地物間と単語間の類似度を考慮した目的をクエリとする地物情報検索

前川　由依†　　荘司　慶行†　　MartinJ. Dürst†

† 青山学院大学 理工学部 情報テクノロジー学科　〒 252 – 5258 神奈川県 相模原市 中央区 淵野辺
E-mail: †maekawa@sw.it.aoyama.ac.jp, ††shoji@it.aoyama.ac.jp, †††duerst@it.aoyama.ac.jp

あらまし　本研究では，入力された目的を達成可能な地物をランキングとして出力する検索アルゴリズムを提案する．現在の地物情報検索では，探したい地物の業種や特徴を入力とする．しかし,「ギターの練習」のできる場所を探す際に，事前知識のない利用者はどこでギターの練習ができるか思いつけず，適切なクエリを入力できない．そこで本研究では，地物に対するレビュー情報を用いることで，近所の公園やカラオケ店など，目的を達成できる地物を直接検索可能にする．そのために，地物とそれに対するレビュー中の語からなる 2 部グラフを作成し，Random Walk with Restart を用いてキーワードクエリと地物の関連度を計算する．この際，地物同士の類似度と目的同士の類似度を考慮してグラフを拡張することで，目的を達成可能な地物をより多く発見する．Google Map の実データを用いた被験者実験を通して，提案手法が有効であることと，特に地物同士の類似度の効果が大きかったことが明らかになった．

キーワード　地物情報検索, レビュー, 情報検索, 目的による検索

1 はじめに

近年，Google Map[1]や Yahoo!ロコ [2]などの様々なサービスに代表されるような，地理情報検索が一般的に普及してきている．幅広い層の人々が，店舗，施設など場所を探すために地理情報検索サービスを利用している．利用者の中には，地物に対する事前知識が十分でない子供や，検索に慣れていない年配者なども存在する．このように幅広い層の人々が地物検索サービスを利用するようになった現在，事前知識が十分でないユーザも満足に検索ができるような地理情報検索アルゴリズムが求められるようになりつつある．

従来の地物情報検索では，目的を達成可能な場所を探すために，ユーザは探したい地物の業種や特徴を入力する必要がある．例えば，本を購入したければ「本屋」，配送サービスを利用したければ「郵便局」というクエリを入力して検索しなければならない．ここで，あるユーザが「ギターの練習」のできる場所を探している場合を考える．本来であれば「音楽スタジオ」といったクエリで検索をする必要がある．しかし，このクエリを思いつくためには，ユーザには「ギターの練習は音楽スタジオでできる」といった事前知識が必要になる．そのため，そのユーザに事前知識がなければ，どのような施設でギターの練習ができるか思いつかず，そもそも適切なクエリを入力することすらできない．この問題は,「ギターの練習」といったように，その場所で行いたい目的をクエリとして検索できれば解決する．しかし，従来の地物検索システムは地物に対する施設名や業種の情報のみに対応しており，その場所で何が実行できるかの情報を持っていないため，目的をクエリとして地物を検索するのは不可能である．

加えて,「ギターの練習は音楽スタジオでできる」という事前知識があった場合でも,「音楽スタジオ」というクエリでは，十分に多くの，ギターの練習が可能な場所を発見できない場合がある．ある目的があった際に，人がとっさに入力できるキーワードは限られている．検索結果のギタースタジオよりも，ユーザの近くにギターの練習が可能な「カラオケ店」や「公園」にあったとしても，ユーザが「音楽スタジオ」というクエリで検索をした場合には，このような，より適切な場所を発見できない．

そこで本研究では，その場所で行いたい行動からなる「目的」を直接入力すると，その目的を達成可能な地物をランキングとして出力する検索アルゴリズムを提案する．例えば,「ギターの練習」と入力した時には,「スタジオ○○中野店」「××カラオケ新宿西口店」など，具体的な地物を順位付けして出力する．このような，目的を入力することができる検索アルゴリズムがあれば，事前知識の有無を問わずに，現在の幅広い利用者層に対応できると考えられる．

本研究では，このような検索アルゴリズムを実現するため，地物へのレビュー情報に着目した．Google Map などの一部の地理情報サービスでは，利用者が地物に対してレビューを投稿できる．このような地物に対するレビューには，その場所でユーザが実際に行った，実行可能な行動が多く含まれる．具体的な地物とその場所で何をしたかの情報が紐づいている情報源として，他に Twitter などの SNS の位置情報つき投稿を使った研究もあるが [1]，SNS の位置情報付き投稿は自動でタグ付けをしているユーザが多いため，たまたまその場所で投稿された，日常的な内容を含むことがある．しかし，地物へのレビュー情報は他人にその場所を紹介するために書かれており，その地物に直接関連する内容である可能性が高い．よって本研究では，正確にその場所で行うことができる情報を得るために，地物へのレビュー情報を用いた．

これらの地物へのレビュー情報は重要な情報源であるが，そのまま用いても，提案する検索アルゴリズムを実現するのに不十分である．原因の 1 つに，レビューの網羅性の限界がある．

1 : https://www.google.com/maps
2 : https://loco.yahoo.co.jp

レビュー情報は表記に揺れがあったり，その場所で行う事ができる目的が全て書かれているわけではない．例えば，ギターの練習が可能な全ての場所に「ここでギターの練習をしました」というレビューが付いているわけではない．また,「楽器の練習をしました」と書いてある場所では多くの場合「ギターの練習」も可能だが，ギターという語はレビュー中に登場しない．よって，単純なクエリとレビューの一致度では，目的を入力しても，それを実現可能な地物を網羅的に発見できない．そこで，入力した目的クエリと，それを実行可能な地物について，より柔軟な適合度計算を行うために，3 つの仮説を立てた．すなわち，

仮説 1　相互再帰による推論
同じ場所でできること同士は似ており，加えて，同じことができる場所同士は似ている，

仮説 2　地物レベルでの拡張
似た地物であれば，同じ行為が達成可能である，

仮説 3　単語レベルでの拡張
意味的に近い行為同士は同じ場所で達成可能である．

単語レベルでの拡張は，例えば，ギターとウクレレは辞書的に意味が近いため,「ギターの練習」ができる場所では「ウクレレの練習」ができる，というように結果を拡張できるという仮説である．地物レベルの拡張は，例えば「カラオケチェーン A 新宿店」と「カラオケチェーン A 渋谷店」は同じ店の別店舗であるため，同じ行為が達成できるという仮説である．そして，これら 2 つの仮説を組み合わせ，相互再帰による推論を行うのが提案するアルゴリズムである．1 つの仮説により拡張された結果をもう一方の仮説で拡張し，これを再帰的に繰り返すことにより，出力結果の再現率を向上可能であると考えた．

