2016 IEEE 16th International Conference on Data Mining Workshops

Website Classification Using Latent Dirichlet Allocation and its Application for Internet Advertising

Sotaro Katsumata
Graduage School of Economics,
Osaka University
1-7 Machikaneyama, Toyonaka, Osaka 565-0043 Japan
Email: katsumata@econ.osaka-u.ac.jp
Telephone: (+81) 6-6850-5246

Eiji Motohashi
Graduate School of International Social Sciences,
Yokohama National University
79-4 Tokiwadai, Hodogaya-ku, Yokohama 240-8501 Japan
Email: motohashi@ynu.ac.jp

Akihiro Nishimoto
School of Business Administration,
Kwansei Gakuin University
1-155 Uegahara Ichiban-Cho, Nishinomiya, Hyogo 662-8501 Japan
Email: anishimoto@kwansei.ac.jp

Eiji Toyosawa
F@N Communications, Inc.
Aoyama Diamond Building (Reception: 2nd Floor)
1-1-8, Shibuya, Shibuya-ku, Tokyo, 150-0002 Japan
Email: e_toyosawa@fancs.com

***Abstract*—This study proposes a model for website classification using website content, and discusses applications for internet advertising (ad) strategies. Internet ad agencies have many ad-spaces embedded in many websites and can choose where to place advertisements. Therefore, ad agencies have to know the properties and topics of each website in order to optimize advertising submission strategy. However, since website content is in natural languages, they have to convert these qualitative sentences into quantitative data if they want to classify websites using statistical models. To address this issue, this study applies statistical analysis to website information written in natural languages. We apply a dictionary of neologisms in order to decompose website sentences into words and create a dataset of $\{0, 1\}$ indicator matrices to classify the websites. From the dataset, we estimate the topics of each website using latent Dirichlet allocation. Finally, we discuss how to apply the results obtained to optimize ad strategies.**

I. Introduction

The internet has dramatically changed our daily lives and business models. As a new medium, consumers can search information whenever they want, while firms can provide product information to consumers who need it. Moreover, as a new channel, consumers can find their favorite products even if the market shares of these products are so small that these rarely are sold in a *brick-and-motar* store. Customization and selection are the key properties of the internet channel and medium [1,2]. These properties are known as *long-tail* [3]. and many emerging business models are closely related to this concept. The internet advertising business also tries to utilize the long-tail property. Consumer preferences are not homogeneous. Traditional advertising (ad) media have to conduct segmentation to try to target consumers as accurately as possible. However, these mass-media marketing efforts have limitations, because mass-media ads are expensive and hence not economic for the long-tail business whose consumers are few and geographically dispersed. On the other hand, internet ads can address this issue. There are many ad-spaces in websites, and it is possible to customize the placement of ads to reach the long-tail customers. These advantages of internet ads are now recognized by many firms, and hence, the market size is increasing substantially. According to Dentsu Inc., the gross expenditures on internet advertising by Japanese firms is 1159.4 billion JPY (roughly 10 billion USD) and the annual growth rate has exceeded 10% from 2014 to 2015 [4].

As shown in Figure 1, internet ad agencies have many advertising spaces embedded in many websites These ad-spaces are known as *media*, the same as for legacy media such as television, magazines, newspapers, and radio. This means that internet ad agencies have vast numbers of media compared with the traditional ad agencies. Since these websites that embed media are managed by many firms and individuals, ad agencies cannot control the content of each webpage. Obviously, ad agencies can browse these webpages because they know the URL of each site; however, there are too many webpages to check manually. Automatic website classification methods are needed in order to optimize ad placement.

Therefore, this study proposes a website classification method using a statistical model developed for natural language analysis. In this paper, we define the model in Section II, explain the dataset and data cleansing in Section III, show the empirical results in Section IV, and discuss applications and limitations in the last section.

