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　４ 年　 ６ 月 １７ 日現在

機関番号：２２６０４
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ１１９８２

研究課題名（和文）語の意味演算のための時空間における偏在性と遍在性に着目したベクトル空間モデル構築

研究課題名（英文）Vector Space Model for Georefferenced Words' Algebra using Its Unevenness and Ubiquity

研究代表者
横山　昌平（Yokoyama, Shohei）

東京都立大学・システムデザイン研究科・准教授

研究者番号：２０４４３２３６

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：本研究期間において、ジオソーシャルビッグデータ分析技術について次の進捗を得た。(1)ジオソーシャルビッグデータ収集基盤技術の実現、(2)地理的説明性の分析のための軌跡に基づいた密度クラスタリング手法の実現、(3)人の行動に基づいた密度クラスタに対するラベル生成手法の実現、(4)テキストタグの地理的な偏在性を明らかにするための、TF-IDFの多次元時空間への応用手法の実現。

研究成果の学術的意義や社会的意義
本研究はジオソーシャルデータに対して、コンテンツのセマンティック(意味空間)とジオタグ(地理空間の統合のための課題である。従来は、時空間分析と意味空間分析は異なるアルゴリズムを適用する事が前提であったが、本研究の成果では、これらを一つのアルゴリズムとして統合した。これは、意味空間と地理空間を跨いだワンストプの分析を行う基盤技術となる。

研究成果の概要（英文）：During this research period, the following progress was made in geosocial big data analysis technology. (1) Realization of geosocial big data collection infrastructure technology (2) Implementation of a trajectory-based density clustering method for geographic explanatory analysis, and
(3) Implementation of a label generation method for density clusters based on human behavior, and
(4) Realization of a method for applying TF-IDF to multidimensional time-space in order to reveal the geographic ubiquity of text tags.

研究分野：データ工学

キーワード：ソーシャルビッグデータ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

本研究計画では、データセットに依らないベクトル空間モデルを構築するために、ジオソーシャルデータに着目する。ジオソーシャルデータとは、SNS 上で共有されている地理情報が付与された、文章・画像・動画等多様なメディアで構成されたビッグデータである。これをコーパスとして用いることで、多様なメディアを一元的な時空間(場所・時間からなる空間)領域に写像し、そのセマンティックを、分布の偏在性(偏って存在)と遍在性(満遍なく存在)から明らかにする手法を研究する。つまり本研究は、自然言語処理や画像処理の課題を、自然言語処理技術や画像処理技術に頼らず、地理情報・時空間データ処理に基づくベクトル空間モデルの課題として一元的な解決を目指している。

２．研究の目的

自然言語処理・画像処理とディープラーニングは親和性が高く、これまでにも様々な提案がされている。有名な事例の一つとして word2vec がある。word2vec では与えられた大規模な文章群を学習する事によって、そこに属する単語のセマンティックをベクトル表現する技術である。自然言語や写真等のデータは本質的に不均質であり、それをベクトル空間モデルとして定量的な表現に変換する事は、それらを分析するための重要なステップである。しかしながら、ここで作られたベクトル空間は、与えるデータセット毎に異なるものが生成される。また、異なるメディアを統一して扱う事は出来ない。

研究代表者はこれまでに、ジオソーシャルデータの分析技術の研究を行ってきた。ジオソーシャルデータとは、ソーシャル・ネットワーキング・サービス(SNS)等、インターネット上で共有されているデータ(ソーシャルデータ)のうち、何らかの地理情報に紐づけられたものを指す。例えば GPS により撮影位置の緯度経度が付与された写真であったり、お店やホテルのレビュー等である。これらは、スマートフォンの普及と共に、爆発的に増加し、ビッグデータを形成している。その多くが、インターネットの一般的なユーザが作成し、誰でも使える形で共有されており、人類の新たな共有財産となっている。その為、ソーシャルデータ分析はマーケティングや観光等様々な応用分野において一つの研究潮流になっている。ひとえにソーシャルデータといっても、文章・画像・動画と多様であり、また文章一つとっても、Twitter のような砕けた表現の短文から、レビューのようなユーザの意見を表す長文等様々であり、言語も多様かつ混在している。