そこで本研究では，これら 3 つの仮説をもとに，目的から直接地物を探す検索アルゴリズムを提案する．提案手法では，地物とそれに対するレビュー中の語からなる 2 部グラフにおいて，Random Walk with Restart（RWR）によるリンク解析を行った．この時，地物同士の類似度と目的同士の類似度を加味するための地物間と単語間の疑似的なリンクを加えた．地物とレビュー関係を 2 部グラフにして処理することで，網羅性の低いレビューから，地物同士の類似度と目的同士の類似度を考慮して推論し，目的を達成可能な地物をより多く検索可能にした．

提案手法の有効性を明らかにするために，実データを用いた実験を行った．まず，Google Map の地物へのレビューデータを用いて提案手法を実現するウェブシステムを実装した．このシステムでは，目的を入力し，その目的が達成可能だとされた，具体的な地物をランキング形式で出力する．出力したランキング内の地物に対して，クエリとして入力した目的を達成可能かどうか，被験者にラベル付けさせた．検索ランキングとして評価することで，手法の有用性を明らかにした．

本論文の構成は次のとおりである．第 2 章では，本研究に関連する研究を紹介する．第 3 章では入力された目的を達成可能な地物をランキングとして出力する検索アルゴリズムを提案する．第 4 章で提案手法の評価を被験者実験により行う．第 5 章では評価実験の結果について考察し，第 6 章で結論と今後の課題を述べる．

2　関 連 研 究

本研究は，目的をクエリとする地物検索アルゴリズムの研究である．そのために，地物名や目的を類義語へと拡張する．よって本研究は，地物の検索についての研究や目的の拡張，地物推薦の研究と深く関連する．

2.1　地 物 検 索

現在の一般的な検索システムでは，主に地物名や地物の種類や，住所を直接クエリとして入力することで地物の検索を行う．そこで，クエリを拡張するなど，より柔軟な入力を可能にするための検索の研究 [2–4] が多く行われている．

Pat ら [5] は，Twitter や Instagram などの SNS から位置情報（ジオタグ）付き投稿を収集し，結果を領域で表す地理情報検索システムを開発した．SNS のジオタグ付き投稿に注目することで，通常は静的である地理情報データベースを動的にすることを試みた．しかし，SNS のジオタグ付き投稿は自動でタグ付けをしているユーザが多いため，実際にその場所に関連した投稿か否かの判別がつかない．そのため，ジオタグ付き投稿の情報の正確性に疑問がある．動的データに着目して地理情報検索を行う点では本研究と類似している．しかし，SNS の投稿はたまたまその場所で投稿された日常的な内容を含むのに対し，地物へのレビュー情報は他人にその場所を紹介するために書かれており，よりその地物に関連する内容を含むことが見込まれる．本研究では，より正確な動的データとして地物のレビューデータを使用した点が異なる．

Bauer ら [6] は，現在オンラインでの通信販売が一般化する中，オフラインでの購入ニーズに関して分析し，実際に購入できる物理的な実店舗の検索手法を提案している．入力された購入したい物を表すキーワードのクエリと，場所をそれぞれベクトル化し，コサイン類似度でランキングを作成することで実現している．実際の地物を検索対象にしている点で本研究と類似しているが，本研究では，単純なベクトル空間モデルにおける類似度計算でなく，2 部グラフ上でのリンク解析を行っている．これにより，より入力の幅が広がった地物検索を目指した点が異なる．また，本研究では何かを購入できる場所販売店だけでなく，その他の地物も結果として表示することができる．

2.2　目的を表す単語の拡張

本研究では同じ場所で実行可能な目的を推論によって拡張することで検索結果の再現率を向上させている．つまり，レビュー中に直接入力されたクエリを含まない地物や同じチェーンの別店舗も検索可能にしている．同様に，代替可能な単語や類似した単語を抽出する研究 [7,8] が多く行われている．行動目的を拡張する例として，Pothirattanachaikul ら [9] は，cQA サイトから同じ目的を達成できる代替案を抽出する手法を提案している．例えば「眠れるようになりたい」という目的に対して「睡眠薬をのむ」と「温かいミルク」を飲むことは，同じ目的を達成できる代替行動である．そのために，cQA サイトから質問と解答情報を抽出し，2 部グラフ化している．これにより，類

似度の順位付けを行うことで，実際に代替行動を発見している．

地物に特化した目的の拡張の例として，松村ら [10] は，行動名をクエリとして，現地でその行動が可能な地物を検索する手法を提案している．例えば「時間をつぶす」というクエリに対して，「喫茶店」と出力するような課題に取り組んだ．その地物属すカテゴリでどの行動ができるか，という情報を cQA（community Question Answering）サイトから抽出し，また，行動が行える地物の類似性に着目した行動情報の拡張を行った．こうして得られた行動情報に基づき，行動名を，その行動が行えるような場所を表す語へと変換した．松村らの研究は，入力された行動名に対して，地物のカテゴリを出力するにとどまっている．本研究ではより発展的に，地物へのレビューに書かれた内容を利用することで，目的を，カテゴリではなく具体的な地物に結び付けることを目標としている．

2.3 地物の特徴化と推薦

本研究は，ユーザの目的が達成可能な地物の検索を目標としている．同様の目的に対して，地物の特徴を抽出し，推薦などのアプローチで解決する研究がある．

地物の特徴を抽出する研究として，Kurashima [11] らは，blog からある地物の情報を抽出し，トピックモデリングによってそこでどのような体験が行われているかを地図上に可視化する手法を提案している．本研究はクエリから地物を発見するため，より網羅性が高いが記述量の少ないレビューを用いて，その場所で何が行えるかを推定している．

