Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 35 (2014) 929 – 936
Procedia Computer Science

18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014

Classifying homographs in Japanese social media texts using a user interest model

Tomohiko Harada[a], Kazuhiko Tsuda*[b,]

[a] Graduate School of Systems and Information Engineering, University of Tsukuba, Bunkyo-ku, Tokyo 112-0012, Japan
[b] Graduate School of Business Sciences University of Tsukuba, Bunkyo-ku, Tokyo 112-0012, Japan

Abstract

The analysis of text data from social media is hampered by irrelevant noisy data, such as homographs. Noisy data is not usable and makes analysis, such as counting estimates, of the target data difficult, which adversely affects the quality of the analysis results. We focus on this issue and propose a method to classify homographs that are contained in social media texts (i.e. Twitter) using topic models. We also report the results of an evaluation experiment. In the evaluation experiment, the proposed method showed an accuracy improvement of 8.5% and a reduction of 16.5% in the misidentification rate compared with conventional methods.

Peer-review under responsibility of KES International.

Keywords: Social media, Twitter, Homographs, Semantic analysis, Topic modelling, Latent Dirichlet allocation

1. Introduction

In recent years, there has been a rapid increase in efforts to generate business strategies, such as marketing strategies and business improvements, by collecting and analysing Big Data. The social network service 'Twitter' has attracted attention as a source of such information. In messages called 'tweets', which can only be 140 characters or less, a user can post thoughts or day-to-day experiences. When a user posts a tweet, the tweet is transmitted from person to person; tweets can be shared among many users. Tweets often include user impressions of purchased products and services, as well as the criteria used to select those products and services. It is increasingly becoming important for businesses to collect and analyse such useful information. However, there is a common problem with noisy data that is contained in results (e.g. tweets) that are collected by keyword searches in the analysis and study of social media, such as Twitter. Such noisy data is unusable for targeted analysis and affects the accuracy of the analytical results. For example, when performing analysis of corporate reputations, if there are homographs, such as the name of another company with the same name, that are included in the results of a keyword search of a company name, this becomes a factor in the analysis, and accuracy is often reduced. In Table 1,847 tweets containing the keyword

* Corresponding author. Tel.: +81-3-3942-6869 ; fax: +81-3-3942-6829.
E-mail address: s1230165@u.tsukuba.ac.jp.

Peer-review under responsibility of KES International.
doi:10.1016/j.procs.2014.08.168

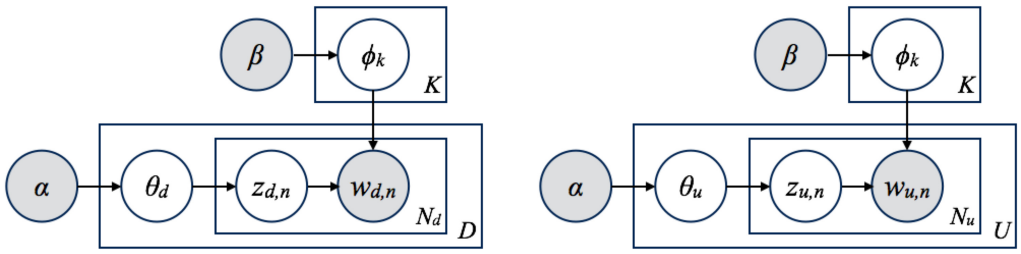

Fig. 1. (a) Simple graphical model of LDA; (b) Graphical model of our LDA.

'apple' in Japanese katakana characters are classified by subject. The Japanese tweets shown were posted from 4 to 11 January 2014. When counting the number of tweets about 'Apple Inc.', the digital consumer electronics and computer manufacturer, after searching for tweets using the keyword 'apple', it is common to encounter tweets with 'apple' used with the intended meaning of fruit, such as 'apple tea' or 'apple juice'. In addition, other companies that may include the word 'apple' will be included in such search results. In general, these unrelated tweets are noisy data. Therefore, it is necessary to classify tweets of interest from search results that include noisy data. In this paper, we focus on this issue and propose a method to classify homographs contained in the text of social media using topic models. We also report the results of evaluation experiments.

Table 1. Examples of the tweets' subjects containing the keyword 'apple'.

Subject	% tweets
Apple Inc.	70
apple tea	4
apple juice	2
other	24
Total	100

2. Topic modeling

Topic modelling [1] has attracted attention as a statistical modelling method that is used to acquire knowledge from large-scale and heterogeneous data. In topic modelling, one document is represented as a mixture of multiple topic information. It has been confirmed that topic modelling can model documents with higher accuracy than a mixture of multinomial distributions represented by one topic one document [2]. In this section, we review latent Dirichlet allocation (LDA) [2], which is a representative topic model that is known to work well. We then review representative studies of applying LDA to the Twitter data analysis.

