Yahoo!知恵袋を利用した施設名の曖昧性解消手法の提案

中川　智也† 　新妻　弘崇†† 　太田　学†

†, †† 岡山大学大学院自然科学研究科　〒 700-8530 岡山県岡山市北区津島中 3-1-1
E-mail: †{nakagawa，ohta}@de.cs.okayama-u.ac.jp, ††niitsuma@cs.okayama-u.ac.jp

あらまし　近年，インターネットの普及に伴い，観光情報を気軽に発信・受信できるようになった．新井らは観光ルート推薦のため，予め収集した施設名を含む Yahoo!知恵袋の質問数を使ってその施設が観光スポットとして適切か判定した．しかし，注目している施設名と同名の他のエンティティがノイズとなり判定を誤ることがあった．そこで，本研究では doc2vec を用いて，Yahoo!知恵袋の質問が当該施設について述べているものかどうかを判定する手法を提案する．提案手法を用いて，岡山県内の施設名を含む質問を，当該施設に関する質問と同名の他のエンティティに関する質問に分類する評価実験を行った．岡山県内の施設名を含む質問 64 件と同名の他のエンティティに関する質問 64 件をテストデータとすると，その 86.72%は正しく判定された．このとき，当該施設に関する質問の 93.75%は正しく判定されたが，同名の他のエンティティに関する質問は 79.69%しか正しく判定されなかった．

キーワード　paragraph vector，Yahoo!知恵袋，観光スポット

1. はじめに

近年，スマートフォンなどの通信機器の普及率が 9 割を超え (注1)，インターネット上の情報にアクセスできる人口が大幅に増加している．こうした多くの人に 利用されているインターネット上 の情報の一つとして，本研究では観光体験情報に注目する．

新井らは観光ルート推薦のため，Twitter(注2) に投稿されたツイートの中から観光スポットに関するものを収集し分析した．収集したツイートの投稿時刻やユーザごとのタイムラインの違いに注目することで，さまざまな観光スポットの訪問時間帯や旅行者の傾向を分析した．この分析結果を利用して，新井らは観光ルートの生成を行った．また，観光ルートを生成するためには，観光スポットのリストを事前に準備する必要がある．新井らは観光スポットのリストを，Google Place API(注3) を使って，観光したいエリア周辺の施設名を収集することで作成した．この時，病院や駅など観光スポットではない施設名も収集されるため，Yahoo!知恵袋の質問検索 API(注4) を用いて収集した施設が観光スポットとして適切か判定した．

しかし，新井らの判定方法では，施設名をクエリとして検索した質問数を用いるため，施設名と同名のエンティティがノイズとなり判定を誤ることがある．例えばレストラン「小樽」を検索したときに北海道の観光地である小樽に関する質問がヒットし，雑貨店「パスポート」を検索した時に旅券に関する質問がヒットするため，判定を誤ることがある．

そこで，本研究では paragraph vector の実装の 1 つである doc2vec [1] を用いて質問が本当に注目している施設に関する質問かどうかを判定する手法を提案する．また，この手法の評価実験を岡山県内の観光スポットについて行う．

2. 関連研究

2.1 インターネット上の情報を用いた観光推薦に関する関連研究

新井ら [2] は観光スポットのリストを予め用意し，観光スポットについて呟いているツイートの中から，実際に観光スポットを訪れていると判別できるツイートを選別し，それらを利用して観光ルートを推薦する手法を提案した．彼らは収集したツイートから各観光スポットについて 3 種類のスコアを求めて，これらのスコアを用いて推薦する観光ルートを決定した．

倉島ら [3] は，写真共有サイトのジオタグ情報を人々の旅行履歴として利用したトラベルルート推薦手法を提案した．倉島らの手法は，場所間の移動しやすさを考慮するマルコフモデルと旅行者の興味を考慮するトピックモデルを確率論的枠組みの中で結合して次に移動する場所を予測し，最良優先探索に基づいて効率的に現在地からの推薦ルートを生成する．