地物の推薦の研究として，Wang ら [12] は，SNS 上の過去の行動，場所の位置情報，ユーザ同士の関係，ユーザの類似度の情報をグラフで表現し，Bookmark-coloring algorithm を拡張することで地物の推薦を行っている．この際，ユーザの類似度を用いることで，従来の推薦よりも高精度で次にそのユーザが行くであろう地物を推薦可能にした．

3 提 案 手 法

目的を直接クエリとして入力すると，その目的を実行可能な地物を順位づけして出力する手法について述べる．本研究では，このような検索アルゴリズムを実現するため，実在する地物とその場所で行える行動を，地物へのレビューから抽出する．ここで，全ての地物に対し，そこでとれる全ての行動がレビューに書かれているわけではない．そのため，レビューに直接目的が書かれていなかったり，レビューのついていない地物を検索するためには，どの場所で何ができるか，推論や拡張が必要になる．そこで，

仮説 1　相互再帰による推論
同じ地物へのレビューに現れる単語同士は似ており，また，同じ語を含むレビューの付く地物同士は似ている

仮説 2　地物レベルでの拡張
メタデータの類似した地物では同じ行為が達成できる，

仮説 3　単語レベルでの拡張
意味的に近い行為同士は同じ場所で達成可能である

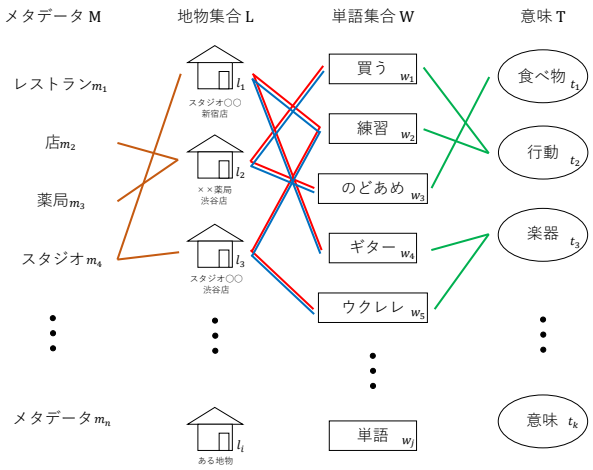

図 1　地物，レビュー中の単語，地物のメタデータと単語の意味からなるグラフ

という 3 つの仮説から，地物とレビュー中の単語の関係からなるグラフとして表すことで，より多様な地物を発見する．

3.1 相互再帰による推論のための地物レビューのグラフ化

本研究では，実在する地物とその場所で行うことができる行動を得られるデータとして，地物へのレビュー情報を利用する．はじめに，同じ地物へのレビューに現れる単語同士は似ており，また，同じ語を含むレビューの付く地物同士は似ている，という仮説を，グラフとして表現する．そのために，レビュー内の単語を地物をノードとし，それらをエッジで繋げたグラフを作成する．地物に対するレビュー全体の模式図を，図 1 に表す．

レビューデータは，ある地物 l_i と地物のレビューに含まれる単語 w_j の関係，地物と地物へのメタデータの関係と，単語と意味の関係として表すことができる．はじめに，地物とそれについたレビューに含まれる単語の関係に注目し，重み付き有向 2 部グラフを作成する．前処理として，それぞれのレビュー文を単語に分割した．この際，目的を表す語のみにするため，ノードとして使用する単語を絞った．例えば「けれど」,「が」などの助詞,「ます」,「そうだ」などの助動詞,「とても」,「はっきり」などの副詞は，目的を表す単語のノードとして不適切である．そこで本研究では，使用する単語の品詞は名詞，動詞，形容詞の 3 つに絞った．これらのうち，動詞には活用形が存在する．例えば,「書く」という動詞は活用して「書か」「書き」「書け」といった形になる．文章を分かち書きしただけでは，動詞は活用形になっており，語幹が同じであっても別の単語として扱われる．日本語の動詞は活用することで大きく意味が変化せず，活用形が別の単語として数えられると 2 部グラフのノードが増大し，類似度の計算に支障が出ると考えられる．そのため，動詞は標準形に統一した．

また，登場頻度をもとに単語数を削減した．一人称などの文章中に多く存在する単語や，ほとんどのレビュー中に登場しない一般的でない単語は，地物の詳細を表す単語として適切でない．多くの地物のレビュー中に登場する単語や，登場頻度が少なすぎる単語は，目的として意味を持たない可能性が高いため，

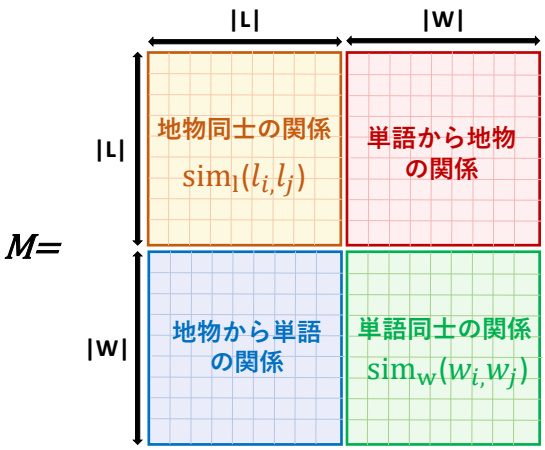

図 2　地物とレビュー中の語の関係を表した行列 $\boldsymbol{M}$ の模式図

使用する単語から省いた．最後に，用意した地物ノードと単語ノードを，地物のレビュー中に単語があればエッジで繋いだ．

作成したグラフを，隣接行列 $\boldsymbol{M}$ として表す．最終的な $\boldsymbol{M}$ の模式図を図 2 に示す．L をグラフ内の全ての地物ノードの集合，W をグラフ内の全ての単語ノードの集合とする．$|L|$ や $|W|$ という記法で集合内の要素の数を表す．つまり，図 2 の行列は，$(|L|+|W|) \times (|L|+|W|)$ の正方行列である．