2375-9259/16 $31.00 © 2016 IEEE
DOI 10.1109/ICDMW.2016.141

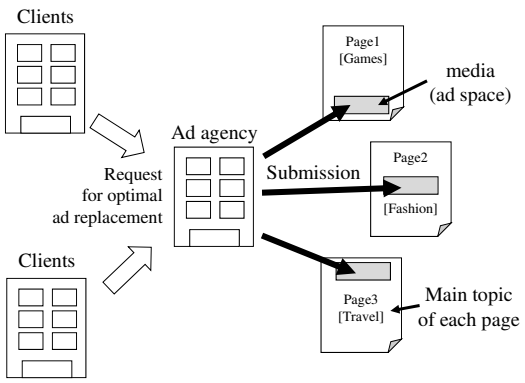

Fig. 1. Business model of the internet ad agencies

II. MODEL

A. Basic Assumptions of Latent Dirichlet Allocation

First, we define the model used in this study. We apply latent Dirichlet allocation (LDA) to classify websites. LDA was first proposed by [5], and an efficient sampling method, called *collapsed Gibbs sampling*, was proposed by [6]. LDA infers the latent topics in each word, and the assigned topics tend to be common among documents and vocabularies.

To define the model, we let the number of *documents* be D, and the number of *vocabularies* be V. w_{di} is the i-th *word* observed in document d, and the number of words in document d is N_d. Note that there are V vocabularies and any one vocabulary is observed in w_{di}; therefore, we describe that w_{di} is a V dimensional vector of $\{0,1\}$ indicator and only one element is 1 and other elements are 0, in other words $\sum_v w_{div} = 1$. For example, if vocabulary $v \in V$ is observed at (d,i), $w_{div} = 1$ and 0 otherwise. In addition, we assume that the words and documents are classified as a certain topic k. Let the number of topics be K and the topic assignment variable of w_{di} be z_{di}, which is a latent variable parameter to be estimated. The topic assignment z_{di} is a K-dimensional vector of $\{0,1\}$ values where $z_{dik} = 1$ if w_{di} is assigned as k and 0 otherwise. In LDA, the observed variable w_{di} and topic assignment z_{di} are assumed to follow the categorical distribution (denoted as *Cat*), which is a specific type of multinomial distribution of a single trial. For w_{di}, we assume that the prior parameter is different depending on the assigned topic z_{di}, and therefore, we let the parameter of w_{di} be $\phi_{z_{di}}$, which is a V-dimensional vector of variables whose elements take the range $(0,1)$ and $\sum_v^V \phi_{z_{di},v} = 1$. For z_{di}, we assume that the assignment latent variable follows the K-dimensional categorical distribution with parameter θ_d, which is different between documents. This means that z_{di} has the same common parameter within the same document. Therefore, these distributions are described as follows:

$$w_{di}|z_{di}, \phi_{z_{di}} \sim Cat_V(\phi_{z_{di}}) \tag{1}$$

$$z_{di}|\theta_d \sim Cat_K(\theta_s) \tag{2}$$

Note that the number of $\phi_{z_{di}}$s incorporated in the model is K and the number of θ_ds is K. In addition, as the prior distribution, LDA assumes the Dirichlet prior, which is denoted as *Dir* for ϕ_k and θ_s, as follows:

$$\phi_k \sim Dir_V(\alpha) \tag{3}$$

$$\theta_d \sim Dir_K(\beta) \tag{4}$$

This study assumes that the Dirichlet distribution is symmetric, and therefore, the parameters α and β are scalar variables in the range $(0,1)$. The parameters of the model are $\{\mathbf{z}, \phi, \theta\}$ where $\mathbf{z}$ is a set of $z_{di}, i \in \{1,\cdots,N_d\}, d \in \{1,\cdots,D\}$, ϕ is a set of $\phi_k, k \in \{1\cdots,K\}$, and θ is a set of $\theta_d, d \in \{1,\cdots,D\}$.

To obtain the parameters, we apply Bayesian inference. Following Bayes' theorem, we obtain the posterior distribution as follows where $\mathbf{w}$ is a set of $z_{di}, i \in \{1,\cdots,N_d\}, d \in \{1,\cdots,D\}$ and we call $\mathbf{w}$ a *corpus*.

$$\pi(\mathbf{z},\phi,\theta|\mathbf{w}) = \frac{\pi(\mathbf{z},\phi,\theta,\mathbf{w})}{\pi(\mathbf{w})} \tag{5}$$

$$= \frac{\pi(\mathbf{w}|\mathbf{z},\phi)\pi(\phi)\times\pi(\mathbf{z}|\theta)\pi(\theta)}{\pi(\mathbf{w})} \tag{6}$$

where, $\pi(\mathbf{w}|\mathbf{z},\phi) = \prod_d^D \prod_i^{N_d} \pi_{Cat}(w_{di}|z_{di},\phi_{z_{di}})$, $\pi(\phi) = \prod_k^K \pi_{Dir}(\phi_k|\alpha)$, $\pi(\mathbf{z}|\theta) = \prod_d^D \prod_i^{N_d} \pi_{Cat}(z_{di}|\theta_d)$, and $\pi(\theta) = \prod_d^D \pi_{Dir}(\theta_d)$.

