情報処理学会論文誌 Vol.60 No.8 1390–1402 (Aug. 2019)

意味空間上の分布表現に基づく Webサイトと閲覧ユーザの統合分析モデル

保坂 大樹[1,a)] 河部 瞭太[1,b)] 山下 遥[2,c)] 後藤 正幸[1,d)]

受付日 2018年10月30日, 採録日 2019年5月9日

概要：近年，消費者の Web サイト閲覧行動は重要なマーケティング分析の対象となっている．サイトの閲覧行動を分析することで，サイト間の関係性やユーザの嗜好を把握し，Web 広告の掲載やメールの配信などのマーケティング施策の最適化や効率化が可能となるためである．しかし，蓄積された Web 閲覧履歴データ中のサイトやユーザの関係性は複雑であり，サイト単位やユーザ単位の分析が困難となる場合が多い．したがって，そのような状況においても適用可能，かつ有効な分析手法が望まれている．本研究では，単語の意味分析において良い性能を示している Word2vec とその拡張モデルに基づき，各サイトや各ユーザをそれぞれ意味空間上の多次元正規分布として表現するとともに，意味空間上のサイトやユーザの関係性に基づいて分析を行うための手法を提案する．学習された意味空間上の表現を用いた分析により，単純な閲覧，被閲覧の関係ではなく，閲覧行動の背後に存在するサイトやユーザの潜在的な特性や関係性を把握することが可能となる．また，提案手法では，サイトやユーザが持つ性質の広がりが多次元正規分布の分散行列として学習される．最後に，提案手法の有効性を検証するために，実際の閲覧履歴データ分析に適用し，得られる結果に関する考察を与える．さらに，分散行列をサイトの閲覧者の多様性，またはユーザの嗜好の多様性として解釈し，より詳細な分析が行えることを示す．

キーワード：Word2vec，Doc2vec，分散的意味表現，Web サイト閲覧履歴データ，ビジネスアナリティクス

A Model for Integrated Analysis of Website and User by Gaussian Embedding

Taiju Hosaka[1,a)] Ryota Kawabe[1,b)] Haruka Yamashita[2,c)] Masayuki Goto[1,d)]

Received: October 30, 2018, Accepted: May 9, 2019

Abstract: Recently, many companies analyze Web browsing history of users for understanding their preferences and relationships between websites. The analysis is useful for optimizing of marketing measures such as the mail distribution or the advertisement on the internet. However, the complex relationship between objects in stored Web browsing history makes the analysis of a certain website or a user difficult. It is, therefore, advisable to provide effective analysis methods even in that circumstance. In this research, we focus on the Word2vec model and extensible models which show high-performance in natural language processing and propose the analysis method based on a multidimensional normal distribution in a semantic vector space. Through the analysis based on distributed representation on semantic vector space, we can understand semantic features and relationships between websites and users behind browsing activity instead of direct relationships based on the number of accesses. Finally, we demonstrate the analysis based on our proposed model by applying a practical data. Furthermore, we interpret a covariance matrix of a website as diversity of viewers, and that of a user as diversity of the preference. We show that these measures can be used to analyze more minutely.

Keywords: Word2vec, Doc2vec, distributed representation, browsing history data, business analytics

[1] 早稲田大学
Waseda University, Shinjuku, Tokyo 169–8555, Japan
[2] 上智大学
Sophia University, Chiyoda, Tokyo 102–8554, Japan

[a)] wasewase@moegi.waseda.jp
[b)] kryota_09@ruri.waseda.jp
[c)] h-yamashita-1g8@sophia.ac.jp
[d)] masagoto@waseda.jp

1. はじめに

近年，インターネット上では，様々なWebサイトを通じて，多くの情報を閲覧できるほか，ユーザが様々なサービスの提供を受けることが可能となっている．これにより，消費者のWebサイト閲覧機会は大幅に増加しており，Webサイト上のマーケティングの重要性が高まっている [1]．多くの企業は，バナー広告の掲載やメールの配信といったマーケティング施策の実行によって売上の向上を図っている．これらの施策の効果については多くの研究において様々な評価がなされており [2], [3]，今後も，これらの施策を最適かつ効率的に実行するための研究が発展すると期待される．

一方，近年，Webサイト閲覧行動履歴データ（以下，閲覧履歴データ）は重要なマーケティング分析の対象となっている．閲覧履歴データには，Webサイトの特性や閲覧ユーザの嗜好が反映されていると考えられているためである．閲覧履歴データの分析を行うことによって，各Webサイトおよび各ユーザの特性や，Webサイト間やユーザ間の関係性を明らかにすることで，マーケティング施策の最適化や効率化が期待できる．実際に閲覧履歴データを対象とした分析については，これまでもいくつかの研究が行われている．たとえば，閲覧履歴からWebページを推薦するためのモデル構築を行う研究 [4], [5], [6] や，閲覧履歴からパターンマイニングを行う研究 [7] などがあり，主に閲覧履歴データを対象とした分析の有用性が示されている．

しかしながら，近年のWebサイト数の爆発的な増加にともない，ユーザのWeb閲覧目的が多様化している状況において，上記の研究で用いられているようなユーザ単位の共起をベースとしたアプローチを適用した場合，ユーザの複数のWeb閲覧目的を区別して分析することはできない．また，複数の目的に基づいて生起する多様なWebサイト間の関係性を表現することは困難であると考えられる．このような背景から，同一ユーザの閲覧に基づいた関係性ではなく，より密接な関係性に着目した分析を行うことが望まれている．

これに対し，近年，自然言語処理分野において高い性能を示している，分散的意味表現に着目した研究 [9] が行われている．この研究では，Doc2vec [16] と呼ばれる分散的意味表現の学習アルゴリズムを，ユーザの閲覧系列に適用することで，ユーザの閲覧行動の性質を比較的低次元のベクトルで表現し，広告割当タスクに活用している．

一方で，Doc2vecによる文書データ分析のアプローチでは，単語や文書の持つ意味を意味空間上の単一の点のみで表現するという制約が存在する．しかしながら，一般的に，単語の意味には広がりが存在し，単語間で意味の広さに差異が存在すると考えられる．それに対し，単語を意味空間上の多次元正規分布で表現することで，単語の意味の広がりを考慮した分散的意味表現の学習モデルが提案されている [19]．このモデルでは，多次元正規分布の分散行列が単語の意味の広さに相当している．単語と同様に，ユーザの閲覧行動の性質にも広がりがあることが想定される．たとえば，様々なWebサイトを広く閲覧するユーザと，似たようなWebサイトのみを閲覧するユーザでは，その嗜好の幅が異なるといえる．これは，Webサイトの観点からも同様である．

本研究では，ユーザの嗜好の幅やWebサイトの性質の幅を多様性と定義する．多様性の高いWebサイトはそれぞれ異なる嗜好を持つユーザ群に閲覧され，逆に，多様性の高いユーザはそれぞれ異なる性質を持つWebサイト群を閲覧する．ここで，Webサイトやユーザの多様性を定量化するために，Tagamiらの手法 [9] を拡張し，各Webサイトと各ユーザを意味空間上の多次元正規分布で表現する手法を提案する．多次元正規分布の平均ベクトルはユーザの平均的な嗜好やサイトの平均的な性質を表現し，分散行列によってその広がりを表現することができる．提案手法では，性質の広がりを分析することで，Webサイト間，ユーザ間，Webサイトとユーザ間の関係性をより的確にとらえることが可能となり，考察により，新たな知見を得ることが期待される．実際に，提案モデルを用いて閲覧履歴データを分析し，提案モデルの有効性を検証するとともに，Webマーケティングを効果的に実行するための知見について考察を与える．

2. 準備

2.1 関連研究

2.1.1 閲覧履歴データを対象とした研究

これまで，閲覧履歴データを対象とした研究は数多く存在する．それらの多くは，特定のWebサイト内のアクセスログに関する研究である．たとえば，Fuら [10] は相関ルールに基づいたWebページ推薦アルゴリズムを提案しているが，このアルゴリズムは，動的に相関ルールを作成するために，計算コストが大きいという問題点がある．Mobasherら [11] も同様に相関ルールに基づく推薦手法を提案しており，これはFuらの研究の問題点を解決し，効率的なWebページ推薦を可能としている．一方，山元ら [4] はユーザ間の最長共通部分列に着目し，頑健なWebページ推薦アルゴリズムを提案した．さらに，松嵜ら [8] は，ECサイト内のWebページ遷移をマルコフ潜在クラスモデルでモデル化し，Webページ間の遷移について詳細な分析を行っている．

上記のような，特定のWebサイト内におけるアクセスログの分析では，分析対象となるページ数が比較的少ないことから，その直接的な遷移に着目して分析することが効果的であると考えられる．しかし，扱うWebサイトに制限を設けずに多くのWebサイトを統合的に分析する場合，