また，$\mathrm{N_w}(l_i)$ を l_i から出るリンクと繋がる W の部分集合，$\mathrm{N_l}(w_i)$ を w_i から出るリンクと繋がる L の部分集合とする．この時，図 2 の行列の左下部分では，$\boldsymbol{M_{ij}}$ は w_i が $\mathrm{N_w}(l_j)$ の要素である場合に 1 になり，それ以外は 0 となる．同様に，図 2 の行列の右上部分では，$\boldsymbol{M_{ij}}$ は l_i が $\mathrm{N_l}(w_j)$ の要素である場合に 1 になり，それ以外は 0 となる．つまり，図 2 の行列の左下と右上の部分では，i 番目のノードと j 番目のノードがエッジで繋がっていた場合には，$\boldsymbol{M_{ij}}$ は 1 となり，それ以外の場合に 0 となる．このように地物へのレビュー情報を整形し，ノードとエッジのデータを準備することで，地物と単語からなる 2 部グラフを隣接行列として表した．

最後に，地物と単語を結ぶエッジに重みをつける．重みを付けない状態では，地物と単語間のエッジの量が増えるほど，エッジの密度の高い部分で，循環的に値が増加してしまう．そこで，エッジの重みを，遷移元ノードの出エッジ数で割ることにより，地物と単語間のエッジの本数に依存せず，地物間の類似度と単語間の関連度による拡張を行う．

実際の地物と単語間のエッジの重みについて，図 3 の例を用いて説明する．図中の「スタジオ○○新宿店」に注目する．地物から出るエッジについて，「スタジオ○○新宿店」からは，単語ノード「練習」と「ギター」へ 2 本のエッジが出ている．単語ノードへ出る 2 本のエッジの重みは，「スタジオ○○新宿店」から単語に対する出エッジ数で割って，それぞれ $\frac{1}{2}$ とする．

逆に，「スタジオ○○新宿店」には，単語ノード「練習」と「ギター」から 2 本のエッジが入っている．この時，「練習」からは「スタジオ○○新宿店」と「スタジオ○○渋谷店」の 2 本のエッジが出ているので，「練習」から「スタジオ○○新宿店」

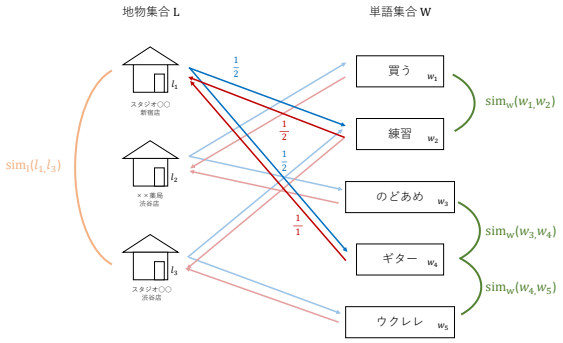

図 3　地物と単語間のエッジを正規化したグラフの例

へのエッジの重みは $\frac{1}{2}$ とする．次に，「ギター」からは地物ノード「スタジオ○○新宿店」へエッジ 1 本が出ており，地物に対する出エッジ数で割ると，エッジの重みは $\frac{1}{1}$ となる．

3.2　地物レベルでの拡張のための地物同士の類似度推定

次に，仮説 2 の地物レベルでの拡張を行うために，地物同士の関係を表す情報をグラフに追加した．地物レベルの拡張は，例えば「カラオケチェーン A 新宿店」と「カラオケチェーン A 渋谷店」のように系列店の別店舗など，類似した地物では，同じ行為が達成できるという仮説である．地物同士の類似度を考慮することで，直接レビューの書かれていない地物も発見可能になると考えられる．そこで，地物同士の類似度を考慮するよう，グラフを拡張する．

Google Map などの一部の地物情報サービスでは，地物にメタデータが付いている．メタデータの 1 つに，「レストラン」や「病院」といった地物のカテゴリ情報がある．本研究では，メタデータとして地物のカテゴリ情報を用いた．

地物同士の関係を表すメタデータを用いて，地物同士の関連度を計算し，グラフに加えた．ここで地物間の関連度を計算する手法は，様々なものが考えられるが，今回はメタデータのカテゴリ情報の一致度を用いた．

地物に対してのメタデータは，ブール値からなるベクトルとみなすことで，ベクトル空間内で地物同士の類似度を計算できる．ここでの地物 l_i のベクトル $\boldsymbol{l_i}$ は，メタデータの種類の総数を n とすると，n 次元のベクトルになる．それぞれの要素は，一例としてメタデータ m_j とのリンクが貼られていたとすると，ベクトルの j 番目の要素は 1 となり，メタデータ m_j とのリンクが貼られていない場合はベクトルの j 番目の要素は 0 となる．

地物 l_i と l_j が与えられたとき，l_i と l_j の関連度 $\mathrm{sim_l}(l_i, l_j)$ は，

$$\mathrm{sim_l}(l_i, l_j) = \frac{\boldsymbol{l_i} \cdot \boldsymbol{l_j}}{|\boldsymbol{l_i}||\boldsymbol{l_j}|} \tag{1}$$

と表す．

ただし，1 つの地物に付加されるカテゴリタグは 1 から 5 個ほどと少なく，種類も 100 種類に満たないほど少ない．タグの種類が少なく，よく用いられるタグ，すなわち「レストラン」などの該当の多いタグは限られるため，すべての地物間の関連度をエッジとして用いると，計算量が増大する．そこで本研究

では，計算の簡略化のため，類似度の低いエッジを間引いた．この際，地物 l_i と l_j がそれぞれ閾値以上のタグを持ち，なおかつ $\mathrm{sim_l}(l_i, l_j)$ の値が 1.0 となる組み合わせのみを有効なエッジとした．

こうしてメタデータの一致度を用いて地物間に仮想的なエッジを貼り，グラフを拡張した．図 2 の行列の左上部分では，$\boldsymbol{M_{ij}}$ は地物 l_i と l_j の複数のタグが一致した場合のみ 1 となり，それ以外の場合は 0 となる．

3.3 単語レベルでの拡張のための単語間の類似度計算

次に，仮説 3 の単語レベルでの拡張を行うために，単語同士の関連度を計算し，グラフに追加した．単語レベルでの拡張は，例えば，ギターとウクレレは辞書的に意味が近いため，「ギターの練習」ができる場所では「ウクレレの練習」ができる，というように結果を拡張できるという仮説である．これにより，直接目的の書かれていないレビューも，地物のランキングに反映可能だと考えられる．本節では，単語同士の類似度を考慮したグラフの拡張方法について述べる．