The proportion of the posterior distribution is easily described as follows:

$$\pi(\mathbf{z},\phi,\theta|\mathbf{w}) \propto \pi(\mathbf{w}|\mathbf{z},\phi)\pi(\phi)\times\pi(\mathbf{z}|\theta)\pi(\theta) \tag{7}$$

$$\propto \pi(\mathbf{w},\phi|\mathbf{z})\pi(\mathbf{z},\theta) \tag{8}$$

As shown above, the posterior distribution is the multiplication of two conditional distributions $\pi(\mathbf{w},\phi|\mathbf{z})$ and $\pi(\mathbf{z},\theta)$ that are obtained respectively as follows:

$$\pi(\mathbf{w},\phi|\mathbf{z}) = \pi(\mathbf{w}|\mathbf{z},\phi)\pi(\phi) \tag{9}$$

$$= \prod_d^D \prod_i^{N_d} \prod_v^V (\phi_{z_{di},v})^{w_{div}} \prod_k^K \frac{\Gamma(V\beta)}{\Gamma(\beta)^V} \prod_v^V (\phi_{kv})^{\beta-1} \tag{10}$$

$$= \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^K \prod_k^K \prod_v^V (\phi_{k,v})^{n_{kv}} (\phi_{kv})^{\beta-1} \tag{11}$$

$$= \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^K \prod_k^K \prod_v^V (\phi_{k,v})^{n_{kv}+\beta-1} \tag{12}$$

where n_{kv} is the number of words assigned as k within the corpus $\mathbf{w}$.

$$\pi(\mathbf{z},\theta) = \pi(\mathbf{z}|\theta)\pi(\theta) \tag{13}$$
$$= \prod_d^D \prod_i^{N_d} \prod_k^K (\phi_{dk})^{z_{dik}} \prod_d^D \prod_k^K \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right)(\theta_{dk})^{\alpha-1} \tag{14}$$
$$= \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right)^D \prod_d^D \prod_k^K \left[\prod_i^{N_d} (\phi_{dk})^{z_{dik}}\right](\theta_{dk})^{\alpha-1} \tag{15}$$
$$= \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right)^D \prod_d^D \prod_k^K (\phi_{dk})^{n_{dk}+\alpha-1} \tag{16}$$

where n_{dk} is the number of words assigned as k within document d.

B. Collapsed Gibbs Sampling

In the model, we usually have to estimate three parameters, $\{\mathbf{z}, \phi, \theta\}$. Using the Markov chain Monte Carlo (MCMC) method , we sequentially generate random variables for each parameter and collect them. However, [6] proposed a new method for parameter estimation called *collapsed Gibbs sampling*. This section provides a detailed explanation of this method, which we use in our analysis.

The basic concept of collapsed Gibbs sampling is to simplify the model through the integration of ϕ and θ; therefore, using this method, we only need to estimate topic assignment parameter $\mathbf{z}$. The mathematical expression is as follows:

$$\pi(\mathbf{z}|\mathbf{w}) \propto \pi(\mathbf{w}|\mathbf{z})\pi(\mathbf{z}) \tag{17}$$
$$\propto \int \pi(\mathbf{w},\phi|\mathbf{z})d\phi \int \pi(\mathbf{z},\theta)d\theta \tag{18}$$

As in the usual definition of the model, collapsed Gibbs sampling can also be divided into two components, obtained from $\pi(\mathbf{z},\phi|\mathbf{w})$ and $\pi(\mathbf{z},\theta)$ defined in previous section.

