~ 町田佳世子:質的研究におけるテキストマイニング活用の利点と留意点, SCU Journal of Design & Nursing Vol. 13, No. 1, pp.47-53 (2019)
[6] 岡部大祐：計量的テキスト分析の基礎．田崎克也編，コミュニケーション研究のデータ解析．ナカニシヤ出版，京都，pp.189-201，(2015)
[7] 今井多樹子，高瀬美由紀，佐藤健一：質的データにおけるテキストマイニングを併用した混合分析法の有用性．日本看護研究学会雑誌 41(4), pp.685-700，(2018)
[8] 久寿居 大,石川 開,奥村 明俊,鳥澤 健太郎,大竹 清敬: WEB データに基づく質問応答システムを用いた企業のデジタルトランスフォーメーション活動の分析と評価,研究報告コンシューマ・デバイス＆システム,2021-CDS-32(15), (2021-08-26)
[9] 日経 BP 総合研究所: DX サーベイ 2 With コロナ時代の実態と課題分析， (Nov. 25, 2020) https://info.nikkeibp.co.jp/nxt/campaign/b/279660/
[10] 矢野経済研究所: 2020 デジタルトランスフォーメーション(DX)市場の現状と展望. (Jul. 30, 2020) https://www.yano.co.jp/press-release/show/press_id/2487
[11] IDC Japan 株式会社 国内企業のデジタルトランスフォーメーション動向調査結果を発表 (De.7 2020) https://www.idc.com/getdoc.jsp?containerId=prJPJ47071820
[12] 一般社団法人日本情報システム・ユーザー協会: 企業 IT 動向調査 2021． (Apr.28, 2021) https://juas.or.jp/library/research_rpt/it_trend/
[13] 一般社団法人情報サービス産業協会: 社会のデジタルトランスフォーメーション(DX)推進に貢献する情報サービス企業のあり方． (May 30, 2019) https://www.jisa.or.jp/publication/tabid/272/pdid/30-J007/Default.aspx
[14] 一般社団法人日本 CTO 協会:DX 動向調査レポート 2021 年度版． (Apr.12,2021) https://cto-a.org/news/2021/04/12/4956/
[15] 東京都: DX 推進実証実験プロジェクト(第 1 期) (Mar.19, 2021) https://www.metro.tokyo.lg.jp/tosei/hodohappyo/press/2021/03/19/05.html
[16] 神奈川県: 神奈川県 DX プロジェクト推進事業, (May 17, 2021) https://www.pref.kanagawa.jp/docs/sr4/dx-project.html
[17] 経済産業省:デジタルトランスフォーメーション銘柄（DX 銘柄）2021． (Jun.7, 2021) https://www.meti.go.jp/policy/it_policy/investment/keiei_meigara/dx-report2021.pdf
[18] 経済産業省: DX 銘柄／攻めの IT 経営銘柄. (June 11, 2021) https://www.meti.go.jp/policy/it_policy/investment/keiei_meigara/keiei_meigara.html
[19] 総務省: 令和 2 年度版情報通信白書 日本企業のデジタル・トランスフォーメーション推進に向けて https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r02/html/nd133440.html
[20] 奥村学監修：質問応答システム自然言語処理シリーズ 2，コロナ社．（Aug. 2009）
[21] 三原英理，藤井敦，石川徹也. World Wide Web を用いたヘルプデスク指向の質問応答システム，第 4 回情報科学技術フォーラム講演論文集, pp. 163-166, (Aug. 2005)
[22] 相濱佑介,土屋誠司,渡部広一:Wikipedia を情報源とした質問応答システムの検討，情報科学技術フォーラム講演論文集 18th, pp.165-166. (Aug.20 2019)
[23] 国立研究開発法人情報通信研究機構（NICT）: WISDOM X (ウィズダム エックス)とは？, https://www.wisdom-nict.jp/#top
[24] Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Ryu Iida, Masahiro Tanaka, and Julien Kloetzer.: A semi-supervised learning approach to why-question answering. In Proceedings of AAAI-16, pp. 3022–3029, (2016).
[25] Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, Istv´an Varga, Jong-Hoon Oh, and Yutaka Kidawara: Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In Proceedings of ACL 2014, pp. 987–997, (2014).
[26] John-Hoon Oh, Kazuma Kadowaki, Julien Kloetzer, Ryu Iida and Kentaro Torisawa: Open domain why-question answering with adversarial learning to encode answer texts. In Proceedings of ACL 2019, pp. 4227– 4237, (2019).
[27] Kazuma Kadowaki, Ryu Iida Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer: Event causality recognition exploiting multiple annotators' judgments and background knowledge. In Proceedings of EMNLP 2019, pp. 5820-5826, (2008).
[28] Honnibal M, Montani I: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, (2017).
[29] 浅原正幸，金山博，宮尾祐介，田中貴秋，大村舞，村脇有吾，松本裕治: Universal Dependencies 日本語コーパス，自然言語処理, vol. 26, no. 1, pp.3-36, (2019).
[30]Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. "Reciprocal rank fusion outperforms Condorcet and individual rank learning methods", SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 758-759, (Jul. 2009).
[31] Zhou, Z.: Ensemble Methods: Foundations and Algorithms, Machine Learning & Pattern Recognition Series, CRC Press, (2012).