その組合せ数は膨大なものとなり，特定の Web サイト内のユーザ行動に対する分析手法をそのまま適用することは困難である．

閲覧履歴データの統合的な分析を行っている研究としては，高須賀ら [5] がユーザごとの閲覧履歴に対して協調フィルタリングのアプローチを用いて Web ページの推薦を行っている．この手法は，ユーザ間で共起している Web サイトに基づいてユーザ間類似度を定義し，Web ページ推薦を行うものである．また，鶴原ら [6] は独立成分分析をユーザの閲覧系列に適用し，ユーザの閲覧の特徴となる合成変数を作成し，典型的な閲覧コンセプトについての分析を行っている．

しかしながら，近年の Web 閲覧目的の多様化にともない，ユーザ単位の共起頻度ベースのアプローチも困難になりつつある．これに対して，Tagami ら [9] は閲覧系列を学習してユーザの分散的意味表現を学習することで，直接的な共起ではなく，共起の背後にあるような Web サイトやユーザ間の関係性をモデル化した．このアプローチでは，ユーザ単位で Web サイトの共起を仮定するのではなく，ある閲覧サイトと前後の閲覧サイトとの関係性を仮定している．Web サイトとユーザの分散的意味表現を用いるこの手法は，広告割当タスクにおいて，頻度ベースの従来手法に比べて良い精度を示している．

2.1.2 分散的意味表現に関する研究

分散的意味表現モデルは，もともとは自然言語データに対する分析モデルとして提案され，様々な方向性へと研究が発展している．その先駆けは，単語の分散的意味表現の学習モデルとして Mikolov らによって提案された Word2vec [12], [13] である．Word2vec は，生起した単語は同じ文脈の単語から予測できるという仮説のもと，ニューラルネットワークのアプローチを用いて単語の分散的意味表現の学習を行う．ニューラルネットワークは，予測の精度を高めるために，同質の入力に対して類似した中間表現を学習する．Mikolov らは，この性質に基づいて，ある単語から文脈中の単語を予測するネットワークの中間表現を分散的意味表現として抽出する方法を提案した．

Word2vec は，自然言語処理のいくつかのタスクで良い精度を示している．たとえば，Xue ら [14] はソーシャルメディアから収集された言語データに対して Word2vec を適用し，感情分析を行った．また，Siencnik [15] は，固有名詞抽出のタスクに Word2vec を適用し，従来の手法よりも良い精度を示すことを確認した．一方で，文書分類などの文書に着目したタスクでは良い精度を示すことができていないという課題が示された．

そこで，Le ら [16] は，単語のベクトル化のアプローチを，文書のベクトル化に拡張した．Doc2vec と呼ばれるこのモデルでは，文書を Word2vec のネットワークに組み込むことで，単語と同様に文書の分散的意味表現に関しても学習を行うことができる．ここで，得られた文書ベクトルは，文書の持つトピック情報を示していると考えられており，文書に関するタスクの精度を向上させる可能性があることが示されている [17], [18].

ほかにも，Word2vec を基礎として多くの研究がなされている．Vilnis ら [19] は，単語を単一のベクトルで表現する Word2vec の制約に着目し，単語をベクトルで表現する代わりに多次元正規分布で表現するモデルを提案した．Word2gauss と呼ばれるこのモデルでは，多次元の正規分布で単語を表現することで，各単語の性質のばらつきを考慮することが可能となる．この手法によって，抽象的な単語や固有名詞などの，単語ごとに異なる意味の広がりを考慮した分散的意味表現の獲得が可能となった．

2.2 単語の分散的意味表現

自然言語処理において，単語を含む文書データを分析する際に，どのように単語を数値で表現するべきであるかについては，数多くの議論がなされている．従来では，Mikolov らの研究 [20] のように，単語を one-hot ベクトルで表現することが多かった．単語の one-hot ベクトルとは，語彙数を次元とし，ある要素のみを 1，他の要素をすべて 0 とすることで，どの単語であるかを表現したベクトルである．one-hot ベクトルによるベクトル表現は構築が容易である一方で，すべての単語を等しく区別して扱うため，同義語や多義語の持つ単語の意味的な性質を反映できないという欠点がある．

one-hot ベクトルに対し，近年，単語の意味をとらえることができる分散的意味表現の重要性が高まっている．分散的意味表現では，あらかじめ設定した次元数の単語ベクトルで各単語を表現する．単語ベクトルの各要素は概念を持っていると考えられており，比較的密なベクトルによって単語間の関係を表現することが可能となっている．ここで，単語が表現されるベクトル空間を意味空間といい，類似した意味を持つ単語群は意味空間上で近傍に布置される．これにより，意味空間上の距離によって単語間の類似性を定量的に測定することが可能となる．また，単語ベクトルは意味的な演算処理が可能であることが示されている．このように，分散的意味表現では表現力の豊かな単語ベクトルによって意味を考慮した単語の活用が可能である．一方，単語ベクトルを獲得するためには，適切な学習アルゴリズムを構築する必要がある．

2.2.1 Word2vec [12]

Word2vec は，単語の分散的意味表現の学習モデルである．Word2vec では，出現する単語は文脈の中で周辺の単語から予測できるという仮説のもと，ニューラルネットワークのアプローチを用いて分散的意味表現における単語ベクトルを学習する．

たとえば，「私は○○を飼っている」という文脈では，

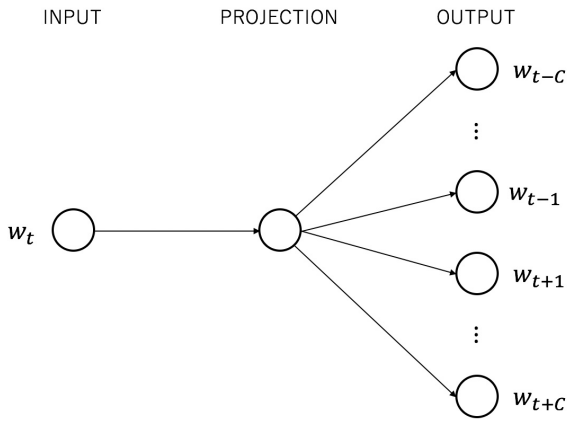

図 1 Continuous Skip-gram モデル
Fig. 1 Continuous Skip-gram model.

“犬” や “猫” といった単語が空欄に入ることが想定されるが，これらの単語は，「ペット」という共通の概念に基づいて文脈に出現する．Word2vec では，このような同じ文脈で出現する単語群に共通の概念を仮定して，学習データの中に含まれる多様な概念を意味空間の軸として学習し，各単語を意味空間上の点として表現する．

Word2vec では，学習データをコーパスの形で与える．学習のあるステップにおいて，コーパス中のある単語 w_t に注目し，注目単語 w_t に対して，その周辺単語群 $w_{t-C},\cdots,w_{t-1},w_{t+1},\cdots,w_{t+C}$ が定義される．ここで C をウィンドウサイズと呼び，周辺単語群の幅を決定するためのパラメータとして定義する．

Word2vec では入力データから教師データを予測するように学習されたネットワークにおいて，射影層におけるベクトルを単語ベクトルとする．Word2vec のモデルとして，次に示す 2 種類のネットワークが提案されている．

Continuous Skip-gram モデル [12]
Continuous Skip-gram モデルでは，注目単語から周辺単語群を予測するためのネットワークを構築する．すなわち，注目単語を入力データとし，周辺単語群を教師データとする．Skip-gram モデルのネットワークを**図 1** に示す．

Continuous Bag-of-Words モデル [12]
Continuous Bag-of-Words モデルでは，周辺単語群から注目単語を予測するためのネットワークを構築する．すなわち，周辺単語群を入力データとし，注目単語を教師データとする．このモデルでは，周辺単語群を入力とするため，各周辺単語の射影層におけるベクトルの和を用いて順伝播計算を行う．

2.2.2 Doc2vec [16]

Word2vec を拡張したモデルの 1 つに，意味空間上の文書ベクトルを学習する，Doc2vec が提案されている．Doc2vec では，Word2vec のネットワークに文書を組み込むことで，文書と単語を同一の意味空間上に表現することができる．このとき，文書ベクトルはトピック情報と呼ばれる，文書の意味を表すことが考えられている．Doc2vec のモデルとして次に示す 2 種類のネットワークが提案されている．

Distributed Memory モデル [16]
Distributed Memory モデルでは，周辺単語群と所属文書から注目単語を予測するためのネットワークを構築する．すなわち，周辺単語群と所属文書を入力データとし，注目単語を教師データとする．このモデルは，Word2vec の Continuous Bag-of-Words モデルの拡張であり，文脈の性質を文書の性質の一部ととらえながら学習を行う．