$\mathrm{sim_w}(w_i, w_j)$ は i 番目と j 番目の単語の関連度を表す．単語から単語へのエッジについて説明する．単語同士の意味的な類似度の計算は，LDA，LSI，Word2Vec などの方法で計算可能である．

本研究では，辞書的な語の類似度に注目するため，Wikipedia から学習したモデルの Word2Vec を用いて単語をベクトル化した．モデルの学習は，実際のレビューデータから計算する方法も考えられたが，レビューデータ中の単語同士は，すでに地物ノードを挟んでエッジで繋がっているため，単語と地物のグラフから算出される関連度と同じような値にしかならない可能性がある．そのため，より一般的な単語の拡張を行うため，今回は Wikipedia を用いた．

単語 w_i を Word2Vec を用いて分散表現化しベクトルとして表したものを，

$$\boldsymbol{w_i} = \mathrm{w2v}(w_i) \tag{2}$$

と表す．また，単語 w_i と w_j が与えられたとき，w_i と w_j の関連度 $\mathrm{sim_w}(w_i, w_j)$ は，

$$\mathrm{sim_w}(w_i, w_j) = \frac{\boldsymbol{w_i} \cdot \boldsymbol{w_j}}{|\boldsymbol{w_i}||\boldsymbol{w_j}|} \tag{3}$$

と表す．

ただし，$\mathrm{sim_w}(w_i, w_j)$ は 0 以上 1 未満の値をとる．全ての単語同士の関連度をエッジとして用いると，計算量が増大する．本研究では，$\mathrm{sim_w}(w_i, w_j)$ に閾値を設け，一定値以上の値を取る組み合わせのみをエッジとして採用した．

こうして計算した関連度をエッジの重みとして，単語同士に仮想的なエッジを貼り，グラフを拡張した．図 2 の行列の右下部分では，$\boldsymbol{M_{ij}}$ は単語 w_i と w_j の関連度 $\mathrm{sim_w}(w_i, w_j)$ が一定値以上となる場合，$\mathrm{sim_w}(w_i, w_j)$ の値となり，それ以外の場合は 0 となる．

3.4 Random Walk with Restart によるランキング作成

ここまでで，地物と地物レビュー中の単語をエッジで結んだグラフを作成し，そのグラフを地物同士の類似度と単語同士の類似度によって拡張した．本研究では，このグラフ内のノード同士の関連度を計算することで，入力された目的を達成可能な地物をランキング形式で出力する．今回は，グラフ内のノード同士の関連度の計算アルゴリズムとして，Random Walk with Restart(RWR) を採用した．本節では，ここまでで作成したグラフを用いて実際に関連度計算をする方法について述べる．

はじめに，RWR での関連度計算を行うため，地物とレビュー中の語を表すグラフ行列を遷移確率行列に変換する．遷移確率行列への変換は，行列を列で正規化，つまり出エッジの重みの和で割ることで行った．よって，地物とレビュー中の語を表すグラフを有向グラフとみなす必要がある．この際，重要視したい拡張がある場合，列で正規化する前に，地物同士または単語同士の関連度へ重み付けすることで対応できる．

L をグラフ内の全ての地物ノードの集合，W をグラフ内の全ての単語ノードの集合とする．$|L|$ や $|W|$ という記法で集合内の要素の数を表す．また，$N_w(l_i)$ を l_i から出るエッジと繋がる W の部分集合，$N_l(w_i)$ を w_i から出るリンクと繋がる L の部分集合とする．$\mathrm{sim_l}(l_i, l_j)$ は i 番目と j 番目の地物の関連度を，$\mathrm{sim_w}(w_i, w_j)$ は i 番目と j 番目の単語の関連度をそれぞれ求める関数である．α, β はそれぞれ重みであり，$0 \leqq \alpha, \beta < 1$ かつ $\alpha + \beta \leqq 1$ である．関連度を考慮し，重み付けした隣接行列を $\boldsymbol{M}$ とする．$\boldsymbol{M}$ を列で正規化し，遷移確率行列としたものを $\boldsymbol{M'}$ とする．$\boldsymbol{M}$ の定義を，

$$\boldsymbol{M_{ij}} = \begin{cases} (\text{if } i > |L|) \begin{cases} (\text{if } j > |L|) & : \beta \mathrm{sim_w}(w_i, w_j) \\ (\text{if } j \leqq |L|) & \begin{cases} (\text{if } w_i \in \mathrm{N_w}(l_j)) & : \frac{1}{|\mathrm{N_l}(w_i)|} \\ (\text{otherwise}) & : 0 \end{cases} \end{cases} \\ (\text{if } i \leqq |L|) \begin{cases} (\text{if } j > |L|) & \begin{cases} (\text{if } l_i \in \mathrm{N_l}(w_j)) & : \frac{1}{|\mathrm{N_w}(w_i)|} \\ (\text{otherwise}) & : 0 \end{cases} \\ (\text{if } j \leqq |L|) & : \alpha \mathrm{sim_l}(l_i, l_j) \end{cases} \end{cases} \tag{4}$$

と表す．また，これを列で正規化した遷移確率行列 $\boldsymbol{M'}$ を，

$$\boldsymbol{M'_{ij}} = \frac{\boldsymbol{M_{ij}}}{\sum_{k=1}^{|L|+|W|} \boldsymbol{M_{kj}}} \tag{5}$$

とする．

RWR は，グラフ上でランダムウォークを行い，ノードに到達する毎に一定確率で開始ノードにランダムジャンプすることにより，ノード間の関連度を計算するアルゴリズムである．通常は，開始ノード q を表すベクトルとして，q 番目の要素が 1 でその他の要素が 0 のベクトル $\boldsymbol{q}$ を用いる．本研究では，この開始ノードを表すベクトルを工夫することで，より多くの地物を検索結果として提示可能にする．そのために，クエリの長さによって，ランダムジャンプ先の扱いを変更する．入力されたクエリが 1 つの単語ノードのみから構成される場合について述べる．例えば，クエリが「歌う」で合った場合，「歌う」の単語ノードを開始ノードとし，通常通りに RWR を行う．

