

多様な Web 情報を活用した地域観光情報基盤の高度化の検討

Analysis of Diverse Tourist Information Distributed across the Internet

土屋 健/Takeshi TSUCHIYA¹・三代沢 正/Tadashi MIYOSAWA²・山田 哲靖/Tetsuyasu YAMADA²・

広瀬 啓雄/Hiroo HIROSE²・澤野 弘明/Hiroaki SAWANO³・小柳 恵一/Keiichi KOYANAGI⁴

¹公立諏訪東京理科大学 工学部 准教授・²公立諏訪東京理科大学 工学部 教授

³愛知工業大学 情報科学部 准教授・⁴早稲田大学 理工学術院 教授

[Abstract]

Herein, we propose and discuss a new method for analyzing various types of tourist information about the Suwa area of Nagano Prefecture, Japan, available on the Internet. This information includes not only long sentences that can be found on web pages and in blogs, but also short sentences comprising a few words posted on social media. In this paper, we propose two novel methods; one is based on a paragraph vector use of neural network, called predict manner, and another is correlative manner for expressing relationships between words included in sentences. Our methods achieve high retrieval accuracy even across social media posts comprising just a few words. Based on our evaluation results, the proposed methods outperform the conventional information retrieval technique wherein sufficient accuracy cannot be achieved as it is based on the occurrence probability of words in sentences. This improvement is achieved by using the word order as an input feature to the neural network model.

[キーワード]

観光情報, SNS, 文脈の類推, ニューラルネットワーク, 単語の相関性

1. はじめに

筆者らは、これまでに所属大学がある長野県諏訪地域を対象とした地域観光情報ポータルサイト(図 1)を構築している [1]。本ポータルサイトでは、諏訪地域の観光に関する情報を blog, SNS などの web 上から収集・解析して提供している。この解析では、観光で訪れた人、観光を検討している人たち、そして観光協会などが発する諏訪地域の観光に関する話題、ホットキーワードが導出され、観光情報として本ポータルサイトで可視化される。情報の内容としては、訪れた人の体験、観光協会の提供するイベント情報、そして山や四季の移り変わりなどが本ポータルサイトで提供されている。これにより、本ポータルサイトは、観光を検討する人々への訪問の促進と、観光で訪れている人には、最新の観光情報を提供することで諏訪の滞在をより有意義になることを期待する。

観光協会、観光名所、ホテルなど観光客を対象として事業を行う事業者(以降、観光事業者)らは、本ポータルサイトの情報提供から観光客と、潜在的な観光客の興味、トピックなどのマーケティング情報の獲得を期待している。同時に、インターネットを活用した観光情報の提供サービスによる他の観光地との差別化をポータルサイトへ期待している。いま、地方の観光地が共通して抱える大きな課題として、これまでと異なる新しい顧客の獲得が求められている。その実現には、新たな広報戦略と、地域のブランド化が求められている。このとき、本ポータルサイトが web から取得する情報の解析への期待は大きい。

現在稼働する地域観光情報ポータルサイトの実装では、情報に含まれる単語の頻度を利用した数値化(ベクトル化)が行われる。そして、情報の類似性の判定、情報の検索を可能としている。各情報の特徴を数値化するためには、一定の単語数に含まれるブログや一般的な web ページコンテンツを前提としている。しかし、若い世代を中心として利用される SNS へ投稿される情報は、写真と関連する短文のコメントしか持たない。これら短文のみで構成される情報では、単語のスパース性に起因して、従来手法では情報の特徴を十分に表現(数値化)できない。結果として、情報のトピック、キーワードの解析、これらに基づく情報の類似性の導出ができない可能性が高い。そして、今後もその手軽さに起因して SNS から取得する情報の増大が予想され、対応が求められる。

諏訪地域の観光情報においても前述した事例があり、諏訪湖をモデルとした大ヒットアニメ映画のシーンと実際のスポット写真の比較、2018 年に 5 年ぶりに発生した御神渡りの様子が SNS 上で非常に多くの情報が発信され

たことがあった。しかし、現在の実装手法では、これら増大する情報の特徴を適切に解析し、可視化することはできなかった。

1.1. 研究目的

本研究 [2] [3] では、従来の地域観光情報ポータルサイトで収集の対象であった web, blog の情報に加えて、新たに SNS へ投稿された情報を対象として、地域観光情報ポータルサイトでの提供を行うための情報の解析手法を検討する。結果として、従来の諏訪地域の観光情報基盤の高度化を目指す。

本稿では、対象とする情報の解析手法として、情報に含まれる語の出現順序に基づく文脈情報をニューラルネットワークによって学習する予測手法と、情報に含まれる単語の集合の相関性を利用する相関性手法を提案する。そして、提案手法の評価として実際にインターネットより取得したデータとなる 2017 年ゴールデンウィーク、シルバーウィーク (以降 GW, SW) に収集した情報に適用し、従来手法との比較評価を行う。



図 1 現在の地域観光情報ポータルサイトのスクリーンショット

以降、2 章において対象となる諏訪地域の観光情報及び、関連研究についてまとめ、3 章において提案手法をまとめる。4 章で提案手法の評価を行い、5 章考察、そしてまとめる。

2. 地域観光情報の特徴と提案手法への要求

本章は、対象とする諏訪地域の観光情報の特徴についてまとめ、本稿の提案手法との関連研究をまとめる。そして、ここまでの検討に基づいて、本稿における仮説を明らかにする。

2.1. 諏訪地域の観光情報と経緯

筆者らの地域観光情報基盤を提案した経緯は文献 [1] [2] [3] において詳細を示している。長野県諏訪地域は、八ヶ岳を中心とした風光明媚な自然環境、諏訪大社などの史跡、そして温泉など様々な観光資源を有している地域である。長野県観光局の分析 [4] によると、諏訪を訪れる観光客の 70% 以上が 40 歳以上の中老年世代である。そのため、観光事業者らはインターネットを活用した情報発信の効果は限定的と考え、これまであまり積極的ではなかった。しかし、パソコンやスマートフォンなどインターネットの利用が生活への浸透、利用する世代の高齢化に伴い、インターネット上に観光客らの体験をブログ、web による情報発信が多く見られるようになった。そして、これら情報を可視化する諏訪地域の観光情報基盤が求められるようになった。

