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

長瀬　永遠

富山県立大学　情報基盤工学講座

June 4, 2021

はじめに

本研究の背景

近年, インターネット上で, 様々な Web サイトを通じ, 多くの情報・サービスの提供が受けられるようになったことにより, 消費者の Web サイト閲覧機会が大幅に上昇.

⇓

Web サイト上でのマーケティングの重要性が高まっている.

現状の課題

- 複数の Web 閲覧目的を区別して分析することが不可能.
- 複数の目的から生起する複雑な Web サイト間の関係性の表現が困難.

本研究の目的

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方法を提案.

関連研究 1

単語の分散的意味表現

- あらかじめ設定した次元数の単語ベクトルを用いて各単語を表現.
- 単語ベクトルはそれぞれが概念を持つ.
- 類似した単語は意味空間上で近傍に布置.

Word2vec

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手法.
＜ Continuous Skip-gram モデル＞
注目単語から周辺単語群を予測
＜ Continuous Bag-of-Words モデル＞
周辺単語群から注目単語を予測

関連研究 2

Doc2vec

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ように Word2vec を拡張した手法.
＜ Distributed Bag-of-Words モデル＞
所属文書から注目単語と周辺単語群を予測.
＜ Distributed Memory モデル＞
周辺単語群と所属文書から注目単語を予測.

Word2gauss

意味空間上において, 各単語を 1 点ではなく, 正規分布で表現するように Word2vec を拡張した手法. 単語間の位置関係だけでなく, 意味的な広がりも表現可能.

関連研究 3

Expected Likelihood

n 次元正規分布 $N_1 = N(\mu_1, \Sigma_1), N_2 = N(\mu_2, \Sigma_2)$ 間の内積

$$
\begin{aligned}
EL(N_1, N_2) &= \int_{s \in R^n} N(s : \mu_1, \Sigma_1) N(s : \mu_2, \Sigma_2) ds \\
&= \frac{\exp\{-\frac{1}{2}(\mu_1 - \mu_2)^\tau (\Sigma_1 + \Sigma_2)^{-1} (\mu_1 - \mu_2)\}}{(2\pi)^{\frac{n}{2}} \det(\Sigma_1 + \Sigma_2)^{\frac{1}{2}}}
\end{aligned}
\tag{1}
$$

$\mu_1 \in R^n, \mu_2 \in R^n$：各々の平均ベクトル
$\Sigma_1 \in R^{n \times n}, \Sigma_2 \in R^{n \times n}$：分散共分散行列

提案モデル 1

モデルの概要

- Word2vec における文書データを閲覧履歴データに置き換える.
- 「文書」→「ユーザ」,「単語」→「Web サイト」,
  「文脈」→「閲覧目的」
- 同一目的のもとで閲覧された Web サイト群に共通の概念を仮定.
- 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.

提案モデル２

変数の定義

- N：全 Web サイト数, M：全ユーザ数
- S：全 Web サイトの集合, U：全ユーザの集合
- d：構築する意味空間の次元
- $\Sigma_{S_n}, \Sigma_{U_m}$：サイト s_n, ユーザ u_m に対応する d 次元の分散行列
- C：ウィンドウサイズ（どこまでを周辺とするか）
- W：入力層から射影層へ写像するときの重み行列
- Z：射影層から出力層へ写像するときの重み行列

提案モデル 3

モデルの定式化

＜射影層のベクトル＞ $w_{S_n} = W x_{S_n}$　x_{S_n}：入力ベクトル

＜入力サイトに対する各 Web サイト, 各ユーザの予測確立分布＞

$$p(o|r_i^{q,h}) = \frac{EL(N(z_o, \Sigma_o), N(w_{S_n}, \Sigma_{S_n})}{\Sigma_{o' \in S \cup U} EL(N(z_o, \Sigma_o), N(w_{S_n}, \Sigma_{S_n}))} \tag{2}$$

＜入力サイト $r_i^{q,h}$ に対する損失関数＞

$$l(r_i^{q,h}, O_i^{q,h}) = -\Sigma_{o \in O_i^{q,h}} \log p(o|r_i^{q,h}) \tag{3}$$

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 $O_i^{q,h} = \{q\} \cup \{r_{i+C}^{q,h} : -C \leq c \leq -1, 1 \leq c \leq C\}$