Distributed Bag-of-Words モデル [16]
Distributed Bag-of-Words モデルでは，所属文書から注目単語と周辺単語群を予測するためのネットワークを構築する．すなわち，所属文書を入力データとし，注目単語と周辺単語群を教師データとする．このモデルは，Word2vec の Continuous Skip-gram モデルの拡張であり，Distributed Memory モデルと比較して，メモリ効率，計算量の面で優れている．一方で，単語系列の語順を無視してしまうという欠点がある．

2.2.3 Word2gauss [19]

Word2vec では各単語を意味空間上の 1 点で表現していた．一方，単語を意味空間上の正規分布で表現することで，単語間の位置関係だけでなく，意味的な広がりも表現したモデルが提案されている．単語の正規分布表現では，各単語に対して，平均ベクトルだけでなく，分散共分散行列も定義されている．その要素である分散により各単語の意味の広さを表現できると考えられている．Word2gauss は，Word2vec を応用し，単語の平均ベクトルと分散行列を学習するモデルである．

Word2vec では単語間の類似度を単語ベクトルの内積で表現していた．一方，Word2gauss では単語を正規分布で表現するため，分布間の類似度を定義する必要がある．Vilnis らの研究 [19] で用いられる分布間の類似度の 1 つに Expected Likelihood Kernel [21] がある．この尺度は，分布間における内積として用いられる．n 次元線型空間 $\mathbb{R}^n$ 上の 2 つの確率分布 $g_1(\boldsymbol{s})$，$g_2(\boldsymbol{s})$ 間の Expected Likelihood Kernel は式 (1) で定義される．

$$EL(g_1(\boldsymbol{s}),g_2(\boldsymbol{s}))=\int_{\boldsymbol{s}\in\mathbb{R}^n} g_1(\boldsymbol{s})g_2(\boldsymbol{s})d\boldsymbol{s} \tag{1}$$

特に，n 次元正規分布 $\mathcal{N}_1=N(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$，$\mathcal{N}_2=N(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ 間の Expected Likelihood Kernel は，式 (2) で計算される．

$$\begin{aligned}EL(\mathcal{N}_1,\mathcal{N}_2)&=\int_{\boldsymbol{s}\in\mathbb{R}^n} N(\boldsymbol{s}:\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)N(\boldsymbol{s}:\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)d\boldsymbol{s}\\&=N(\boldsymbol{0}:\boldsymbol{\mu}_1-\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)\\&=\frac{\exp\left\{-\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right\}}{(2\pi)^{\frac{n}{2}}\det(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{\frac{1}{2}}}\end{aligned} \tag{2}$$

表 1 閲覧数と被閲覧数の基本統計量
Table 1 Fundamental statistics of browsing and view.

	平均値	中央値	最大値	最小値
ユーザごと閲覧数	733	458	16,730	10
サイトごと被閲覧数	260	29	720,269	1

ただし，$\boldsymbol{\mu}_1 \in \mathbb{R}^n$，$\boldsymbol{\mu}_2 \in \mathbb{R}^n$ は各々の平均ベクトル，$\boldsymbol{\Sigma}_1 \in \mathbb{R}^{n\times n}$，$\boldsymbol{\Sigma}_2 \in \mathbb{R}^{n\times n}$ は分散共分散行列である．

式 (2) は 2 つの正規分布の重なりの度合いを定量化している．複数の意味を持つような単語や抽象的な単語は多様な文脈のもとで生起するため，意味空間上で互いに離れている単語群との類似度が高くなるように正規分布表現が学習され，その分散が大きくなる．対して，特定の文脈においてのみ用いられるような単語は，意味空間上で近傍に位置する単語群との類似度が高くなるように学習が行われるため，その分散は小さくなる．

2.3 分析対象データ

今回分析対象とするデータは，各ユーザのセッションごとの閲覧サイト系列の形で与えられている．学習データ中の Web サイト数は 27,789 個，ユーザ数は 9,851 人である．また，ユーザごとの閲覧数および Web サイトごとの被閲覧数の基本統計量を**表 1** に示す．

表 1 から，多くのユーザが数百回以上 Web サイトを閲覧していることが分かる．この結果からも，閲覧履歴データは短期間であっても大量に蓄積され，ユーザあたりのデータ数が十分にあるため，ユーザ単位の分析は有用であることが想定される．一方で，ほとんどの Web サイトは数回〜数十回しか閲覧されていない．このように，本研究で扱うデータは，ユーザ単位のデータ量は比較的多いが，Web サイト単位のデータ量は少ないことが 1 つの特徴である．したがって，2 つの Web サイトをともに閲覧しているユーザ数によってこれらの Web サイト間の関係性を定量化してしまうと少数のデータがほとんどになってしまい，対象データ構造の精度の良い表現を得ることが困難である．

3. 提案モデル

3.1 モデル概要

先述したとおり，本研究の目的は，Web サイトおよび閲覧ユーザの複雑な関係性を柔軟に分析することである．これまで，ユーザ単位，Web サイト単位の共起を考慮したモデルが多く提案されているが，事前分析の結果にも示されているように，ユーザ単位で Web サイトの共起を仮定するモデルは対象データに有効ではない．

そこで，対象データが文書データと類似した構造を持つ点に着目し，Word2vec に基づくアプローチを考える．Word2vec では，単語の生起の時系列性に着目し，文書単位で単語の共起を仮定する代わりに，文脈単位で単語の共起を仮定する．対象データについても，Web サイトの閲覧の時系列性に意味があると仮定する．

たとえば，旅行を計画しているユーザが飲食店を探すという目的で Web サイトの閲覧を行うとき，その閲覧系列が（宿泊予約サイト A）→（グルメサイト A）→（グルメサイト B）であるとする．このとき，これらの Web サイト群は「旅行の計画」という同一の閲覧目的のもとで共起している．本研究では，このような同一の目的のもとで閲覧される Web サイト群に共通の概念を仮定する．また，共通の概念を持つ Web サイト群に，意味的な類似性を仮定する．同様に，このユーザがニュースをチェックするという目的で Web サイトの閲覧を行うとき，閲覧されたニュースサイトはその前後に閲覧される Web サイトと類似性が仮定される．一方で，ニュースサイトとグルメサイトは同一のユーザに閲覧されているが，異なる閲覧目的のもとで閲覧されているため，この 2 つの Web サイトには類似性が仮定されない．

また，ユーザごとに，閲覧目的そのものだけでなく，閲覧目的の幅も異なると仮説を立てる．上記の例で示したユーザは「ニュースのチェック」から「旅行の計画」まで幅広い閲覧目的に基づいた閲覧を行っている．一方，「動画の閲覧」という目的でのみ Web サイトを閲覧するユーザの存在も考えられる．同様に，SNS のような複数のコンテンツを含む Web サイトは様々な閲覧目的のもとで閲覧されるのに対して，専門的なアイテムの通販サイトは，特定の閲覧目的でのみ閲覧される．このような，閲覧目的の多様性の差異を考慮してユーザや Web サイトの関係を分析するために，Word2gauss に基づいた表現が有効であると考えられる．

上記で例示したような，閲覧系列の局所的な共起関係を学習するために，Word2vec を基礎としたモデルが有効であると考えた．ここで，本研究で扱う閲覧履歴データを，Word2vec が対象とする文書データと比較すると，「文書」→「ユーザ」，「単語」→「Web サイト」，「文脈」→「閲覧目的」というような対応づけがなされる．

本研究では，各ユーザのセッションごとの閲覧サイト系列を学習データとし，意味空間上の各 Web サイトと各ユーザの正規分布表現を獲得するためのモデルを提案する．これは，Tagami らの研究 [9] で得られる分散的意味表現を拡張し，閲覧の多様性を考慮した分析を可能とする．このモデルでは，意味空間上の正規分布表現を学習するために，Doc2vec と Word2gauss のアイデアを援用する．

具体的には，Doc2vec における Distributed Memory モデルを拡張すると同時に，ユーザ q の h 番目のセッションにおける i 番目の閲覧サイト $r_i^{q,h}$ に対して，**図 2** のようにネットワークを構築する．分布間の類似度は，Word2gauss のモデルに基づき，Expected Likelihood Kernel を用いて計算を行う．

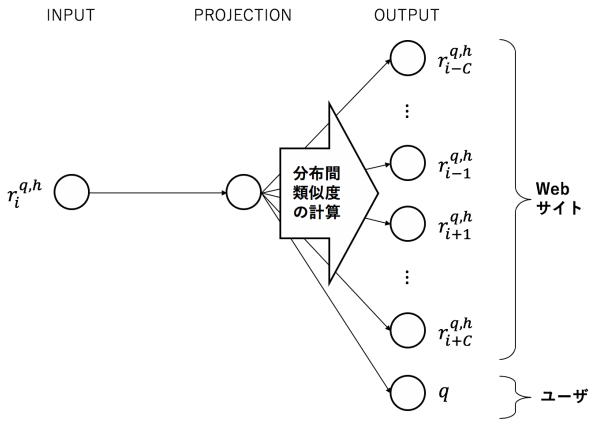

図 2 提案モデルのネットワーク
Fig. 2 Proposal model network.