藤坂ら [4] は位置情報付きツイートを用いて観光情報の抽出を試みている．彼らは，K-means 法により分割した日本の各領域に対して，ツイート数，Twitter ユーザ数，Twitter ユーザの移動量からノーマルパターンを規定し，ある時間におけるツイート数，Twitter ユーザ数，Twitter ユーザの移動量をそれぞれのノーマルパターンと比較することで地域イベントが行われている領域を検知した．

石野ら [5] は，ANPI NLP(注6) で提供される震災情報に関わるツイートを利用して，被災時における避難経路を自動抽出す

(注1)：総務省 ICT サービスの利用動向，http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h28/html/nc252110.html
(注2)：Twitter，http://twitter.com/
(注3)：Google Place API，https://developers.google.com/places/?hl=ja
(注4)：Yahoo!デベロッパーネットワークトップ，http://developer.yahoo.co.jp/webapi/chiebukuro/chiebukuro/v1/questionsearch.html
(注6)：ANPI NLP:メインページ，http://trans-aid.jp/ANPI_NLP/index.php/メインページ

る手法を提案した．石野らは機械学習を用いて，移動元，移動先，移動手段のタグを東日本大震災に関連するツイートに自動付与することによって，被災者の行動経路を抽出した．

郡ら [6] は行動計画の立案支援として，ブログからユーザの旅行時の代表的な経路とその文脈を抽出し，地図上にマッピングして提示するシステムを提案した．郡らの手法では，ブログ内に出現する各地名に対して，旅行者が実際にその場所を訪れたかどうかを文脈から判定し，訪れたと判定した場合はその地名をルート要素とする．その後，ルート要素に対して順序付けを行い，地名の系列パターンを抽出した．

Lee ら [7] は目的関数として電気自動車の充電待ち時間を最小化する観光ルートを遺伝的アルゴリズムを用いて推薦した．Lee らの手法は旅行者の観光ルートの中で代表的な観光スポットを出発点として観光ルートの突然変異や進化を行ない，より良いルートを探索した．

三富ら [9] は Twitter と Flickr の投稿の位置情報を比較することで，FreeWiFi がない観光スポットを発見する手法を提案した．ツイートを発信するときはデータ通信が必要なため，ツイートに位置情報が付与されている場所は FreeWiFi がある．Flickr に投稿されている写真に位置情報を付与するときはデータ通信は特に必要なく，観光スポットで撮られる写真に位置情報が付与されやすいことから位置情報が付与されている写真が多い場所は観光スポットとした．以上のことから位置情報が付与されている写真が多い観光スポットの中からツイートの少ない観光スポットを探して，FreeWiFi がない観光スポットを発見した．

2.2 曖昧性解消問題に関する関連研究

木村ら [10] は検索エンジンを用いて人名を検索した際に，同姓同名の人物がヒットする問題において，検索対象の人物の組織名や肩書に注目して検索結果を人物単位に分類した．木村らは，検索タイトルとスニペットから組織名や肩書などのタームを抽出し，それぞれのタームに重み付けを行い，文書ベクトルを作成した．そして，文書間の類似度を利用してクラスタリングを行った．

片岡ら [11] は木村らと同じ問題において，2 段階のクラスタリングを行う手法を提案した．1 段階目は Web ページの中に現れる他の人物を比較してクラスタリングを行い，2 段階目は Web ページに含まれる組織名や地名等の特徴語を用いてベクトルを作成して，再度階層型クラスタリングを行う．

落合ら [12] はマイクロブログを対象に地名の曖昧性解消手法を提案した．落合らは，観光案内や Wikipedia(注7) から抽出した季節変動などに依存しない静的特徴語と，マイクロブログから抽出した既に曖昧性が解消された季節変動などに依存する動的特徴語，これらと地名の共起を用いて地名の曖昧性を解消した．