本研究計画では、冒頭で述べたディープラーニングにおける、データセット毎に分断されたベクトル空間の問題、そして次に述べたソーシャルデータの不均質性に関する問題を解決するために、ジオソーシャルデータに紐づけられた地理情報を用いて様々なメディアを一元的な時空間(場所と時間からなる空間)領域に写像し、その分布（偏在性と遍在性）とセマンティックの関係を明らかにする。

３．研究の方法

本研究課題では、地理情報に紐づいたソーシャルデータの、時空間分析の側面に着目し、これをコンテンツ、すなわち Tweet や口コミのようなテキストや、写真・イラストのような画像と、同空間にて分析するための、ベクトル空間モデルの検討を行った。

データの遍在性と偏在性を分析する事によって得られる『地理的説明性』という新しい概念の定式化を行った。またビッグデータを高効率で収集するための基盤技術にも取り組んだ。具体的には次にあげる項目について研究を行った。

(1) ジオソーシャルデータ収集基盤

ソーシャルデータは一般的には Web API を通じて検索・取得されるものである。これはインターネットを介してデータベースへアクセスするため、ローカルにあるデータベースへのアクセスと比べ制約が大きい。この問題に対して、高効率でデータを収集する検索アルゴリズムを研究した。

(2) 地理的説明性

コンテンツに紐づいた地理情報を集約すると、そのデータがランダムでない限り、一様分布とはならず、偏在性が認められる。しかしながら、この偏在の度合いはコンテンツのセマンティックを反映したものとなる。そこで、広い空間に存在するデータの偏在性をクラスタとして分類するアルゴリズムの研究を行った。

(3) 語の地理空間から意味空間への射影

データの地理空間における分布は、そのセマンティックによって異なる偏在性を有する。そこでジオソーシャルビッグデータをクラスタリングし、できたクラスタに対して、意味空間におけるラベル付けを行う手法を研究した。

(4) 語の意味空間から地理空間への射影

一方で、逆方向、すなわち語の意味空間を地理空間における分析へと射影する事も重要である。口コミ等のテキストデータを地理空間において集約し分析する手法を研究した。

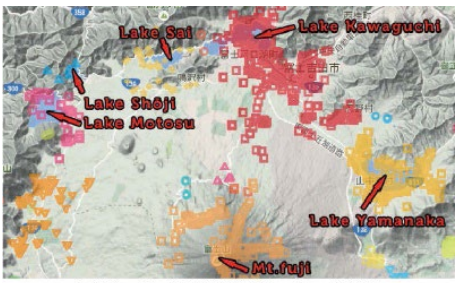
４．研究成果

(1) ジオソーシャルビッグデータ収集基盤技術

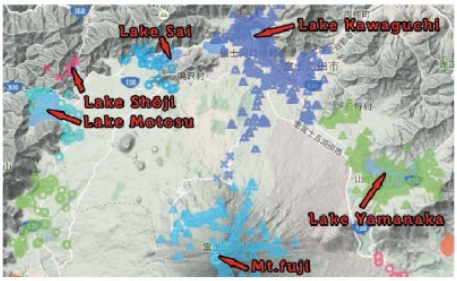
本研究計画の付随する課題として、SNS 等のソーシャルビッグデータからの高効率なデータ収集手法の実現がある。本研究期間では特に Google Maps から口コミを含む Venue データの収集において、顕著な成果があった。これは Google Maps だけでなく、密度の統計情報を予めもたない時空間データに対して k-近傍探索を用いて網羅的な検索を行う課題においても有効な検索命令構成法である。