入力されたクエリが 2 つ以上の単語ノードから構成される際に，入力された複数のクエリすべてにランダムジャンプすると，クエリと関係しない地物が検索される可能性が高い．これは，入力された複数のクエリについて独立に考えてしまうためである．例えば，「ギターの練習」とクエリに入力した際，クエリは「ギター」「練習」と 2 つの単語に分割される．しかし，この 2

つの単語を独立で開始ノードとし，ランダムジャンプ先として扱うと,「ギター」のノードと類似度の高いノードと「練習」のノードと類似度の高いノードが混在した検索結果になる．この場合,「練習」と類似度の高いとされるノードは必ず「ギターの練習」に適しているとは限らず,「アーチェリーの練習」に適した場所や「華道の練習」に適した場所である可能性がある．このように，複数の単語のノードを独立に処理しようとすると，「ギター」と「練習」に適した場所の「OR 検索」に近い結果となる．これを解決するため，地物を「AND 検索」する手法として，全てのクエリ中の語を含んだレビューに注目する．クエリとして入力された，複数の単語の全てにリンクが繋がっている地物を開始ノードとし，ランダムジャンプ先にする．これにより，クエリとして入力された単語を，互いに独立で登場した場合と，同時に登場した場合で区別して扱える．例えば,「ギターの練習」とクエリに入力した際，何らかの「ギター」と類似度の高い地物や，何かの「練習」をした地物ではなく，より正確に誰かが「ギターの練習」をした地物と類似度の高い地物を提示できる．図 3 の例では,「ギターの練習」とクエリが入力された場合,「スタジオ○○新宿店」が開始ノードとなる．こうした地物が複数あった場合には，その全てをランダムジャンプ先にする．以上のように開始ノードを設定することにより，クエリの長さに左右されず，より正確な検索が可能になる．

開始ノードを表す $|L|+|W|$ 次元ベクトル $\boldsymbol{r}$ は，$|R|$ が 1 のとき，r_1 番目の要素のみが 1，それ以外の要素が 0 とする．また，$\boldsymbol{r}$ は $|R|$ が 2 以上のとき，$N_l(R)$ 内の要素番目を 1，それ以外の要素を 0 とし，正規化したベクトルとする．以上で述べた記号を使用すると，提案するアルゴリズムの RWR は，

$$\boldsymbol{p} = (1-c)\boldsymbol{M'p} + c\boldsymbol{r} \tag{6}$$

で表される．$\boldsymbol{p}$ の初期値には $\boldsymbol{r}$ を用いる．$\boldsymbol{p}$ の要素が収束するまで，この計算を再帰的に繰り返す．最後に得られたベクトル $\boldsymbol{p}$ の要素 $\boldsymbol{p_u}$ をそれぞれ，クエリに対してのノード u の関連度とする．

こうして計算したベクトル $\boldsymbol{p}$ を用いて，すべての $l_i \in L$ を $\boldsymbol{p_i}$ の値で降順で順位付けし，地物ノードのうち上位から数件分を，検索結果のランキングとして出力する．

4 評価実験

Google Map の実データを用いた被験者実験で，手法の有用性を評価した．あらかじめ用意した 9 個の目的クエリについて，提案手法を含めた 5 手法で検索を行い，人手で評価した．評価の際には，ランキング上位に登場した地物に対して，人手で 1 件ずつウェブ検索を行い，公式サイトなどの情報から目的を達成可能かどうか判定した．

4.1 データセット

実験のため，Google Map の Places API で収集した地物のレビューデータを用いた．まず，Google Map の Places API を用いて地物とそれに紐づいたレビュー情報を収集した．Google Map では API の制限により，一定期間で収集できるデータには量的な限界がある．そこで，検索対象となる地域を，新宿区，渋谷区，千代田区を中心とした約 80km^2 に限定し，そこに含まれるすべての地物を収集した．収集の際には，あるエリアに含まれる地物と，それぞれの地物のレビューについて，個別の API を経由して収集する必要があった．Google Find Place API でエリアに含まれる地物を収集する際に，1 つの緯度経度で指定したエリア内では，上位 32 件までの結果しか取得できない．そのため，含まれる地物の件数が上限に達した範囲に対して，その範囲を 4 分割して再帰的に API を呼び出した．最終的に，25m 四方まで範囲を狭めることで，261,492 件の地物が得られた．これらの地物へのレビューを，Place Details API で収集した．API の制限により，それぞれの地物に対するレビューは上位 5 件のもののみ得られた．収集した情報のうち，文章を伴ったレビューが 1 件以上付いた地物は 85,942 件であった．

4.2 実　　装

収集した Google Map 内の地物 85,942 件分のレビューデータについて，MeCab と分かち書き辞書である mecab-ipadic-NEologd を用いて単語に分割した．この際，使用する品詞は動詞，名詞，形容詞に限定し，動詞は標準形に統一した．また，登場頻度をもとに，めったに使われない語と，多く登場しすぎる語を除去した．今回は，全 85,942 件分の地物のレビュー中，登場地物数が 50 件に満たない単語と，4 割以上の地物のレビューに登場した単語を削除した．最終的に 9,816 件の単語をグラフのノードとした．

次に，地物同士の類似度計算をあらかじめ計算した．地物同士の類似度の計算には，Google Map の地物に最大で 5 個まで付与されるメタデータである，カテゴリタグを用いた．収集した地物に付与された全 99 種類のカテゴリのうち，頻出のもの（すなわち，establishment と point_of_interest）を除く 97 カテゴリを使用して，ブール値からなるベクトルを生成した．このベクトルから，ベクトル空間内でのコサイン類似度を求めた．本研究では，計算量の問題から，使用する類似度は 3 つ以上カテゴリを持った地物同士で，そのベクトルが完全に一致するもののみを用いた．

同様に，単語動詞の類似度についても事前に計算した．提案手法では，意味的に似た目的を考慮するため，グラフ内の単語同士を仮想的なエッジで結んだ．そのために，9,816 件の単語ノードについてすべての組合せの単語同士の類似度を計算した．今回は，Wikipedia のデータで学習した Word2Vec のモデルを用いた．Word2Vec は Python のトピック分析ライブラリである gensim による実装を用いた．計算量の問題から，行列を疎行列として保つため，今回は類似度 0.5 以上の組み合わせのみを採用し，それ以外の組み合わせの類似度は 0 として扱った．