3. 新井らの観光スポット収集手法

本節では新井らの観光スポット収集手法 [2] について説明する．新井らの手法では，まず最初に注目する観光地の中心となる場所を決める．例えば京都を観光したいなら京都駅，岡山を観光したいなら岡山駅など，注目する観光地の中心スポットを決める．次に，Google Places API を用いて，選択した中心スポット，例えば岡山駅の周辺の施設の情報を 20 件取得し，キューに入れる．次に，キューの先頭の施設を取り出し，この施設の周辺の施設を新たに収集し，未収集の施設をキューに追加する．これをキューが空なるまで繰り返す．こうして収集した施設名には岡山市立市民病院のような観光地として適当ではない地名も含まれているため，このような施設名を以下で説明する方法で Yahoo!知恵袋を使って除外する．

Yahoo!知恵袋はカテゴリを指定して質問を投稿したり検索することが出来る．そこで観光スポットとして適切かどうか調べる施設名をクエリとして「地域，旅行，おでかけ」のカテゴリ下の「国内」カテゴリでの質問数を数えることで，観光スポットかどうかの判断基準とした．このとき，質問数が 10 件以下の施設は観光スポットではないとみなす．次にカテゴリを指定せず全カテゴリでの質問を数える．そして式 (1) のようにして関連度を求める．

$$\text{関連度} = \frac{\text{「国内」カテゴリの質問数}}{\text{全カテゴリでの質問数}} \tag{1}$$

この関連度が高いほど注目している施設が観光スポットとしての人々の関心が高いとみなし，この関連度が閾値以上のものを観光スポットと判定する．

しかし，同名の施設が他の場所にも存在したり，一般名詞と同名の施設もある．このような施設名を含む質問数を数えると，同名のエンティティがノイズとなり，観光スポットではない施設を観光スポットと判定してしまうことがある．

4. 施設名の曖昧性解消手法

本研究では 3 節で説明した同名エンティティの曖昧性を解消するために，paragraph vector [15] の実装の 1 つである doc2vec を用いて質問の特徴を数値ベクトル化する．さらに，そのベクトルを SVM で 2 値分類して質問が注目している施設に関する質問か判定する．

本節では 4.1 節で doc2vec について説明し，4.2 節で具体的に分類手法を説明する．

4.1 word2vec と doc2vec

word2vec とは Mikolov ら [13] [14] が提案したニューラルネットワークで，単語の特徴を低次元な数値ベクトルで表現する手法である．word2vec は，文章中のある単語を周辺の単語の前後関係から予測する問題をニューラルネットワークで学習し，そのニューラルネットワークの中間層を単語の特徴ベクトルとして抽出する手法である．また，word2vec の生成する単語ベクトルはベクトル同士の演算ができ，その演算によって単語間の意味の関係を表わすことができる．例えば以下のような計算式が成り立つ．

$$king - man + woman = queen \tag{2}$$

式 (2) は king から man を引き，woman を足すと queen のベ

(注7)：Wikipedia，https://ja.wikipedia.org/wiki/メインページ

クトルとなるということを示す．また，似たような意味を持つ単語をそれぞれベクトル化した場合，互いに似たようなベクトルが生成される．

paragraph vector は word2vec のベクトル化の対象を単語ではなく，文章に拡張したものである [15]．似たような文章からは似たようなベクトルが生成され，文章の特徴をベクトル化することが出来る．例えば，渋谷の待ち合わせ場所として有名なハチ公像のレビュー文書をベクトル化したものは，渋谷のモヤイ像や上野公園の西郷隆盛像のレビュー文書をベクトル化したものとのコサイン類似が高くなる [16]．モヤイ像はハチ公像と同じ渋谷にあり，モヤイ像も待ち合わせ場所として知られており，渋谷の地名や待ち合わせに関する単語がレビュー中で共通するため類似度が高くなる．西郷隆盛像は所在地は渋谷ではないが，こちらも待ち合わせ場所として知られており，待ち合わせに関する単語がレビュー中で共通するために類似度が高くなる．

