An analytical model of website relationships based on browsing history embedding considerations of page transitions

Taiju Hosaka*

Graduate School of Creative Science and Engineering,
Waseda University,
Shinjuku-ku, Tokyo, Japan
Email: wasewase@moegi.waseda.jp
*Corresponding author

Haruka Yamashita

Faculty of Science and Technology,
Sophia University,
Chiyoda-ku, Tokyo, Japan
Email: h-yamashita-1g8@sophia.ac.jp

Masayuki Goto

School of Creative Science and Engineering,
Waseda University,
Shinjuku-ku, Tokyo, Japan
Email: masagoto@waseda.jp

Abstract: In recent years, obtaining a large amount of information and receiving various services through websites has become possible. Consequently, web browsing activities are increasing and numerous companies are conducting their businesses online. In this scenario, users' web browsing behaviour is one of the important topics in web marketing analysis. Several studies have been conducted with distributed representation model to analyse web browsing history. This method can capture the relationships between the websites continuously browsed as similar representations. However, the browsing behavior without a clear browsing purpose in web browsing history can deteriorate the representations. In this paper, we propose the sparse skip gram model using the regularised online learning approach for analysing website relationships robustly. In addition, we apply our method to actual browsing history data and discuss the findings acquired from the analytical results. We show that our proposed model represents the characteristics of websites with subspaces in the embedding space.

Keywords: distributed representation; word2vec; sparse regularisation; browsing history data; business analytics.

Reference to this paper should be made as follows: Hosaka, T., Yamashita, H. and Goto, M. (2021) ‘An analytical model of website relationships based on browsing history embedding considerations of page transitions’, *Asian J. Management Science and Applications*, Vol. 6, No. 1, pp.1–16.

Biographical notes: Taiju Hosaka is a worker in Sansan, Inc., Japan. He received his Bachelor’s degree from the Waseda University in 2018 and his Master’s degree in 2020. His main research interests are machine learning and statistical modelling.

Haruka Yamashita is an Assistant Professor at the Department of Information and Communication Sciences, Sophia University, Japan. She received her PhD from the Keio University in 2015. Her research interests include machine learning, applied statistics, and multivariate data analysis.

Masayuki Goto is a Professor of School of Creative Science and Engineering, Waseda University, Japan. He received his DrEng from the Waseda University in 2000. His research interests include data science, machine learning, business analytics, information theory and its application, and marketing analysis. He is a member of IEEE, INFORMS, the Institute of Electronics, Information and Communication Engineers, Japan Industrial Management Association, the Japan Society for Management Information, the Japanese Society for Artificial Intelligence, Business Model Association, and the Operations Research Society of Japan.

This paper is a revised and expanded version of a paper entitled ‘An analytical model of web sites relationship based on browsing history embedding considering page transitions’ presented at 2019 Asian Conference of Management Science & Applications, Penglai, Shandong, China, 11–14 October 2019.

1 Introduction

In recent years, various services are being provided to consumers through various websites over the internet. For example, one can purchase a variety of items on electronic commerce sites (EC sites) and ask for the purchased items to be delivered to their homes. When going for dinner, gourmet websites allow us to search for preferable restaurants and make a reservation. Internet has therefore become an indispensable tool in our daily life. The accumulated web browsing history generated by the browsing behaviour of consumers reflects their preferences and several studies have been conducted on this topic from various viewpoints. For example, Liu et al. (2015) employed a convolutional neural network model to predict the tendency of consumers to click on advertisements based on the web browsing behaviour as input. Marketing analysis based on web browsing history data is a well-investigated approach in this field. Marketing analysis is performed to optimise marketing measures such as advertisement on the internet or mail distribution, and the consumer preferences are frequently explained in terms of the websites browsed by them. For instance, it is assumed that a consumer who has browsed several fashion sites is interested in fashion and will visit

other fashion sites in the future as well. However, directly interpreting and analysing the characteristics of the large number of browsed websites on the internet is difficult.

Hence, a method for conducting an effective analysis that estimates the characteristics of each website through web browsing history data is required. In previous studies, such as by Tsuruhara et al. (2008), the authors performed an independent component analysis of browsing history data and extracted the browsing concepts. Schroder et al. (2019) segmented the consumer data based on the topic model and analysed the type of websites the consumers of each segment were interested in. Notably, these approaches assumed that all the websites browsed by a single consumer are largely similar.

However, considering the web browsing activities performed by each user, although there are strong relationships between the websites sequentially browsed before and after a particular website, not all pairs of websites browsed by the same consumer are similar. For example, if a consumer planning a trip browses hotel-booking sites and gourmet sites continuously, these sites are browsed for planning the trip. Therefore, it should be considered that there might be multiple browsing purposes in a consumer's web browsing history data. In this situation, if we apply the methods that define the similarities between websites by simple co-occurrence at the consumer level, distinguishing between multiple browsing purposes can be challenging. Therefore, it is desirable that an analytical model that considers the similarities between local browsing actions instead of within the complete browsing history is applied.