ここで取り組んだ問題は、一回の k-近傍探索でカバーされる検索範囲はデータの密度に依存して大きさが異なり、予め密度の統計情報を持たないソーシャルビッグデータに対する検索の場合は、k-近傍探索のクエリ中心点の配置戦略を立てるのが困難である点である。この問題に対し、Open Street Maps の地図データを用いて、Google Maps の Venue が高密度に存在するエリアは、道路データも高密度であるという仮説に基づいて、道路の交差点にクエリを配置する手法を提案した。この手法では、グリッドベースの収集方法に比べて、数倍から数十倍のクリエ効率を達成する事を実験にて示した。

(2) 地理的説明性について



(a) Result of EBSCAN (tooFar=0.013)

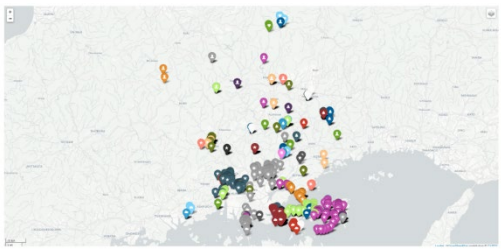


(b) Result of DBSCAN (Eps=0.012/MinPts=30)

地図空間において、地理情報が紐づいたソーシャルデータが高密度に存在している場所を発見する事で、ソーシャルデータの地理的説明性を分析が可能となる。DBSCAN 等、古典的な密度クラスタリング手法を用いてもそのような高密度地域の発見は可能である。地理情報が紐づいたソーシャルデータは、荒い人の行動軌跡であり、空間だけでなく、時系列データであり、既存の密度クラスタリングでは、データが内包する時系列セマンティックを無視し、単に空間内の密度だけしか着目しない。これに対し、我々は、人の移動に着目した密度クラスタリングである EBSCAN を提案した。

EBSCAN は軌跡の交点に着目し、その交点を共有する四端点に対して、その距離に応じて、クラスタを形成するという全く新しいアイデアに基づいた密度クラスタラング手法である。DBSCAN との比較実験において、EBSCAN で生成されたクラスタは、実社会的に意味のある纏まり（例えば、商店街という単位でクラスタが形成される）をビッグデータから抽出できる事を示した。

(3) 語の地理空間から意味空間への射影



本課題の成果の一つとして、ジオタグ付き写真の密度クラスタリングの結果に対して、ユーザの行動に基づいたセマンティックを付与する手法を実現した。

前項目にて、EBSCAN が実社会的に意味のある纏まりを抽出できると説明したが、その実験的な検証も行った。Flickr から、ジオタグ付き写真を収集し、それに対して EBSCAN を用いて、クラスタを生成する。理想的には、このクラスター一つ一つ

が、なんらかの意味を持っているはずである。そこで、事例研究として、広島県全域の写真群をクラスタリング氏、その結果から広島への玄関口となるクラスタを機械的に探す事に取り組んだ。

これは本研究計画で目指す時空間と意味空間の融合課題である。画像の物体認識技術とクラスタリング技術を併用し、画像(意味空間)のデータを用いて、クラスタ(地理空間)のラベリングを行う手法を実現した。

(4) 語の意味空間から地理空間への射影

本課題の成果の一つとして、Google Maps API を用いて、口コミからその地域特有の美味しいものを発見するための手法を実現した。

地図上のとある領域において、そこに存在するレストランへの口コミの多寡は、様々な要因による。もちろんその土地の名物(浜松の鰻など)であれば多くの口コミが集まるが、それ以外に、有名チェーン店(ファストフードなど)、大規模レストラン(居酒屋やファミリーレストランなど)にも口コミが集まりやすい。つまり口コミの量や、そこに寄せられた内容からだけでは、その土地特有の料理が何で、それに対する感想がどのようなものかという分析は難しい、そこで TD-IDF と同様の概念に基づき、これを多次元へと拡張子、他業態との比較、他地域との比較を重畳する事で、その土地特有の料理とその料理への口コミを機械的に抽出するアルゴリズムを実現した。

実験では、例えば高知市において、カツオが名物である事は有名であるが、ウツボ料理等、また全国的に知られていない名物も、Google Maps にある口コミデータのみで機械的に明らかにできる事を示した