4.2 曖昧性のある施設に関する質問の判定手法

本研究では，paragraph vector が似たような単語が含まれている文章からは似たようなベクトルを出力することに注目する．まず，paragraph vector を用いて質問をベクトル化する．さらに，曖昧性のない施設に関する質問のベクトルを用いて paragraph vector を分類する SVM を学習することで，観光したいエリアの施設の質問に共通する特徴を取得する．次に，学習した SVM で曖昧性のある施設に関する質問を 2 値分類して質問が注目している施設に関する質問か判定する．

具体的には，岡山県内にある施設について説明する文であるかないかを分類する SVM を学習し，この学習した SVM を使って曖昧性のある施設の分類を行なう．本研究では，正例のみで学習を行う One Class SVM と正例と負例を学習に使う標準的な SVM の 2 種類で分類を行った．まず，3 節で説明した手法で収集した岡山県内にある施設についての Yahoo!知恵袋の質問を doc2vec を用いてベクトル化する．次に，岡山県内の施設の中から曖昧性の無い観光スポットを人手で決める．さらに，岡山県以外にある曖昧性の無い観光スポットを人手で決め，同様に Yahoo!知恵袋の質問を集めてベクトル化する．

人手で決めた曖昧性の無い岡山県内の観光スポットに関する質問のベクトル 1,000 件を正例とし，通常の SVM を用いる場合は岡山県以外にある曖昧性の無い観光スポットに関する質問のベクトル 1,000 件を負例として SVM を学習する．このとき，最終的に岡山県内の観光スポットに関する質問とそれ以外の都道府県の観光スポットに関する質問に分類したいので，負例は岡山県以外の複数の観光スポットの質問のベクトルを含むようにする．これらの SVM を用いて曖昧性のある施設に関する質問のベクトルを 2 値分類することで，その質問が岡山県にあるその施設に関するものかを判定する．

このようにしてその施設に関するものであると分類された質問のみを使って式 (1) の関連度を求めれば，観光スポットの判定がより正確になる．

表 1　市町村別のスポット数 (S_1) と新井らの方法で判定した観光スポット数 (S_2) と人手で判定した観光スポット数 (S_3)

	S_1	S_2	S_3
岡山市	37,455	179	37
倉敷市	20,610	85	22
津山市	6,342	26	7
真庭市	2,902	36	18
玉野市	2,698	18	5
その他	23,378	156	65
計	93,385	500	154

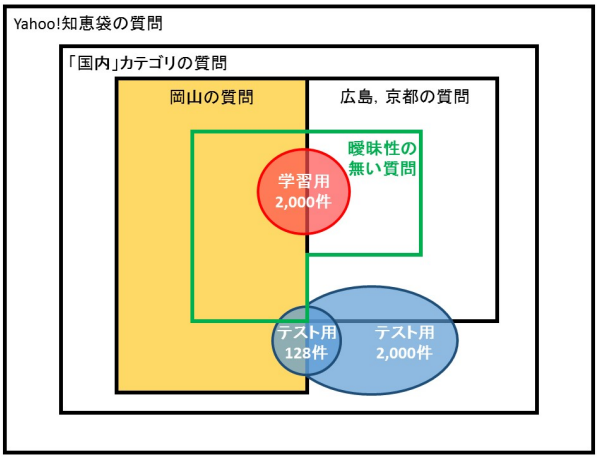

図 1　実験に用いる質問

5. 評価実験

5.1 新井らの手法により収集した観光スポット

3 節で示した新井らの方法で施設名の収集と観光スポットの判定を行った．岡山駅を中心スポットとして Google Places API で収集した岡山県内のスポット数 (S_1)，3 節で説明した Yahoo!知恵袋を用いて観光スポット候補か判定する方法で，観光スポットと判定されたスポット数 (S_2)，S_2 のスポット候補からさらに人手で観光スポットと判定したスポット数 (S_3) を岡山県内の市ごとに表 1 にまとめる．なお，観光スポット候補の判定に用いる閾値は新井ら [2] と同じ 0.27 とした．

