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　６ 年　 ６ 月 ２０ 日現在

機関番号：３２６７８
研究種目：研究活動スタート支援
研究期間：2022～2023
課題番号：２２Ｋ２１２８８

研究課題名（和文）メタデータのテキスト情報を利用したデータサイエンス自動化プラットホームの開発

研究課題名（英文）Development of a platform for automatic data science by using text information in metadata

研究代表者
増田　聡（Masuda, Satoshi）

東京都市大学・メディア情報学部・教授

研究者番号：６０９４７９２７

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：本研究では，データ項目名のテキスト情報から特徴量抽出を自動化する新たなアプローチを取った．具体的には，既存のデータサイエンスにおけるデータ項目名およびソースコードに対して，自然言語処理やソースコード分析技術を利用し，特にdatetime特徴量に着目した知識データベースを作成した．さらに，その知識データベースを利用し，新たに与えられるテキスト情報からdatetime特徴量を推薦するシステムを開発した．また，単語ベクトル化をone-hotベクトルや単語埋め込みの手法を用いて精度の向上を図った．実験では，その知識データベースの分類精度を確認し，予測実タスクに適用し予測精度の向上を確認した．

研究成果の学術的意義や社会的意義
膨大なデータから新たな知見を得る分析はデータサイエンスと呼ばれ，その普及が推進されている．データの特徴量を抽出する作業は，特徴量エンジニアリングと呼ばれ，データサイエンスの作業ステップの一つである．現在，特徴量エンジニアリングの作業は，エキスパートの経験に頼っているため，その作業の自動化の研究が行われている．本研究は，テキスト情報からdatetime特徴量を推薦する方法を提案し，システムを開発し，有効性を確認した．これにより，自動特徴量エンジニアリングの学術的領域に貢献した．

研究成果の概要（英文）：In this research, we took a new approach to automate feature extraction from textual information of data item names. Specifically, we created a knowledge database focusing on time series (datetime) features by using natural language processing and source code analysis techniques for data item names and source codes in existing data science. Furthermore, we developed a system that recommends datetime features from newly provided text information using the knowledge database. For the feature recommendation mechanism, we improved the accuracy of word vectorization by using one-hot vector and word embedding methods. In experiments, we confirmed the classification accuracy of the knowledge database and applied it to actual forecasting tasks, such as house price forecasting, to confirm the improvement in forecasting accuracy.

研究分野：ソフトウェア工学

キーワード：データサイエンス　自動特徴量エンジニアリング　自然言語処理　時系列特徴量

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

データを分析することにより有益な知見を引き出すことはデータサイエンスと呼ばれ，現在，社会で多くの取り組みがあり，研究開発も盛んに行われている．

データサイエンスの作業ステップは一般的に(1)データの前処理，(2)特徴量の抽出，(3)機械学習モデルの適用からなっている．これら作業ステップの作業内容はそれぞれ下記となっている．例として，「既存の住宅の取引データから，ある住宅の将来の取引価格を予測すること」を挙げている．

(1) データの前処理：データが与えられ，データの欠損処理や正規化などを行う．例えば，住宅取引データとして物件名(Residence)，取引月日(Date)，価格(Price)などのデータが与えられる．しかし，データは完全ではなく欠けている部分もあるのでそれらの削除または補完を行う．

(2) 特徴量の抽出：分析精度を上げるため，データの特徴を抽出する．この一連の作業を特徴量エンジニアリングと言う．これは，与えられたデータだけでなくデータの特徴を追加すると予測精度が向上することが知られているからである．例えば，取引月日から月名という特徴を抽出し，データ項目名に追加する．現在，特徴量の抽出は，エキスパートがデータ分析を繰り返すことによって行われており，専門的な経験が必要とされている．

(3) 機械学習モデルの適用：回帰モデルや決定木などの機械学習モデルを適用し，予測を行う．例えば，ある広さと間取りの住宅の将来の取引価格を予測する．

各作業を容易にするよう，各作業ステップで自動化の研究が行われている．この作業ステップの中で，特徴量の抽出は経験が必要とされているため，より自動化が望まれている領域である．現在，特徴量抽出の自動化は，膨大な量のデータの数値情報に着目し，数値の相関関係を分析することにより，特徴量を抽出することが行われている．

この特徴量抽出の結果として，データサイエンスで使われている様々なソースコードには，「住宅価格なら月単位」,「コンビニエンスストアの売り上げならば時間単位」などが常套手段として見られる．本研究の開始当初は，この特徴量抽出の常套手段をテキスト情報から再利用できれば，わざわざ膨大な量のデータの数値の相関を調べなくとも特徴量の抽出が容易になると考えた．