最後に，実際の Random Walk with Restart の計算を行い，クエリと地物の適合度を計算した．地物と単語からなる 95,758 次元の正方行列の計算を高速に行うため，Python のライブラリである Scipy を用いた．

実験の際には，地物同士の類似度の重み α と，単語同士の類似度の重み β を，手法ごとにそれぞれ人手で設定した．また，

計算量の問題で，再帰回数は 10 回に固定した．また，開始ノードへランダムジャンプする確率を 0.25 とした．

4.3 比較対象

提案するアルゴリズムは，3 つの仮説から成り立つ．各仮説の有用性を評価するために，

- **提案手法** （仮説 1+2+3）
  3 つの仮説すべてを適用した提案手法
  ($\alpha = 0.1, \beta = 0.1$)，
- **地物のみ拡張** （仮説 1+2）
  相互再帰に地物の拡張を加えた比較手法
  ($\alpha = 0.1, \beta = 0$)，
- **単語のみ拡張** （仮説 1+3）
  相互再帰に単語の拡張を加えた比較手法
  ($\alpha = 0, \beta = 0.1$)，
- **拡張なし** （仮説 1）
  単語と地物の相互再帰のみによる比較手法
  ($\alpha = 0, \beta = 0$)
- **クエリの完全一致**
  レビュー中に直接目的が登場するもののみ発見するベースライン手法

の 5 つの手法で検索結果をあらかじめ作成した．

4.4 結果のラベル付け

あらかじめ用意した 9 個のクエリ（表 1 に記載）について，提案手法を含めた 5 手法で検索を行い，結果として出力されたランキングを評価した．検索結果の 20 位以内ににランクされた地物に対して，クエリとして入力した目的を達成可能かどうか，人手で 2 値によるラベル付けを行った．クエリとして入力した目的を達成可能である地物を正解地物として 1 を，クエリとして入力した目的を達成可能である地物を不正解地物として 0 を，それぞれ評価値としてラベル付けした．この際，クエリの完全一致のみによるベースライン手法はランク付けされていないため，ランダムに抽出した 20 件を評価した．

用意したクエリの目的が達成可能か否かは，人によって評価が変わるものではないため，ラベル付けは一人の被験者が行った．また，本研究では時期や時間の考慮をしていないため，ある地物で，入力された目的を達成可能である時期や時間が限られる場合にも，その地物は正解とした．例えば,「泳ぐ」というクエリで検索を行って，検索結果の上位に現れた「プール」が夏季にのみ開放されていたとして，そのプールは正解として扱う．また，データを収集してから評価するまでの間に閉店した店舗についても，かつてその場所で目的を達成可能だった場合には正解とした．

4.5 実験結果

手法ごと，クエリごとの適合率とランキング評価，実際の出力について述べる．表 1 に，実験に使用した 9 つのクエリによって得られた適合率 (p@k) と nDCG を示す（ただし，ベースラインである**クエリの完全一致**はランキングではなくブール検索であるため，nDCG は計算できない）．全クエリの平均結果について，適合率と nDCG の両方で，**地物のみ拡張**が最も高い評価となった．**提案手法**は，適合率において**クエリの完全一致**の値を上回り，nDCG において**拡張なし**の値を上回った．

提案手法が最も高精度と評価されたのは，クエリが「アフタヌーンティーをする」,「ピザを買う」のときであった．これらのクエリでは，**クエリの完全一致**と**拡張なし**による適合率を大きく上回った．

「パソコンの購入」を入力とした際の結果では，全ての手法について軒並み適合率が低かった．相互再帰を用いた 4 つの手法では，**提案手法**と**地物のみ拡張**で，**拡張なし**と**単語のみ拡張**の適合率，nDCG，正解数のすべてを上回った．**単語のみ拡張**は，拡張しなかった場合よりも，全ての評価指標で下回った．

5 考　　察

得られた結果から，提案手法の性質について議論する．全体を通して，適合率と nDCG の両方で，**地物のみ拡張**が最も有効であった．また，地物レベルでの拡張に単語レベルでの拡張を追加した**提案手法**でも，適合率において**クエリの完全一致**の値を上回り，nDCG において**拡張なし**の値を上回った．その上で，**クエリの完全一致**により提示された地物以外の正解を新たに発見したクエリも多数あり，**地物のみ拡張**や**提案手法**にもそれぞれ有用性があることが明らかになった．ベースラインである**クエリの完全一致**の適合率が低すぎるものや高すぎるクエリに対して，**提案手法**は相対的に有効でなかった．

地物レベルでの拡張が有効に働いたクエリとして「パソコンの購入」などがあった．適合率が向上した理由は，家電量販店であれば PC の購入が可能というカテゴリによる推論が行われ，上位に多くの家電量販店が登場したためだと考えられる．ほかに「BBQ をする」のクエリでも，**地物のみ拡張**の適合率が最も高かった．このクエリではメニュー名に「BBQ ソース」などが入った飲食店が多く，ほかの手法では適合率が下がったが,「BBQ 場」というカテゴリ名による地物での拡張が有効に働き，不正解の地物の順位を相対的にを下げることができた．

クエリによっては，適合率が下がっていても，拡張前には発見できなかった目的を達成可能な地物を発見できている場合があった．クエリ「ギターの練習」では，レビュー中に直接（ギター，練習）の語を含む地物ノード内の正解地物は 3 件であり，その全てが音楽教室であった．**拡張なし**では，より多くの，ギターの練習ができる音楽教室がヒットした．ここから，**提案手法**では，適合率は下がったものの,「島村楽器 新宿 PePe 店」などのギターのレッスンを行っている，または弾くスペースが併設されている楽器店を発見可能であった．これらの地物は，単語レベルと地物レベル両方での拡張が合わさってはじめて上位に順位付け可能だったと考えられる．

そもそも目的クエリを全て含むレビューが多くあり，その中に正解が多数含まれるようなクエリに対しては，ベースライン手法が十分に有効であり，提案手法が相対的に有効でなかった．