閲覧履歴データに提案モデルを適用することで，各 Web サイトと各ユーザをそれぞれ同一の意味空間上に布置することが可能となる．学習される意味空間上の正規分布表現において，Web サイトやユーザの平均ベクトル間の距離は，オブジェクト間の類似性を表現する．また，Web サイトやユーザの分散行列の大きさは，オブジェクトの多様性を表現する．本研究では，提案モデルに基づき，Web サイトやユーザの性質について，多角的な分析を行う．

3.2 変数の定義

まず，提案モデルで用いる変数を定義する．全 Web サイト数を N，全ユーザ数を M とし，全 Web サイト集合を $\mathcal{S} = \{s_n : 1 \leq n \leq N\}$，全ユーザ集合を $\mathcal{U} = \{u_m : 1 \leq m \leq M\}$ とする．また，構築する意味空間の次元を d とし，サイト s_n，ユーザ u_m に対応する d 次元分散行列をそれぞれ $\boldsymbol{\Sigma}_{s_n}$，$\boldsymbol{\Sigma}_{u_m}$ と記述する．また，ウィンドウサイズを C とする．入力層から射影層へ写像する際の重み行列を $\boldsymbol{W}$，射影層から出力層へ写像する際の重み行列を $\boldsymbol{Z}$ とする．

3.3 モデルの定式化

ユーザ $q = u_m$ の h 番目のセッション内で i 番目に閲覧した Web サイトを $r_i^{q,h} = s_n$ とする．また，あるサイト $r_i^{q,h}$ に対し，その前後 C サイトの計 $2C$ サイト $r_{i+c}^{q,h} = s_{n_c}$ $(-C \leq c \leq -1, 1 \leq c \leq C)$ を周辺サイトと定義する．

本モデルは，ニューラルネットへの入力を，s_n に対応する one-hot ベクトル $\boldsymbol{x}_{s_n}$ とする．ここで，射影層のベクトル $\boldsymbol{w}_{s_n}$ は式 (3) で計算され，s_n に対応する $\boldsymbol{W}$ の列ベクトル，s_n に対応する多次元正規分布の平均ベクトルに一致する．

$$\boldsymbol{w}_{s_n} = \boldsymbol{W}\boldsymbol{x}_{s_n} \tag{3}$$

出力層では活性化関数としてソフトマックス関数を用いて，入力サイトに対する各 Web サイト，各ユーザの予測確率分布 $p(o|r_i^{q,h})$（$o \in \mathcal{S} \cup \mathcal{U}$）を，式 (4) で計算する．ここで，$o$ は予測対象とする Web サイトまたはユーザであり，$\boldsymbol{z}_o$ は，o に対応する $\boldsymbol{Z}$ の行ベクトル，o に対応する多次元正規分布の平均ベクトルに一致する．ただし，Web サイトやユーザの分散的意味表現として抽出するパラメータは $\boldsymbol{w}_o$，$\boldsymbol{\Sigma}_o$（$o \in \mathcal{S} \cup \mathcal{U}$）である．

$$p(o|r_i^{q,h}) = \frac{EL\big(N(\boldsymbol{z}_o, \boldsymbol{\Sigma}_o), N(\boldsymbol{w}_{s_n}, \boldsymbol{\Sigma}_{s_n})\big)}{\sum_{o' \in \mathcal{S} \cup \mathcal{U}} EL\big(N(\boldsymbol{z}_{o'}, \boldsymbol{\Sigma}_{o'}), N(\boldsymbol{w}_{s_n}, \boldsymbol{\Sigma}_{s_n})\big)} \tag{4}$$

入力サイト $r_i^{q,h}$ に対して，その周辺サイト，および閲覧ユーザを教師集合 $\mathcal{O}_i^{q,h} = \{q\} \cup \{r_{i+c}^{q,h} : -C \leq c \leq -1, 1 \leq c \leq C\}$ として定義する．入力サイト $r_i^{q,h}$ に対する損失 $l(r_i^{q,h}, \mathcal{O}_i^{q,h})$ は式 (5) で定義される．

$$l(r_i^{q,h}, \mathcal{O}_i^{q,h}) = -\sum_{o \in \mathcal{O}_i^{q,h}} \log p(o|r_i^{q,h}) \tag{5}$$

ここでは，確率的降下法に基づいた誤差逆伝播法によって損失関数を最小化する．

3.4 計算の効率化

式 (4) では，Web サイト数やユーザ数が膨大である場合に，その計算量も膨大となる．本研究では，計算量削減のために，ネガティブサンプリング [13] を用いたアプローチに基づき，パラメータを推定する．

ネガティブサンプリングでは，任意の確率値を割り当てたそれぞれの Web サイトの存在確率分布をノイズ分布とする．そして，周辺サイト $r_{i+c}^{q,h}$ に対して，K 個の Web サイト $r_k^{q,h,i+c}$ $(1 \leq k \leq K)$ をノイズ分布から独立にサンプリングする．このサンプリングされたサイトをネガティブサイトと定義する．

ネガティブサンプリングでは，$r_{i+c}^{q,h}, r_1^{q,h,i+c}, \cdots, r_K^{q,h,i+c}$ について，周辺サイトとネガティブサイトのどちらであるかの二値分類を行う．入力サイト $r_i^{q,h}$ のもとで出力 o が周辺サイトである確率 $p_N(o|r_i^{q,h})$ は式 (6) で計算される．

$$p_N(o|r_i^{q,h}) = \exp\left\{-\frac{1}{2}(\boldsymbol{z}_o - \boldsymbol{w}_{s_n})^\top (\boldsymbol{\Sigma}_o + \boldsymbol{\Sigma}_{s_n})^{-1}(\boldsymbol{z}_o - \boldsymbol{w}_{s_n})\right\} \tag{6}$$

Web サイトだけでなく，ユーザについても同様の操作を行い，ネガティブユーザとする．ここで，出力 $o \in \mathcal{O}_i^{q,h}$ に対するネガティブサイトもしくはネガティブユーザを δ_k^o $(1 \leq k \leq K)$ と定義すると，入力サイト $r_i^{q,h}$ に対する損失 $l_N(r_i^{q,h}, \mathcal{O}_i^{q,h})$ は式 (7) で計算される．

$$l_N(r_i^{q,h}, \mathcal{O}_i^{q,h}) = -\sum_{o \in \mathcal{O}_i^{q,h}} \left\{\log p_N(o|r_i^{q,h}) + \frac{1}{K}\sum_{k=1}^{K} \log\big(1 - p_N(\delta_k^o|r_i^{q,h})\big)\right\} \tag{7}$$

ユーザ q の閲覧セッション数を H_q，h 番目のセッション内での Web サイト閲覧数を $I_{q,h}$ とすると，学習データセット全体の損失 l_{all} は式 (8) で定義される．

$$l_{all} = \sum_{q \in \mathcal{U}} \sum_{h=1}^{H_q} \sum_{i=1}^{I_{q,h}} l_N(r_i^{q,h}, \mathcal{O}_i^{q,h}) \tag{8}$$

3.5 学習アルゴリズム

提案モデルにおけるパラメータ $\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\Sigma}$ は，確率的降下法によって式 (8) を極小化することで最適化を行う．各パラメータにおける式 (7) の勾配は式 (9)〜(14) で計算される．

$$\frac{\partial l_N}{\partial \boldsymbol{z}_o} = -(\boldsymbol{\Sigma}_o + \boldsymbol{\Sigma}_{s_n})^{-1}(\boldsymbol{w}_{s_n} - \boldsymbol{z}_o) \tag{9}$$

$$\frac{\partial l_N}{\partial \boldsymbol{z}_{\delta_k^o}} = -\frac{p_N(\delta_k^o | r_i^{q,h})(\boldsymbol{\Sigma}_{\delta_k^o} + \boldsymbol{\Sigma}_{s_n})^{-1}(\boldsymbol{z}_{\delta_k^o} - \boldsymbol{w}_{s_n})}{K(1 - p_N(\delta_k^o | r_i^{q,h}))} \tag{10}$$

$$\frac{\partial l_N}{\partial \boldsymbol{w}_{s_n}} = -\sum_{o \in \mathcal{O}_i^{q,h}} \left\{ \frac{\partial l_N}{\partial \boldsymbol{z}_o} + \sum_{k=1}^{K} \frac{\partial l_N}{\partial \boldsymbol{z}_{\delta_k^o}} \right\} \tag{11}$$

$$\frac{\partial l_N}{\partial \boldsymbol{\Sigma}_o} = -\frac{1}{2}\left(\frac{\partial l_N}{\partial \boldsymbol{z}_o}\right)\left(\frac{\partial l_N}{\partial \boldsymbol{z}_o}\right)^{\top} \tag{12}$$

$$\frac{\partial l_N}{\partial \boldsymbol{\Sigma}_{\delta_k^o}} = \frac{K(1 - p_N(\delta_k^o | r_i^{q,h}))}{2(p_N(\delta_k^o | r_i^{q,h}))}\left(\frac{\partial l_N}{\partial \boldsymbol{z}_{\delta_k^o}}\right)\left(\frac{\partial l_N}{\partial \boldsymbol{z}_{\delta_k^o}}\right)^{\top} \tag{13}$$

$$\frac{\partial l_N}{\partial \boldsymbol{\Sigma}_{s_n}} = \sum_{o \in \mathcal{O}_i^{q,h}} \left\{ \frac{\partial l_N}{\partial \boldsymbol{\Sigma}_o} + \sum_{k=1}^{K} \frac{\partial l_N}{\partial \boldsymbol{\Sigma}_{\delta_k^o}} \right\} \tag{14}$$