$$\pi(\mathbf{w}|\mathbf{z}) = \int \pi(\mathbf{w},\phi|\mathbf{z})d\phi \tag{19}$$
$$= \int \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^K \prod_k^K \prod_v^V (\phi_{k,v})^{n_{kv}+\beta-1} d\phi \tag{20}$$
$$= \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^K \prod_k^K \left(\int \prod_v^V (\phi_{kv})^{n_{kv}+\beta-1} d\phi_k\right) \tag{21}$$
$$= \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^K \prod_k^K \frac{\prod_v^V \Gamma(n_{kv}+\beta)}{\Gamma\left(\sum_v^V n_{kv} + V\beta\right)} \tag{22}$$

where the last equation is obtained using the property of the Beta function such that $B(a_1,\cdots,a_L) = \int \prod_l^L q_l^{a_l-1} dq = \prod_l^L \Gamma(a_l)/\Gamma\left(\sum_l^L a_l\right)$. In a similar way to $\pi(\mathbf{w}|\mathbf{z})$, we can also integrate out θ.

$$\pi(\mathbf{z}) = \int \pi(\mathbf{z},\theta)d\theta \tag{23}$$
$$= \int \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right) \prod_d^D \prod_k^K (\phi_{dk})^{n_{dk}+\alpha-1} d\theta \tag{24}$$
$$= \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right)^D \prod_d^D \int \prod_k^K (\phi_{dk})^{n_{dk}+\alpha-1} d\theta \tag{25}$$
$$= \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right)^D \prod_d^D \frac{\prod_k^K \Gamma(n_{dk}+\alpha)}{\Gamma\left(\sum_k^K n_{dk} + K\alpha\right)} \tag{26}$$

The posterior distribution of $\mathbf{z}$ is proportional to multiplications of the above equations.

$$\pi(\mathbf{z}|\mathbf{w}) \propto \pi(\mathbf{w}|\mathbf{z})\pi(\mathbf{z}) \tag{27}$$
$$\propto \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^K \prod_k^K \frac{\prod_v^V \Gamma(n_{kv}+\beta)}{\Gamma\left(\sum_v^V n_{kv} + V\beta\right)} \tag{28}$$
$$\times \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right)^D \prod_d^D \frac{\prod_k^K \Gamma(n_{dk}+\alpha)}{\Gamma\left(\sum_k^K n_{dk} + K\alpha\right)} \tag{29}$$
$$\propto \prod_k^K \frac{\prod_v^V \Gamma(n_{kv}+\beta)}{\Gamma(n_k + V\beta)} \prod_d^D \frac{\prod_k^K \Gamma(n_{dk}+\alpha)}{\Gamma(N_d + K\alpha)} \tag{30}$$

where n_k is the number of words assigned as k within the whole corpus and N_d is the length of document d, as defined in the previous section.

The posterior distribution of $\mathbf{z}$ becomes simpler to decompose. The posterior probability that $z_{di} = j$ is

$$\pi(z_{di} = j|\mathbf{z}_{-di},\mathbf{w}) \propto \pi(\mathbf{w}|\mathbf{z})\pi(\mathbf{z}) \tag{31}$$
$$\propto \prod_k^K \frac{\prod_v^V \Gamma(n_{kv}+\beta)}{\Gamma(n_k + V\beta)} \prod_d^D \frac{\prod_k^K \Gamma(n_{dk}+\alpha)}{\Gamma(N_d + K\alpha)} \tag{32}$$
$$\propto \prod_k^K \frac{\prod_v^V \Gamma(n_{kv}+\beta)}{\Gamma(n_k + V\beta)} \Gamma(n_{dk}+\alpha) \tag{33}$$
$$\propto \frac{\prod_v^V \Gamma(n_{jv}+\beta)}{\Gamma(n_j + V\beta)} \Gamma(n_{dj}+\alpha) \tag{34}$$
$$\times \prod_{k\neq j} \frac{\prod_v^V \Gamma(n_{kv}+\beta)}{\Gamma(n_k + V\beta)} \Gamma(n_{dk}+\alpha) \tag{35}$$
$$\propto \Gamma(n_{dj}+\alpha) \frac{\Gamma(n_{w_{di},j}+\beta)\prod_{v\neq w_{di}} \Gamma(n_{jv}+\beta)}{\Gamma(n_j + V\beta)} \tag{36}$$
$$\propto \Gamma(n_{dj}+\alpha) \frac{\Gamma(n_{w_{di},j}+\beta)}{\Gamma(n_j + V\beta)} \tag{37}$$

where, $n_{w_{di},j}$ is the number of word $v = w_{di}$ which assigned as topic j.