2.1. LDA: Latent dirichlet allocation

Blei et al. [2] proposed LDA, a technique in which the Dirichlet prior distribution is taken as a prior distribution of the multinomial distribution that represents the topic of a document. The potential of topic modeling has recently attracted attention, and LDA is known to work well. Based on the idea that a document is represented as a random mixture over latent topics, where each topic is characterized by a distribution over words, LDA infers the probability distribution of the topic.

Fig.1.(a) shows the graphical model of LDA, where random variables and parameters are represented by a vertex; their dependencies are represented by a directed edge. The shaded vertex indicates observed variables; the other vertices indicate latent parameters or latent variables. The number written at a rectangle's corner indicates the iterations of the variable in the rectangle. D is the number of documents, K is the number of topics, and N_d is the word count in

document d. θ and ϕ are multinomial distribution parameters of the topic and of the word in each topic, respectively. α and β are Dirichlet hyperparameters in θ and ϕ, respectively. The document set $\boldsymbol{W}$ generative processes used in this graphical model are as follows:

1. For each of the topics $k = 1, \cdots, K$:
   (a) Choose $\boldsymbol{\phi}_k \sim Dir(\beta)$,
2. For each of document $d = 1, \cdots, D$:
   (a) Choose $\boldsymbol{\theta}_d \sim Dir(\alpha)$,
   (b) For each of the words $w_{d,n}$ where $n = 1, \cdots, N_d$,
      i. Choose a topic $z_{d,n} \sim Multi(\boldsymbol{\theta}_d)$,
      ii. Choose a word $w_{d,n} \sim Multi(\boldsymbol{\phi}_{z_{d,n}})$,

where $\boldsymbol{\phi}_k$ is the word distribution for topic k, $\boldsymbol{\theta}_d$ is the topic distribution for document d, $z_{d,n}$ is the topic for the n-th word in document d, and $w_{d,n}$ is the n-th word in document d. $Dir(\cdot)$ is the Dirichlet distribution for parameter α, and $Multi(\cdot)$ is the multinomial distribution for parameter β. According to this LDA model, the total probability of the model is given as

$$P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\phi}|\alpha, \beta) = \prod_{k=1}^{K} P(\boldsymbol{\phi}_k|\beta) \prod_{d=1}^{D} P(\boldsymbol{\theta}_d|\alpha) \prod_{n=1}^{N} P(z_{d,n}|\boldsymbol{\theta}_d) P(w_{d,n}|\boldsymbol{\phi}_{z_{d,n}}). \tag{1}$$

To infer the unknown parameters, various methods have been proposed. Collapsed Gibbs sampling proposed by Griffiths et al.[3] is known to model estimation with high accuracy if a sufficient number of iterations have been obtained. Though there are $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ in the LDA model, the Collapsed Gibbs sampler collapses (integrates out) these and derives the updating formula of a form without $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. The updating formula derived by Collapsed Gibbs sampling is given below:

$$P(z_i = k|\boldsymbol{Z}_{\setminus i}, \boldsymbol{W}) \propto \frac{N_{dk\setminus i} + \alpha}{N_{d\setminus i} + \alpha K} \cdot \frac{N_{kw_i\setminus i} + \beta}{N_{k\setminus i} + \beta V}, \tag{2}$$

where z_i is the topic for the n-th word in document d, and $\boldsymbol{Z}_{\setminus i}$ indicates that it does not include the current assignment z_i from topic set $\boldsymbol{Z}$. N_{dk} is the count of words that is assigned topic k in document d, and N_d is a summation over that dimension k. N_{kw} is the count of word w in topic k, and N_k is a summation over that dimension w. In the both cases, $\setminus i$ indicates that does not include the current assignment z_i. We can calculate a predictive distribution of a topic distribution $\boldsymbol{\theta}$ for each document and a predictive distribution of a word distribution $\boldsymbol{\phi}$ for each topic using samples obtained by Gibbs sampling. Equation (3) estimates quantity $\hat{\theta}_d^k$ of the probability that topic k is generated in document d, and equation (4) estimates quantity $\hat{\phi}_k^w$ of the probability that word w exists when topic k is chosen.