提案モデル４

計算の効率化

提案モデルでは, 計算の効率化のためにネガティブサンプリングを用いて計算量を削減する.
＜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＞

$$p_N(o|r_i^{q,h}) = \exp\{-\frac{1}{2}(z_o - w_{S_n})^\tau(\Sigma_o + \Sigma_{S_n})^{-1}(z_o - w_{S_n})\} \quad (4)$$

＜上記を考慮した損失関数＞

$$\begin{aligned} l_N(r_i^{q,h}, O_i^{q,h}) = &- \Sigma_{o\in O_i^{q,h}}\{\log p_N(o|r_i^{q,h}) \\ &+ \frac{1}{K}\Sigma_{k=1}^{K}\log(1 - p_N(\delta_k^o|r_i^{q,h}))\} \end{aligned} \quad (5)$$

＜学習データセット全体の損失関数＞

$$l_{all} = \Sigma_{q\in U}\Sigma_{h=1}^{H_q}\Sigma_{i=1}^{l_{q,h}} l_N(r_i^{q,h}, O_i^{q,h}) \quad (6)$$

提案モデル５

学習アルゴリズム

＜各パラメータにおける勾配＞

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

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

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

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

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

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

分析条件

＜データ＞

- 閲覧履歴データ（株式会社ヴァリューズ）
- 期間：2017 年 8 月 1 日から 2017 年 10 月 31 日
- 総閲覧数：7,224,737 回
- ユーザ数：9,851 人
- Web サイト数：27,789 個

＜条件＞

- 多次元正規分布の次元数：$d = 50$
- ウィンドウサイズ：$C = 10$
- ネガティブサイト・ユーザのサンプリング数：$K = 20$

分析結果 1

表 1: "食べログ" と類似度の高い Web サイト

	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

分析結果 2

表 2: "ZOZOTOWN" と類似度の高い Web サイト

	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

分析結果 3

表 3: 対象ユーザと類似度の高い Web サイト

	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

分析結果 4

表 4: 対象ユーザによる閲覧数が多い Web サイト

	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

分析結果 5

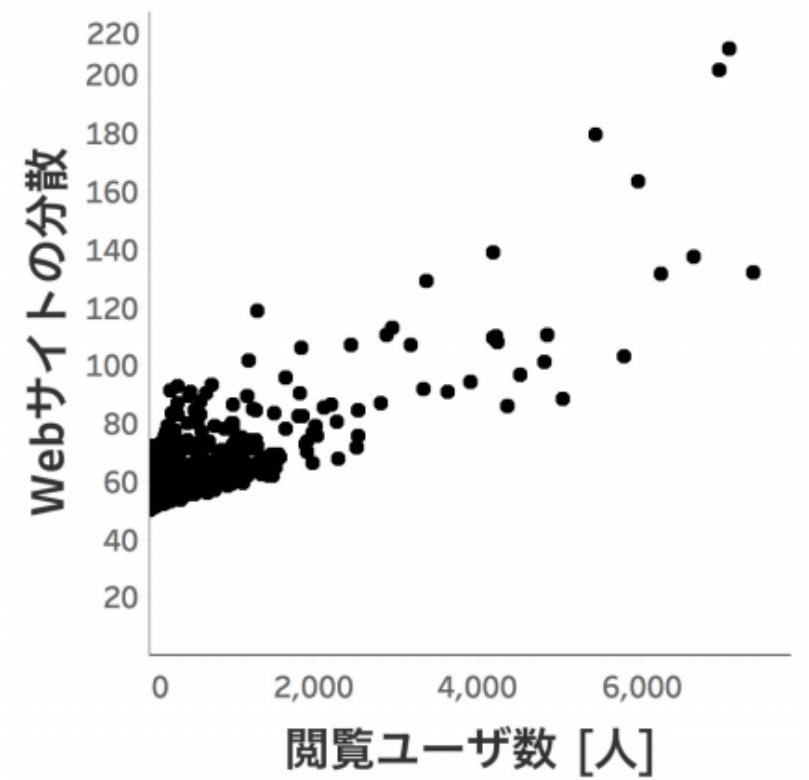

図 1: 閲覧ユーザ数と分散の関係

分析結果 6

表 5: 「車」カテゴリ中で分散の大きかった Web サイト

	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

考察

実現したこと

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新たなモデルの提案.

今後の課題

- 有用な分析結果を効率的に抽出する方法の検討.
- 提案モデルの定量的評価