We can further simplify the above equation applying the relationship that $n_{dj} = n_{dj}^{-(di)} + 1$, $n_j = n_j^{-(di)} + 1$, $n_{w_{di},j} =$

$n_{w_{di,j}}^{-(di)}+1$, and a property of gamma function $\Gamma(a+1) = a\Gamma(a)$ [1].

$$\pi(z_{di} = j|\mathbf{z}_{-di}, \mathbf{w}) \propto \pi(\mathbf{w}|\mathbf{z})\pi(\mathbf{z}) \tag{38}$$

$$\propto \Gamma(n_{dj}^{-(di)} + 1 + \alpha)\frac{\Gamma(n_{w_{di},j}^{-(di)} + 1 + \beta)}{\Gamma(n_j^{-(di)} + 1 + V\beta)} \tag{39}$$

$$\propto (n_{dj}^{-(di)} + \alpha)\Gamma(n_{dj}^{-(di)} + \alpha) \tag{40}$$

$$\times \frac{(n_{w_{di},j}^{-(di)} + \beta)\Gamma(n_{w_{di},j}^{-(di)} + \beta)}{(n_j^{-(di)} + V\beta)\Gamma(n_j^{-(di)} + V\beta)} \tag{41}$$

$$\propto (n_{dj}^{-(di)} + \alpha)\prod_{k=1}^{K}\Gamma(n_{dk}^{-(di)} + \alpha) \tag{42}$$

$$\times \frac{(n_{w_{di},j}^{-(di)} + \beta)}{(n_j^{-(di)} + V\beta)}\prod_{k=1}^{K}\frac{\Gamma(n_{w_{di},k}^{-(di)} + \beta)}{\Gamma(n_k^{-(di)} + V\beta)} \tag{43}$$

$$\propto (n_{dj}^{-(di)} + \alpha)\frac{(n_{w_{di},j}^{-(di)} + \beta)}{(n_j^{-(di)} + V\beta)} \tag{44}$$

A posterior random sample of z_{di} is generated from following categorical distribtuion for $d = 1, \cdots, D$, $i = 1, \cdots, N_d$ sequentially, we obtain the posterior samples of $\mathbf{z}$ after enough iterations.

$$z_{di}|\mathbf{z}_{-(di)}, \mathbf{w} \sim Cat_K(\psi^*) \tag{45}$$

where ψ^* is a K-dimensiontal vector parameters. Its element $\psi_j^* = \psi_j / \sum_{k=1}^{K}\psi_k$, where $\psi_j \propto \pi(z_{di} = j|\mathbf{z}_{-di}, \mathbf{w})$.

In addition, the parameter ϕ and θ is obtained from the following equation. These equations are used to obtain predictive distributions of new vocabularies or new documents.

$$\phi_{kv} = \frac{n_{kv} + \beta}{n_k + V\beta} \tag{46}$$

$$\theta_{dk} = \frac{n_{dk} + \alpha}{n_d + K\alpha} \tag{47}$$

III. Data

As mentioned, our research purpose is to classify websites based on the text content of each page. Therefore, we first have to decompose the text content into words to incorporate into quantitative models.

The website dataset is provided by an internet ad-agency located in Japan. We select 1000 websites for analysis. To obtain a quantitative dataset, we apply the Japanese word decomposition engine. In this study, MeCab which is the Japanese morphological analyzer developed by [7] and its dictionary ipadic-neologed is used for analysis [8]. This engine refers to a language dictionary to decompose words; the accuracy of the decomposition is highly dependent on the dictionary. The standard dictionary provided by MeCab developers has enough performance to decompose regular sentences such as in newspaper articles or academic papers. However, it is difficult to apply the built-in dictionary to decompose website test sentences, because there are many neologisms such as abbreviated words and proper nouns in these sentences. In contrast to English sentences, Japanese sentences do not use space between independent words; therefore, without a neologism dictionary it is impossible to decompose website texts appropriately.

Using MeCab with ipd-neologed, we obtain a decomposed dataset of the website content, and we extract nouns, verbs, and adjectives in analysis. In our case, the number of documents $D = 1000$, and the number of vocabularies $V = 31448$. The sum of the length of documents, in other words, the corpus length $N = \sum_d N_d = 1228630$. Among the 1000 websites, the average length of the website is 767.7 words and the longest website has 36,030 words on a page.