On the contrary, in the field of natural language processing, extensive research has been performed on the distributed representation of text data that shows high performance in the semantic analysis of words. Distributed representation is a technique that represents the meaning of a word through its semantic vector; word2vec proposed by Mikolov et al. (2013a) is a typical model for learning distributed representation. In this model, each word in the corpus is assumed similar to the context words that tend to appear in the same sentence and the representation learning is accomplished by an online learning algorithm. With the same logic, objects in web browsing history data also show similarity to surrounding objects. Later, Tagami et al. (2018) and Hosaka et al. (2019) demonstrated the effectiveness of applying distributed representation models such as word2vec to web browsing history data as well.

However, there is a significant difference between web browsing history data and text data. The words in the text data are based on clear contextual meanings, whereas browsing behaviour does not always reflect the consumers' browsing purpose. For example, during multi-purpose browsing, there is a possibility that websites sequentially browsed before and after a particular one are based on different purposes. Moreover, if one gets a phone call during web browsing, the browsing behaviour can change suddenly because of the content being discussed on the phone. Thus, there are several website relationships that cannot be extracted in a straightforward manner from web browsing history.

In this paper, we propose a distributed representation learning method of strong relationships in the web browsing history data considering the novel sparse regularisation for the estimation. The browsing behaviour of users is not always generated by a rational process and websites sequentially viewed by a user may have no relationship among themselves. In a few cases, a user may browse websites randomly in his/her spare time. However, most users usually browse websites on the internet sequentially with a certain context or meaning. The relationship between the websites

that we focused on in this study is in the meaning of the relative frequency with which they are browsed by each user in the same context. If the websites that tend to be browsed after or before a target website can be identified, we can form a hypothesis about the contexts and intentions behind the browsing behaviour. Such hypotheses can be useful in considering appropriate promotion strategies and marketing measures. Occasionally, reasonably interpreting the relationships between websites generated by the browsing behaviours of the users may be difficult. However, such findings may be useful for marketing planning as it is difficult for an individual marketer to imagine them based on just his/her own experiences or logical thinking.

Website relationships can be appropriately analysed using our approach. In particular, we introduce follow the regularised leader-proximal (FTRL-proximal), an online learning algorithm suggested by McMahan et al. (2013) that enables regularised learning of the word2vec model and the estimation of sparse distributed representation of the websites. The algorithm allows the extraction of strong relationships between the websites by using sparse distributed representations constructed using the proposed method. Finally, we apply the proposed method to actual web browsing history data and analyse the relationships between websites based on the results of distributed representation. We demonstrate the effectiveness of the proposed model by performing an analysis of the real data.

2 Related work

2.1 Distributed representation

In the field of natural language processing, one of the most important issues is the numerical representation of words. Previously, two major methods were used to represent words in a numerical manner: one-hot encoding method that treats each word as a dummy variable, and the term frequency and inverse document frequency (TF-IDF) method that assigns weights to the word characteristics. However, in recent years, distributed representation methods that provide numeric representations in an embedding space after considering the meaning of the words have gained prominence. This technique represents each word as a numerical large dimensional vector in the constructed embedding space. Words with similar meanings have similar distributed representations and the distance between the word vectors quantifies the similarity between the meanings of the words. However, to use the distributed representation efficiently, an appropriate algorithm is necessary to estimate the word vector accurately. In this context, word2vec, suggested by Mikolov et al. (2013a) is the most typical model to learn the distributed representation and construct an embedding space.

The word2vec model estimates the distributed representation using neural networks based on the hypothesis that a word in the corpus can be predicted from the context words. There are two learning approaches used for word2vec. In the skip gram model, a one-hot vector of the words in the corpus is received as input data and two-layer neural network learning is employed to predict the context words. Then, the weight vector from the input unit to the hidden layer is extracted as the word vector. In practice, from the viewpoint of computational complexity of the learning algorithm, negative sampling, which approximates a multiclass classification problem with a few binary

classification problems, as suggested by Mikolov et al. (2013b), is widely used. The learning algorithm of the skip gram model with negative sampling is shown below.

We denote the number of words by I and the size of the vocabulary by V . Further, let the width of the window size W before and after be regarded as the context. Moreover, let the negative sample size K be the number of negative examples to be sampled for each word. Let $\mathbf{u}_v$ and $\mathbf{c}_v$ be the D -dimensional word vector and context vector for the word v ($1 \leq v \leq V$) respectively. Then, we build a noise distribution θ as follows:

$$\theta_v = \frac{(freq_v)^{3/4}}{\sum_{v'=1}^V (freq_{v'})^{3/4}} \quad (1)$$

Here, $freq_v$ is the frequency of the word v in the corpus. In the skip gram model, K negative words w_k ($1 \leq k \leq K$) are sampled for the combination of a word w_i in the corpus. The context word w_{i+j} ($-W \leq j \leq W, j \neq 0$) and loss $L(w_{i+j}|w_i)$ is calculated as follows:

$$L(w_{i+j}|w_i) = -\log \sigma(\mathbf{c}_{w_{i+j}}^\top \mathbf{u}_{w_i}) - \frac{1}{K} \sum_{k=1}^K \log \sigma(-\mathbf{c}_{w_k}^\top \mathbf{u}_{w_i}) \quad (2)$$

Here, $\sigma(\cdot)$ is the sigmoid function. Moreover, the loss for the corpus L_{SGNS} is calculated as follows:

$$L_{\text{SGNS}} = \sum_{i=1}^I \sum_{-W \leq j \leq W, j \neq 0} L(w_{i+j}|w_i) \quad (3)$$

The model parameters $\mathbf{u}_v$, $\mathbf{c}_v$ ($1 \leq v \leq V$) are estimated by minimising equation (3) according to the stochastic gradient descent.