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 伊藤光太郎，横山昌平	４．巻 14-3
２．論文標題 移動軌跡の交点を用いた密度クラスタリングアルゴリズム	５．発行年 2021年
３．雑誌名 情報処理学会論文誌データベース	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	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 －

１．著者名 井嶋 蒼，横山 昌平	４．巻 18-J
２．論文標題 ロードマップを用いたジオソーシャルデータに対する効率的なクローリング手法	５．発行年 2020年
３．雑誌名 DBSJ Japanese Journal	６．最初と最後の頁 1-6
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計14件（うち招待講演　1件／うち国際学会　2件）

１．発表者名 柿本 航太郎（都立大），井嶋 蒼(都立大)，横山 昌平（都立大）
２．発表標題 クチコミのジオリファレンスを用いた地域における特徴語の抽出
３．学会等名 9月DBS/IFAT/DE合同研究会，オンライン参加
４．発表年 2020年

１．発表者名 夏 思浩（都立大），井嶋 蒼(都立大)，横山 昌平（都立大）
２．発表標題 大規模ソーシャルデータを用いた寄り道ルート推薦手法
３．学会等名 ARG 第16回Webインテリジェンスとインタラクション研究会(Wi2)
４．発表年 2020年

１．発表者名
Shohei Yokoyama

２．発表標題
Efficient Crawling of Georeferenced Documents from Large Web Map Services

３．学会等名
Korea-Japan (Japan-Korea) Database Workshop 2020（招待講演）

４．発表年
2020年

１．発表者名
大平 翼（都立大），横山 昌平（都立大）

２．発表標題
ポーカーの統計を用いたプレイヤーの実力評価

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(DEIM2021)

４．発表年
2021年

１．発表者名
笠原 悠樹（都立大），横山 昌平（都立大）

２．発表標題
全天球カメラを用いたリアルタイム行列検知

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(DEIM2021)

４．発表年
2021年

１．発表者名
藤原 夏姫（都立大），横山 昌平（都立大）

２．発表標題
360度動画中における商品認識

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(DEIM2021)

４．発表年
2021年

１．発表者名
渋川 大樹（都立大），横山 昌平（都立大）

２．発表標題
関係データベースおよびMIDIデータを用いた演奏支援システム

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(DEIM2021)

４．発表年
2021年

１．発表者名
Sou Ijima, Masaharu Hirota, Shohei Yokoyama

２．発表標題
A Crawling Method with No Parameters for Geo-Social Data based on Road Maps

３．学会等名
The 21st International Conference on Information Integration and Web-based Applications & Services(iiWAS2019)（国際学会）

４．発表年
2019年

１．発表者名
Shohei Yokoyama, Sou Ijima

２．発表標題
Towards Efficient Crawling of Georeferenced Documents from Location-based Social Networks

３．学会等名
The 11th International Conference on Management of Digital EcoSystems (MEDES ' 19)（国際学会）

４．発表年
2019年

１．発表者名
増田 純也（首都大），横山 昌平（首都大）

２．発表標題
全天球カメラを用いた机上物体に対する位置推定

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(DEIM2020)

４．発表年
2020年

１．発表者名
林田 和磨（首都大），横山 昌平（首都大）

２．発表標題
全天球カメラにより配信される正距円筒図法動画からのリアルタイム人物検出

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(DEIM2020)

４．発表年
2020年

１．発表者名
伊藤 光太郎（首都大），横山 昌平（首都大）

２．発表標題
移動軌跡の交点を用いたジオソーシャルデータに対するクラスタリング手法

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(DEIM2020)

４．発表年
2020年

１．発表者名
江口 航野（首都大），横山 昌平（首都大）

２．発表標題
著作権法第10条2項に該当しうる部分文章の抽出

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(DEIM2020)

４．発表年
2020年

１．発表者名
若狭 孟（首都大），横山 昌平（首都大）

２．発表標題
Twitterを用いたサッカー選手採点のための感情値辞書構築に向けて

３．学会等名
ARG 第14回Webインテリジェンスとインタラクション研究会(Wi2)

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