表 1 に、筆者らが 2015 年 3 月に諏訪地域の観光事業者 (ホテル・旅館 計 30 件) を対象に行ったアンケートの一部結果を示す。結果では、客室規模に関わらず旅行系サイトにおける口コミ・ブログは集客に影響することを示している。そのため、観光事業者らもインターネット上での口コミでの評判、インターネットを利用した集客

を意識した情報発信を行うようになった。しかし、観光事業者も継続して観光客に求められる情報を提供し続けることは難しい。また、訪れた観光客の体験、感想を綴った web ページ、blog も訪れる観光客の世代を考慮するとその数は限定される。従って、諏訪への観光を検討している人は、検索エンジンから到達できる過去のイベントのプロモーション情報や、現在と異なる古い情報に基づく観光体験を参照して旅行プランを構築せざるを得ない状況であった。つまり、地域観光情報ポータルでは、観光客が求める情報、事業者らが提供したい情報へ到達させることが目的であった。

表 1 旅行系サイトの口コミ・ブログの集客への影響

客室数	強く感じる	少し感じる	あまり感じない	全く感じない
0～15	50%	50%	0%	0%
15～30	40%	60%	0%	0%
30～50	100%	0%	0%	0%
50～90	50%	33%	17%	0%
90～	100%	0%	0%	0%

以上の経緯から、現在の地域観光情報ポータルサイトは、最新の諏訪地域の観光情報を可視化できるよう設計されている。観光事業者が発する各 web サイト情報、旅行記の投稿情報サイト、個人ブログのポータル、SNS などの web サービスを中心とした数ホップ内をクロウリング・スクレイピングしている。そして、これらを解析することで観光情報のトピック、ホットワードを導出している。この処理を定期的に行うことで、地域観光情報基盤として最新の観光情報の取得、可視化を実現している。

筆者らが上記手法でポータルサイトを提供してから数年が経過した。諏訪の観光情報に関して、この間 SNS から収集される情報は徐々に増加したが、大きな割合を占めるものではなかった。しかし、昨年これまでの状況と一変した。人気アニメ映画が諏訪湖周辺をイメージして作成されたことに起因して、若者が SNS 上でアニメに登場する諏訪湖周辺のスポットへ言及する情報が増大した。これら情報は、聖地巡礼と呼ばれるスポットへ訪れるための場所に関する情報共有、現地とアニメシーンとの比較など様々な形で諏訪の情報が爆発的な増加を見せた。この傾向と連動して、若い世代が“インスタ映え”するスポットとしてパワースポットとして有名な諏訪大社に興味を持ち、多くの情報が SNS に投稿されている。しかし、前章でも触れたが、多くの若者が利用する twitter, Instagram の投稿では写真と短文で構成される。投稿あたりの文量が少ないことに起因して、各情報に対して、十分な特徴の抽出を期待できない。結果として、地域観光情報観光基盤として、観光情報の可視化を提供できない。

2.2. 関連研究

地域観光情報ポータルサイトで現在実装されている分析手法は、各情報に含まれる単語の頻度を考慮した Term Frequency・Inverse Document Frequency (TF・IDF) [5] によるベクトル化が採用されている。そして、トピック推定と次元削減を行う Latent Dirichlet Allocation (LDA) [6] が適用されている。この手法は出現単語の頻度に基づき情報の特徴を導出することから、計測モデルに分類される。

現在、議論されている解析手法として、情報の文脈に基づいて出現する単語を予測する予測モデルが提案されている [7] [8]。この文脈とは、情報に含まれる単語の出現順序のパターンをニューラルネットワーク上で学習する、つまり、単語の前後関係から文脈に応じた単語を予測する。いま、文献 [7] では、情報に含まれる各語の前後の一定の範囲内(ウィンドウと呼ばれる固定範囲)に含まれる語の前後関係を学習する。そして、このウィンドウをスライドさせることでモデルを構築していく。このとき、前述した文脈とは各語に対してウィンドウに含まれる語を指している。そして、予測フェーズには、文脈から次の単語を予測する手法(CBoW:Continuous Bag of Word)と、単語から文脈を予測する手法(skip-gram)がある。CBoW では、ウィンドウに含まれている語と、ウィンドウ内に存在しない語の条件付き確率が最大となる単語を予測する。Skip-gram では逆に、特定の単語からウィンドウに含まれる条件付き確率の高い単語、つまり文脈となる周辺の語を予測する。

文献 [12] では、twitter から特定の製品の評判情報を文献 [8] の予測モデルを利用して抽出するための検討を行なっている。提案手法では、本稿と同様に投稿される tweet に含まれる語のスパース性が課題となっている。

この課題に対し、あらかじめ評判情報に利用が想定される基本語を辞書として管理することにより、単語拡張する手法を提案している。評価では、単語のスパース性が高いほど予測モデルが優位性を持つことを示している。そして、単語の語順は、文量に比例して予測モデルの性能に大きく影響することを示している。本稿で対象とする情報は、SNS だけでなく、web, blog が含まれることから単語数は大きく異なるものが混在し、その存在割合も大きく異なる。従って、本稿の提案手法では語順が与える影響を考慮した手法を検討する必要がある。

文献 [9] では、予測モデルの提案手法と、LSA (Latent Semantic Analysis) を適用した計測モデルの比較を行っている。この評価では、構文と意味解析について行い、この観点では予測モデルは計測モデルよりも優れた性能を有していることを明らかにしている。同時に、この文献では LSA を適用した計測モデルの詳細について記述されている。