Word2vec performs excellently in several tasks of natural language processing. For example, Xue et al. (2014) applied word2vec to text data collected from social media and conducted a sentiment analysis. They calculated a score to represent the positivity or negativity of a word based on the similarity between the words. In addition, Siencnik (2015) applied a clustering method to the word vector learned with word2vec and used the clusters for named entity recognition. The results confirmed that the model shows higher accuracy than the linear support vector machine for classification baseline.

2.2 Online learning with regularisation

Using the stochastic gradient descent to update the parameters in the negative gradient direction of loss per data is the most basic optimisation method used in online learning. However, the stochastic gradient descent cannot be directly applied to a loss function that includes nondifferentiable points such as the L1 norm. Although Duchi et al. (2011) adapted the subgradient method for online learning and estimation of parameters even in such a situation, it is difficult to learn sparse parameters from a loss function with L1 norm using the subgradient method.

However, the effectiveness of a learning algorithm that combines the stochastic gradient descent and the proximal gradient method has been demonstrated in previous

studies. Duchi and Singer (2009) suggested an online learning algorithm called forward backward splitting (FOBOS). This algorithm can optimise the loss function, including the regularisation term, efficiently by dividing it into differentiable and nondifferentiable terms and alternately applying the stochastic gradient descent and the proximal gradient method. Xiao (2009) suggested an online learning algorithm called regularised dual averaging that sparsifies a parameter based on the past gradients. Furthermore, McMahan et al. (2013) suggested FTRL-proximal, a hybrid of FOBOS and regularised dual averaging algorithms. This algorithm can efficiently estimate a parameter with the desired structure according to the regularisation term because it learns the parameter by referring to not only the past gradients but also the past parameter values. The learning algorithm of FTRL-proximal with L1 regularisation is presented below.

We denote the loss function for parameter $\mathbf{x}$ by $\rho(\mathbf{x})$. $\rho(\mathbf{x})$ can be divided into the differentiable function and the L1 regularisation term. We denote the differentiable function by $f(\mathbf{x})$ and the regularisation strength by β . Then, $\rho(\mathbf{x})$ is represented as follows:

$$\rho(\mathbf{x}) = f(\mathbf{x}) + \beta \|\mathbf{x}\|_1 \quad (4)$$

Here, let $\mathbf{x}_t$ be the value of $\mathbf{x}$ after t^{th} update. Then, $\mathbf{x}_{t+1}$ is calculated as follows:

$$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x}} \left\{ \sum_{\tau=1}^t \nabla f(\mathbf{x}_{\tau})^{\top} \mathbf{x} + t\beta \|\mathbf{x}\|_1 + \frac{1}{2} \sum_{\tau=1}^t \sigma_{\tau} \|\mathbf{x} - \mathbf{x}_{\tau}\|^2 \right\} \quad (5)$$

$$\sigma_{\tau} = \begin{cases} \frac{1}{\eta_{\tau}} - \frac{1}{\eta_{\tau-1}} & \tau \geq 2 \\ \frac{1}{\eta_{\tau}} & \tau = 1 \end{cases} \quad (6)$$

$$\mathbf{y}_t = \sum_{\tau=1}^t \nabla f(\mathbf{x}_{\tau}) - \sum_{\tau=1}^t \sigma_{\tau} \mathbf{x}_{\tau} \quad (7)$$

$$x_{t+1,i} = \begin{cases} 0 & |y_{t,i}| \leq \beta \\ -\eta_t(y_{t,i} - \text{sign}(y_{t,i})\beta) & \text{otherwise} \end{cases} \quad (8)$$

Here, η_t is the learning rate of t^{th} update and $x_{t,i}$ is the i^{th} factor of $\mathbf{x}_t$.

2.3 Sparse distributed representation

In this study, our goal was to estimate the sparse distributed representation of websites. Several previous studies have also reported the construction of a model to estimate such sparse distributed representation from the point of view of interpretability.

Faruqui et al. (2015) proposed a method to map the estimated word vectors linearly to the higher-dimensional and sparse vector space using the matrix factorisation approach. In their research, the loss function L_{SMF} was defined as follows:

$$L_{\text{SMF}} = \sum_{v=1}^V \|\mathbf{u}_v - \mathbf{Q}\mathbf{s}_v\|_2^2 + \beta \|\mathbf{s}_v\|_1 + \gamma \|\mathbf{Q}\|_2^2 \quad (9)$$

Here, $\mathbf{Q}$ is the dictionary of base vectors and $\mathbf{s}_v$ is the sparse word vector for the word v ($1 \leq v \leq V$). β is the L1 regularisation strength of $\mathbf{s}_v$ and γ is the L2

regularisation strength of $\mathbf{Q}$. Faruqui et al. (2015) optimised equation (9) by using regularised dual averaging. They showed that it was possible to learn word vectors with high interpretability and representability by sparsifying them while increasing the dimensions of the embedding space.