表 1 に示す通り，新井らの方法で観光スポットと判定した 500 件を人手で観光スポットかどうか判定するとそのうちの 154 件しか正しくない．この適合率は 0.31 で，正確な選定とはいえない．

5.2 観光スポット名の曖昧性解消実験

図 1 の赤色の質問を用いて SVM の学習を行い，青色の質問が黄色エリアに含まれるか分類するテストを行う．青色のデータを提案手法により判定した結果と人手で判定した結果を比較して，式 (3) で分類精度を測る．

$$\text{分類精度} = \frac{\text{提案手法で正しく分類した質問数}}{\text{人手で分類した質問数}} \tag{3}$$

5.2.1 SVM の学習

本研究ではまず，SVM を学習するために曖昧でない観光スポット名を含む質問を表 2 の通り収集した．この質問のうち，

表 2　分類実験のために収集した質問

岡山県		広島県		京都府	
スポット	質問件数	スポット	質問件数	スポット	質問件数
岡山城	282	原爆ドーム	1,000	八坂神社	1,000
岡山後楽園	57	厳島神社	1,000	北野天満宮	1,000
美観地区	637	宮島水族館	136	天橋立	1,000
三井アウトレットパーク倉敷	19	広島城	557	清水寺	1,000
大原美術館	121	広島平和記念資料館	42	金閣寺	1,000
計	1,116	計	2,735	計	5,000

表 3　One Class SVM を用いた 128 件の質問の分類結果

		人手による判定	
		当該施設	その他
SVM による分類	当該施設	25	13
	その他	39	51
計		64	64

表 4　One Class SVM を用いた 2,000 件の質問の分類結果

		人手による判定	
		当該施設	その他
SVM による分類	当該施設	25	543
	その他	39	1393
計		64	1936

表 5　SVM を用いた 128 件の質問の分類結果

		人手による判定	
		当該施設	その他
SVM による分類	当該施設	60	13
	その他	4	51
計		64	64

表 6　SVM を用いた 2,000 件の質問の分類結果

		人手による判定	
		当該施設	その他
SVM による分類	当該施設	60	479
	その他	4	1,457
計		64	1,936

正例として岡山県の観光スポットに関する質問 1,116 件の中から無作為に選んだ 1,000 件を用意した．また，負例として広島県の観光スポットに関する質問 2,735 件の中から無作為に選んだ 500 件，京都府の観光スポットに関する質問 5,000 件の中から無作為に選んだ 500 件を用意した．

5.2.2　実験に用いるテストデータ

本項では，評価実験に用いるテストデータについて説明する．表 1 の岡山県内の施設 (S_1)93,385 件のうち,「国内」カテゴリの質問数が 10 件以上ある施設 1,593 件に関する質問を分類する．なお，この「国内」カテゴリの質問の中には同名エンティティに関する質問が含まれていることもある．まず，1,593 件の施設に関する全質問 369,132 件から 2,000 件の質問を無作為に選出し，その 2,000 件を人手で判定した．その結果，当該施設の質問は 64 件，同名のエンティティの質問は 1,936 件と偏りがあった．そのため，さらに同名のエンティティの質問 1,936 件から 64 件を無作為に選出し，これらと当該施設の質問 64 件の計 128 件の質問集合を用意した．実験ではその 2,000 件の分類とこの 128 件の分類の 2 種類を行う．

5.2.3　One Class SVM によるテストデータの分類

本項では学習データの正例 1,000 件のみを使って学習した One Class SVM を用いて，5.2.2 項で説明したテストデータの分類を行う．

最初に，128 件の質問を 2 値分類し，人手による判定と比較して分類精度を測る．分類結果を表 3 に示す．128 件のうち 76 件が正しく分類され，式 (3) で全体の分類精度を求めると 59.38%となった．また，当該施設の質問は 39.06%，同名エンティティの質問は 79.69%の分類精度となった．