文献 [10] では、文献 [9] で述べた上記の 2 つのモデルをベースとした様々な実装をコーパスに適用することで体系的な性能評価を行っている。評価では、設定可能なパラメータの組み合わせを考慮し、性能への影響を検討している。結果として、多くのコーパスに対して従来の計測モデルよりも予測モデルを採用した実装の方が優れた性能を有していることを明らかにしている。そして、予測モデルでのパラメータチューニングによる性能の変化の影響は小さく、最適化されなくとも大きく性能が低下しないことを明らかにしている。しかし、計測モデルでは多様なパラメータを柔軟に設定できることから、性能を劇的に向上できるパラメータを発見する可能性が存在する。しかし、パラメータの自由度の高さから発見は困難であるとまとめている。以上のことから、情報の解析では、計測モデルから予測モデルへシフトすることが予想されると指摘している。文献では、評価用の複数のコーパスを用いて評価しているが、このコーパスに含まれる文書はいずれも十分な長さを持つ文書であり、予測モデルの採用に起因してこれまでの評価結果の改善を実現している。

2.3. 本研究の仮説

ここまでの検討と、前節の関連研究に基づいて、筆者は以下の仮説を立てる。本稿で対象とする観光情報は、従来の web ページ、ブログの情報に加えて、SNS への投稿された情報を対象とする。このとき、情報に依存して含まれる文量が大きく異なる。しかし、短文で構成される情報でも、トピックの特徴を示す単語(以降では特徴語)は含まれる、同一のトピックなど類似情報であれば、特徴語の周辺にはどの情報においても類似する単語の出現が期待できる。従って、短い文章でも特徴語とその周辺の単語から文脈を類推が可能とする手法であれば、本稿の課題を解決改善できると仮説を立てることができる。

以上の仮説に基づいて、本稿では地域観光情報の解析に、2 つの手法を検討する。1 つは、文献 [8] と同様にニューラルネットワークを利用した予測モデルである予測手法を検討する。もう一方の手法は、情報に含まれる単語の集合(以降では BoW: Bag of Words)の相関性と、各語の出現率を併用して情報の類似性を導出する相関性手法を提案する。

3. 提案手法

本章では、前章に示した仮説に基づき、取得した情報からその特徴を解析する 2 つの手法；情報の文脈に基づいた単語の予測手法(以降では“予測手法”)と、情報に含まれる単語の集合としての相関性に基づく手法(以降では“相関性手法”)を提案し、それぞれを以下に示す。

3.1. 収集対象となる地域観光情報

提案手法は、これまでと同様にインターネットからのスクレイピングによって、諏訪地域の観光情報を取得する。対象は、web サイト、blog を中心にそこからリンク先 3 ホップ内を対象として、諏訪地域に関する観光情報のみを抽出する。このとき、筆者らの先行研究 [13], No SQL 型などビッグデータに対応可能な分散型 DB [14] の利用により、対象とする情報を制限なく収集・蓄積は可能である。しかし、関連性の低い情報を 3.2 節に示す内容と同様の処理負荷・管理コストの観点から、諏訪地域に関する情報のみを収集・抽出する。そのため、あらかじめ起点となる web サイトを中心として収集する。このとき、諏訪地域の“地名”と、“代表的な観光地”、“観光スポット名”、“正式名ではないが呼称”の情報を辞書として管理・参照している。辞書の利用は、市町村名など広域情報を含まない粒度の情報を取得対象とすることが可能となる。

SNS に投稿される情報は、その更新頻度の高さに起因して、各サービスが投稿情報を取得できる API (Application Programable Interface) を提供している。従って、情報の取得には、投稿を定期的に API 経由で監視し、Web サイトと同様に辞書を利用して諏訪地域と関連する投稿、観光情報をフィルタリングして取得する。

2.1 節で事例として述べたが、SNS に投稿される情報は状況によって爆発的に拡大する。このとき、1 章で述べた地域観光情報ポータルサイトの役割を鑑みると、増大する情報も可能限り観光客の興味、ホットトピックの解析と、可視化を行うことが求められると考えている。

現在の実装では、観光事業者が提供する web ページ [15]、訪れた観光客のブログ・旅行記の 4travel [16]、旅に関する情報サイト trip advisor [17]、観光キュレーションサイトである retrip [18] を起点としてスクレイピングを行なっている。加えて、本稿で新たに対象となる SNS は、twitter, Facebook のサービスを中心として、投稿内に含まれるリンク先も情報取得の範囲対象としている。若い女性を中心に利用される Instagram は魅力的なコンテンツではあるが、利用規約においてプログラムによるデータ取得を禁止していることから情報取得の対象から外している。

3.2. 地域観光情報の解析の前処理

インターネットから取得した様々な情報は、特徴を抽出するために、あらかじめ前処理が必要となる。これら情報は、日本語で記述されていることを前提とすると、テキスト情報を単語への分解が必要となる。しかし、日本語は品詞ごとに活用されることから、英語のようにそのまま分解することはできない。そのため、単語を抽出するためには、形態素解析を実行する必要がある。本稿では、オープンソースソフトウェアである MeCab [19] を適用し、対象の情報を単語へ分解と、それぞれの単語へ品詞をラベリングする。

その後、情報の文脈、コンテキストとは無関係と思われる単語、助詞、接続詞、数詞といった情報の内容に関わらない単語は、特徴抽出の雑音となる可能性、処理負荷の低減の観点からあらかじめ削除する。逆に、情報の内容の特徴を示す語、例えば観光地、イベントなどの固有名詞、地名は 3.1 節で述べた観光情報用の辞書を利用して識別する。

SNS で新たに創造される単語、新たに一般利用されるようになった単語は、迅速に辞書に登録する必要がある。このような新しい単語は対象とならない場合、観光情報ポータルサイトの提供する情報に大きく影響する可能性がある。従って、正しく形態素解析が行われるようこの辞書の登録管理は本稿の提案に関わるだけでなく、地域観光情報ポータルサイトのサービス品質において重要な要素であると言える。