Besides Faruqui et al. (2015), several other researchers have also successfully transformed the word vector to a sparse word vector. Park et al. (2017) proposed a vector rotation algorithm that decreases the complexity of the representation. Moreover, Subramanian et al. (2018) proposed a method to nonlinearly transform the distributed representation of each word to high-dimensional sparse vectors by applying a k-sparse autoencoder suggested by Makhzani and Frey (2014). They evaluated the interpretability of each axis in the embedding space by using the word intrusion detection test and demonstrated the improvement in interpretability.

In contrast, Sun et al. (2016) proposed a method to learn the sparse word vector by introducing regularised dual averaging to the continuous bag-of-words (CBoW) model, one of the structural models of word2vec. They built an algorithm to sparsify and estimate the distributed representation of each word simultaneously, unlike in the other studies mentioned above.

3 Proposed model

In this paper, we propose a model to learn the website vector for extracting strong relationships from web browsing history data by introducing the sparse regularisation method to the distributed representation model. Previously, Sun et al. (2016) had proposed a similar algorithm that combined regularised dual averaging and the CBoW model. However, the parameters of infrequent words estimated by the CBoW model are unstable because it predicts the word from the context and there are representation conflicts between words occurring in similar contexts. Moreover, it has been shown that FTRL-proximal can estimate parameters more accurately than regularised dual averaging because it updates the parameters while referring to their past values. Furthermore, online learning is compatible with web browsing history data not only from the perspective of the convergence of the calculations, but also because numerous websites are created or shut down every day, and the semantic structure of the websites keeps changing constantly.

Thus, we propose a new model to estimate the sparse website vector with higher accuracy by introducing FTRL-proximal to the skip gram model. Our proposed model is formulated as follows.

Let I and V be the number of total views and websites, respectively. Additionally, we denote the window size and the negative sample size with W and K , respectively. Further, we denote D -dimensional website vector as $\mathbf{u}_v$ and context vector as $\mathbf{c}_v$ for website v ($1 \leq v \leq V$). We build noise distribution θ according to equation (1). Furthermore, let λ , η , and α be the regularisation strength, and the learning rates of website vector and context vector, respectively. We adjust the learning rate according to Adagrad suggested by Duchi et al. (2011) and the adjusted parameter is denoted as ϵ .

The loss for each combination of the website w_i and the context website w_{i+j} ($-W \leq j \leq W$, $j \neq 0$) is calculated according to equation (2). Then, the loss for learning dataset L_{SSGNS} is calculated as follows:

$$L_{\text{SSGNS}} = \sum_{i=1}^I \sum_{-W \leq j \leq W, j \neq 0} L(w_{i+j}|w_i) + \lambda \sum_{v=1}^V \|\mathbf{u}_v\|_1 \quad (10)$$

Moreover, the gradient of equation (2) at each parameter is calculated as follows:

$$\frac{\partial L(w_{i+j}|w_i)}{\partial \mathbf{u}_{w_i}} = -\sigma(-\mathbf{c}_{w_{i+j}}^\top \mathbf{u}_{w_i}) \mathbf{c}_{w_{i+j}} + \frac{1}{K} \sum_{k=1}^K \sigma(\mathbf{c}_{w_k}^\top \mathbf{u}_{w_i}) \mathbf{c}_{w_k} \quad (11)$$

$$\frac{\partial L(w_{i+j}|w_i)}{\partial \mathbf{c}_{w_{i+j}}} = -\sigma(-\mathbf{c}_{w_{i+j}}^\top \mathbf{u}_{w_i}) \mathbf{u}_{w_i} \quad (12)$$

$$\frac{\partial L(w_{i+j}|w_i)}{\partial \mathbf{c}_{w_k}} = \frac{1}{K} \sigma(\mathbf{c}_{w_k}^\top \mathbf{u}_{w_i}) \mathbf{u}_{w_i} \quad (13)$$

The calculation procedure for the t^{th} update of website vector $\mathbf{u}_{w_i}$ is shown below. Let $\mathbf{u}_{w_i,t}$ be the value of $\mathbf{u}_{w_i}$ before the t^{th} update. In addition, let $\mathbf{g}_{w_i,t}$ and $\mathbf{g}_{w_i,t}^2$ be sum of the gradients at $\mathbf{u}_{w_i}$ until the t^{th} update and the square sum of the gradients at $\mathbf{u}_{w_i}$ until the t^{th} update, respectively. In other words, $\mathbf{g}_{w_i,t}$ and $\mathbf{g}_{w_i,t}^2$ are updated as follows:

$$\mathbf{g}_{w_i,t} = \mathbf{g}_{w_i,t-1} + \frac{\partial L(w_{i+j}|w_i)}{\partial \mathbf{u}_{w_i,t}} \quad (14)$$

$$\mathbf{g}_{w_i,t}^2 = \mathbf{g}_{w_i,t-1}^2 + \left(\frac{\partial L(w_{i+j}|w_i)}{\partial \mathbf{u}_{w_i,t}} \right)^2 \quad (15)$$

Then, the learning rate of the t^{th} update $\eta_{w_i,t}$ is calculated as follows:

$$\eta_{w_i,t} = \frac{\eta}{\epsilon + \sqrt{\mathbf{g}_{w_i,t}^2}} \quad (16)$$

