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 ２２ 日現在

機関番号：５２６０５
研究種目：基盤研究(C)（一般）
研究期間：2019～2023
課題番号：１９Ｋ１２１１０

研究課題名（和文）誤分類に基づいたクラスタ間の関連性を分析するための枠組み構築に関する研究

研究課題名（英文）Relation Analysis among Clusters based on the Miss-Classifiers

研究代表者
横井　健（Yokoi, Takeru）

東京都立産業技術高等専門学校・ものづくり工学科・准教授

研究者番号：４０４６９５７３

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：　本研究課題では、「クラスタリング結果において誤分類された対象に基づいて、クラスタ間の関連性を分析するための枠組み」の提案を目指した。そのために、1.)クラスタリング結果において誤分類された対象が表現するクラスタ間の関連性とはいかなるものか？、2.)クラスタリング結果において誤分類された対象が示す関連性の度合いをどのように測るか？、という2つの研究課題に取り組んだ。
　その結果、1.)に関しては、クラスタ間における関連性の要因分析の枠組み、2.)に関しては、クラスタ間の関連性の度合いを測る尺度の作成に一定の目処を立てることができた。

研究成果の学術的意義や社会的意義
　昨今の情報爆発時代の到来を受け、有用な情報を効率よく選別するための情報選別技術の重要性はますます増している。その中の代表的な技術のひとつに、似た対象を自動的に集約する「クラスタリング」がある。クラスタリングされた結果において有用と考えられる情報は大きく分けて2種類存在すると考えられる。ひとつは、そのクラスタを代表する対象の集合、もうひとつは、クラスタ間の関連性を表現する対象の集合である。
　本研究課題は、この「クラスタリング」という簡便な方法を用いて、後者のクラスタ間の関連性を表現する対象の集合を抽出し、その関連性を分析するための新たな枠組みを構築を目指した点において、有意であると考えられる。

研究成果の概要（英文）：This study has proposed a novel framework for analying the relations among the clusters based on the miss-clustered subjects. In order to achieve the purpose, two main research questions have been addressed. One is what is the relation among the clusters identified by the miss-clustered subjects. The other is how to measure the degree of relation among the clusters identified by the miss-clustered subjects.
As a result of our research, regarding the 1st research question, the framework of factor analysis for the relations was constructed and the analysis of the relations was conducted. Regarding the second research question, some measures of the degree of the relations among clusters were established, i.e., based on the number of miss-clustered subjects, the probability of miss-clustering, and the focus on the proper nouns.
The experiments were carried out on news articles and the usefulness of the results was verified.

研究分野：データマイニング

キーワード：テキストマイニング　データマイニング　自然言語処理

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

研究代表者らは、従前の研究において、さまざまな国で発行されたニュース記事に対するクラスタリング結果における誤分類されたニュース記事の分析結果から、それら誤分類されたニュース記事が、もとの国と誤分類された国に共通する点を示しており、それが、国と国との関連性を示唆している可能性を明らかにしてきた。

対象間の関連性抽出は大きな研究タスクであり、これまでも多くの研究がなされている。多くの関連性抽出の研究では、一文中に存在する対象間の関連性に焦点を当てており、本研究課題のように文書集合を対象とし、その中から関連性を発見するという枠組みとは異なっていた。

２．研究の目的

これまでの研究代表者らの研究結果から導き出された「クラスタリング結果において誤分類された対象はその間違ったクラスタ間の関連性を示唆している。」という仮説のもと、「クラスタリング結果において誤分類された対象に基づいて、クラスタ間の関連性を分析する枠組み」を構築することを目指す。本研究課題では、上記目的を達成するために、大きく分けて、以下の２つの研究課題に取り組んだ。

【研究課題１】クラスタリング結果において誤分類された対象が表現するクラスタ間の関連性とはどのようなものか？

【研究課題２】クラスタリング結果において誤分類された対象が示す関連度の度合いをどのように測るか？

以上の２つの研究課題を軸に、クラスタ間の関連性を分析する枠組みの構築を目指した。

３．研究の方法

本研究課題では、「クラスタリング結果において誤分類された対象」に着目する。誤分類された対象とは、図１の▲や●が示すような対象である。▲は、本来、クラスＡに、●はクラスＢにそれぞれ属すると考えられるが、クラスＢ、または、その逆に間違ってクラスタリングされた対象である。なお、×は各クラスのセンタを表している。