3.3. 予測手法

2.3 節に示した本研究の仮説に基づき、情報の特徴を導出するためとして、文脈を得るために含まれる語順を利用する手法を提案する。語順の学習に関しては、ニューラルネットワークを利用する [8] の手法を利用する。このとき、本稿では、2.2 節で示した 2 種のモデルのうち、PV-DM (Paragraph Vector—Distributed Memory) モデルを適用する。この手法では、図 2 左部に示すように、入力はウィンドウ内の単語、出力がウィンドウに隣接する次の語となるように学習を進める。図 2 では、パラメータ 'window size' を 3 word が設定された環境において、'target word D' の予測するために 'word A' ~ 'word C' を文脈として学習する。その後、順にウィンドウをずらし、情報の文脈を学習していく。図 2 では対象を右にスライドさせることから、'word E' を予測するために 'word B' ~ 'word D' を文脈として学習する。加えて、本稿では、逆方向つまり、図 2 右部入力をウィンドウ内の単語、出力をウィンドウの直前の語とした学習も行う。つまり、ウィンドウの両側に存在する語を予測可能とするようにニューラルネットワークを構築する。これにより、単語数の少ない情報でも効率的な学習を期待できる。

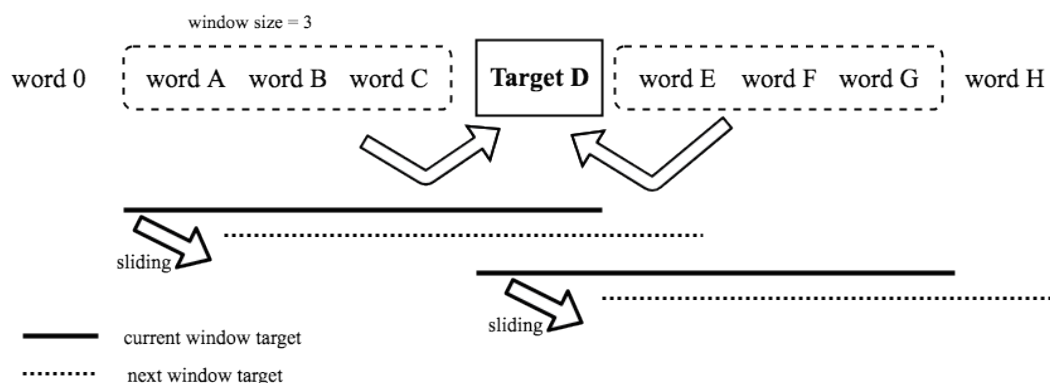


図 2 文脈の学習

文脈学習の結果として、情報を示すベクトル空間と、文脈から導出された各単語のベクトル空間が構築される。情報の類似性、単語の類似性の判定には、構築したそれぞれのベクトル空間から取得した2ベクトル間のなす角の導出し、その大きさによる判定(cos 類似度)を行う。

3.4. 相関性手法

2.3 節で示した仮説に基づいて、情報の特徴語は文脈に含まれる単語の集合として表されるといえる。このとき、情報に含まれる単語の集合(BoW)の相関性に基づいて情報間の類似性を判定する手法を提案する。しかし、単語集合の相関性だけでは前節の予測手法の語順と比較して、十分な類似性の判定が困難である。そのため、この相関性による特徴の判定に加え、各語の出現率による判定を組み合わせる手法を検討する。

本手法は以下の2つのフェーズに分けて類似性を導出することになる。(1)比較する情報に含まれる単語の相関性の導出、(2)これらから相関性の高い情報に対し、出現確率に基づいて類似性を判定する。

(1) 相関性の導出

単語の相関性は、以下の式(1)で導出する Jaccard 係数の値に基づいて判定する。このとき、相関性を求める2つの情報に含まれる単語の集合を X, Y とする。Jaccard 係数は、2つの集合における共通する要素の割合を導出する。つまり、集合が完全に一致する場合は Jaccard 係数の値は1、全く一致しない場合は0となり、2つの集合の類似性を0~1の範囲で定量化させることになる。

$$Jaccard = \frac{|X \cap Y|}{|X \cup Y|} \quad \text{式(1)}$$

式(1)では、分母が和集合であることから、各集合の大きさ、つまり BoW の大きさに反比例する。従って、web や blog のような文量があるものよりも、SNS の投稿は数語で構成されるため BoW の大きさは小さくなり、式(1)の分母も小さくなる。従って、同じ単語を含んでいても語数の少ない情報、SNS の投稿される情報の方が web など語数が多い情報よりも高い相関性を容易に得られることになる。しかし、このケースでは BoW が小さい情報の方が各単語の重要度が高いと判断できるため、現在の実装では相関性の判定基準を緩めることで対応している。この課題は本稿に限らず、異なるサイズの情報の比較時には発生する事象である。一般的には、正規化し単語あたりの相関率で判定することで対応する。

(2) 出現率の導出

式(1)で導出した相関性が閾値 α (現在の実装では $\alpha=0.75$)以上の BoW を相関がある情報としている。この値は、75%の BoW の一致で類似すると判定されるため、類似度としては緩いと考える。特に、SNS からの情報は小さい BoW が同一となることに起因して、類似判定の可能性が高くなる。しかし、前述したように文量が異なる多様な情報が含まれることから、閾値を緩くして大きく類似度の異なる BoW を除外している。そして、本ステップにおいて、詳細に類似判定を行う。現在の実装している閾値は本稿の対象 BoW から処理負荷と精度の観点から経験的に決定している。本来であれば対象となる情報に適応的に決定すべきと考えている。

同一の相関度を持つ情報が複数存在する場合も想定される。本ステップに含まれる単語の出現率(出現数/語数)を対象である情報ごとに導出してベクトル空間を構築する。これら各情報のベクトル間の cos 類似度を導出して順序を決定する。つまり、BoW で相関性を判定できない場合には、出現率を導出することで判定している。このとき、比較元である情報(集合 X)と各情報の相関性(類似性)だけでなく、導出した各語の出現率から、集合に出現が予測される語も導出できる。