Finally, $u_{w_i,t,d}$ that is the d^{th} element ($1 \leq d \leq D$) of $\mathbf{u}_{w_i,t}$ is updated as follows:

$$\phi_{w_i,\tau} = \begin{cases} \frac{1}{\eta_{w_i,\tau}} - \frac{1}{\eta_{w_i,\tau-1}} & \tau \geq 2 \\ \frac{1}{\eta_{w_i,\tau}} & \tau = 1 \end{cases} \quad (17)$$

$$\mathbf{z}_{w_i,t} = \mathbf{g}_{w_i,t} - \sum_{\tau=1}^t \phi_{w_i,\tau} \mathbf{u}_{w_i,\tau} \quad (18)$$

$$u_{w_i,t+1,d} = \begin{cases} 0 & |z_{w_i,t,d}| \leq \frac{\lambda}{\eta_{w_i,t}} \\ -\eta_{w_i,t} \left(z_{w_i,t,d} - \text{sign}(z_{w_i,t,d}) \frac{\lambda}{\eta_{w_i,t}} \right) & \text{otherwise} \end{cases} \quad (19)$$

The context vector $\mathbf{c}_v$ is estimated by applying the stochastic gradient descent based on Adagrad similar to the skip gram model.

4 Analytical experiments with actual data

In this study, we applied the proposed model to real web browsing history data and obtained sparse distributed representation of websites. In addition, we analysed the relationships between the websites based on the learned representation to verify the effectiveness of the proposed model.

4.1 Experimental setup

We employed the web browsing history data provided by VALUES, Inc., Japan, to demonstrate the suitability of our proposed model in real-world data analysis. The data comprise the host names of websites browsed on a PC or smartphone captured by a monitor who agreed to register for the survey. The data include $I = 78,508,580$ total views, 49,787 users, and $V = 47,854$ websites. The model parameters of our model are described below.

We set the dimensions of the website and context vectors as $D = 200$, the window size as $W = 4$, and the negative sample size as $K = 20$. In addition, we set the regularisation strength as $\lambda = 2.5$, the learning rate of the website and context vectors as $\eta = \alpha = 0.5$, and the adjusted parameter of Adagrad as $\epsilon = 1.0$, based on preliminary analysis.

We extracted websites that displayed high similarities with a particular target website and verified the reasonability of the resultant estimated website relationships. Moreover, as the evaluation of the model was based on how much the model can suppress the extraction of weak relationships, we defined the average sparse rate of the website vector S_{vec} and that of cosine similarity matrix generated by the website vector S_{sim} as follows:

$$S_{\text{vec}} = \frac{1}{VD} \sum_{v=1}^V \sum_{d=1}^D \mathbb{I}(u_{v,d} = 0) \quad (20)$$

$$S_{\text{sim}} = \frac{2}{V(V-1)} \sum_{v=1}^{V-1} \sum_{v'=v+1}^V \mathbb{I}(\cos(\mathbf{u}_v, \mathbf{u}_{v'}) = 0) \quad (21)$$

Here, $\mathbb{I}(\cdot)$ is an indicator function to evaluate if the relationship in parentheses is correct or not and $\cos(\mathbf{a}, \mathbf{b})$ represents the cosine similarity between $\mathbf{a}$ and $\mathbf{b}$.

We compared the proposed model with the models proposed by Mikolov et al. (2013a) (skip gram model), Faruqui et al. (2015), and Sun et al. (2016)¹ For the model of Sun et al. (2016), we used FTRL-proximal for the optimisation of the CBoW model instead of regularised dual averaging, considering the comparability with the proposed model. Moreover, we adjusted the parameters of the models proposed by Faruqui et al. (2015) and Sun et al. (2016) to bring their results closer to that of the proposed model. Further, we set the website vector learned by the model of Mikolov et al. (2013a) as the input data for the model of Faruqui et al. (2015).

4.2 Comparison with previous study

In this section, we discuss the characteristics of the representation learned by the proposed model by comparing them with the representations obtained in previous studies. To measure the effectiveness of the representation, we extracted websites displaying a high similarity with the target website and analysed the website relationships on the embedding space. Here, as the extent of similarity was measured by considering the number of shared axes, we defined the sign-dot similarity between the D -dimensional vectors $\mathbf{a}$ and $\mathbf{b}$ as follows:

$$\text{sign-dot}(\mathbf{a}, \mathbf{b}) = \sum_{d=1}^D \text{sign}(a_d) \text{sign}(b_d) \quad (22)$$

Because of the imaginability, we set ‘Tabelog (tabelog.com)’, a typical Japanese gourmet site, as the target site of the analytical experiments in this study.