学習データセット中の閲覧について順に式 (9)〜(14) の勾配を用いてパラメータを更新する．これを十分な回数繰り返すことで最適なパラメータを獲得する．

4. 分析事例

本章では，提案モデルを実際の閲覧履歴データに適用して，Web サイトと閲覧ユーザの分散的意味表現を獲得する．また，獲得した分散的意味表現を用いて，Web サイト閲覧行動に関する多角的な分析を行う．

4.1 分析条件

モデルの学習には，株式会社ヴァリューズ提供の閲覧履歴データを用いる．このデータは，登録に同意したモニタが PC またはスマートフォンから閲覧した Web サイトのホスト名を記録したものである．対象データのデータ期間は，2017 年 8 月 1 日から 2017 年 10 月 31 日，総閲覧数は 7,224,737 回，ユーザ数は 9,851 人，Web サイト数は 27,789 個である．また，Web サイトには人手でカテゴリが付与されているものがあり，そのカテゴリ情報は，4.2.3 項で利用する．以下では，モデルのパラメータについて述べる．

各 Web サイトおよび各ユーザに対応づける多次元正規分布の次元数は $d = 50$ であり，ウィンドウサイズを $C = 10$，ネガティブサイトおよびネガティブユーザのサンプルサイズを $K = 20$ とした．ノイズ分布は，Web サイトとユーザについて，学習データセット内の生起頻度を確率分布として正規化したものをそれぞれ用いた．また，計算の効率化と過学習の回避を目的として，分散行列にスカラー行列を仮定した．

モデルパラメータは，確率的降下法に基づいて学習を行う．学習率は，Adagrad [22] に基づき，式 (15) でその値を更新する．ここで，学習ステップ t において，パラメータ θ について $\eta_{\theta,t}$ を学習率とする．また，$g_{\theta,t}$ は学習ステップ t までの θ の勾配の平方和，ϵ は閾値パラメータであり，$\epsilon = 1.0$ と設定した．Adagrad では，十分に学習されたパラメータに関する学習率を小さくすることで，パラメータの過学習を防止する．ただし，学習率の初期値を $\eta = 0.05$ とした．

$$\eta_{\theta,t} = \frac{\eta}{\sqrt{\epsilon + g_{\theta,t}}} \tag{15}$$

また，4.2.1 項では，比較手法として，意味空間上の各ユーザと各 Web サイトのベクトル表現を学習する Doc2vec を用いる．Doc2vec のパラメータ設定は，提案モデルと同じ値を用いた．

4.2 分析結果

4.2.1 Web サイト間の関係分析

学習された多次元正規分布の平均ベクトルに基づいて，Web サイト間の関係分析を行う．類似した性質を持つ Web サイトは，意味空間上で近傍に布置されるため，意味空間上の距離で Web サイトの類似性を定量化することができる．本研究では Web サイトに対応する平均ベクトル間の cos 類似度を Web サイト間の類似度として扱う．

提案モデルの有効性について考察するために，分析対象とする Web サイトは，その性質がある程度想定できる Web サイトが望ましい．そこで，本研究では，代表的なグルメサイトである “tabelog.com（食べログ）” と代表的な旅行サイトである “www.jalan.net（じゃらん net）”，代表的なファッション EC サイトである “zozo.jp（ZOZOTOWN）” を対象事例として分析結果を示す．

上記の 3 サイトとの平均ベクトルとの cos 類似度が高い Web サイトの上位 10 件をそれぞれ**表 2**，**表 4**，**表 5** に示す．また，対象データに Doc2vec を適用することで得られた “食べログ” に対応するベクトルとの cos 類似度が高い Web サイトの上位 10 件を**表 3** に示す．

表 2 より，“食べログ” と同様に代表的なグルメサイトである “retty.me（Retty グルメ）” や “gnavi.co.jp（ぐるなび）” といった Web サイトが上位であることが分かる．この結果から，飲食店を検索・予約する際に “食べログ” を含む複数のグルメサイトを併用するというユーザの閲覧背景がうかがえる．また，百貨店のサイトである “www.jr-takashimaya.co.jp（ジェイアール名古屋タカシマヤ）” や宿泊予約のサイトである “www.yado-sagashi.jp（宿

表 2 提案モデルで “食べログ” と類似度の高い Web サイト
Table 2 Websites similar to “tabelog.com” on proposal model.

	ホスト名	類似度	被閲覧数
1	retty.me	0.928	1,769
2	member.s-pt.jp	0.903	47
3	www.newotani.co.jp	0.890	113
4	www.jr-takashimaya.co.jp	0.884	43
5	www.yado-sagashi.jp	0.879	44
6	gogo.gs	0.876	151
7	selfs.dai-ichi-life.co.jp	0.874	22
8	gnavi.co.jp	0.874	4,279
9	topisyu.hatenablog.com	0.869	42
10	www.persona.co.jp	0.865	40

表 3 Doc2vec で “食べログ” と類似度の高い Web サイト
Table 3 Websites similar to “tabelog.com” on Doc2vec.

	ホスト名	類似度	被閲覧数
1	web.tenmaya.co.jp	0.926	4
2	sso.meiji.ac.jp	0.923	14
3	talent.theatre.co.jp	0.911	20
4	retty.me	0.908	1,769
5	www.u-aroma.com	0.894	20
6	pepy.jp	0.894	74
7	agenthub.jetstar.com	0.892	18
8	www.hana300.com	0.879	37
9	www.favy.jp	0.878	98
10	matome.miil.me	0.873	71

表 4 提案モデルで “じゃらん net” と類似度の高い Web サイト
Table 4 Websites similar to “www.jalan.net” on proposal model.

	ホスト名	類似度	被閲覧数
1	rurubu.travel	0.847	1,148
2	www.trivago.jp	0.783	814
3	www.eonet.ne.jp	0.777	154
4	www.tavigator.co.jp	0.774	172
5	www.poke.co.jp	0.758	21
6	kobaton-mileage.jp	0.755	11
7	www.c-nexco.co.jp	0.744	325
8	tabi-moni.com	0.740	13
9	www.nonhoi.jp	0.739	36
10	www.tabitora.co.jp	0.727	44

探し.com)” などが上位であることから，旅行や遊びに出かける際に，飲食店の検索も同時に行うような閲覧行動の可能性を指摘することができる．

表 3 では，表 2 と同様に，“Retty グルメ” を類似サイトとして検出した．そのほかにも，“www.favy.jp（favy)” や “matome.miil.me（ミイルまとめ)” のようなグルメサイトが見受けられる．この結果から，提案手法で学習される意味空間は，Doc2vec で学習される意味空間と同等の構造にあることがうかがえる．

表 5 提案モデルで “ZOZOTOWN” と類似度の高い Web サイト
Table 5 Websites similar to “zozo.jp” on proposal model.

