様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 １９ 日現在

機関番号：３２６４２
研究種目：基盤研究(C)（一般）
研究期間：2018～2022
課題番号：１８Ｋ１１３１８

研究課題名（和文）コンテンツ指向のデータモデルライフサイクルを支援するデータベースの研究開発

研究課題名（英文）Research on Database Systems for Supporting Life Cycles of Data Models based based on conntents

研究代表者
中野　美由紀（Nakano, Miyuki）

津田塾大学・学芸学部・教授

研究者番号：３０２２７８６３

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：コンテンツを主体とした効率のよい「データモデルのライフサイクル」を支援するデータベース環境を確立するために、機械学習を用いたデータモデルのライフサイクルを典型的なデータ解析処理の上で評価した。オープンデータ（音楽DBやチェスログ等）や人工的データを用いて、データの時間推移による変化を表す特徴量等を抽出し、データドリフトが生じる際に保持すべきデータを削減しつつ、学習モデルの精度を担保、向上させられることを示し、「データモデルのライフサイクル」の支援の有効性について明らかにした。

研究成果の学術的意義や社会的意義
本研究では、コンテンツを主体とした効率のよい「データモデルのライフサイクル」を支援するデータベース環境を確立するために、データモデルのライフサイクルを典型的なデータ解析処理と具体的事例（オープンデータ）を用いてコンテンツ主体のデータ管理手法を設計した。
音楽配信サイトのデータと感情空間上にマッピングした音楽データベースの構築、感情空間上における音楽データベースの特徴量の有用性について解析を行った。また、機械学習コンテストで多く用いられる人工的なデータセットおよびオンラインチェスゲームのログと実データとして取り上げ、時間的な変化指標としてのモデル精度がデータ分布変化の指標として検討を行った。

研究成果の概要（英文）：In our data-driven society, it is highly expected that database systems support the "Life-cycle of data models" for the these days' efficient data analysis. Some typical machine learning algorithms with open data are considered as a first-step prototype. Concept drift data such as the music database, chess log, and synthetic datasets are investigated in order to extract indicators for a novel database function for supporting the data model life-cycle. Consequently, it is revealed that the contents oriented indicators are useful to keep the precision of service or models.

研究分野：情報学

キーワード：データ解析　データベース・システム　データ流通　データのライフサイクル　機械学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

現在、収集されるデータの蓄積と迅速なデータ解析は情報処理技術において不可欠な要素である。データの蓄積自体はデータベース技術による永続的な管理の枠組みが提供されているが、蓄積されたデータの解析は、データの特性、データ解析手法に加え、解析の利用目的という社会的な観点が必須ということもあり、必要に応じてデータ解析の専門家による個別の処理であることが多い。つまり、現存のデータ解析処理は蓄積されたデータの器であるデータベースと必ずしも連携していない。IoT 等の利用によりデータは量的にも質的にも時々刻々と変化している状況において、自律的にデータ解析を適宜に行う、すなわち、自律的な「データモデルのライフサイクル」（ここでは、情報基盤システムとして適切な予測・推定モデルを構築・維持すること）の適用は解決されておらず、データ解析処理を利用する様々な情報処理システムにおいて大きな課題となっている。

データ解析技術として、近年、機械学習、特に深層学習が大きく注目され、膨大なデータから教師データを生成、利用する技術、大規模データにおける高品質かつ効率のよい機械学習アルゴリズムは多く研究され、実用化されつつあるが、「データモデルのライフサイクル」に関してはようやく研究の端緒に至った状況にある。現在、音声認識、画像認識、機械翻訳等の様々な分野において大規模データを用いた機械学習（深層学習）の利用により飛躍的に認識率が向上することが示されている。その結果、多くの分野でデータ解析の重要性が認識されてきたが、データ解析モデルの構築は人の力が不可欠であり、近い将来数万人の規模でデータサイエンティストが不足するという指摘もある。このような状況において、蓄積されるデータの変化を量的、質的双方の観点から捉え、**「データモデルのライフサイクル」の効率化**が可能となるデータ解析環境の構築が必要不可欠である。