$$\hat{\theta}_d^k = \frac{N_{dk} + \alpha}{N_d + \alpha K}, \tag{3}$$

$$\hat{\phi}_k^w = \frac{N_{kw} + \beta}{N_k + \beta V}. \tag{4}$$

2.2. Applying LDA to analysis of twitter data

Many studies have applied LDA to the analysis of Twitter data[4]. Weng et al.[5] proposed a method to detect influential users of Twitter using LDA. In addition, Pennacchiotti et al.[6] proposed a user classification model based on tweet information using LDA. However, the text of a tweet is shorter than a report or letter, and it is not possible to adequately capture the meaning using common topic models, such as LDA. Therefore, many of these studies handled all the tweets that a user posted as a single document based on the Author-Topic model[7] rather than using LDA to handle one tweet as a single document. Zhao et al.[8] proposed a Twitter-LDA model based on the hypothesis that one tweet comprises one topic. They solved the problem that topic models cannot be estimated properly due to the

shortness of tweets and demonstrated that their model is superior to conventional models. Furthermore, Sasaki et al.[9] focused on the fact that it is not possible for Twitter-LDA to consider how topics change over time in the same way as conventional LDA, even though the interest of the user may have changed. They added a topic tracking model mechanism[10] to Twitter-LDA and proposed a method that can be modelled effectively using the dynamics of the topic and user interest.

We follow the idea of the Author-Topic model and handle all tweets that each user posts as a single document. This document is used to model the user's interest distribution, which we then use to estimate whether a given tweet should be collected. Furthermore, to consider the dynamics of user interest and topics, we compared several models using tweets during different periods of time and selected a model that showed good performance. To the best of our knowledge, no study has previously tackled the problem of classifying homographs contained in tweets collected, using a topic model.

3. Estimation of homographs using a user interest model

When you attempt to collect a set of tweets for a certain company by performing a keyword search, it is often the case that homographs, such as the name of another company with the same name, are included in the results. These tweets, which are unrelated to the intended purpose, are noisy data. We propose a method to classify desired tweets from search results that include noisy data. Our method uses both information contained in a tweet and the original poster's user interest information to determine whether that user frequently tends to tweet about the target topic. For example, as mentioned in Section 1, to collect 'Apple Inc.' tweets, when you search for tweets using the keyword 'apple', the search results commonly contain tweets with 'apple' used with the intended meaning of fruit, such as 'apple tea' or 'apple juice', as well as other companies that include the word 'apple'. The proposed method uses both information about the tweet and information about whether the user who posted the tweet tends to post about 'Apple Inc.' or about other 'apple' topics.

First, the proposed method estimates a 'user interest model'. This models both the probability of a user being interested in a topic and the appearance probability of words for each topic using the LDA that handles the set of tweets a user has posted in the past as a single document. Here, we assume that each user from the user set $\boldsymbol{U}$ has a specific proportion of topics $\boldsymbol{\theta}_u$ (representing the probability that user u is interested in each topic). We consider that the word w tweeted by user u was generated from the word distribution $\boldsymbol{\phi}_k$, which is specific to the given topic k, after topic k was chosen from $\boldsymbol{\theta}_u$. The graphical model of LDA is illustrated in Fig.1.(b) The tweet generative processes used in this graphical model are as follows:

1. For each of the topics $k = 1, \cdots, K$:
   (a) Choose $\boldsymbol{\phi}_k \sim Dir(\beta)$,
2. For each of user $u = 1, \cdots, U$:
   (a) Choose $\boldsymbol{\theta}_u \sim Dir(\alpha)$,
   (b) For each of the words $w_{u,n}$ where $n = 1, \cdots, N_u$,
      i. Choose a topic $z_{u,n} \sim Multi(\boldsymbol{\theta}_u)$,
      ii. Choose a word $w_{u,n} \sim Multi(\boldsymbol{\phi}_{z_{u,n}})$,

Next, using the 'user interest model' estimated by LDA, we identify whether the tweets match the target topic. However, it is difficult to identify whether a tweet matches the target topic using only latent user interest information. The proposed method uses a combination of user interest information and the surface information of a tweet. The following three steps are used in combination to identify a target topic.

Step1 Identified by a classifier using the surface information of the tweet.

Step2 Identified by a classifier using the 'user interest model' of the user who posted the tweet.

Step3 Identified by scores calculated from the probability between Step 1 and Step 2 using linear interpolation.