	ホスト名	類似度	被閲覧数
1	intlsystem-2017.nippon-rad.co.jp	0.719	41
2	www.ipat.jra.go.jp	0.709	2,105
3	shop67.makeshop.jp	0.706	9
4	www.thegearpage.net	0.706	37
5	photo.gazo.space	0.705	25
6	fdoc.jp	0.705	22
7	geinou-news.jp	0.704	63
8	passport-web.soc.shukutoku.ac.jp	0.701	30
9	www32.jvckenwood.com	0.701	16
10	www.yonden.co.jp	0.699	85

表 4 では，“じゃらん net” と類似したサイトとして，旅行関係のサイトを多く検出した．その中には，日帰り旅行の情報サイトである “www.poke.co.jp（ポケカル)” や宿泊予約のサイトである “www.trivago.jp（トリバゴ)” および “www.tavigator.jp（たびゲーター)” が存在しているため，“じゃらん net” は様々な旅行のニーズを持ったユーザ群に閲覧されていることが想定できる．また，類似サイトには宿泊事業者向けの法人サービスのサイトである “www.tabitora.co.jp（株式会社たび寅)” も見受けられる．このことから，旅行を計画しているユーザだけでなく，旅行に携わる事業者による閲覧も存在すると考えられる．

一方で，表 5 では，“ZOZOTOWN” と同様のファッション関係の Web サイトを検出していない．“ZOZOTOWN” と類似しているとして検出された Web サイトの例として，医院の検索サイトである “fdoc.jp (EPARK クリニック・病院)” や地方競馬の投票サイトである “www.ipat.jra.go.jp (JRA 日本中央競馬会)” があげられる．ファッション関係の Web サイトを検出しなかったことから，ファッションアイテムの購入を考えているユーザが “ZOZOTOWN” と他のファッションサイトを同時に閲覧することは少なく，“ZOZOTOWN” 内でファッションアイテムの購入を完結する閲覧傾向の存在が考えられる．

4.2.2 ユーザの閲覧嗜好に関する分析

本項では，多次元正規分布の平均ベクトルに基づいて，ユーザの閲覧嗜好に関する分析を行う．提案モデルでは，ユーザが実際に閲覧した Web サイトとの類似性が高くなるように，閲覧ユーザと Web サイトのパラメータ更新を行う．そのため，特定のユーザと類似した Web サイト群はユーザの閲覧嗜好を反映していることが想定される．

ここでは，閲覧数が平均的なユーザを対象として分析を行う．閲覧回数が中央値と等しいユーザ[*1]を抽出し，“対象ユーザ” と定義する．対象ユーザの平均ベクトルと cos 類似度が高い Web サイトの上位 10 件を**表 6** に，対象ユーザによる閲覧数が多い Web サイトの上位 10 件を**表 7** に

[*1] ユーザ間の閲覧数の偏りが大きいため，中央値を採用した．

表 6 提案モデルで対象ユーザと類似度の高い Web サイト
Table 6 Websites similar to target user on proposal model.

	ホスト名	類似度	被閲覧数	大別される Web サイト群の性質
1	rlx.jp	0.962	126	高級グルメ・高級旅館
2	www.bestcarton.com	0.942	14	その他
3	www.aniplexplus.com	0.939	23	アニメ情報
4	vipper-trendy.net	0.934	64	その他
5	www.tohoho-web.com	0.933	32	Web コンテンツ制作支援
6	saruwakakun.com	0.927	53	Web コンテンツ制作支援
7	www.fate-sn.com	0.924	59	アニメ情報
8	www.aniplex.co.jp	0.919	23	アニメ情報
9	sumapotibm.xsrv.jp	0.918	21	Web コンテンツ制作支援
10	clubmichelin.jp	0.918	25	高級グルメ・高級旅館

表 7 対象ユーザによる閲覧数が多い Web サイト
Table 7 Websites frequently viewed by target user.

	ホスト名	閲覧数	被閲覧数	大別される Web サイト群の性質
1	websearch.rakuten.co.jp	107	96,642	代表的な EC サイト・Web サービス
2	www.rakuten.co.jp	66	144,534	代表的な EC サイト・Web サービス
3	www.4gamer.net	33	1,330	ゲーム関連の Web サイト
4	jp.finalfantasyxiv.com	27	1,133	ゲーム関連の Web サイト
5	www.amazon.co.jp	23	123,806	代表的な EC サイト・Web サービス
6	ff14wiki.info	22	126	ゲーム関連の Web サイト
7	books.rakuten.co.jp	11	9,794	代表的な EC サイト・Web サービス
8	www.square-enix.co.jp	11	624	ゲーム関連の Web サイト
9	appmedia.jp	10	1,135	ゲーム関連の Web サイト
10	my.rakuten.co.jp	10	11,528	代表的な EC サイト・Web サービス

示す．

まず，表 7 をもとに対象ユーザの閲覧嗜好を考察してみる．対象ユーザが頻繁に閲覧しているサイトは 2 種類の Web サイト群に大別することができる．片方の Web サイト群には，“www.rakuten.co.jp（楽天市場）” や “www.amazon.co.jp（Amazon）” といった代表的な EC サイト，Web サービスが分類される．もう片方の Web サイト群には，“jp.finalfantasyxiv.com（FINAL FANTASY XIV）” といったオンラインゲームや，“www.4gamer.net（4gamer）” といったゲームの情報サイトのようなゲーム関連の Web サイトが分類される．2 種類の Web サイト群は，Web サイトの性質が大きく異なり，また，被閲覧数の観点からも顕著な差が見受けられる．

一方，表 6 において，対象ユーザと類似しているとして抽出された Web サイトは「アニメ情報」「Web コンテンツ制作支援」「高級グルメ・高級旅館」「その他」といった性質を持つ 4 種類の Web サイト群に大別することができる．一般的に，ゲームとアニメはコンテンツを共有することが多く，ジャンル間に高い類似性を持つことが想定される．したがって，対象ユーザのゲーム関係のサイトの閲覧に起因して，アニメ関係のサイトとの類似性が高くなったと考えられる．ここで留意すべきことは，他のユーザも含む全ユーザの閲覧によって，ゲーム関係のサイトとアニメ関係のサイトの強いつながりを学習したという点である．同様に，そのような Web サイトを閲覧するユーザは「Web コンテンツ制作支援」「高級グルメ・高級旅館」といった Web サイトを閲覧する傾向にあることを指摘することができる．

続いて，表 2，表 4，表 6 を比較することで，Web サイトの被閲覧数と類似性に関する考察を行う．提案モデルでは，被閲覧数の多い Web サイトがネガティブサイトとして抽出されやすく，類似性が低くなるようにパラメータの更新が行われる．しかし，表 2，表 4 では，被閲覧数の多い Web サイトが類似サイトとして検出されている．このことから，表中に出現した被閲覧数の多い Web サイトには，“食べログ” や “じゃらん net” とのより強い関係性があると考えられる．対して，表 6 では，被閲覧数の多い Web サイトを検出していない．一般的に，特定のユーザの Web 閲覧に着目すると，ニュースのチェックや飲食店の検索など，閲覧の目的は複数にわたる場合が多く，特定の Web サイトと強い関係性が抽出されることは少ない．したがって，ユーザに着目した場合，被閲覧数の多い Web サイトについて，閲覧回数よりもネガティブサイトとしてサンプリングされる回数の方が相対的に多くなり，ユーザと Web サイトは高い類似性を持ちにくいと考えられる．被閲覧数

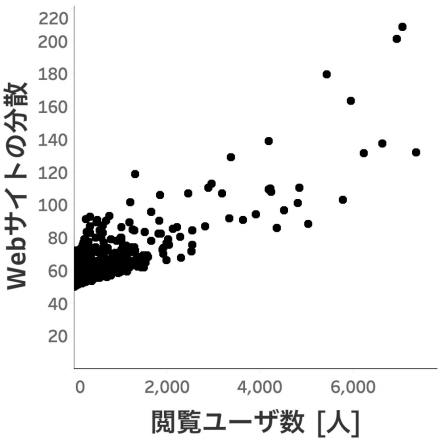

図 3 閲覧ユーザ数と分散の関係

Fig. 3 Relationship between number of browsing user and variance.

表 8 「車」カテゴリ中で提案モデルにおいて分散の大きかった Web サイト

Table 8 Websites with large variance in the “car” category on proposal model.

	ホスト名	分散	閲覧ユーザ数
1	carview.co.jp	67.13	602
2	carsensor.net	65.63	518
3	goo-net.com	62.10	522
4	carview.yahoo.co.jp	60.55	478
5	response.jp	59.86	474
6	car.watch.impress.co.jp	59.81	220
7	www.honda.co.jp	57.88	688
8	toyota.jp	57.42	694
9	auto.rakuten.co.jp	57.40	363
10	autoc-one.jp	56.66	402

の多い Web サイトは，多くのユーザが日常的に利用する Web サイトである場合が多く，その閲覧には個人の閲覧嗜好が反映されにくい．対して，被閲覧数の少ない Web サイトの閲覧には，個人の閲覧嗜好が反映されていると考えられる．このことから，提案モデルでは，ユーザの閲覧嗜好を反映した Web サイトを検出することができるといえる．

4.2.3 Web サイトの多様性に関する分析

本項では，提案モデルの適用によって得られた分散に基づき，Web サイトの多様性について分析を行う．Word2gauss [19] において，意味空間上の分散は，単語の意味の広がりを表現していた．同様に，提案モデルにおいても，意味空間上の分散はサイトの性質の広がりを表現していると考えられる．具体的には，様々な性質を持ったユーザ群による閲覧や，様々な性質を持った Web サイト群との共起によって分散パラメータが大きくなるように学習される．