4. 評価

本章では、インターネットから取得された実データに基づいて、提案手法の評価を行う。

4.1. 評価に用いるコンテンツ

本評価に用いる情報は、3.1 節で示した手法によって2017年のGW (Golden Week: 4月下旬~5月上旬)の期間、SW (Silver Week: 9月下旬)の期間においてインターネット上で取得された web, blog そして、SNS への投

稿された情報である。本評価では、期間内外に蓄積されたコンテンツは対象とせず、両期間で取得された諏訪地域の観光に関する情報のみを対象としている。このとき、取得された情報の内訳を表 2 に示す。

表 2 取得データの構成

	GW	SW	Word Average
Web & blog	2432	1543	128.4
SNS (twitter)	約 7900	約 13000	14.2
Total	10332	14543	

評価期間は、どちらの期間も諏訪地域では季節の変わり目 (GW は春, SW は秋) で観光客が増加する時期であること、休日が短期間に集中するため観光に関する情報も集中すると考え、本評価はこの時期を選定した。

対象となる情報は、3.1 節で検討した観光情報サイト [16] [17] [18]、観光客、観光事業者が提供する Web, blog といった文量に制限のない情報と、twitter に代表される SNS から取得した情報で構成されている。Web と blog のコンテンツは、当該期間に取得されたコンテンツで構成されている。また、SNS からの情報も、期間内で取得できた諏訪の観光に関するツイートから、リツイート、シェアなど複数回登場する投稿を極力削除した投稿数である。このとき、一部削除できなかった投稿も存在することから概算となっている。従って、本評価では両期間とも 10,000 超程度のコンテンツを評価に用いる。また、3.2 節で示した前処理後に含まれる情報の平均単語数を“Word Average”として示している。web 系と SNS の平均単語数が大きく異なることが自明となっている。

4.2. 評価方法

本評価では、3.3 節、3.4 節で示した 2 つの提案手法を用いて、両期間で取得した情報の解析を行う。それぞれ情報検索の性能の比較により、評価を行う。この評価では検索性能の観点から、各情報の持つ特徴を表現できているのかを明らかにする。つまり、各情報のトピック、情報間の類似性、情報の特徴を示すキーワードに代表される情報の特徴が表現できているかを評価することとなる。

このとき、表 2 に示した情報種別と同程度の割合となるように、評価用となる情報 50 件を用いて類似となる情報の検索を行う。また、提案手法との比較として、現在の実装手法である単語の頻度をベースとした手法である tf・idf 法に LDA [5] を適用した手法と比較する。

表 3 評価環境

Environment	Specification
前処理	gensim [20]
予測手法	Tensor Flow 1.8 [21] , Python 3.6.2
相関性手法	Python 3.6.2
従来手法	gensim [20] tfidf, LDA, Python 3.6.2
OS	ubuntu 16.10
System	Docker 17.09

表 4 評価用パラメータ

Environment	Specification
予測手法	PV-DM, Dim:200, Window size:4words,
	learning rate:0.01, iter:200
相関性手法	Jaccard:0.75
従来手法	tf·idf, topics:70,

評価環境は表 3 に示す環境で構築されており、共に Python で開発されている。予測手法の実装には、簡潔にトピック分析を開発できる gensim [20] の API と tensorflow [21] を用いて実装を行なっている。Doc2Vec をベースに手法はそのまま利用し、取得パラメータ等の改変を行なった。

表 4 に本評価において、各手法で用いたパラメータを示す。予測手法のパラメータであるウィンドウサイズは、図 2 に示した学習に影響する。特に、本稿では結果を示さないが、本評価では学習時間と性能のバランスからウィンドウサイズを 3 ワードとした。相関性手法のジャッカード係数の閾値は、3.4 節で示した類似性を 0.75 としている。従来手法では、現在の実装の値を利用している。

定性的な評価であるが、地域観光情報ポータルとして、利用が想定される代表的な 3 つのキーワードに対して共起される単語に関して比較評価を行う。その語と意味的に類似する単語を明らかにし、適切に対象となる情報が表現できているかを評価する。

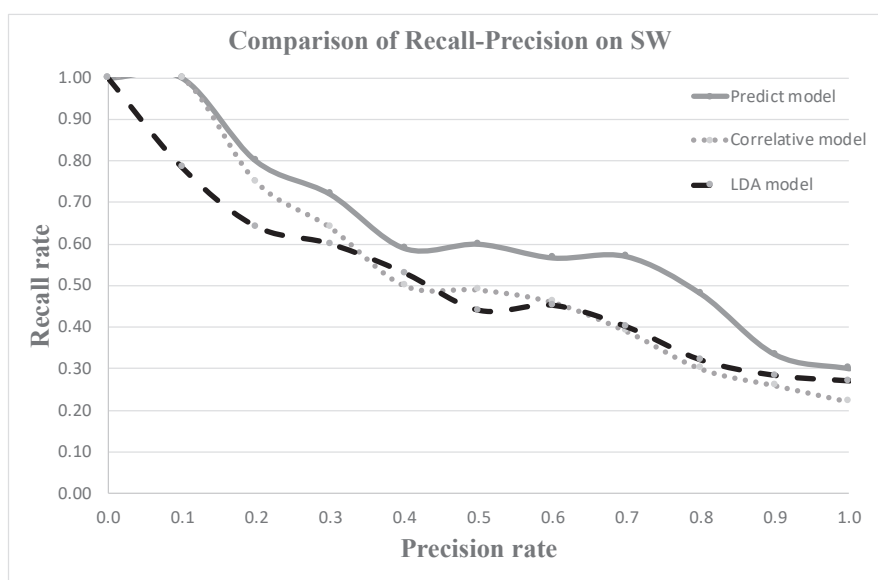
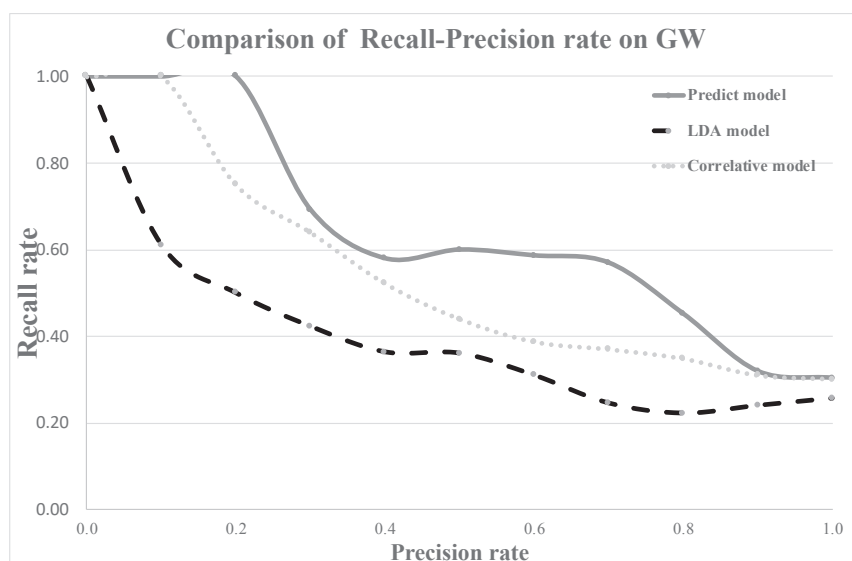