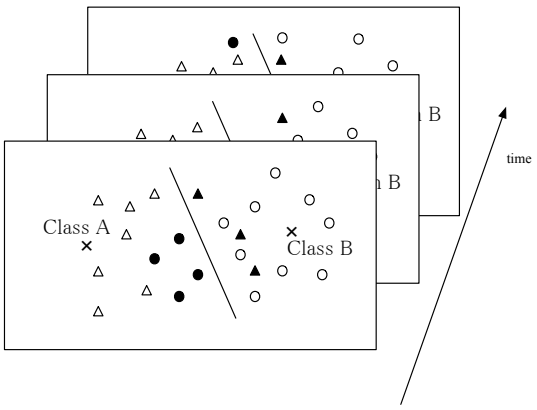

図1　間違ったクラスタリング結果と関連性

しかしながら、見方を変えて、これらの対象は、クラスＡとクラス Ｂ をつなぐ重要な役割を担った対象であると本研究課題では考える。つまり、これらの対象は、クラスＡとクラスＢの関連性を表現している対象であると考える。

まず、【研究課題１】のこれらの対象が表現しているクラスタ間の関連性とはいかなるものかという点に対して、分析の枠組み構築を行う。具体的には、1.）誤分類された対象とクラスタリング対象を表現した特徴量の正準相関分析、2.）Latent Dirichlet Allocation (LDA) 等のトピック分析手法の適用を検討した。特に、これまで、クラスタリング結果における誤分類された対象を人手で確認し、関連性の内容を分析していた部分について、より客観的かつ数値的にその関連性を分析する枠組みについて検討を行った。まず、トピック分析を行うことで、関連性の内容についての分析を行った。さらに、誤分類によって得られる関連性とそれらのトピックとの相関について、正準相関分析を用いて分析する手法を検討し、関連性がどのような要因から構成されているか客観的な指標で分析を行った。クラスタリングを行う対象は、Embedding されているとし、本研究課題で取り扱ったニュース記事の場合では、Doc2Vec を用いて Embedding を行った。

また、関連性の時間的な変化の分析を行った。特に、この分析では、誤分類結果の時間的変化に焦点をあて、その誤分類結果に含まれる関連性のトピックの時間的な変化を分析した。より具体的には、図１の Time 軸における変化に着目し、誤分類された対象を１ヶ月程度の期間ごとでスナップショットとしてまとめ、それのデータを、DTM (Dynamic Topic Model) を用いて分析した。

次に、【研究課題２】のこれらの関連性を測る尺度の構築について検討を行った。引き続き、経済に関するニュース記事を対象に、関連性の度合いを測る尺度として以下の 1.)誤分類された対象の数の平均値、2.)誤分類される確率を用いて確率分布間の擬距離、の２つの尺度を提案した。１つめの尺度においては、クラスタリング手法として用いた k-means が、初期値によってクラスタリング結果が変わるため、複数回クラスタリングを実行し、各クラスタリングにおいて誤分類された記事数の平均値を関連性の尺度として定義した。２つめの尺度は、各国の収集した記事数がそれぞれ異なることに着目し、誤分類される確率を用いたものである。ここでは、確率間の擬

距離を JSD（Jensen-Shannon Divergence） を用いて測ることで、関連性の度合いを定義した。また、単純な確率の差による尺度についても検討を行った。

さらに、この関連性を分析するフレームワークをさまざまな国で発行されたトピックを限定しないニュース記事に適用し、国家間の関連性の度合いを測ることを試みた。より具体的には、誤分類されたニュース記事に含まれる固有表現の出現頻度と各固有表現に割り当てられた重要度から、統計的な国家間の関連性の度合い、ある国からある国への興味スコア、を算出する手法を提案した。ここで用いた固有表現は、人物名、地名、組織名、企業名の４つのカテゴリの固有表現である。なお、各固有表現に対する重要度は、クラスタリングにおいて正しく分類される記事中の出現頻度が高く、一方で、誤分類記事中における出現頻度が低い固有表現に高い重要度が割り当てられるように設計した。