[1] In addition, note that $n_{dk} = n_{dk}^{-(di)}$, $n_k = n_k^{-(di)}$, $n_{w_{di},k} = n_{w_{di,j}}^{-(di)}$ for $k \neq j$.

IV. Results

A. Number of Topics

To establish the number of topics, we use the marginal likelihood and compare the model fitness. The marginal likelihood is easily obtained from the posterior sample likelihood. The simplest way is to calculate harmonic mean of the posterior samples proposed by [9]. As comparison models, we compare 12 models whose topics are $K = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200\}$. The number of iterations is 2000.

Figure 2 shows the log-marginal likelihood of models. We find that the model fitness increases rapidly from the 10-topic model to the 40-topic model. However, after that the marginal improvement of model fitness is small. Since our research purpose is to classify websites efficiently, the models with fewer topics are preferable, and so, we choose the 40-topic model.

B. Topic Keywords

In this section, we further discuss the results of the 40-topic model. The number of iterations are 2000, and we collect the last sample of $\mathbf{z}$ for classification [6]. At first, we examine the vocabularies to interpret each topic. The probability of vocabularies belonging each topic is easily obtained from the posterior samples. We calculate the probability for all vocabularies and sort them for each topic in decreasing order of ability to interpret each topic characteristic. The top 100 words are examined and we label each topic manually. Since there are 40 topics, we have to see words for each topic. Similar to factor analysis, analysts have to label each topic for easier interpretation. In addition, we further classify the topics into hypertopics from words and labels.

Next, the websites are classified for each topic. We simply classify websites based on the topic assignment probability parameter θ. Each document has a K-dimensional parameter θ that implies the topic belongs to the document. For document d, if $\theta_{dk} > \max(\theta_{d,-k})$, we classify the document as belonging to topic k.

Table I shows the labels, hypertopics, and number of sites for each topic. Since some websites cannot be classified into

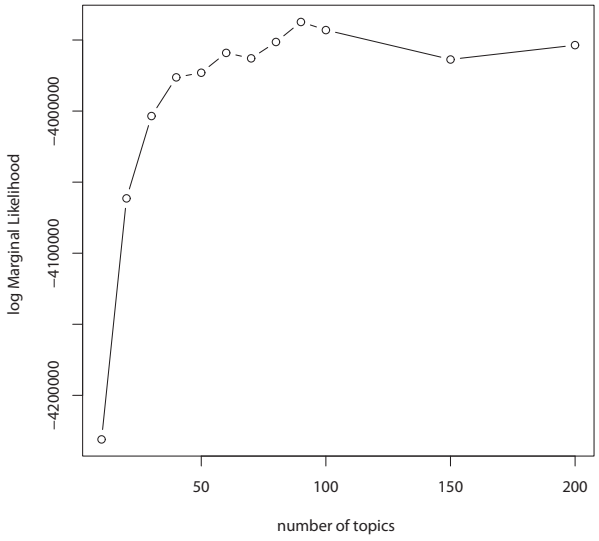

Fig. 2. Fitness of models

one topic, the sum of sites for all 40 topics is less than 1000. From these 40 topics, we assume eight hypertopics named Adults, BBS (Bulletin Board Systems), Contents, Diary, Entertainments, Gamble, Games, and News. We find some similar topics and we manually classify these into the eight hypertopics as shown in Table I. In the hypertopic "Adults," there are particular words used in the website; therefore, it is easy to identify the characteristic of the topics. Hypertopic "BBS" indicates websites summarizing the major BBS websites in Japan. The raw BBS content has a lot of noise and fragmented words; therefore, it is not easy to quickly read the raw BBS directory. Moreover, some large BBS websites have too many threads and it is hard to find the information. Reflecting the need to find and read valuable information rapidly, many summary websites act as curators in Japan. We find that the main subject of the BBS topic websites varies: the websites are on subjects such as Baseball, Football, and Home and Family. Therefore, in the hypertopic BBS, the website visitors for each topic may differ. The third hypertopic, "Contents" contains websites that provide information on content such as games, comics, and novels. In the Japanese content market, many consumers exchange information related to TV on-airs, DVD releases , and other related subjects through the internet. Therefore, there are many content websites and their visitors overlap to some extent. However, some content sites are highly segmented and focus on niche target markets; their visitors may unique compared with other websites. The fourth hypertopic "Diary" contains individual diaries and blogs, for which it is hard to find the main subjects. The high-ranking vocabularies are mainly general nouns such as dates and other general vocabularies used in blogs. Since we rarely find the subjects of these websites, it is not easy to find the points in common among them. The fifth hypertopic "Entertainments" only contains a single topic, number 23 in the Table. In the topic analysis results, we find very few mass-media entertainment subjects such as TV programs and movies. We find that these mass-media topics are not very significant in internet media. The sixth hypertopic is "Gamble" and it contains two topics, horse races and Pachinko. These two activities are well-known legal gambling topics, and the websites provide winning tips and techniques. The seventh hypertopic is "Games." Although similar to "Contents," there are