データの収納庫であるデータベースにおいて、データへのアクセス頻度など従来の情報のみならず、機械学習、深層学習における学習過程の情報、類似したデータで利用された解析手法の履歴等、データの蓄積に加え、新たなメタデータ情報を蓄積し、データ解析処理との連携が容易にできる環境が求められている。

２．研究の目的

本研究では、**コンテンツを主体とした効率のよい「データモデルのライフサイクル」を支援するデータベース環境の確立**を目的とする。データ解析では、主としてモデル構築のためのアルゴリズムの研究がその中心となっており、例えば、深層学習の Google の TensorFlow、PFN の Chainer に代表されるようにライブラリィとの実行環境の実現が大きく着目されている。また、統計解析パッケージにみられるようにデータ解析結果の視覚化も様々なツールが提供されるに至っている。一方で、いかにデータクレンジングが行われ、どのようなアルゴリズムを用いられ、どのような過程で特徴量の抽出や学習フェーズにおけるパラメタの調整が行われたか等の過程の多くは個人の経験にとどまり、共有する状況にはない。大規模データの解析においては、特徴量の抽出、深層学習における学習フェーズ等、適切な予測モデルを構築するために繰り返し行わなければならない処理は負荷が高い。本研究では、解析に利用されるデータに着目する。データ解析処理の過程は現在それを行う「人」に付随しているが、本来は対象となるデータと組みになる情報であると捉え、データ解析の過程をデータに紐づける。

３．研究の方法

本研究では、コンテンツを主体とした効率のよい「データモデルのライフサイクル」を支援するデータベース環境を確立するために、データモデルのライフサイクルを典型的なデータ解析処理（決定木、クラスタリング、深層学習等）と具体的事例（オープンデータ）を用いてコンテンツ主体のデータ管理手法を設計し、「データモデルのライフサイクル」の支援の有効性について具体的事例を基に検討する。

1)「データモデルのライフサイクル」および「知識発見プロセス」の過程からコンテンツと共にあるべきログ情報（クレンジングの過程、特徴量抽出や学習過程等）およびライブラリィ情報（処理アルゴリズム、最終の予測・推定モデル等）に関する検討を行い、2)データ解析処理と連携するために、コンテンツと共にあるべきデータ（ログ情報等）の形式およびライブラリィ情報を保持するためのモデルリポジトリ等のデータベース環境の設計をし、支援機能を付加したデータベース上にて、典型的なデータ解析処理（決定木、クラスタリング、深層学習等）と具体的応用例（コンテンツ推薦、データ選択、ユーザ判別等）を用いて、定性的および定量的な評価を行う。

４．研究成果

1）利用者に適応的な音楽推薦システムのための音楽データベースの構築と評価

近年、音楽配信サービスの普及により、インターネット上で数百万曲の楽曲にアクセスできるようになった。しかし、音楽は単なる娯楽ではなく、ショッピングセンターの BGM に始ま

り、音楽療法など、私たちの生活の中で様々な用途に利用されている。特定の利用シーンに基づき、複数の楽曲から１つの楽曲を効率的に選択することは困難である。そこで、高度な技術を使いこなすことが困難な高齢者に適した音楽推薦手法を提案する。感情空間にマッピングされた音楽データベースや、対話型ロボット・エージェントを導入し、適切な雰囲気やユーザの望む感情を実現する。また、高齢者の個人嗜好に合わせた適切な印象（感情）語を取得するための仕組みについて考察した。

音楽データベースは、曲の energy と valence によって表される感情空間上にマッピングされており、同じアーティストでも曲調によって感情空間上に広く分布する。さらに、異なるアーティストの楽曲群はそれぞれ、異なる分布を示す。図１は感情空間上に三人のアーティストの楽曲を緑、赤、青の点で示しており、同色の円はおおよそ 80%以上の楽曲を含む範囲を示している。この図から、ユーザの感情に合わせて楽曲を推薦する場合、アーティストごとの楽曲分布をあらかじめ情報として得なければならないことが明らかになった。

