分布的表現に基づくサイト閲覧行動分析モデルに関する研究

1X14C114-2　保坂 大樹
指導教員　後藤 正幸

1　研究背景と目的

近年，消費者の Web サイト閲覧行動は大変重要なマーケティング分析の対象となっている．Web サイトの閲覧行動を分析することで，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，蓄積されたサイトの閲覧履歴データを分析する際，分析対象となるサイト数やユーザ数は膨大となるため，そのような状況下で有効な分析手法を整備する必要がある．

一方，自然言語処理の分野において，Word2vec [1] と呼ばれる手法が，大量の文書データを学習して膨大な種類の単語の意味を分析する方法として注目されている．Word2vec とは，単語を数十〜数百次元の意味空間上の点で表現する手法であり，得られたベクトルを用いて単語間の意味的な関係を分析できる点が注目されている．また，Word2vec を基礎として，Doc2vec [2] や Word2gauss [3] などの様々な拡張モデルが提案されている．Doc2vec では，単語と同一意味空間上で文書をベクトル化し，文書のトピック情報の表現を可能とする．Word2gauss では，複数の単語を意味空間上の正規分布として同一空間上に表現し，各単語の持つ意味の広がりを表現することを可能とする．

本研究では，文書データにおける単語を Web サイト，文書を閲覧ユーザに置き換えることで，Doc2vec と同様に，意味空間上で，Web 閲覧履歴データを分析することを考える．さらに，Word2gauss のモデルの考え方を応用し，サイトとユーザを意味空間上の正規分布により表現する学習モデルを提案する．これにより，ユーザ間，サイト間，およびユーザとサイト間の類似性だけでなく，各ユーザや各サイトの意味の広がりや多様性を考慮したモデル化が可能となる．また，実際の Web サイト閲覧履歴データに提案手法を適用し，提案モデルの有効性を示すとともにその分析結果を示す．

2　準備

2.1　単語の分散的意味表現

自然言語処理において，従来では与えられた単語に対して，単語を one-hot ベクトルとして表現していた．これは語彙数を次元数とし，ある要素のみを 1 とし他の要素を 0 とすることでどの単語であるかを表現するベクトルである．一方，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．単語の分散的意味表現では，単語ベクトルの各要素が概念を持っていると考えており，比較的密な単語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，単語を射影する先の比較的低次元の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この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．

2.2　従来のモデル

意味空間上の単語ベクトルの学習法として，基本的な手法は Word2vec [1] である．Word2vec では，単語の出現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，学習コーパスにおいて注目単語ベクトルと周辺単語ベクトルとの内積を大きくするように各単語ベクトルを更新する．また，単語ベクトルの過学習を防ぐために，全単語中から確率的に単語をサンプリングするノイズ分布をあらかじめ定義しておき，注目単語に対してその文脈とは関係なく獲得される単語（ネガティブ単語）をこのノイズ分布に従ってサンプリングする．そして，ネガティブ単語と注目単語との内積を小さくするように各単語ベクトルを更新する．学習された単語ベクトルは，類似した意味の単語の検出や他手法の入力ベクトルとするなど，様々な場面で汎用的に用いられ，良い性能が示されている．

一方，単語を意味空間上の正規分布として表現することで，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も提案されている．この正規分布の分散により各単語の多様性を表現できると考えられている．複数の意味を持つような単語や抽象的な単語は分散が大きくなり，特定の文脈でのみ用いられるような単語は分散が小さくなる．Word2gauss [3] では，単語の平均ベクトルと分散行列を学習する．その際，学習コーパス内の各単語の正規分布とその周辺単語の正規分布との類似度が高くなるように各パラメータを更新する．類似度は分布間の内積として知られている Expected Likelihood Kernel（以下，EL）を用いる．EL は，分布同士の積を実数全体で積分した値として定義される．

3　提案モデル

3.1　概要

本研究では，自然言語処理と同様に膨大な種類の Web サイトを扱うという点，データがコーパスと同様にシーケンスデータを含む Web 閲覧履歴データであるという点に着目し，Word2vec が持つ分散的意味表現のアイデアを，Web サイトとユーザのモデル化に導入することを考える．各サイトをベクトルで表現することによって，サイト間の類似性を定量的に測定できると期待される．しかし，このモデルではユーザは意味空間上に表現されず，ユーザの情報を分析に用いることができない．また，ベクトル表現では，サイトは意味空間上の点として表現されるため，サイトの多様性を表現することができない．