次に，2,000 件の質問を 2 値分類し，人手による判定と比較して当該施設と同名エンティティの質問それぞれの分類精度を測る．分類結果を表 4 に示す．2,000 件のうち 1,418 件が正しく分類され，式 (3) で全体の分類精度を求めると 70.9%となった．当該施設の質問は 64 件のうち 25 件が正しく分類され，分類精度は 39.06%となった．同名のエンティティの質問は 1,936 件のうち 1,393 件が正しく分類され，分類精度は 71.95%となった．

5.2.4　標準的な SVM によるテストデータの分類

本項では学習データの正例 1,000 件と負例 1,000 件を使って学習した SVM を用いて，5.2.2 項で説明したテストデータの分類を行う．

最初に，128 件の質問を 2 値分類し，人手による判定と比較して分類精度を測る．分類結果を表 5 に示す．128 件のうち 111 件が正しく分類され，式 (3) で全体の分類精度を求めると 86.72%となった．また，当該施設の質問は 93.75%，同名エンティティの質問は 79.69%の分類精度となった．

次に，2,000 件の質問を 2 値分類し，人手による判定と比較して当該施設と同名エンティティの質問それぞれの分類精度を測る．分類結果を表 6 に示す．2,000 件のうち 1,517 件が正しく分類され，式 (3) で全体の分類精度を求めると 75.85%となった．当該施設の質問は 64 件のうち 60 件が正しく分類され，分類精度は 93.75%となった．同名のエンティティの質問は 1,936 件のうち 1,457 件が正しく分類され，分類精度は 71.95%となった．

5.2.5　考　　察

5.2.3 項において，正例のみで学習した One Class SVM による分類では岡山県内の施設の質問の分類精度は 39.06%，同名のエンティティの質問の分類精度は 71.96%となった．岡山県内の施設の質問の分類精度は 39.06%と低く，One Class SVM では岡山県内の施設に関する質問の特徴を上手く取得できな

かった．

一方，5.2.4 項において，正例と負例で学習した SVM による分類では岡山県内の施設の質問の分類精度は 93.75%，同名のエンティティの質問の分類精度は 75.26%となった．どちらも One Class SVM による分類より高い精度となり，特に岡山県内の施設の質問については大きく上昇し，岡山県内の施設に関する質問の特徴を取得できた．また，SVM による分類実験では岡山県以外の施設の質問は京都府と広島県の施設に関する質問であるため，それ以外の都道府県ある施設の質問は考慮されない．例えば,「円通寺」は岡山県の他に京都府に存在するが,「円通寺」を含む質問 25 件を人手で判定すると，岡山県の「円通寺」についての質問が 1 件，京都府の「円通寺」についての質問が 24 件で，提案手法を用いると 100%の分類精度となった．しかし,「妙立寺」は岡山県の他に石川県に存在するが,「妙立寺」を含む質問 150 件を人手で判定すると，150 件すべて石川県の「妙立寺」についての質問で．提案手法を用いると 52.7%の分類精度となった.「円通寺」は高い精度で分類できたことから，SVM の学習に石川県の施設に関する質問を追加すれば「妙立寺」を含む質問の分類精度の向上が期待できる．

6. ま と め

本研究では，観光スポットとなる施設名の曖昧性を解消する手法を提案した．具体的には，Yahoo!知恵袋に投稿された質問からそのような曖昧な施設名を含む質問を選択し，岡山県内の施設についての質問とその他の質問に分類した．提案手法では doc2vec を用いて質問をベクトル化し，正例のみで学習する One Class SVM，正例と負例で学習する SVM を用いて質問が岡山県内の当該施設に関する質問かどうかをそれぞれ判定した．

提案手法について評価実験を行った結果，正例のみで学習した One Class SVM は 59.38%，正例と負例で学習した SVM は 86.72%の精度で分類された．One Class SVM は岡山県内の施設の質問の分類精度が 39.06%と低く，正しく分類出来なかった．SVM は岡山県内の施設の質問のみ，同名のエンティティの質問のみの分類精度はそれぞれ，93.75%，75.26%と，どちらも One Class SVM より高くなった．また，同名のエンティティの質問に関しては，SVM の学習に負例として使用した施設と同じ都道府県にある施設に関する質問は高い精度で分類できた．

今後の課題として，SVM の学習時のパラメータを調整することで，同名のエンティティの質問の分類精度を上げることが挙げられる．

文　献

[1] Radim Řehůřek , Petr Sojka,“Software Framework for Topic Modelling with Large Corpora”，Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45-50，2010.

