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　３ 年　 ６ 月 １７ 日現在

機関番号：１３９０１
研究種目：若手研究
研究期間：2018～2020
課題番号：１８Ｋ１８０５６

研究課題名（和文）大規模データ分析のための多視点分析管理システムの研究開発

研究課題名（英文）Multi-dimensional Analysis and Management for Large and Various Data

研究代表者
駒水　孝裕（Komamizu, Takahiro）

名古屋大学・情報基盤センター・助教

研究者番号：３０７５６３６７

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：オープンデータ化が推進される中で，公開されたデータを以下に活用するかが未だに課題である．本研究では，複雑に構造化されたデータを効率的に検索する技術，独立に作成されたデータを横断的に扱うために統合する技術を開発した．これらにより，活用のための分析技術を用いるためのデータの抽出が可能となった．また，関連する情報を結びつけることでより高度で精緻な分析が可能となった．

研究成果の学術的意義や社会的意義
オープンデータやデジタルトランスフォーメーションが進行している現状において，デジタル化・オープン化したデータを活用することは重要である．一方で，データを作成する組織は別々であることも多く，横断的な活用には障害が残る．本研究では，異なる組織が公開したデータの関連性に基づいたデータ統合や複雑化したデータから必要な情報を一般的な検索方法を用いて検索できるようにした．これらは今後のオープンデータ活用における基礎的な技術である

研究成果の概要（英文）：In the digital transformation era, utilizing open data for various applications is an important issue. In this research, two techniques are developed; one is efficient search entities from graph-structured data consisting of relationships between data, and the other is an integration technique for data published by different organizations. On the basis of these techniques, efficient extracting data of interest becomes possible, and precise data analysis using data of multiple granularity.

研究分野：データベース

キーワード：Linked Open Data　データ統合　データ分類　検索　不均衡データ分類

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

蓄積されたデータを多角的に分析する技術に多視点分析技術がある．多視点分析では，データ分析者が分析のための基準を属性の組合せでもって与えることで，データを分析する．データ分析者は，データについての仮説に基づき分析のための基準を設定する．その分析結果から仮説を評価し，必要に応じて仮説と分析基準の変更を行う．この作業を繰り返すことで，データ分析者は大量のデータから知見を得ることができる．

大規模なデータを対象とした場合，データ分析のための仮説を立て検証するプロセスを効率化することは重要な課題のひとつである．データが多様な属性を持つことでデータの表現力が上昇し，より精確な状態を記録できるようになった．そのため，データを分析する際に設定する基準の可能性が膨大になり，人手で“良い”分析基準を設定することが難しくなっている．本来，データ分析自体は目的ではなく，分析結果を経営戦略や施策に活かすことが目的である．故に，データ分析にかける労力は極力減らすべきである．

多視点分析の結果についての要因解析をサポートすることで要因解析にかかる労力の削減が期待される．多視点分析が自動化されるにつれ，要因分析を行いたい分析結果の数も増えることが予想される．よって，要因分析をサポートするシステムの需要が見込まれる．このシステムを開発するために，本研究では上記問いに対する解を得るための研究開発を行う．

２．研究の目的

本研究は，多視点分析の結果に対して実世界イベントを紐付け，要因分析を行うためのシステムを構築することで，分析結果に対する要因分析を支援することを目的とする．

電子商取引の活発化やセンサーデータ取得の容易化により，大量のデータが取得可能となった．取得したデータを活用したサービスの向上や環境分析は重要なデータ活用方法である．大量のデータを活用するための分析を従来の技法に頼ると依然として労力が掛かるため，目的を達成するのに時間がかかる．本研究の目的が達成されることで，(1) 分析にかかる労力と (2) 分析結果の要因分析に要する労力の削減が期待できる．

(1) 本研究は最新研究に倣い，多視点分析の自動化を取り入れる．これにより，分析者との対話で行う分析に比べ，分析にかかる労力を抑えられる．最新研究にはまだまだ効率化の可能性があり，本研究ではさらなる効率化を達成することで，データ分析にかかる労力の低減を実現する．

(2) 本研究の最大のポイントは分析結果に対する要因分析を補助する機能である．これまでの分析者は，背景となる事象を推測・調査しその要因分析を行ってきた．一般的に，分析結果の背景事象の推測や調査には膨大な時間がかかる．これに対して，本研究はニュースや SNS などの記事から関連する情報を探し出し，分析者に提示することで分析結果に対する背景や事象を把握する手助けとする．その結果，分析者の推察や調査に関わる時間を低減する．

３．研究の方法

データ分析の自動化を実際に運用するためには，大規模データを高速に処理をすることが重要である．その際に課題となるのは，取りうる分析基準の列挙や各分析基準について分布の傾向の評価を行う時に取りうる組合せ数が膨大になることである．既存手法では，同じ値を二度計算しないためのデータ構造や計算順序を提案することにより，無駄な計算を省いている．しかしながら，この方法は大規模データに対応するためには不十分で，さらなる効率化のためにはアルゴリズムの効率化や並列化による高速化が必要である．