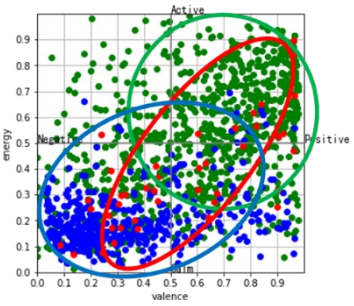

図 1　感情空間上の楽曲分布

また、人の感情における曖昧さに対する考慮も必要となる。複数のユーザによるアンケート結果と感情空間上における一般的な感情語について比較を行った。図２では、「楽しい」「明るい」等のキーワードの絶対値（図２の橙色円）とユーザが曲を聞いて感じる感情（橙色の感情に該当すると選んだ楽曲群の位置）を緑色の円で示している。図２から分かるように、ユーザの感情に合わせた楽曲推薦においては、あらかじめ得られた感情空間上の絶対値だけではなく、ユーザごとに異なる感情に対しての補正を行うことで楽曲推薦に対する精度が向上することが明らかになった。

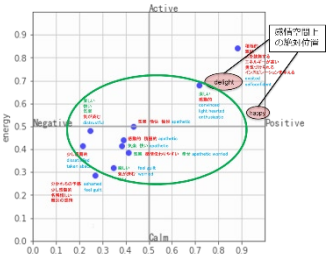

図 2　ユーザ感情の曖昧さ

2）コンセプトドリフト対処のための、Adversarial Validation を用いた学習データ選択手法

機械学習モデルの利活用が進み、あるタスクを行うモデルが長期的に利用されるシナリオが想定されるようになった。しかし一度学習を行い、良い精度を得られたモデルであっても、使い続けているうちに精度が低下していくことがある。そのような現象の原因として、コンセプトドリフトが知られている。コンセプトドリフトは必ずしも時間経過によって不可逆的に悪化するものではなく、古いデータの一部は未来の予測にとっても価値のあるデータである。

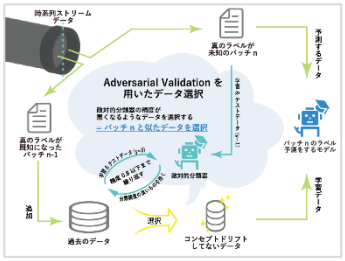

図 3　データ選択手法アルゴリズム

ここでは、大規模ストリームデータにおいて、時間経過などによりコンセプトドリフトが発生する場面で、特定のタスクを行う教師あり機械学習モデルを継続的に入手し、自動的に時系列データの予測を行うアルゴリズムに関する提案（図３）を行った。本研究のメリットとして、過去のデータから予測したいバッチに似たデータを選んで学習データとすることで、過去の有用なデータを活用できる点が挙げられる。Adversarial Validation を用いて特徴量選択を行う方法を含む 3 通りのコンセプトドリフト適応手法に適用し、比較・検討し、学習において十分な精度を保つために過去の有用なデータのみを利用するための指標を明らかにした。表１に示すように、有用と思われるデータの選択により精度がデータ選択を行わない場合と比較して向上していることが確認できた。

表 1 評価結果

exit condition of selection loop(AUC)	50%	60%	70%	no data selection
Prediction performance (AUC)	80.75%	80.54%	77.20%	77.20%
Num of using data (/10,000)	8890.00	8945.67	10000.00	10000.00

3）オンラインチェスログを用いたユーザランクの推定

オンラインチェスサーバ Lichess では過去の棋譜が数多く公開され、オープントーナメントには 4000 人の参加者がいる。数千人の参加者がいる大会などで対戦を調整するための自動化は必須である。また、Lichess 参加者のレーティングは試合結果、新しい参加者などにより常に変動し、データドリフトの代表的なデータの一つである（図４）。棋譜の分析を行い、レーティングのない選手に対し、適切なレーティングを決定できる高速判別器の生成をおこなった。