図 3 GW(上) SW(下) における性能比較

表 5 特定語に類似する語の比較

date	GW(2017/5/8)			SW(2017/9/25)		
words	' 大社'	' 温泉'	' 聖地'+ ' 映画'	' 大社'	' 温泉'	' 聖地'+ ' 映画'
予測手法	春宮 オーラス 下社 上社 男神	湯 源泉 旅館 足湯 風呂	アニメ 咲 アート 記事 ダンス	参拝 本宮 上社 お参り 諏訪湖	湯 露天風呂 風呂 入浴 旅館	巡礼 君 東方 アニメ 忍び
相関性手法	諏訪 諏訪湖 観光 宮 上社	諏訪湖 勧め 上諏訪 ホテル 蓼科	諏訪湖 聖地 映画 名 君	茅野 こころ 旅 火野 につぼん	長野 諏訪 諏訪湖 茅野 組合	原村 和 局 咲 誕生
従来手法	宮 柱 春宮 上社 下社	勧め 車山 車山 白樺湖 峰	君 名 聖地 富士見 巡礼	下社 春宮 宮 秋 上社	蓼科 柱 祭り 神社 縄文	ミーティング 姫 ファンクラブ さくら 映画

4.3. 評価結果

提案手法の評価結果を図 3 に示す。

図 3 において、上部は GW のデータ、下部は SW のデータに対する結果を示している。このとき、どちらも x 軸は適合率(Precision rate), y 軸は再現率(Recall rate)を示している。グラフ内では、予測手法(Prediction model), 相関性手法(Correlative model), 単語頻度を用いた従来手法(LDA model)の結果を示している。

一般的に、情報検索に用いられる Precision-Recall グラフでは、グラフの右上側に表されるほど性能が優れている。つまり、その手法は、情報の特徴をよく表現できているといえる。図では GW, SW のどちらの期間においても、予測手法は他の手法よりも右上側に現れている。従って、情報の特徴を表現できていることがわかる。

相関性手法は、GW では従来手法よりも良い性能を示している。しかし、SW 期間においては、従来手法とほぼ同程度の性能を示している。しかし、どちらの提案手法も、従来の単語の頻度に基づく手法と同程度以上の性能を有していることがグラフより読み取ることができる。

次に、表 5 に諏訪地域の観光において利用が想定される 3 つの単語(' 大社', ' 温泉', ' 巡礼'+ ' 映画')に対して、類似する特徴を持つ語の上位 5 つを示している。それぞれ手法により、重み付け手法が異なることから、同一のデータに対しても類似する語は異なる。また、期間によって情報のトピックが異なることから、類似する語も大きく異なることがわかる。

以下は主観的となるが、予測手法では、想定される語のいずれに対しても同一の意味を持つ別の語、詳細に言い換えた語、連想される語が類似する語として判定されている。相関性手法では、類似する後として同一トピックやコンテキストに登場が予想される語(共起する語)(' 大社' → 地名, ' 温泉' → 地名, ' 聖地+映画' → 登場人物の名前)が登場する。従来手法においても、相関性手法と同様に共起する語が類似する語として評価されている。特に、予測手法では他の手法と比較して意味的に類似する語、近い語を抽出することができている。

5. 考察

本章では、前節の評価結果に基づき、各手法の考察を行う。

5.1. 評価から

4.3 節に示した評価結果より、2 つの提案手法は、従来手法と比較して、同等以上の性能を有している手法と判断できる。

予測手法は、GW, SNS の投稿割合の多い SW どちらの評価結果においても提案手法による性能向上を確認でき

る。特に、対象が短期休暇の期間中であることから、この期間に観光を考えている人々により SNS の話題が特定のトピック（諏訪大社、開花、紅葉などの自然情報）に集中して増加している。この結果は、3.1 節で示した地域観光情報ポータルサイトとして、SNS からの情報を解析し、観光客の興味を可視化できる性能を持つと考えることができる。そして、予測手法は本稿の目的である SNS 情報の解析を実現しているといえる。

しかし、図 3 のどちらの結果においても他の 2 手法と比較してグラフの示す結果は安定しない。これは、本稿の提案手法に起因するものではなく、評価に用いた情報の網羅性に起因すると類推する。本評価では 10,000 件超の情報を対象としてモデル構築を行なっている。このとき、SNS から収集した情報が特定のトピックに情報が偏ることに起因して、安定した結果を得られていないと推測している。従来であれば、評価用コレクションを利用することで、コンテンツの網羅性に対する定量的な評価を行うことが一般的である。しかし、本研究で対象としている web や blog からの情報と、SNS からの情報が混在する評価用コレクションは存在しないこと、元来 SNS から取得される情報は特定トピックへ集中する可能性が高いことから、本稿の評価ではインターネットから実データによる評価を優先した。