そこで，本研究では，各サイトおよび各ユーザを同一意味空間上の正規分布で表現し，閲覧の特性を考慮して全サイトと全ユーザの正規分布の空間配置の獲得を可能とする学習モデル（Site2gauss）を提案する．このモデルは，あるサイトの閲覧について，その前後の閲覧サイトだけでなく閲覧ユーザとも EL が大きくなるようにパラメータを更新することで学習される．これにより，ユーザ間，サイト間およびユーザとサイト間の類似性を分析することや各ユーザや各サイトの閲覧，被閲覧の多様性を分析することが可能となる．

3.2　変数の定義

全サイト数を N，全ユーザ数を M とし，全サイト集合を $\mathcal{S} = \{s_n : 1 \leq n \leq N\}$，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 次元正規分布をそれぞれ $\mathcal{N}_{s_n}$，$\mathcal{N}_{u_m}$ と記述する．ウィンドウサイズを W，学習する周辺サイト数を c，ネガティブサンプル数を k とし，サイト，ユーザのノイズ分布をそれぞれ $noise_S$，$noise_U$ と表す．ウィンドウサイズ W は，前後の閲覧サイトとの関係性を学習するために，注目サイトの前後の W サイトを周辺サイト候補と設定するパラメータである．

3.3　提案モデルの定式化

ユーザ $q \in \mathcal{U}$ が i 番目に閲覧したサイトを $p^{q,i} \in \mathcal{S}$ とし，q と $p^{q,i}$ に対応する正規分布をそれぞれ $\mathcal{N}_q$，$\mathcal{N}_{p^{q,i}}$ とする．$p^{q,i}$ に対し，その前後 W サイトの計 $2W$ サイトから c 個

の周辺サイト $h_j^{q,i} \in \mathcal{S}$ $(1 \leq j \leq c)$ をランダム抽出する．同様に，$h_j^{q,i}$ に対し，$noise_S$ から k 個のネガティブサイト $t_x^{q,i,j} \in \mathcal{S}$ $(1 \leq x \leq k)$ を，$p^{q,i}$ に対し，$noise_U$ から k 人のネガティブユーザ $v_y^{q,i} \in \mathcal{U}$ $(1 \leq y \leq k)$ をそれぞれ抽出する．このとき，注目サイト $p^{q,i}$ に対する，周辺サイトおよび閲覧ユーザとの類似性に対する損失は式 (1) で定義される．

$$
\begin{aligned}
&Loss(p^{q,i}) = \\
&\sum_{j=1}^{c} max\Big\{0, z + \log(\sigma(EL(\mathcal{N}_{p^{q,i}}, \mathcal{N}_{h_j^{q,i}}))) \\
&\qquad\qquad - \sum_{x=1}^{k} \log(\sigma(EL(\mathcal{N}_{p^{q,i}}, \mathcal{N}_{t_x^{q,i,j}})))\Big\} \\
&+ max\Big\{0, z + \log(\sigma(EL(\mathcal{N}_{p^{q,i}}, \mathcal{N}_q))) \\
&\qquad\qquad - \sum_{y=1}^{k} \log(\sigma(EL(\mathcal{N}_{p^{q,i}}, \mathcal{N}_{v_y^{q,i}})))\Big\} \qquad (1)
\end{aligned}
$$

ただし，$\sigma(\cdot)$ はシグモイド関数であり，$EL(\mathcal{N}_a, \mathcal{N}_b)$ は正規分布 $\mathcal{N}_a$，$\mathcal{N}_b$ 間の Expected Likelihood Kernel を表す．また，損失関数の閾値を z とする．

ユーザ q の閲覧サイト数を I_q とすれば，学習データセット全体における損失は以下の式 (2) で定義される．

$$
Loss_{All} = \sum_{q \in \mathcal{U}} \sum_{i=1}^{I_q} Loss(p^{q,i}) \qquad (2)
$$

3.4 パラメータの学習

注目サイトの正規分布を $\mathcal{N}_a = N(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$，周辺サイトもしくは閲覧ユーザの正規分布を $\mathcal{N}_b = N(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b)$ とした場合，学習率を η とすると，各パラメータの更新量 $\Delta\boldsymbol{\mu}_a$，$\Delta\boldsymbol{\Sigma}_a$，$\Delta\boldsymbol{\mu}_b$，$\Delta\boldsymbol{\Sigma}_b$ は以下の式で定義される．これは正規分布同士の EL を微分することにより導かれる．

$$
\Delta\boldsymbol{\mu}_a = -\Delta\boldsymbol{\mu}_b = \eta \cdot \left\{-(\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_b)^{-1}(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)\right\} \quad (3)
$$

$$
\Delta\boldsymbol{\Sigma}_a = \Delta\boldsymbol{\Sigma}_b = \eta \cdot \frac{1}{2}(\Delta\boldsymbol{\mu}_a \Delta\boldsymbol{\mu}_a^{\mathrm{T}} - (\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_b)^{-1}) \quad (4)
$$