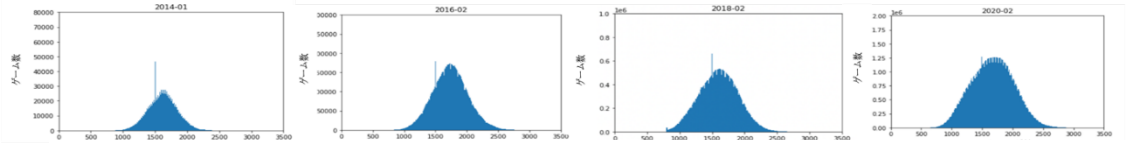

図 4　Lichess のレーティング値分布（2014、 2016、 2018、 2020 年の分布）

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計25件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 韓 語佳，中野 美由紀，小口 正人
２．発表標題 利用者の印象に基づく音楽レコメンドサービス－適切な印象語を得るためのユーザインタフェースの考察－
３．学会等名 ルチメディア，分散，協調とモバイル（DICOMO2021）シンポジウム
４．発表年 2021年

１．発表者名 YuJia Han, Miyuki Nakano, and Masato Oguchi
２．発表標題 Music recommendation service based on user impressions: Study of user interface acquiring appropriate impression words
３．学会等名 the 15th IEEE International Conference on Ubiquitous Information Management and Communication (IMCOM2022)（国際学会）
４．発表年 2022年

１．発表者名 韓 語佳，中野 美由紀，小口 正人
２．発表標題 利用者の印象に基づく高齢者向け音楽推薦システム「元気フクロウ」－嗜好に合わせた適応的推薦手法の提案－
３．学会等名 第 13 回データ工学と情報マネジメントに関するフォーラム（DEIM2022）
４．発表年 2022年

１．発表者名 韓 語佳，中野 美由紀，小口 正人
２．発表標題 利用者の印象に基づく音楽推薦手法の研究
３．学会等名 情報処理学会第 84 回全国大会
４．発表年 2022年

１．発表者名
今野由麻 , 中野美由紀, 小口正人

２．発表標題
コンセプトドリフト対処のための、Adversarial Validationを用いた学習データ選択に関する検討

３．学会等名
第 13 回データ工学と情報マネジメントに関するフォーラム (DEIM2022)