２．研究の目的

本研究の目的は，データサイエンスにおけるデータのテキスト情報から特徴量を自動抽出する技術の開発である．この特徴量の抽出の自動化という課題に対するアプローチとして，従来はデータの数値に対する相関分析の技術であったものを，データのテキスト情報に対する自然言語処理の技術とする点が本研究の独自性である．テキスト情報であれば，膨大なデータを分析することなく，軽量な処理でより容易に特徴量を抽出することが可能である．また，自然言語処理の技術を適用するにあたり，単語ベクトル化の手法を用いて新たな仕組みを創造する．

３．研究の方法

特徴量の抽出の作業ステップに着目し，構造化データの項目名のテキスト情報から特徴量の推薦システムの研究開発を行う．特に datetime の特徴量を取り上げ，テキスト情報から datetime 特徴量の推薦システムを提案した．単語のベクトル化方法を，先行研究では one-hot ベクトルを用いていたところを，本論文では単語埋め込み(word　embedding)を用いて精度や効果の改善を目指した．

テキスト分析では自然言語処理を利用し，単語のベクトル化またはルール化の研究が広く行われている．これらの自然言語を対象とした研究はデータの項目名や説明文であるデータ記述などに応用が可能である．例えば，上記のステップ(2)の特徴量の抽出の例では，「住宅取引は季節性があり前年同月比を抽出すると良い」という分析結果および知識に基づくものである．本研

究では，これらの知識を既存のデータ記述やソースコードの情報から抽出し再利用する技術を開発している．

図 1 は本論文が対象とするメタデータのテキスト情報から特徴量抽出を行う例を示し，本研究のアプローチを表している．

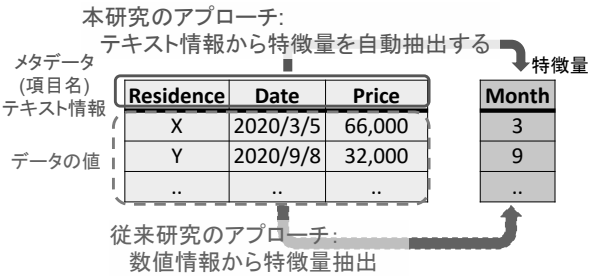

図１．本研究のアプローチ

図２は，本研究のアプローチを具体的な例を用いて示している．提案する仕組みでは，まず，データサイエンスに関連する可能性のあるソースコードは，Github や Kaggle などのソースコードリポジトリに数千万本以上存在している．これらの大量のソースコードから，構造化データのデータ項目名と datetime 特徴量が追加されている命令文を抽出し，データ項目名と datetime 特徴量名から成る「datetime 知識データベース」(DateTime Knowledge dataBase，DTKB)を作成する．次に，新たに与えられるデータの項目名を入力し，DTKB に対するテキスト類似度から datetime 特徴量の候補を出力する．このように，DTKB を利用し，新たに与えられるデータの項目名に対しても datetime 特徴量の候補を推薦するシステムである．

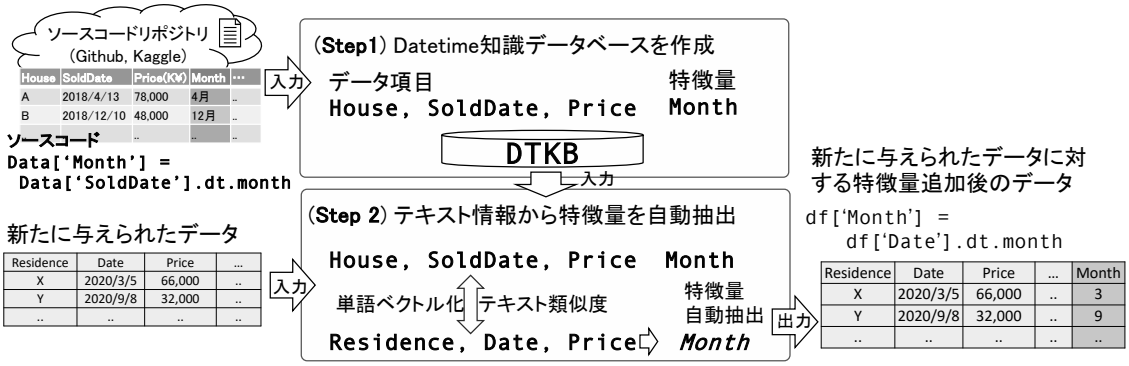

図２．テキスト情報を利用した特徴量抽出の自動化の仕組み

４．研究成果

図３は，システムの出力例を示している．データ項目名：incident_datetime，incident_cord_x，incident_cord_y，num_victims location_type，Crime_Type を入力として，提案手法により DTKB を利用して detetime API を推薦している．これは，与えられたデータ項目名と DTKB 内のテキストの類似度を Score として値を出力していて，year 0.32343 などのように出力される．