Table 1 Top five websites with high cosine similarity with the website vector of ‘Tabelog’

	<i>Mikolov et al.</i>	<i>Faruqui et al.</i>	<i>Sun et al.</i>	<i>Proposed model</i>
No. 1	Gourmet site A	Gourmet site A	Gourmet site G	Gourmet site H
No. 2	Gourmet site B	Restaurant site A	Restaurant site C	Fashion site A
No. 3	Gourmet site C	Hotel site A	Gourmet site A	Regional site B
No. 4	Gourmet site D	Gourmet site F	Regional site A	Gourmet site C
No. 5	Gourmet site E	Restaurant site B	Restaurant site D	Restaurant site E

Table 2 Top five websites with high sign-dot similarity with the website vector of ‘Tabelog’

	<i>Sun et al.</i>		<i>Proposed model</i>
No. 1	Hotel booking site A	No. 1	Coffee shop site A
No. 2	Gourmet site G	No. 2	Gourmet site C
No. 2	Gourmet blog site A	No. 2	Gourmet site I
No. 2	Cloud service A	No. 3	Hotel booking site A
No. 2	Hotel booking site B	No. 3	Gourmet site J

Table 3 The sparsity of each model

	S_{vec}	S_{sim}
Mikolov et al.	0.0%	0.0%
Faruqui et al.	93.2%	0.0%
Sun et al.	94.7%	1.0%
Proposed model	95.2%	22.1%

Table 1 shows the top five websites with a high cosine similarity with the website vector of ‘Tabelog’ for each model. According to Table 1, the model of Mikolov et al. (2013a) can extract gourmet sites similar to ‘Tabelog’. However, the models of Faruqui et al. (2015) and Sun et al. (2016) extracted restaurant and hotel sites in addition to gourmet sites as these sites are expected to have strong relationships with gourmet sites. In the case of the proposed model, fashion sites and regional sites were also extracted in addition to the gourmet sites. It is considered that fashion sites, similar to gourmet sites, are browsed by young consumers. Next, Table 2 shows the top five websites with high a sign-dot similarity with the website vector of ‘Tabelog’. According to Table 2, the model of Sun et al. (2016) and the proposed model extracted gourmet sites and websites with high relevance to gourmet sites. Therefore, it is evident that the proposed model also learns representation by considering the meaning as observed in the previous studies. In addition, a clear concept can be assigned to each axis of the embedding space where the website is mapped by the proposed model.

Next, we discuss the sparsity of the parameters learned by the proposed model. Table 3 shows the average sparse rate of the website vector S_{vec} and the cosine similarity matrix of the website vector S_{sim} for each model. According to Table 3, the model of Mikolov et al. (2013a) constructed a dense vector and almost all the elements of the

word vector are non-zero. On the contrary, the average sparse rate of the website vector learned by the model of Faruqui et al. (2015) is more than 90% owing to the adjusted parameter. However, in this method, the cosine similarity with the website vector is zero. The model of Faruqui et al. (2015) transforms the learned distributed representation to sparse vector space while maintaining the relationships. Thus, this model cannot sparsify the close relationships constructed by the model of Mikolov et al. (2013a). On the contrary, the proposed model can sparsify not only the website vector but also the cosine similarity matrix because this model learns and sparsifies the distributed representation simultaneously. This result implies that websites have a structure that can be distributed on multiple subspaces in the embedding space. However, the model of Sun et al. (2016) cannot learn the sparse relationships. There are two possible reasons for obtaining this result. First, the CBoW model averages the context vectors and the website vectors are assumed to have relationships with more abstract representations. Thus, the CBoW model has difficulty in learning the characteristics of websites with low frequency. Second, the CBoW model considers the website as a negative sample for the context websites whereas the skip gram model samples the context websites positively. Therefore, the CBoW model emphasises learning the characteristics of the context websites instead of that of the website.

Table 4 Top five websites with high values in certain dimensions of the website vector space

	$d = 118$	$d = 133$
No. 1	Hot spring navigation site A	App game walkthrough site A
No. 2	Theme park official site A	App game walkthrough site B
No. 3	Hotel booking site B	App game walkthrough site C
No. 4	Hotel booking site C	App game walkthrough site D
No. 5	Hotel booking site D	App game walkthrough site E

Figure 1 The scatter of websites on the certain subspace in embedding space
(see online version for colours)

4.3 Analysis by proposed model

In this section, we give an example of the analysis performed by the proposed model. First, we interpret the subspace in the embedding space learned by the proposed model. Table 4 shows the top five websites with a high value in certain dimensions of the website vector learned by the proposed model. According to Table 4, the 118th and 133rd axes of embedding space learned by the proposed model are assigned as interpretable topics. The websites with a high value on the 118th axis of the embedding space are relevant to travel or leisure. Moreover, a high value on the 133rd axis indicates relevance with app games walkthrough. Moreover, Figure 1 shows the scatter of websites on the subspace comprising the 28th, 49th, and 157th axes of the embedding space learned by the proposed model. Here, we select the automotive websites in the dataset arbitrarily. In Figure 1, the automotive websites are represented by a star sign and the websites in subspace² are represented by a darker color. The percentage of automotive websites in the subspace in all automotive websites is 60.5% while that of non-automotive websites is 7.2%. Thus, it can be said that this subspace represents the characteristic of automotive websites. In summary, according to Table 4 and Figure 1, the proposed model represents the characteristics of a website by using a subspace in the embedding space.