予測手法の安定と、対象となる情報の網羅性の関連性は、今後学習の対象となる期間を限定するのではなく、学習の対象となる情報を増大させて評価を行うことを検討している。年間を通して学習する対象となる情報を取得することにより、情報のトピックを網羅的に増加させ、性能の安定を期待することができると考える。類似する語に関しては、定量的な評価ではないが前述したように利用が想定される語に近い意味を有する語が得られており、定性的にはこの語の特徴を抽出できていると考えることができる。

相関性手法では、GW と比較して SW の結果では従来手法と同程度の性能しか得られていなかった。これは、GW よりも SW の方が SNS から取得データ割合が大きいこと、この SNS では類似する情報に関するトピックが多い事に起因すると類推している。このとき、前処理後の各 BoW には、地名、イベントなどの固有名詞、観光、自然、季節を指す名詞、動詞に限定されることになる。従って、情報を表す語が一致する可能性が高いといえる。そのため、SW の評価では、3.4 節で示した手法の (1) フェーズの情報の相関性は高く導出される。そのため、(2) フェーズの各語の出現率によって類似性が決定されていたと類推できる。結果として、単語の出現頻度を利用する従来手法とほぼ同一の手法で類似する情報が導出されていたため、十分な性能を得られなかったと考える。

また、相関性手法では 3.4 節で示したように毎回 2 つのステップが必要となる可能性がある。比較する情報との相関性を都度ごとに導出することから、予測手法と比較すると対象情報の増加に比例して計算量が増加する。結果として、対象情報が増加するとユーザビリティも低下するという固有の課題を抱えている。

SNS の情報は、その重要性や話題など情報の増加に比例して“リツイート”つまり、同一の内容を示す情報も増大する。そのため、相関性手法では BoW がほぼ同一となる可能性が高くなる。このとき、あらかじめ類似する内容を完全にフィルタリングすることは困難であることから、前述した SW における相関性手法の性能低下と同じ事象が発生することになる。従って、2 つの提案手法のうち、SNS からの情報が増大する場合では、予測手法が相関性手法よりも、有効であると考えられる。情報に含まれる語順を考慮することは、語数の少ない情報かつ、類似するトピックが増大する場合では文脈の導出に有効であると類推している。

いま、予測手法では、各情報を十分に表すことができているのかを統計の観点から示すために、SW のデータに対して主成分分析を行う。評価において、類似している上位 10 件の情報に対して、第 1 因子、第 2 因子に対する累積寄与率を求め、図 4 に示す。この図 4 の結果では、大多数部分の累積寄与率が 0.7 前後から 0.8 前後の範囲に収束していることから、統計的な観点からみても提案手法ではコンテンツの特徴を捉えていると判断することができる。

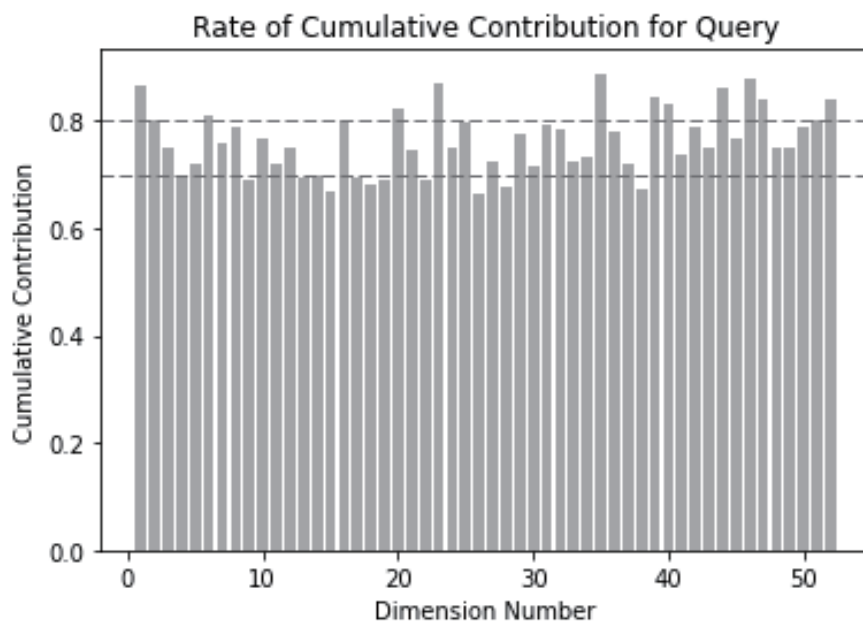


図 4 各クエリの上位 10 コンテンツに対する累積寄与率

本稿の評価では、表 4 に示したパラメータを用いた評価結果のみを示している。実際には、複数のパラメータの組み合わせの変更による評価を行っており、最も良い評価結果を本評価で示している。しかし、2.2 節で述べた文献 [10] によると、本稿の予測手法はパラメータの変更による性能への影響は小さいとされていることから本稿ではパラメータ変更による議論は行わなかった。しかし、予測手法では、文脈の学習範囲を決定する 'window size'、ニューラルネットワークにおける学習の試行回数を示す 'iteration' は、学習の性能に影響するだけでなく、処理負荷とのトレードオフの関係が存在する。従って、予測手法を地域観光情報ポータルサイトで採用するためには、対象となる情報に適応して、ニューラルネットワーク構築に関するパラメータチューニングが必要となると考える。