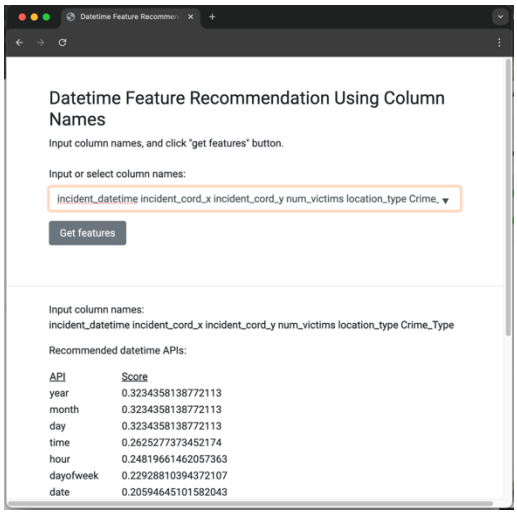

図３．提案手法を用いたシステムによる datetime API の推薦例

表１は，予測実タスクに対して，各比較実験において順に datetime 特徴量を追加していく様子を示している．元データの予測の予測精度は 0.4890 である．all feat.は，元データの datetime カラムから 26 個の datetime 特徴量を全て追加したデータを用いた予測精度が 0.4843 であり元データの予測精度から 0.9756%下がっている．random_1 は表の 26 個の API からランダムに 7 個を抽出したものを順にひとつずつ追加して予測精度の変化を確認している．random_1 の#(1)の行は元データに is_month_end を特徴量として追加したデータを用いて予測した結果，予測精度が 0.4873 で元データの予測精度から 0.3462%下がっている．random_1 の#(2)の行で

は#(1)のデータに is_year_start を追加している．つまり，元データから is_month_end と is_year_start の特徴量が追加されている．このデータを用いて予測精度を確認している．random_1 の#(2)の場合，予測精度が 0.4875 で元データの予測精度から 0.3226%下がっている．以下同様に 7 個の特徴量を順に追加して予測精度の変化を確認している．popular は DTKB 内で出現回数の多い datetime API から 7 個を順に追加していき予測精度の変化を確認している．popular の中では#(4)で予測精度が 5.3659%向上し最も予測精度が良い結果となっている．one-hot ベクトル化手法を適用した，prev. work(one-hot)は，図３で出力された API を Confidence の大きい datetime API から 7 個を選択し，1 番目から順に元データの datetime データに対する特徴量を作成し追加している．その追加されたデータを用いて予測実タスクを実行し予測精度の変化を確認している．prev. work (one-hot)の中では，#(4)が予測精度が 5.3895%向上し最も予測精度が良い結果となっている．本論文で採用した word embedding によるベクトル化手法を適用した，proposes (word embed.)も同様に，datetime API から 7 個を選択し，1 番目から順に元データの datetime データに対する特徴量を作成し追加している．proposes (word embed.)の中では，#(5)が予測精度が 4.740%向上し最も予測精度が良い結果となっている．

表２は実際の予測タスクへの適用実験の実験結果である予測精度の向上率を示している．表２は表１に示した実験を表 A-1 に No.1 から 10 まで 10 個の予測実タスクに対して行い，7 個ずつ追加していく中での最大値をそれぞれ挙げたものである．No.の行の中で最大値を太文字下線で示している．また，最下段に平均値を示している．本論文で採用した word embedding によるベクトル化は proposed (word embed.)は，精度の向上率が比較対象の中で最大となる予測タスクは，prev. work (one-hot)と同じ値であるが 1 個であった．prev. work (one-hot)が 10 個の実予測タスクの中で 5 個のタスクで最も最大値の多い方法であり，平均値でも最大である．次に popular が実予測タスクの中で 3 個のタスクで最も最大値の多い方法であり，平均値でも２番目に大きい方法である．