４．発表年
2022年

１．発表者名
今野由麻, 中野美由紀, 小口正人

２．発表標題
コンセプトドリフト対処のための、Adversarial Validationを用いた学習データ選択に関する考察

３．学会等名
情報処理学会第 84 回全国大会

４．発表年
2022年

１．発表者名
山田　飛, 小口正人, 中野美由紀

２．発表標題
プロセスマイニングを用いたチェスレーティングシステムにおけるプロセスウィンドウの変化とその評価

３．学会等名
情報処理学会第 84 回全国大会

４．発表年
2022年

１．発表者名
野元麻由, 石田紗弓, 小林千尋, 佐藤沙耶, 西島　菫, 中野美由紀

２．発表標題
大学生のための新しい生活様式におけるコミュニティ生成サポートアプリ ivy : 趣味に基づくマッチング方式とその評価

３．学会等名
情報処理学会第 84 回全国大会

４．発表年
2022年

１．発表者名
鈴木会子，松井美結，浦川紗貴，中野美由紀

２．発表標題
レビュー投稿における不快表現の通知機能システム

３．学会等名
情報処理学会第 84 回全国大会

４．発表年
2022年

１．発表者名
小松麻子，今野由麻，平山理美子，中野美由紀

２．発表標題
既存予定を生かした効率的なスケジュール推薦システム　- 時間と位置によるスケジュール提案と嗜好に合わせた推薦機能-

３．学会等名
情報処理学会第８３回全国大会，７L-03

４．発表年
2021年

１．発表者名
韓　語佳，中野美由紀，小口正人

２．発表標題
利用者の印象に基づく音楽レコメンドサービス-音楽の感情空間における文化的要因の影響に関する考察-

３．学会等名
情報処理学会第８３回全国大会，１P-05

４．発表年
2021年

１．発表者名
韓　語佳，中野美由紀，小口正人

２．発表標題
利用者の印象に基づく音楽レコメンドサービス

３．学会等名
DEIM2021，D25-3

４．発表年
2021年

１．発表者名
河澄菜々夏，坂口璃実佳，西村奈那子，湯浅　郁，中野美由紀

２．発表標題
複数路線を持つ大規模駅の待ち合わせ場所推薦システムの提案

３．学会等名
情報処理学会第８２回全国大会，5N-2

４．発表年
2020年

１．発表者名
徳丸瑞季，岡村奈々花，川村華峰，仲山友海，中野美由紀

２．発表標題
ルールベースに基づくビジネスシーンにおける敬語変換手法の検討 ○

３．学会等名
情報処理学会第８２回全国大会，７S-7

４．発表年
2020年

１．発表者名
Habuki Yamada; Nobuko Kishi; Masato Oguchi; Miyuki Nakano

２．発表標題
A Method for Estimating Online Chess Game Player Ratings with Decision Tree

３．学会等名
2023 IEEE International Conference on Big Data and Smart Computing (BigComp), DOI =10.1109/BigComp57234.2023.00066, 2023.2（国際学会）

４．発表年
2023年

１．発表者名
Yuma Konno, Miyuki Nakano, and Masato Oguchi

２．発表標題
Efficient Data Selection Indicators for Updating Models under Data Drifted Environment

３．学会等名
ig Data 2022, Volume: 1, Pages: 6724-6726, 2022.Dec.（国際学会）

４．発表年
2022年

１．発表者名
山田 飛，小口 正人，中野 美由紀

２．発表標題
オンラインチェスログを用いたチェスプレイヤーランキングの推定に関する考察

３．学会等名
FIT 2022，F002，2022.9

４．発表年
2022年

１．発表者名
今野 由麻 ，中野 美由紀 ，小口 正人

２．発表標題
コンセプトドリフト対処のための，Adversarial Validationを用いた学習データ選択アルゴリズムの方式検討

３．学会等名
DICOMO 2022，pp.28-35，1C-1，2022.6

４．発表年
2022年

１．発表者名
今野 由麻，中野 美由紀，小口 正人

２．発表標題
データドリフト対処のためのAdversarial Validationを用いたデータ選択手法の評価

３．学会等名
DEIM 2023，1-a2-1，2023.3

４．発表年
2023年

１．発表者名
今野由麻，中野美由紀，小口正人

２．発表標題
データドリフト対処のためのAdversarial Validationを用いたデータ選択指標の評価

３．学会等名
情報処理学会第８５回全国大会，6q-06，2023.3

４．発表年
2023年

１．発表者名
山田　飛，小口正人，中野美由紀

２．発表標題
オンラインチェスにおけるRNNを用いたチェスプレイヤーランキング判別方式とその評価

３．学会等名
情報処理学会第８５回全国大会，2N-03，2023.3

４．発表年
2023年

１．発表者名
服部まどか，小松夏紀，馬場美羽，外園奈那，中野美由紀

２．発表標題
学生生活を快適に過ごせる大学情報アプリ「てtsudaっぷ。」の開発と評価

３．学会等名
情報処理学会第８５回全国大会，5N-08，2023.3

４．発表年
2023年

１．発表者名
曽田円香，志風美雨，辻愛美紗，中野美由紀

２．発表標題
Spotify音楽データを用いたユーザの感情に基づく音楽推薦手法の提案

３．学会等名
情報処理学会第８５回全国大会，6N-07，2023.3

４．発表年
2023年

１．発表者名
中野美由紀，石山悟志，君野史明，小久保勇気，斉藤大介，鈴木智尚，宮崎晃一，吉村慎祐

２．発表標題
利用者の印象に基づく音楽レコメンドサービス： 画像を利用した利用者指向の音楽レコメンド手法の考察

３．学会等名
情報処理学会2019年度第81回全国大会，1C-04，2019.3

４．発表年
2019年

１．発表者名 中野美由紀，石山悟志，君野史明，小久保勇気，斉藤大介，鈴木智尚，宮崎晃一，吉村慎祐†
２．発表標題 画像選択インタフェースを用いた音楽推薦サービス”Diggin”：画像選択における感情空間と印象値に関する考察
３．学会等名 電子情報通信学会2019年度総合大会
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関