[2] 新井晃平，新妻弘崇，太田学，“ Twitter を利用した観光ルート推薦の一手法”，第 7 回データ工学と情報マネジメントに関するフォーラム，pp. 1-8，2015.

[3] 倉島健，岩田具治，入江豪，藤村考，“ 写真共有サイトにおけるジオタグ情報を利用したトラベルルート推薦 (不均質なライフログからのデータマイニング及び一般)”，電子情報通信学会技術研究報告．LOIS，ライフインテリジェンスとオフィス情報システム，Vol. 109，No. 450，pp. 55-60，2010.

[4] 藤坂達也，李龍，角谷和俊，“ 地域イベント発見のためのジオタグ付マイクロブログを用いたノーマルパターン検出手法”，平成 22 年度情報処理学会関西支部大会，Vol. 2010，2010.

[5] 石野亜耶，小田原周平，難波英嗣，竹澤寿幸，“ Twitter からの被災時の行動経路の自動抽出および可視化”，言語処理学会 第 18 回年次大会，pp. 907-910，2012.

[6] 郡宏志，服部峻，手塚太郎，田島敬史，田中克己，“ブログからのビジターの代表的な行動経路とそのコンテキストの抽出”，電子情報通信学会技術研究報告，Vol. 106，No. 149，pp. 29-34，2006.

[7] Lee, j., Kim, S.-W. and Park, G.-L, “ A tour recommendation service for electric vehicles based on a hybrid orienteering model”，Proceedings of the 28th Annual ACM Symposium on Applied Computing, SAC ’13，New York, NY，USA，ACM，pp. 1652-1654，2013.

[8] 永田祐一，“ 多点探索の最前線 -巡回セールスマン問題に対する遺伝的アルゴリズムの適用-”，日本オペレーションズ・リサーチ学会 2014 年秋季シンポジウム”，2014.

[9] 三富恵佑，遠藤雅樹，江原遥，廣田雅春，横山昌平，石川博，“外国人にアクセシブルな FreeWiFi がない観光スポットの発見”，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，pp. 1-6，2016.

[10] 木村塁，戸田浩之，田中克己，“ 検索結果スニペットのクラスタリングによる同姓同名人物の特定”，電子情報通信学会第 17 回データ工学ワークショップ，pp. 1-8，2006.

[11] 片岡真一，上田洋，村上晴美，辰巳昭治，“ 人物名に着目した二段階クラスタリングによる Web 上の同姓同名人物の分離”，人工知能学会全国大会論文集，pp. 1-2，2008.

[12] 落合桂一，鳥居大祐,“ 時間変化する特徴語によるマイクロブログ地名曖昧性解消”，情報処理学会論文誌，Vol. 7，No. 2，pp. 51-60，2014.

[13] Mikolov，T.，Sutskever，I.，Chen，K.，Corrado G. and Dean，J.，“Distributed representations of words and phrases and their compositionality”，Advances in Neural Information Processing Systems，pp. 3111-3119，2013.

[14] Mikolov，T.，Chen，K.，Corrado G. and Dean，J.“Efficient estimation of word representations in vector space”，arXiv preprint arXiv:1301.3781，pp. 1-12，2013.

[15] Quoc V, Le and Tomas Mikolov,“ Distributed Representations of Sentences and Documents”，In Proceedings of The 31st International Conference on Machine Learning, pp. 1188-1196，2014.

[16] 工学院大学インタラクティブメディア研究室，https://kitayamalab.wordpress.com/2016/11/14/python-と-gensim-で-doc2vec-を使う/