以上のように，本研究は，テキスト情報から datetime 特徴量を推薦する方法を提案し，システムを開発し，有効性を確認した．

表 1. 表 A-1 の No.1 実タスクにおける予測精度の向上率

orig.	all feat.			random_1			popular			prev. work (one-hot)			proposed (word embed.)		
acc	acc	gain(%)	#	added	acc	gain(%)	added	acc	gain(%)	added	acc	gain(%)	added	acc	gain(%)
			(1)	is_month_end	0.4873	-0.3462	month	0.4873	-0.3619	year	0.4885	-0.1180	year	0.4885	-0.1180
			(2)	(1)+is_year_start	0.4875	**-0.3226**	(1)+year	0.4860	-0.6136	(1)+month	0.4867	-0.4798	dayofweek	0.4934	0.8970
			(3)	(2)+daysinmonth	0.4825	-1.3375	(2)+hour	0.5056	3.3911	(2)+day	0.4882	-0.1651	weekofyear	0.4904	0.2830
0.4890	0.4843	**-0.9756**	(4)	(3)+quarter	0.4836	-1.1172	(3)+dayofweek	0.5153	**5.3659**	(3)+time	0.5154	**5.3895**	month	0.4927	0.7400
			(5)	(4)+day_of_year	0.4815	-1.5420	(4)+day	0.5117	4.6264	(4)+hour	0.5068	3.6350	hour	0.5122	**4.7440**
			(6)	(5)+weekofyear	0.4800	-1.8410	(5)+date	0.5073	3.7451	(5)+dayofweek	0.5099	4.2723	day	0.509	4.0830
			(7)	(6)+is_year_end	0.4779	-2.2737	(6)+weekday	0.5061	3.4855	(6)+date	0.5060	3.4776	date	0.5047	3.2020

表２．表 A-1 の各タスクにおける予測精度の向上率 (%)

No.	all feat.	random	popular	prev. work (one-hot)	proposed (word embed.)
1	-0.9756	0.8760	5.3659	**5.3895**	4.7443
2	-11.1060	7.0240	**9.5923**	6.1118	7.7988
3	-33.3312	-9.6112	18.0845	**21.3068**	16.0899
4	-43.8899	**-1.0839**	-5.3964	-5.3964	-5.3964
5	**8.9731**	3.5948	7.1785	4.5032	4.4865
6	-9.4483	2.6336	2.0529	**5.6222**	**5.6222**
7	12.8099	14.2106	13.5103	**14.5761**	13.4304
8	-36.9133	7.8031	**8.0478**	6.0179	6.0179
9	-61.0312	-0.2954	**0.8772**	0.3697	0.3697
10	-47.3690	-1.3943	-10.2844	**-0.8669**	-1.221687
Average	-22.2282	2.3757	4.9029	**5.7634**	5.1942

表 A-1. 予測精度の確認に使用したタスク

No.	Name	URL
1	Atlanta Crime	https://github.com/siddhantmaharana/atlanta-crime-prediction/blob/master/Feature_Engineering.ipynb
2	LearnX Sales Forecasting	https://github.com/romanlents/Bain-And-Company-Hackathon-2019/blob/master/Bain%20Hackathon%202019.ipynb
3	Bikesharing	https://www.kaggle.com/fatmakursun/bike-sharing-feature-engineering/comments
4	CoronaTracker	https://github.com/docligot/CoronaTracker-simulteModel/blob/master/baseline-regression/Baseline_RidgeRegression.ipynb
5	Crime_FIR	https://github.com/CrimePrediction/blob/master/preTrainedModels.ipynb
6	King County House Price	https://github.com/anupjsebastian/King_County_House_Price_Prediction/blob/master/Redfin_Model_Train.ipynb
7	Sales and Marketing Analytics	https://github.com/blurred-machine/Sales-and-Marketing-Analytics/blob/master/6-%20Predicting%20Sales/Sales%20Prediction.ipynb
8	World-cup predictions	https://github.com/itsmuriuki/FIFA-2018-World-cup-predictions/blob/master/Predicting%20Fifa%202018.ipynb
9	rossmann sales	https://github.com/gabeferrari/rossmann_sales_prediction_project/blob/main/m07_v01_store_sales_prediction.ipynb
10	stock-market prediction	https://github.com/AndersonJo/stock-market-prediction/blob/master/03%20Simple%20S%26P%20500%20Prediction%20%28More%20Stable%29.ipynb

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 増田 聡, 武田 友宏
２．発表標題 データ項目名を利用した時系列特徴量の推薦システム
３．学会等名 電子情報通信学会 知能ソフトウェア工学研究会(KBSE)
４．発表年 2024年

１．発表者名 Satoshi Masuda, Takaaki Tateishi, Toshihiro Takahashi
２．発表標題 Datetime Feature Recommendation Using Textual Information
３．学会等名 27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2023)（国際学会）
４．発表年 2023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