Figure 2 Browsing sequence of a consumer and the sign-dot similarity to previous websites (see online version for colours)

Figure 2 shows the browsing sequence of a consumer in a session³ and the average of the sign-dot similarities to previous W websites. According to Figure 2, this consumer checked the weather on a weather forecast site. Next, he/she viewed various news sites and a baseball team site, and used the mail and social networking services. Afterward, the consumer browsed multiple automotive websites and finally viewed a few bank sites. According to the figure, the partial browsing sequence from the 9th to the 16th browse has high general similarities among the visited websites. In other words, this partial browsing sequence can be interpreted as browsing on the same subspace in the embedding space with an interest in automotive websites. Meanwhile, a low similarity means changing the subspace in the embedding space and the partial browsing sequences

with overall low similarities with immediately previous websites can be interpreted as browsing without a certain purpose, similar to internet surfing. Hence, we can detect a consumer's interests while web browsing by superimposing the similarities between the website vectors learned by the proposed model and of the actual web browsing. Furthermore, as the similarities can be calculated simultaneously with browsing, we expect that this analysis will be applied to web marketing measures in real-time.

5 Discussion

5.1 Distributed representation model for browsing history data

In Subsection 4.2, we verified that distributed representations learned by word2vec and the expansion models can extract websites similar to a given website. This is due to the structural similarity between the text data and the browsing history data. Both the types are time series data and include a large number of objects. Moreover, Piantadosi (2014) and Adamic and Huberman (2002) have shown that frequencies of the words and websites follow Zipf's law.

In this study, we focused on the possibility that similarities may exist between proximately visited websites, based on the characteristic assumed by word2vec. This is a cogent hypothesis because we browse the internet with a single browsing purpose in the short-term but with a variety of browsing purposes in the long-term. Consequently, each method can extract reasonable relationships by calculating the cosine similarity between learned distributed representations.

The distributed representation of websites can be applied to machine learning tasks related to the internet. Actually, Tagami et al. (2018) learned distributed representation of browsing consumers from the web browsing history data and applied it to advertisement click prediction.

5.2 Regularisation of distributed representation

Among the previous studies on sparse distributed representation, the majority proposed a model that transformed the representation learned by other models to the sparse representation. However, the model of Sun et al. (2016) and our proposed model learn and sparsify the representations simultaneously.

According to the results presented in Subsection 4.2, the sparsity of the learned representations' similarity of the proposed model differs greatly from that of the previous models although all of them can learn the sparse representation. According to Table 1, all the models in previous studies and the proposed model can extract strong relationships between websites. However, the results for websites with weak relationships are different. The proposed model estimates the similarity between these websites as zero whereas the models in previous studies still display a small value.

The models that transform the representation learned by other models to the sparse representation retain the dense relationships constructed by the original representation. In addition, the model of Sun et al. (2016) cannot learn the sparse similarity because of the CBoW model feature that averages the context vectors and samples the websites as negative samples. Therefore, our proposal model demonstrates a novel method to sparsify the similarity between the websites. Furthermore, the difference between our

proposed model and that of Sun et al. (2016) proves the advantage of the skip gram model over the CBoW model from the perspective of capturing the characteristics of the websites.

5.3 *Significance of removing weak relationships*

While considering the analysis of relational data, most studies focus on finding the high similarity between objects, such as items in the purchase history, because several tasks pertaining to relational data aim to extract strong relationship from big data. In other words, from a micro perspective, the purpose of learning is to create a high similarity between objects with strong relationship. For example, on an EC site, the users are recommended items that are similar to items they purchased in the past.

However, extracting strong relationships is insufficient from a macro perspective. If you want to analyse the multiple relationships between objects comprehensively, simply extracting the strong relationships can cause confusion as well as difficulties in interpreting the analysis results. Therefore, limiting the extraction of non-strong relationships is as important as extracting strong relationships while analysing multiple relationship objects simultaneously.

In actual analysis, we often manage relationships such as similarity between objects. In such a situation, the proposed model is effective because it estimates not only the strength of a relationship but also its existence. For example, in Subsection 4.3, browsing activity with zero similarity between browsed websites can be interpreted as browsing with variable interest whereas browsing activity with high similarity between browsed websites represents browsing with a strong interest. Moreover, the network structure of websites can be simplified by decreasing the links between them through an introducing sparsifying method and it is expected that the network analysis of the representation learned by the proposed model is effective.

6 **Conclusions and future work**

In this paper, we introduced FTRL-proximal, a tool for analysing web browsing history data. FTRL-proximal considers both the estimation accuracy and the sparsity of the parameters in the word2vec model, which shows excellent performance in word semantic text analysis, to extract stronger relationships between websites. Additionally, we proposed a representation learning algorithm to analyse the relationships between websites while ignoring the weak relationships present in the web browsing history data. We also applied the proposed model to real-time web browsing history data and compared the performance of the proposed model with that of existing ones from the perspective of semantic validity and sparsity of representation. Finally, we analysed the web browsing behaviour of a certain consumer using the proposed model. As a result, we verified that the representation learned by the proposed model combines with semantic expressiveness to not only extract the strongly related websites, similar to that achieved by previous models, but also the sparsity to mark the similarity between weakly related websites as zero. Furthermore, we demonstrated the capability of the proposed model to extract the consumers' interests from their web browsing behaviour.