ネガティブサイトもしくはネガティブユーザの正規分布を $\mathcal{N}_b$ とした場合は式 (3)，(4) を (-1) 倍したものを更新量とする．式 (2) が収束するまでパラメータを更新する．

4 分析

提案モデルの有用性を示すため，株式会社ヴァリューズ提供のサイト閲覧履歴データに提案手法を適用して分析を行う．

4.1 分析条件

データ期間は 2017 年 8 月 1 日から 2017 年 10 月 31 日であり，総閲覧数は 31,099,962 件，ユーザ数は $N = 10,000$ 人，サイトラベル数は $M = 529,646$ 個である．正規分布の次元数は $d = 50$，ウィンドウサイズは $W = 4$，学習に用いる周辺単語は $c = 2$，ネガティブサンプル数は $k = 2$ とした．ノイズ分布には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用いた．学習には誤差逆伝播法と確率的降下法を用いる．学習率 η の初期値は 0.01 とし，Adagrad [4] に基づいてその値を更新する．また，分散行列は対角成分の値が全て等しい対角行列であると仮定し，その値をオブジェクトの分散として扱う．

4.2 分析結果と考察

分散パラメータの有用性について検討するために，サイトの多様性との関係性について分析を行う．

Word2gauss と同様，分散はサイトの多様性を表していると考えられる．すなわち，分散と閲覧ユーザ数はある程度の正の相関関係にあることが想定される．しかし，全サイトにおける閲覧ユーザ数と分散の相関係数は 0.09 であった．一方で，被閲覧数が 300 以上のサイトに限定するとその相関係数は 0.41 であり，より強い相関を確認した．これは，被閲覧数の小さいサイトはパラメータ更新における初期値への依存が強くパラメータを十分に学習できないためであると考えられる．そこで，分散の考察や解釈では，被閲覧数が 300 以上のサイトのみを対象とする．

次に，具体的にどのようなサイトで分散が大きくなっているかを確認する．また，分散と閲覧ユーザ数の関係について検討する．ここでは，サイトカテゴリ「車」に属する全 48 サイトのうち分散の上位 4 件と下位 4 件について，サイトの分散および全ユーザのうちサイトを閲覧したユーザの割合を表 1 に示す．

表 1: 「車」カテゴリにおけるサイトの分散

サイト名	分散	閲覧ユーザ割合
オートックワン	8.04	8.46%
Autoblog	7.86	1.88%
ダイハツ工業	7.61	2.86%
MAZDA	7.55	6.79%
⋮	⋮	⋮
Honda Cars 総合	4.03	1.18%
Audi Japan	3.76	1.41%
本田技研工業	3.68	13.33%
スズキ	3.47	5.13%

表 1 より，分散上位のサイトは複数のメーカーの車の情報を掲載するポータルサイトであることが確認できる．これらのサイトは，多様なユーザに閲覧される傾向のあるサイトである．よって，多様なユーザに閲覧されている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．一方で，特定のメーカーのサイト間では分散の差異が見受けられ，多様性の高いサイトと低いサイトの双方が存在していることがわかる．これは，分散が大きいことと単純に閲覧ユーザが多いことは等価でないことを示している．

以上より，学習した分布表現によってオブジェクトの多様性を定量的に測ることができたと考えられる．

5 まとめと今後の課題

本研究ではサイト閲覧行動の多角的な分析のために，閲覧履歴データからサイトとユーザの正規分布表現を獲得する学習モデルを提案し，実際の閲覧履歴データを用いて提案モデルの有効性を示した．

今後の課題として，ノイズ分布の再考や最適なパラメータの探索といったモデルの改良や，得られた知見を用いた具体的な施策の提案といったビジネスへの応用などが挙げられる．

参考文献

[1] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality”, *Advances in NIPS26,* pp. 3111-3119, 2013.

[2] Q. Le, and T. Mikolov, “Distributed Representations of Sentences and Documents” *Proc. of the ICML2014,* pp.1188-1196, 2014.

[3] L. Vilnis, and A. McCallum, “Word representations via gaussian embedding”, *arXiv preprint arXiv:1412.6623,* 2014.

[4] J. C. Duchi, E. Hazan, and Y. Singer, ”Adaptive Subgradient Methods for Online Learning and Stochastic Optimization”, *Jounal of Machine Learning Research,* Vol.12, pp.2121-2159, 2011.