5.2. 今後の課題

提案手法を短期的な視点で実用化するための課題として、対象となる情報に適応したニューラルネットワークパラメータの決定手法の検討が必要である。このパラメータの最適化の課題に関しては、文献 [22] において、本稿と同様の Paragraph Vector に対して、二分法とランダムサーチの組み合わせにより、約 5000 通りのパラメータを低検索コストかつ効率的に発見する手法を明らかにしている。本稿の環境でも、類似する手法の検討により、実行時点におけるパラメータの最適値を導出する手法の検討できると考える。今後の課題として、これら手法の検討を行う予定である。

長期的な視点に立つと、本稿での評価では、短期的な期間に取得できた情報のみを対象とした。特に、予測手法は期間に急激に増加するこれら情報に対しても、優れた性能を示した。今後、実システムとして利用することを考えると、予測手法は、期間、ユーザの興味の頻繁な移り変わり、爆発的に発せられる情報に対応する必要がある。つまり、安定した性能を得るためには、対象となる情報を網羅的にあらかじめ収集する必要があると考える。そのために、特定の期間だけではなく、地域観光情報ポータルの対象となる情報の収集を継続的に取得する必要があると考える。同時に、この情報を増大させるためには事前処理、モデル化のための計算処理のリソース、導出した情報の管理データベースといったシステム構成も含めて検討が必要となる。今後、これらを準備し、継続的な情報の取得による提案手法の評価が必要であると考えている。

6. まとめ

本稿では、長野県諏訪地域の観光情報を対象とした情報の解析手法を検討した。従来は、web, blog 情報に含まれる単語の出現頻度に基づき重みづけられる解析手法が用いられてきた。しかし、twitter などの SNS に投稿

される情報は、その文量が短く、単語数が少ないためその特徴を表現することが困難であった。

これら多様な情報を解析する手法として、本稿では2つの手法：語順から文脈を学習して出現語を予測する予測手法、BoWの相関から類似性を導出する相関性手法を提案した。

提案手法の評価として、2017年のGW及び、SWに取得した諏訪地域の観光に関する情報を用いて、提案手法と従来手法を情報検索の観点で比較による評価を行った。結果として、予測手法は比較したいずれの手法よりもSNSからの情報が増加しても優れた性能を有していること、相関性手法は、従来手法と同等以上の性能を有していることを明らかにした。提案手法の比較として、予測手法の文脈の学習は、SNSのような文量の少ない情報を対象としても有効であることを考察した。今後の課題として、対象となる情報に適応したパラメータ最適値の取得方法の検討と、情報の網羅性を得ることで、予測手法の性能を安定させるために対象となる情報を継続的に収集することが必要なことを明らかにした。

謝辞

本研究の一部は、科学研究費基盤研究(C) JP17K01149の助成にて行われました

参考文献

- [1] 土屋健, “インターネットから収集する情報 ～地域の観光防災情報の高度化を目指して～”, 理大 科学フォーラム, Vol. 32 No. 6, pp. 34-39 2015年6月
- [2] 土屋健, 青柳志穂里, 三代沢正, 山田哲靖, 広瀬啓雄, 澤野弘明, 小柳恵一, “多様な情報を活用したParagraph Vectorによる地域観光情報基盤の高度化の検討”, 信学会 情報ネットワーク研究会, IN2017-74, pp. 25-30, 2018
- [3] TSUCHIYA Takeshi, HIROSE Hiroo, MIYOSAWA Tadashi, YAMADA Tetsuyasu, SAWANO Hiroaki, and KOYANAGI Keiichi, “Research on Improvement of Information Platform for Local Tourism by Paragraph Vector”, Proc. of 10th European Symposium on Computational Intelligence and Mathematics. Riga (Latvia), October, 2018
- [4] 長野県観光部観光企画課, “統計から見る長野県観光の現況”, 2013年9月
- [5] 徳永健伸, “情報検索と言語処理”, 東京大学出版会, 1999
- [6] D. Blei, A. Ng, and M. Jordan, “Latent Dirichlet Allocation”, in Journal of Machine Learning Research, 2003, pp. 1107-1135.
- [7] Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of the International Conference on Learning Representations (ICLR 2013), pp. 1-12, (2013)
- [8] Quoc Le, Tomas Mikolov, “Distributed Representations of Sentences and Documents”, Proc. of Int. XConf. on Machine Learning, Beijing, China, 2014 vol. 32.
- [9] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig, “Linguistic regularities in continuous space word representations”, Proc. of NAACL, pp. 746-751, Atlanta
- [10] G. Marco Baroni, Georgiana Dinu, “Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors”, 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference (2014)
- [11] 納村, 沼尾, 福井, “語順を基にした分散的意味表現による観光文書表現の検証”, 第30回人工知能学会全国大会論文集, June 2016
- [12] 芥子, 鈴木, 吉野, Neubig, 大原, 向井, 中村, “単語意味ベクトル辞書を用いたTwitterからの評判情報抽出”, 信学論 Vol. J100-D, No. 4, pp. 530-pp. 543, Apr. 2017.
- [13] 土屋健, 吉永浩和, 小柳恵一, “Peer-to-Peer 型クラウドコンピューティング技術による拡散データを統合する分散データベースサービスの検討”, 日本知能情報ファジィ学会誌, Vol. 25.No. 1, pp. 583-597, 2013年2月
- [14] Avinash Lakshman, Prashant Malik, 'Cassandra: a decentralized structured storage system', ACM SIGOPS, Vol. 44 Issue 2, April 2010, pp. 35-40
- [15] 諏訪市観光ガイド, <http://www.suwakanko.jp/>

- [16] 4travel, <http://4travel.jp/>
 - [17] Trip Advisor, <https://www.tripadvisor.jp/>
 - [18] Retrip, <https://retrip.jp/>
 - [19] MeCab, <http://taku910.github.io/mecab/>
 - [20] topic modelling for humans, <https://radimrehurek.com/gensim/>
 - [21] TensorFlow, <https://www.tensorflow.org/>
 - [22] 原, 新妻, 太田, 'Paragraph Vector のための効率的なパラメータサーチの検討', DEIM Forum 2017
-

(2018 年 12 月 20 日受理)