As a future work, it is necessary to consider a method to analyse the complete relationship in a group of plural websites. Based on the distributed representation model,

few studies have focused on the overall relationship of each site, each user, and each site and user, whereas numerous studies have been performed to analyse one-to-one relationships between two objects. Therefore, it is possible that the representation learned by the proposed model is effective in these analyses as well. In addition, we need to verify whether the proposed model is effective on the text data analysed in previous studies. Interpretability is an important research subject in the field of distributed representation of language and the proposed model is expected to learn the representation with good interpretability, as achieved in previous studies.

Moreover, we proposed a new model to estimate the sparse website vector with high accuracy and applied the model to actual data. The knowledge extracted from the proposed method is dependent on the actual case and the purpose of this study is not to find new common facts. Therefore, additional experiments are required for examining the rightfulness of the extracted hypotheses as future work.

Acknowledgements

We are very grateful to all the staff of VALUES, Inc. for providing data to us and for conducting beneficial discussions with us.

References

- Adamic, L. and Huberman, B. (2002) ‘Zipf’s law and the internet’, *Glottometrics*, Vol. 3, pp.143–150.
- Duchi, J. and Singer, Y. (2009) ‘Efficient online and batch learning using forward backward splitting’, *Journal of Machine Learning Research*, Vol. 10, pp.2899–2934.
- Duchi, J., Hazan, E. and Singer, Y. (2011) ‘Adaptive subgradient methods for online learning and stochastic optimization’, *Journal of Machine Learning Research*, Vol. 12, pp.2121–2159.
- Faruqui, M., Tsvetkov, Y., Yogatama, D., Dyer, C. and Smith, N.A. (2015) ‘Sparse overcomplete word vector representations’, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pp.1491–1500.
- Hosaka, T., Kawabe, R., Yamashita, H. and Goto, M. (2019) ‘A model for integrated analysis of website and user by Gaussian embedding’, *Information Processing Society of Japan*, Vol. 60, No. 8, pp.1390–1402.
- Liu, Q., Yu, F., Wu, S. and Wang, L. (2015) ‘A convolutional click prediction model’, *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp.1743–1746.
- Makhzani, A. and Frey, B. (2014) ‘k-sparse autoencoders’, *Proceedings of the 2nd International Conference on Learning Representations*.
- McMahan, B., Holt, G., Sculley, G., Young, M., Ebner, D., Grady, J., Nie, L., Phillips, T., Davydov, E., Golovin, D., Chikkerur, S., Liu, D., Wattenberg, M., Hrafnkelsson, A.M., Boulos, T. and Kubica, J. (2013) ‘Ad click prediction: a view from the trenches’, *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.1222–1230.

- Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) 'Efficient estimation of word representation in vector space', *Proceedings of the 1st International Conference on Learning Representations Workshop*.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013) 'Distributed representations of words and phrases and their compositionality', *Advances in the 26th Conference on Neural Information Processing Systems*, pp.3111–3119.
- Park, S., Bak, J.Y. and Oh, A. (2017) 'Rotated word vector representations and their interpretability', *Proceedings of the 14th Conference on Empirical Methods in Natural Language Processing*, pp.401–411.
- Piantadosi, S.T. (2014) 'Zipf's word frequency law in natural language: a critical review and future directions', *Psychonomic Bulletin & Review*, Vol. 21, No. 5, pp.1112–1130.
- Schroder, N., Falke, A., Hruschka, H. and Reutterer, T. (2019) 'Analyzing the browsing basket: a latent interests-based segmentation tool', *Journal of Interactive Marketing*, Vol. 47, pp.181–197.
- Siencnik, S.K. (2015) 'Adapting word2vec to named entity recognition', *Proceedings of the 20th Nordic Conference of Computational Linguistics*, pp.239–243.
- Subramanian, A., Pruthi, D., Jhamtani, H., Kirkpatrick, T.B. and Hovy, E. (2018) 'SPINE: sparse interpretable neural embeddings', *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pp.4921–4928.
- Sun, F., Guo, J., Lan, Y., Xu, J. and Cheng, X. (2016) 'Sparse word embeddings using ℓ_1 regularized online learning', *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pp.2915–2921.
- Tagami, Y., Kobayashi, H., Ono, S. and Tajima, A. (2013) 'Representation learning for users' web browsing sequences', *IEICE Transactions on Information and Systems*, Vol. E101.D, No. 7, pp.1870–1879.
- Tsuruhara, T., Takasuga, K., Maruyama, K. and Terada, M. (2008) 'Analysis of web browsing history using independent component analysis and its application to web page recommendation', *Proceedings of the 19th Data Engineering Workshop*.
- Xiao, L. (2009) 'Dual averaging methods for regularized stochastic learning and online optimization', *Journal of Machine Learning Research*, Vol. 11, pp.2543–2596.
- Xue, B., Fu, C. and Shaobin, Z.A. (2014) 'A study on sentiment computing and classification of Sina Weibo with word2vec', *Proceedings of the 2nd IEEE International Congress on Big Data*, pp.358–363.

Notes

- 1 We set $D = 50$ for the model of Mikolov et al. (2013a), which is assumed to construct a dense vector and $D = 200$ for the model of Faruqui et al. (2015) and Sun et al. (2016), which are assumed to construct a sparse vector similar to the proposed model.
- 2 We describe a website with non-zero values on two or more axes of the subspace as a website in the subspace.
- 3 We define the sequence browsing of websites within 30 min as one session.