従来のエンティティリンキングは知識ベース上での近接性と語が出現した文における文脈を利用して紐付けるべきエンティティを決定する．しかしながら，従来法では同じ語で表される複数のエンティティがあった場合に，より著名なエンティティに紐付いてしまう傾向にある．この問題に対して，既存手法ではエンティティの流行り度合いを Entity Recency 指標で評価することで対処した．本研究では，このアイデアを取り入れたエンティティリンキングを用いる．Entity Recency を用いたエンティティリンキング手法は，データの時系列性に依存するため，処理の分散が難しく処理効率が悪くなる．本研究では膨大な量のニュースなどを想定するため，効率化が不可欠である．そのために，Entity Recency を導入したエンティティリンキングの並列化について取り組む．

得られた分析結果とエンティティリンキングされたリソースをもとに，分析結果を裏付ける実世界イベントを検出する．本研究の対象とするイベント検出では，分析結果を裏付ける事柄の組合せを検出する．例えば，SUV の売上が年々増加している，という分析結果に対して，SUV をプロモートするような事柄（人気メーカが高機能な SUV を販売した，など）を年ごとに検出する．従来のイベント検出を用いる場合，各々のイベントを予め発見し，後に関連付けることになる．この方法の問題点としては，　先のイベント検出時に適当なイベントを検出できない可能性

があること，および検出されるイベントの数が多くなり全体の処理に時間がかかることである．本研究が取り組む課題は，分析結果に適したイベントをピンポイントで検出し，効率的に分析結果に統合することである．

4. 研究成果

エンティティリンキングおよび紐付けたデータを活用するための基礎技術として，紐付けられたデータに対する検索技術を開発した．特に，一般的に使いやすい検索を実現するために，キーワード検索を用いて，ユーザの興味のあるエンティティを発見する手法を開発した．[論文 2，学会発表 10，11，12]

実データへの応用として，日本法令を対象として Linked Open Data (LOD) を構築し，関連データとの紐付け技術を開発した．さらに，LOD と文書データが統合されたデータにおいて，文書間の関連付けや LOD と文書に対する同時検索に関する技術を開発した．加えて，ユーザインターフェースとして，VR 空間で提示することで複雑なデータにおける探索的な検索をサポートするシステムを開発した．[学会発表 1，3，4，6，7，9]

エンティティリンキングにおいて，紐付ける情報の不均衡性が性能低下の原因となるため，不均衡データに対して頑健な手法のための基礎技術を開発した．[論文 1，学会発表 2，5，8]

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　1件／うちオープンアクセス　0件）

１．著者名 Takahiro Komamizu	４．巻 0
２．論文標題 Random walk-based entity representation learning and re-ranking for entity search	５．発行年 2020年
３．雑誌名 Knowledge and Information Systems (KAIS)	６．最初と最後の頁 1-25
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10115-020-01445-4	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

〔学会発表〕　計9件（うち招待講演　0件／うち国際学会　5件）

１．発表者名 Takahiro Komamizu, Yushi Uchida, Yasuhiro Ogawa, Katsuhiko Toyama
２．発表標題 Analyzing Japanese Law History through Modeling Multi-versioned Entity
３．学会等名 the 2nd International Workshop on Contextualized Knowledge Graphs (CKG@ISWC 2019)（国際学会）
４．発表年 2019年

１．発表者名 植原 リサ，駒水 孝裕，小川 泰弘，外山 勝彦
２．発表標題 弱分類器の調整に基づく不均衡データ向けアンサンブル・フレームワーク
３．学会等名 第12回Webとデータベースに関するフォーラム
４．発表年 2019年

１．発表者名 植原 リサ，駒水 孝裕，小川 泰弘，外山 勝彦
２．発表標題 不均衡データ分類フレームワークにおけるサンプリング比率の最適化
３．学会等名 第12回データ工学と情報マネジメントに関するフォーラム
４．発表年 2020年

１．発表者名
Takahiro Komamizu

２．発表標題
Graph Analytical Re-ranking for Entity Search

３．学会等名
the 1st International Workshop on EntitY REtrieval (EYRE 2018)（国際学会）

４．発表年
2018年

１．発表者名
Takahiro Komamizu

２．発表標題
Learning Interpretable Entity Representation in Linked Data

３．学会等名
the 29th International Conference on Database and Expert Systems Applications (DEXA 2018)（国際学会）

４．発表年
2018年

１．発表者名
駒水 孝裕

２．発表標題
グラフ構造を利用したエンティティ検索

３．学会等名
第11回データ工学と情報マネジメントに関するフォーラム

４．発表年
2019年

１．発表者名
Takahiro Komamizu , Risa Uehara, Yasuhiro Ogawa, Katsuhiko Toyama

２．発表標題
MUEnsemble: Multi-ratio Undersampling-Based Ensemble Framework for Imbalanced Data

３．学会等名
The 31st International Conference on Database and Expert Systems Applications（国際学会）

４．発表年
2020年

１．発表者名 Takahiro Komamizu
２．発表標題 SPARQL with XQuery-based Filtering
３．学会等名 The 19th International Semantic Web Conference（国際学会）
４．発表年 2020年

１．発表者名 駒水 孝裕 , 小川 泰弘, 外山 勝彦
２．発表標題 法令沿革 LOD 構築のための DBpedia における法令エンティティの同定
３．学会等名 第51回人工知能学会セマンティックウェブとオントロジー（SWO）研究会
４．発表年 2020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