４．研究成果

まず、【研究課題 1】の成果について述べる。

本研究課題開始前に収集した日本、イギリス、アメリカ、カナダの４カ国の経済に関する英語のニュース記事を対象とし、経済というキーワードにおける国家間の関連性に着目をして分析を行った。

その結果、【研究課題 1】の問いに対して、関連性を構築している要因に関する分析手法を提案することができた。しかしながら、対象 A と対象 B の間の関連性があった際に、その A と B の直接の関連性について詳細な分析ができるようになった一方で、対象 C を介した、A と C、B と C といった偏相関のような関連性については、今の枠組みでは分析が難しいことが判明した。

DTM を用いた、関連性の時系列変化に対する分析では、関連性の内容の時系列的な変化を把握することが可能となった。例えば、分析を行った期間は、アメリカ大統領選挙に関するニュース記事が各国で報道されていた時期であり、アメリカ大統領選挙に対する関心度の変化から、各国間の関連性の変化を分析することができたと考えられる。

次に、【研究課題 2】の成果について述べる。

誤分類された記事数の平均値による関連性を測る尺度では、各国のニュース記事数の差異によって、大きく結果が異なってしまうため、なんらかの正規化を行う必要があることが改めて分かった。その正規化の手段としては、確率として取り扱うことが有効であると考えられる。

研究期間前半における大きな懸念点として、これまで検討してきた枠組みで抽出できるのは、国家というような抽象的な概念の関連性ではなく、ある事象やトピックの関連性に留まっているという点が挙げられた。この点については、各国の代表的な出版元によるニュース記事をトピックや事象などの制限を設けず収集し、分析を進めた場合に顕著に現れる傾向であった。分析に使用するニュース記事を経済や特定の事象などのある特定のトピックなどに制限した場合、トピックや特定の事象の関連性という観点ではなく、別の観点で分析ができるのではないかという着想に至った。

この着想に基づいた予備実験として、ある特定の著者による X（旧 Twitter） の投稿を用いて事前学習済みの BERT のモデルをファインチューニングし、その学習させた BERT を用いて、他のユーザの投稿とその学習させた著者の投稿の混合集合から、投稿の内容によったトピックとは別次元の「著者」という軸による分類を試みた。なお、分類に使用した投稿集合は、ある特定のキーワード集合で検索して集めたものであり、類似したトピックを持った投稿であると考えられる。

その地域と関連性が深いと考えられる固有表現に着目し、その固有表現を用いて関連性を測る手法については、実験を、アメリカ、イギリス、カナダの 3 カ国の新聞記事を用いて実施した。その結果、国際影響力の高いアメリカとイギリスの 2 カ国の興味スコアの順位が高いという結果が得られた。なお、使用する固有表現のカテゴリを変更することで、興味スコアの順位が変動することも判明した。

また、研究期間を通して、データの拡充についても随時行い、分析結果に反映できるようにした。Media Cloud というニュース記事を集めたサイトを利用して、さまざまなニュース媒体やウェブサイトから、より効率的にニュース記事やテキスト情報を収集できるようなツールの作成についても検討した。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 葛野航希，横井　健
２．発表標題 ニュース記事に含まれる固有表現を用いた国家間の関連性分析の検討
３．学会等名 情報処理学会第86回全国大会
４．発表年 2024年

１．発表者名 波多野陸、横井　健
２．発表標題 GPT-2が生成した文章に対する書き手の個性に着目した自動分類
３．学会等名 情報処理学会第85回全国大会
４．発表年 2023年

１．発表者名 Takeru Yokoi, Ryota Ikushima, Roliana Ibrahim
２．発表標題 Quantitative Analytic Framework of Relations among Unstructured Data
３．学会等名 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC2020)（国際学会）
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