単語レベルでの拡張が有効であった例と有効でなかった例について考察する．全手法の中で，全てのクエリの平均値では，

表 1　9 つのクエリを各 5 つの手法で検索した評価結果

	提案手法		地物のみ		単語のみ		拡張なし		完全一致	
	p@20	nDCG	p@20	nDCG	p@20	nDCG	p@20	nDCG	p@20	発見数
ギターの練習	0.30	0.40	0.35	0.43	0.40	0.54	**0.54**	**0.57**	0.15	4
パソコンの購入	**0.45**	**0.59**	**0.45**	**0.59**	0.35	0.38	0.40	0.41	**0.45**	49
パソコンの修理	0.70	0.76	**0.75**	**0.79**	0.70	0.76	**0.75**	**0.79**	0.65	13
ピザを食べる	0.75	0.64	0.80	0.68	0.85	**0.84**	0.80	0.81	**0.95**	466
ピザを買う	**0.80**	**0.87**	0.75	0.84	0.70	0.68	0.70	0.68	0.65	32
魚を釣る	**0.25**	0.27	**0.25**	0.28	**0.25**	**0.35**	**0.25**	0.32	**0.25**	23
BBQ をする	0.70	0.66	**0.75**	**0.68**	0.60	0.58	0.50	0.48	0.30	124
アフタヌーンティーをする	**0.90**	**0.94**	**0.90**	**0.94**	**0.90**	0.79	0.80	0.76	0.75	91
泳ぐ	0.05	0.03	0.20	0.14	0.15	0.10	**0.25**	**0.21**	0.20	78
平均	0.54	0.57	**0.58**	**0.60**	0.54	0.56	0.54	0.56	0.48	-

単語のみ拡張による適合率と nDCG は，**拡張なし**による値からほぼ変化がなかったが，クエリによって効果に差があった．単語レベルでの拡張が有効に働いた例について述べる．「ピザを食べる」のクエリで検索した結果では，**クエリの完全一致**を除いて，**単語のみ拡張**を使用した手法で最も高い適合率となった．これは，「パスタ」などの「ピザ」と意味的に近い単語で拡張した場合に，「パスタを食べる」ことのできる場所では高確率で「ピザを食べる」こともできるので，単語レベルでの拡張が有効に働いたと考えられる．このように，意味の近い単語でも目的を達成できる地物が大きく変化しない場合には，単語レベルでの拡張が有効であることがわかった．

6　まとめと今後の課題

本研究では，目的を達成可能な地物をランキングとして出力する検索アルゴリズムを提案した．従来の地物情報検索では，探したい地物の業種や特徴を入力としており，「ギターの練習」のできる場所など，目的から地物を発見することが難しかった．そこで Google map などの地物レビュー情報を用い，3 種類の仮説からなる拡張を行うことで，目的を達成できる地物を直接検索可能にした．Random Walk with Restart によるグラフ処理を用いた手法に基づくウェブアプリケーションを実装し，被験者実験を行うことで，提案手法がより多くの地物を発見可能であることを示した．

今後の課題として，高精度化が必要である．また，本手法はクエリを入力するたびに，収束計算を伴うグラフ処理を行う必要がある．実際にサービスとして運用するためには，似た目的や似た地物をあらかじめまとめるなどの高速化が必要である．今後，このような検索を実際のウェブサービスとして実現するための，より発展的な研究を行う必要があると考えられる．

謝　　辞

本研究は JSPS 科研費 18K18161（代表:荘司慶行），18H03243（代表：田中克己）の助成を受けたものです．ここに記して謝意を表します．

文　　献

[1] Yoshiyuki Shoji, Katsurou Takahashi, Martin J Dürst, Yusuke Yamamoto, and Hiroaki Ohshima. Location2vec: Generating distributed representation of location by using geo-tagged microblog posts. In *International Conference on Social Informatics*, pp. 261–270. Springer, 2018.

[2] 加藤誠, 大島裕明, 小山聡, 田中克己. 地域コンテキストを考慮した動的な特徴空間に基づく地理情報例示検索. 情報処理学会論文誌, Vol. 52, No. 12, pp. 3448–3460, 2011.

[3] 安田宜仁, 戸田浩之. 検索位置のごく周辺を対象とした地理情報検索. 人工知能学会論文誌, Vol. 23, No. 5, pp. 364–373, 2008.

[4] 廣嶋伸章, 安田宜仁, 藤田尚樹, 片岡良治. 地理情報検索におけるクエリ入力支援のための特徴語の提示. 人工知能学会全国大会論文集 第 26 回全国大会 (2012), pp. 1C1R56–1C1R56. 一般社団法人 人工知能学会, 2012.

[5] Barak Pat, Yaron Kanza, and Mor Naaman. Geosocial search: Finding places based on geotagged social-media posts. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 231–234. ACM, 2015.

[6] Sandro Bauer, Filip Radlinski, and Ryen W White. Where can i buy a boulder?: Searching for offline retail locations. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 1225–1235. International World Wide Web Conferences Steering Committee, 2016.

[7] 笠原要, 松澤和光, 石川勉. 国語辞書を利用した日常語の類似性判別. 情報処理学会論文誌, Vol. 38, No. 7, pp. 1272–1283, 1997.

[8] 王玉馨, 清水伸幸, 吉田稔, 中川裕志. 単語類似度ネットワークを通じた自動同義語獲得. 情報処理学会研究報告音声言語情報処理 (SLP), Vol. 2008, No. 46 (2008-SLP-071), pp. 7–14, 2008.

[9] Suppanut Pothirattanachaikul, Takehiro Yamamoto, Sumio Fujita, Akira Tajima, Katsumi Tanaka, and Masatoshi Yoshikawa. Mining alternative actions from community q&a corpus. *Journal of Information Processing*, Vol. 26, pp. 427–438, 2018.

[10] 松村優也, 大島裕明, 田中克己. 行動名をクエリとした地理情報検索. 第 8 回データ工学と情報マネジメントに関するフォーラム (DEIM 2016) 会議録, pp. H5–6, 2016.

[11] Takeshi Kurashima, Taro Tezuka, and Katsumi Tanaka. Blog map of experiences: Extracting and geographically mapping visitor experiences from urban blogs. In *International Conference on Web Information Systems Engineering*, pp. 496–503. Springer, 2005.

[12] Hao Wang, Manolis Terrovitis, and Nikos Mamoulis. Location recommendation in location-based social networks using user check-in data. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 374–383. ACM, 2013.