まず，様々な性質を持ったユーザ群による閲覧と，単純に数が多いユーザ群による閲覧との関係性について考える．ここで，分散行列にはスカラー行列を仮定しているため，各 Web サイトの分散は単一の値で表され，この値を用いて分析を行う．各 Web サイトの分散と閲覧ユーザ数の関係を**図 3** に示す．2 つの変数間の相関係数は 0.85 であった．このことから，Web サイトの分散と閲覧ユーザ数の間には強い正の相関があるといえる．この結果は，より多くのユーザに閲覧されることで，Web サイトの分散が大きくなるという本提案モデルの性質を示している．しかし，意味空間上で互いに離れた位置に表現されたユーザ群，すなわち，様々な性質を持ったユーザ群の閲覧により，さらに分散は大きなものとなるため，完全な線形関係ではないことが分かる．

次に，具体的にどのような Web サイトで分散が大きくなっているかを確認する．類似した性質の Web サイト間で結果を比較するために，ここでは「車」カテゴリに属する Web サイトのうち分散の大きい Web サイトの上位 10 件を**表 8** に示す．先述したことからも分かるとおり，表 8 中の Web サイトはいずれも多くのユーザに閲覧されている．また，分散の値と閲覧ユーザ数の大小が完全に一致していないことも確認できる．たとえば，6 番目に分散が大きい “car.watch.impress.co.jp（Car Watch）” では，表中のほとんどの Web サイトに比べて半分以下のユーザしか閲覧していないが，分散が大きな値となっている．これは，それぞれ異なる閲覧の性質を持つユーザ群が “Car Watch” を閲覧しているということを示唆していると考えられる．対して，7 番目に分散が大きい “www.honda.co.jp（本田技研工業）” や 8 番目に分散が大きい “toyota.jp（トヨタ自動車）” は，閲覧ユーザ数が他の Web サイトよりも多いにもかかわらず，その分散は表中の他の Web サイトと比較して大きいとはいえない．このような結果が得られた理由として，それぞれ類似した閲覧の性質を持つユーザ群がこれらの Web サイトを閲覧している可能性を指摘することができる．また，以上の結果から，複数のメーカの車を扱うポータルサイトは閲覧ユーザ数に対して分散が大きく，自社の車のみを扱う特定のメーカのサイトでは閲覧ユーザ数に対する分散が小さいことが分かる．一般的に，ポータルサイトは多様なユーザに閲覧される傾向のある Web サイトであるため，多様なユーザに閲覧されている Web サイトは分散が高くなるという解釈が与えられ，これは提案モデルにおける分散に多様性の意味があることを支持する結果となっている．

5. 考察

5.1 ハイパーパラメータの決定

提案モデルは，従来の Word2vec と同様に，結果の解釈性の観点から探索的に正規分布の次元数 d，ウィンドウサイズ C，ネガティブサンプルサイズ K およびノイズ分布を決定する必要がある．本研究では，従来研究の論文で用

いられている設定[*2]を含めて，ある程度の広範囲にわたり，結果の解釈性の高いパラメータを探索的に決定した．

正規分布の次元数 d はパラメータ数を定めるハイパーパラメータであるため，学習データに応じてその値を調整する必要がある．次元数が適切かどうかの客観的な基準としてオブジェクト間の類似度があげられる．次元数が小さい場合には，データの意味的な構造を表現することができず，オブジェクト間の類似度は大きな値をとらなくなる．次元数が大きい場合には，パラメータの過学習が発生し，オブジェクト間の類似度は過度に大きな値をとる．本研究においては，先行研究の次元数を用いて対象データを学習した場合に，Web サイト間の cos 類似度が極端に大きくなったため，Web サイト間の cos 類似度が 1 に偏らない結果になるまで次元数を段階的に小さくして，$d = 50$ と決定した．また，この結果から，対象データの意味的な構造は，文書データに比べて単純であることがうかがえる．

ウィンドウサイズ C は共起を仮定するサイトの幅を決定するハイパーパラメータであり，学習されるパラメータの性質に影響する．閲覧履歴データにおいては，ウィンドウサイズが小さいとき，Web サイト間の直接的な遷移を表現したパラメータが学習される．逆に，ウィンドウサイズが大きいとき，同一ユーザによる閲覧を表現したパラメータが学習される．分析者は，分析目的に応じてウィンドウサイズを調整する必要がある．本研究の対象データでは，データの抽出の仕方から，複数の Web サイトを交互に閲覧するような傾向が多く見受けられたため，より離れた位置にある Web サイトとも類似性を仮定し，先行研究よりも大きな値として，$C = 10$ と設定した．

ネガティブサンプルサイズ K は，ネガティブサンプルとして学習する Web サイトまたはユーザの数を決定するハイパーパラメータである．学習データ構造を表現したパラメータを学習するという観点から，ネガティブサンプルサイズは大きいほど望ましい．ただし，ネガティブサンプルサイズに比例して計算量は大きくなるため，実行可能な学習時間の範囲内で大きくするべきであるといえる．本研究においても，実行が可能な学習時間の範囲内で大きな値として $k = 20$ を採用した．この値は Tagami らの研究 [9] のネガティブサンプルサイズ $k = 5$ よりも大きな値である．Tagami らの研究では AUC を用いてモデルの評価を行っており，パラメータの最適化には，評価指標を局所最大化するためにパラメータの探索を多く行う必要がある．一方，本研究では，定性的にパラメータの最適性を判断しているため，パラメータの細かな調整の効果は低いと考えられ，行われていない．そのため，本研究はモデルの学習時間を多く確保することが可能であり，このような理由から，ネガティブサンプルサイズは Tagami らの研究よりも大きな値となると考えられる．

ノイズ分布について，Word2vec では生起頻度の 3/4 乗を正規化した確率分布が経験的に良いことが報告されている [13]．Web サイト閲覧履歴データにおいてその結果は同様とはいえないが，その発想を援用し，生起頻度の指数を大きくすることで，一般的に閲覧されるサイトをネガティブサイトとしてサンプリングされやすくして，より特徴的な Web サイトを検出することが可能となると想定される．逆に，指数を小さくすることで，より閲覧履歴を重視した分析が可能になる．本研究においては，学習を効率的に行うために，生起頻度を正規化した確率分布を用いた．

5.2 対象データの解析と他データへの適用

提案モデルによって，Web サイトや閲覧ユーザの特性を反映した正規分布表現を学習し，類似性，多様性といった観点から，Web サイトや閲覧ユーザを分析することが可能になった．特に，多様性の定量化については，他手法で行うことは困難であり，意味空間上で Web サイトやユーザの素性を考慮した表現を学習する提案モデルならではの分析の観点であるといえる．

しかしながら，単語の意味の分析 [12] と比較すると，表 5 のように，提案モデルでは関係ない Web サイトどうしの類似性が高くなっている場合が多い．この問題の原因としては，対象データの性質が考えられる．Web サイト閲覧履歴データが文書データと大きく異なる点として，解釈が困難な Web サイトが非常に多いことや，文法やいい回しなどのルールが存在しないことがあげられる．そのために，分析結果を解釈する際に，対象データ中の前提となる関係性に基づいて解釈をすることが望ましいといえる．

本研究の提案モデルは，大量のデータを学習し，複数種類のオブジェクトについて，その類似性と多様性を定量的に測定するためのモデルである．したがって，関係性を仮定できるデータであれば，様々な分野に適用可能であると考えられる．特に，Word2gauss では分散は単語の語義の広さや階層関係を表していたが，本提案モデルにおいては分散は閲覧ユーザ数だけでは測定できないような Web サイトの多様性を表している．このことからも，他分野においても本提案モデルを適用することで，オブジェクトの多様性を検出し，従来手法では得られなかった新たな知見を得ることができるのではないかと期待される．たとえば，顧客の商品の購買履歴データに提案モデルを適用した場合，顧客および商品間の関係性だけでなく，どれだけ多様なユーザに購買されているかといった多様性を，オブジェクトの正規分布によって表現することが想定される．

5.3 マーケティング施策への応用

本研究の提案モデルによる分析結果は Web マーケティ

[*2] Mikilov らの研究 [16] では $d = 400$，$C = 8$ である．また，Tagami らの研究 [9] では $d = 400$，$C = 5$，$K = 5$ であり，ノイズ分布は生起頻度の 3/4 乗を正規化した確率分布である．

ング施策の最適化や効率化への活用が期待される．代表的なWebマーケティング施策として，バナー広告の掲載やメールの配信などがある．以下では，具体的な例を用いて分析結果の活用場面について考える．

ここでは一例として，マーケティング施策を行おうとしている企業Aが存在すると仮定する．企業Aはバナー広告の掲載やメール配信によって自社サイトの閲覧ユーザ数を増加させたいと考えている．バナー広告の掲載は，Webサイトを対象とした施策であり，対象サイトを閲覧したユーザがバナー広告を通じて自社サイトに興味を持ち，自社サイトへアクセスすることを目的とする．対して，メールの配信は，ユーザを対象とした施策であり，対象ユーザへの直接的なアクションによって自社サイトへアクセスすることを目的とする．