TABLE I
TOPICS AND HYPERTOPICS

Topic	Topic Label	Hypertopic	# of websites
1	News	News	8
2	Hatena	Diary	60
3	Adults 1	Adults	16
4	Games (*Title A*)	Games	14
5	Pachinko	Gamble	11
6	BBS:Baseball	BBS	19
7	Games (*Title B*)	Games	36
8	Adult Comics 1	Adults	22
9	Adults 2	Adults	8
10	Horse Races	Gamble	16
11	Games (*Title C*)	Games	9
12	Novels 1	Content	5
13	Diary 1	Diary	14
14	BBS:2ch	BBS	25
15	Diary 2	Diary	88
16	Novels 2	Content	5
17	Games (Social)	Games	17
18	Free Topic Talk	BBS	8
19	Adults 3	Adults	43
20	News	News	16
21	Animations	Content	26
22	Adults 4	Adults	1
23	Entertainment (Girl Groups)	Entertainment	19
24	Diary 3	Diary	22
25	Adult Comics 2	Adults	1
26	Comics	Content	14
27	News (International)	News	14
28	Comics (*Magazine A*)	Content	120
29	Games	Games	12
30	Games (*Title D*)	Games	18
31	BBS:Football	BBS	11
32	Adult Comics 3	Adults	24
33	News	News	17
34	BBS:Local	BBS	27
35	News BBS	BBS	48
36	Adult Novels	Adults	10
37	Games (*Title D*)	Games	36
38	BBS:Home and Family	BBS	38
39	Comics	Content	18
40	News	News	68

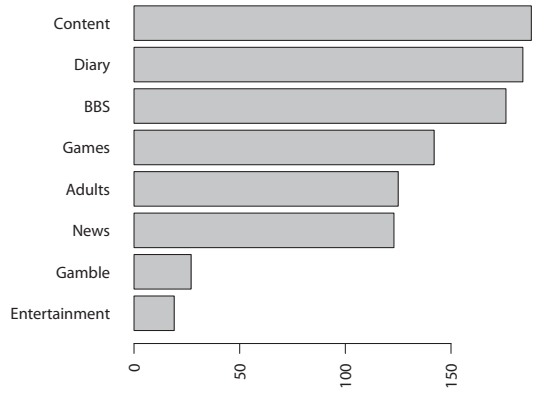

Fig. 3. Hypertopics and number of websites

many games websites compared with other topics. Therefore, we separate games into an independent hypertopic. The main subjects of games are different for each topic. For example, websites classified as topic 4 provide information on *Title A*, and topic 7, on *Title B*. Since the consumers of these titles overlap, we expect that these websites have common visitors. The last hypertopic "News" contains websites of the news and related discussions. The characteristics of these news websites are similar to those of the hypertopic BBS that curates information on certain subjects. In websites classified into this hypertopic, the content is mainly related to recent news articles such as domestic news, accidents, international news, and politics. The subjects differ depending on the current news.

Figure 3 shows the number of websites classified into each hypertopic. From the figure, we find that 19.1 % of websites are classified as "Contents," 18.7% are classified as "Diary," 17.8% are classified as "BBS," 12.8% are classified as "Adults," and 12.5% are classified as "News." The websites classified as "Gamble" and "Entertainments" are few, at 2.7% and 1.9% respectively. This implies that there are many content websites on the internet. Note that these content subjects are closely related to comics (Manga) and Animation. Although the market size and number of consumers in Japan are large compared with other countries, the share of these websites is still remarkably high. This implies that the content market is strongly supported by internet media.

C. Advertising Submission Strategies

Based on these discussions, we discuss the ad-submission strategies of the internet ad agencies and client firms. First, from the topic labels and website classifications, ad agencies can avoid websites that may damage clients' reputation. For example, some clients do not want to place their ads on websites for adults or gambling. Without website classification, there is a risk of placing ads on these websites. If the ad agencies know the website topics, they can customize ad placement. Since, according to the previous subsection, the share of adult websites is roughly 12% of internet content, using random ad-placement, more than 1 in 10 ads will be displayed on adult websites. In addition, the ad agencies can conduct more detailed ad targeting through the classification information. For example, there are few entertainment websites on the internet; however, if ad agencies are able to classify these minor topic websites, their clients could send efficient ads to minor but enthusiastic consumers. The share of websites is not a major concern, because there are many websites on the internet.

V. Discussion and Conclusion

In this study, we propose a method for classification of websites using text data. We focus on the text content of websites and decompose them into individual words using a neologism dictionary developed for internet content. After decomposition, we analyze the word dataset using LDA, which was developed as a statistical model for natural languages. We classify each website based on the results of LDA and discuss the consequent impact on ad-submission strategies for ad agencies and client firms. Our analysis framework enables ad agencies to classify and detect media characteristics automatically.

In this study, we extract data on 1000 websites for analysis. The size of the sample is rather small. However, we can easily expand the number of websites in the analysis because LDA has some useful advantages for big-data analysis. As shown in the previous section, we only need to obtain the latent topic assignment $\mathbf{z}$ using collapsed Gibbs sampling. Moreover, the posterior parameters become simple, compared to other statistical models such as linear regression models that need to obtain inverse matrices or other linear algebraic calculations. The posterior distribution of $\mathbf{z}$ is only the sum of variables and simple ratios. In addition, we can easily perform additional analyses using predictive distributions of ϕ and ψ. We can supply information through these parameters to the new documents.

There are some issues to be addressed, however. First, in order to examine the impact of our analysis on firm performance, we need to conduct a field test. For example, if we try to examine the performance of target ad display, we should place ads based on our model in parallel with a random control, and measure and compare performance indicators such as click-through and conversion rates. If these performance indicators are significantly higher for our model than random placement, we can conclude that our model has enough capability to contribute to firm performance. Second, we need to improve our model. For example, in this study, we classify topics into hypertopics manually. However, we have to improve LDA and classify topics into hypertopics simultaneously in a statistical way. Some studies propose a multi-level model that assumes hypertopics [10]. It is easy

to expand the model because LDA can be estimated using MCMC; however, a more complex model will need more time to estimate parameters. Complex models may not be appropriate for practical use; we have to examine the trade-off between theoretical and practical improvement.

ACKNOWLEDGMENT

The authors would like to thank Kazuki Oomori and members of *F@N Communications Data-Mining team*, and anonymous reviewers for helpful comments and suggestions.

REFERENCES

[1] A. Ansari, S. Essegaier, and R. Kohli, "Internet Recommendation Systems." *Journal of Marketing Research*,vol. 37, no. 3, pp. 363-75, 2000.

[2] A. Ansari, and C. Mela. "E-Customization." *Journal of Marketing Research*, vol. 40, no. 2, pp. 131-46, 2003.

[3] C. Anderson. *The Long Tail: Why the Future of Business is Selling Less of More*, revised ed. Harlow, Hachette Books, 2008.

[4] Dentsu, Inc. *2015 Advertising Expenditures in Japan*, [Online] Available: http://www.dentsu.com/knowledgeanddata/ad_expenditures/pdf/expenditures_2015.pdf [Accessed 12- Aug.- 2016].

[5] D. M. Blei,, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[6] T. L. Griffiths, and M. Steyvers "Finding Scientific Topics," *PNAS*, vol. 101, no. 1, pp. 5228-5235, 2014.

[7] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying Conditional Random Fields to Japanese Morphological Analysis," *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pp. 230-237, 2004.

[8] T. Sato, *mecab-ipadic-NEologd : Neologism dictionary for MeCab*, [Online] Available: https://github.com/neologd/mecab-ipadic-neologd [Accessed 12- Aug.- 2016].

[9] M. A. Newton, and A. E. Raftery "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 56, no. 1, pp. 3-48, 1994.

[10] W. Li, and A. McCallum, "Pachinko Allocation: DAG-structured Mixture Models of Topic Correlations," *ICML '06 Proceedings of the 23rd international conference on Machine learning*, pp. 577-584, 2006.