このとき，提案モデルによって得られる意味空間上の表現を用いて，自社サイトと類似したWebサイトやユーザに対して施策を実行することで，効率性の実現が考えられる．ここで，実際の閲覧ではなく，意味的に類似している対象に施策を実行する利点として，自社サイトの認知に関係なく施策を実行できることがあげられる．たとえば，若者に人気のスポーツメーカのサイトBが，50代の男性ユーザにも好まれる商品を揃えている場合，少数の50代の男性ユーザのサイトBの閲覧によって，他の50代の男性ユーザともサイトBは類似しているという結果が得られる可能性があり，これは実際の閲覧のみを考慮した場合には無視されてしまう関係性である．

また，バナー広告をWebサイトに掲載する場合，類似しているWebサイト群の中でも，より多数のユーザに閲覧されているWebサイトに掲載した方が施策の効果が得られることが想定される．ここで，本研究で定義した多様性も閲覧ユーザ数と同様に意思決定の基準になりうると考えられる．より多様性の大きいWebサイトに対して施策を実行することで，様々な性質を持ったユーザ群を自社サイトに誘導できる可能性を指摘することができる．ソーシャルメディア上の口コミなどを考慮すれば，様々な性質を持ったユーザ群の顧客化は，さらなる新規顧客の獲得につながることが想定される．

以上の理由から，提案モデルに基づいて定量化されるオブジェクト間の類似性や多様性は，Webマーケティング施策を効果的に実行するために有効であると考えられる．

6. まとめと今後の課題

本研究では，自然言語処理の分野で膨大な数の単語を分析可能とするWord2vecを基礎とし，Webサイト閲覧データに基づきWebサイトと閲覧ユーザの関係性を表現するモデルを新たに提案した．さらに，提案したモデルを実際の閲覧履歴データに適用し，Webサイトおよび閲覧ユーザの関係性の分析を行った．その結果として，Webサイトやユーザの平均ベクトルによって，オブジェクト間の類似性を，Webサイトやユーザの分散によって，オブジェクトの多様性を定量的に表現することができた．本研究の提案モデルにより，インターネット上の各Webサイトの特徴をユーザの閲覧行動という観点から分析可能とし，同時に各ユーザの特徴を意味空間上で把握することも可能となった．

今後の課題として，有用な分析結果を効率的に抽出する方法を検討する必要があるといえる．本研究の分析では，特定のWebサイトやユーザに着目して分析を行ったが，実際にはWebサイト数やユーザ数が膨大であるため，すべてのオブジェクト間の関係性を比較することは難しい．そのため，全体のオブジェクト間の関係性を俯瞰して分析を行う必要があり，それらを考慮した分析手法の拡張が望まれる．また，提案モデルに定量的な評価を与えることも課題としてあげられる．たとえば，Webサイトの推薦タスクに提案モデルを適用することで，その有用性を検討することができると考えられる．

謝辞 本研究にあたり，熱心な議論と貴重なデータの提供をいただいた株式会社ヴァリューズの皆様に深く感謝いたします．

参考文献

[1] 中桐大寿：Webマーケティングと消費者行動，日本情報経営学会誌，Vol.29, No.3, pp.23–28 (2008).
[2] 松田 憲，平岡斉士，杉森絵里子，楠見 孝：バナー広告の単純接触が商品評価と購買意図に及ぼす評価，認知科学，Vol.14, No.1, pp.133–154 (2007).
[3] 千田康弘，渡辺裕明：大規模メール配信とパフォーマンス，情報処理学会研究報告，2003-EVA-006，Vol.2003, No.61, pp.7–12 (2003).
[4] 山元理絵，小林 大，吉原朋宏，小林隆志，横田治夫：アクセスログに基づくWebページ推薦におけるLCSの利用とその解析，情報処理学会論文誌：データベース，Vol.48, No.Sig11(TOD 34), pp.38–48 (2007).
[5] 高須賀清隆，丸山一貴，寺田 実：閲覧履歴を利用した協調フィルタリングによるWebページ推薦とその評価，情報処理学会研究報告，2007-DBS-143，Vol.2007, No.65, pp.115–120 (2007).
[6] 鶴原翔夢，高須賀清隆，丸山一貴，寺田 実：独立成分分析を用いたWeb閲覧履歴の解析とWebページ推薦への応用，DEWS2008，B2-3 (2008).
[7] 石井久治，市川裕介，佐藤宏之，小林 透：Webアクセスログからのパターンマイニングによる購買行動の推定，電気情報通信学会技術研究報告，Vol.109, No.272, pp.89–94 (2009).
[8] 松嵜祐樹，三川健太，後藤正幸：マルコフ潜在クラスに基づくECサイトにおける施策実施効果分析に関する一考察，情報処理学会論文誌，Vol.58, No.12, pp.2034–2045 (2017).
[9] Tagami, Y., Kobayashi, H., Ono, S. and Tajima, A.: Modeling User Activities on the Web using Paragraph Vector, WWW Companion (2015).
[10] Fu, X., Budzik, J. and Hammond, K.J.: Mining navigation history for recommendation, *Proc. 5th Intelligent User Interfaces*, pp.106–112 (2000).
[11] Mobasher, B., Dai, H., Luo, T. and Nakagawa, M.: Effective personalization based on association rule discovery

from Web usage data, *Proc. 3rd Intl. Workshop on Web Information and Data Management*, pp.9–15 (2001).
[12] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *ICLR Workshop* (2013).
[13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, *Advances in NIPS26*, pp.3111–3119 (2013).
[14] Xue, B., Fu, C. and Shaobin, Z.A.: Study on Sentiment Computing and Classification of Sina Weibo with Word2vec, *2014 IEEE International Congress on Big Data*, pp.358–363, IEEE (2014).
[15] Siencnik, S.K.: Adapting word2vec to named entity recognition, *Proc. NODALIDA2015*, pp.239–243 (2015).
[16] Le, Q. and Mikolov, T.: Distributed Representations of Sentences and Documents, *Proc. ICML2014*, pp.1188–1196 (2014).
[17] Phi, V.T., Chen, L. and Hirate, Y.: Distributed representation based recommender systems in e-commerce, 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論文集，C8-1 (2016).
[18] Lai, S., Xu, L., Liu, K. and Zhao, J.: Recurrent Convolutional Neural Networks for Text Classification, *Proc. AAAI15*, pp.2267–2273 (2015).
[19] Vilnis, L. and McCallum, A.: Word representations via gaussian embedding, *ICLR2015* (2015).
[20] Mikolov, T., Karafiat, M., Burget, L., Cernocky, J. and Khudanpur, S.: Reccurent Neural Network Based Language Model, *Proc. INTERSPEECH-2010*, pp.1045–1048 (2010).
[21] Jebara, T. and Kondor, R.: Bhattacharyya and Expected Likelihood Kernels, *Proc. COLT*, Vol.2777 of *LNCS*, pp.57–71 (2003).
[22] Duchi, J.C., Hazan, E. and Singer, Y.: Adaptive Subgradient Methods for Online Learning and Stochastic Optimization, *Jounal of Machine Learning Research*, Vol.12, pp.2121–2159 (2011).

保坂 大樹

1996 年生．2018 年早稲田大学創造理工学部経営システム工学科卒業．現在，同大学大学院創造理工学研究科経営システム工学専攻修士課程在学中．機械学習を用いたデータ分析に関する研究に興味を持つ．

河部 瞭太

1993 年生．2017 年早稲田大学創造理工学部経営システム工学科卒業．2019 年同大学大学院創造理工学研究科経営システム工学専攻修士課程修了．現在，株式会社 NTT データ勤務．在学時，グラフマイニングを用いたビジネスデータ分析に従事．

山下 遥

1987 年生．2010 年東京理科大学理工学部経営工学科卒業．2012 年慶應義塾大学大学院修士課程修了．2015 年慶應義塾大学大学院博士課程修了．博士（工学）．2015 年早稲田大学創造理工学部助手．2017 年より上智大学理工学部助教．品質管理，統計学，情報工学を融合させた新たなデータ解析方法に関する研究に従事．応用統計学会，日本経営工学会，日本品質管理学会等，各会員．

後藤 正幸 （正会員）

1969 年生．1994 年武蔵工業大学大学院修士課程修了．2000 年早稲田大学博士課程修了．博士（工学）．1997 年早稲田大学理工学部助手．2000 年東京大学大学院工学系研究科助手．2002 年武蔵工業大学環境情報学部助教授．2008 年早稲田大学創造理工学部経営システム工学科准教授．2011 年同大教授．情報数理応用とデータサイエンスの研究に従事．著書に，『入門パターン認識と機械学習』，コロナ社 (2014)，『ビジネス統計 統計基礎とエクセル分析』，オデッセイコミュニケーションズ (2015) 等．IEEE，INFORMS，電子情報通信学会，人工知能学会，日本経営工学会，経営情報学会等，各会員．