In Step 1, we handle one tweet as a single document and create a document vector using word importance. Then, we employ a classifier that uses the document vector as features and estimate the probability that each tweet is a tweet of the target topic using machine learning. In Step 2, we use the 'user interest model' estimated by LDA. We employ a classifier that uses the topic proportions of each user as features and estimates the probability that each tweet is a tweet of the target topic using machine learning. In Step 3, we calculate the scores of each tweet by linear interpolation using Equation (5) between the probability $P_{baseline}$ estimated by the classifier in Step 1 and the probability P_{lda} estimated by the classifier in Step 2. We then estimate the probability that each tweet is a tweet of the target topic using this score.

$$Score = \lambda P_{baseline} + (1 - \lambda) P_{lda} \tag{5}$$

The probabilities of Step 2 and Step 1 take values between 0 and 1. The scores calculated in Step 3 also take a value between 0 and 1. For identification of the binary, we determine that a tweet is related to 'Apple Inc.' when the probability or the score that the tweet is related to 'Apple Inc.' is greater than 0.5.

4. Evaluation experiments using Japanese tweets

In the evaluation experiment, we collected tweets containing the keyword that means 'apple' in Japanese katakana characters from tweets that were posted from 4 to 11 January 2014. We performed an experiment to identify whether each collected tweet is related to 'Apple Inc.' or another 'apple' topic.

4.1. Experimental details

The experimental data was collected using the Twitter API, and 10,000 tweets were randomly sampled from 179,079 tweets that matched the given keyword. Then, after bot and public relations tweets were removed from the 10,000 tweets, we were left with 904 tweets by 855 users that could be gathered from a one year period. We labelled the two types of tweets according to their relation to 'Apple Inc.' or another 'apple' topic for each tweet in advance. Then, we excluded 57 tweets that could not be identified as belonging to either category and created test data out of the remaining 847 tweets. We also extracted 802 users from the test data.

Training data for LDA was 1,151,739 tweets posted by the 855 users over a one year period. From this data, we created four sets of training data from the last month, the last three months, the last six months and the last year. We used the MeCab[11] morpheme analyser to extract common nouns and proper nouns. To deal with new words and buzzwords, we added the title words of Japanese Wikipedia entries as common nouns to the morphological analysers user dictionary in advance. Following the method of Iwata et al.[12], LDA learning used collapsed Gibbs sampling[3], and the hyper-parameters α and β were estimated using a fixed-point iterative method each time sampling was performed. We determined the number of topics $k = 150$ relative to the computation time and the stability of the model by the perplexity value compared with the experiment. Then, using each of the four models (last one month, last three months, last six months and last one year) created by LDA learning, we generated the classifiers that were used in the topic proportions of each user for each model as features. We performed a preliminary experiment to compare the identification ability of the four models by each classifier for the test data that labelled the answer. In this experiment, we used the WEKA[13] data mining software to implement the classifier and used the sequential minimal optimization (SMO) algorithm, which showed high accuracy in a preliminary experiment. In addition, we used the 10-fold cross method for evaluation and the default value for other options. The results of the preliminary experiment are shown in Table 2.

Table 2. Comparison of four models using training data from different periods.

	Last 1 month	Last 3 monthes	Last 6 months	Last 1 year
% Correct	75.2	75.9	78.3	77.0
Avg. #words	651	1,769	3,512	6,698

Relative to the dynamics of topic and user interests, it is considered that newer tweets show higher identification ability. However, Table 2 shows that this trend is only true for the comparison between the last one year and the last six months. On the other hand, by comparing the last one month, the last three months and the last six months, it can be seen that the accuracy rates improve as the period becomes longer. This experiment shows that the effect of learning from tweets that were collected over a longer period of time is greater than the effects of topic and user interest dynamics because the vocabulary is increased and the occurrences of many events that characterize the distribution of topics are better covered. Therefore, we used the LDA model for the last six months in this experiment.

For Step 1, we used a classifier using the surface information of the tweet. We extracted only proper nouns and common nouns from the 847 experimental data tweets as one document per tweet. In addition, we employed a document vector using the term frequency-inverse document frequency values of the word extracted as the features. We also reduced the dimension to 150, which is the same as the number of the topics using singular value decomposition. Then, we employed a classifier using the document vector as the features, and we estimated the probability that each tweet is related to 'Apple Inc.' using machine learning. Here, we used the same SMO algorithm for the classifier and estimated the probability of each instance using an option that fits logistic regression models. In addition, the estimation used a random number; therefore, we have averaged the results using 10 different seeds. In Step 2, we used the 'user interest model' estimated by LDA. We employed a classifier that used the topic proportions of each user as the features and estimated the probability that each tweet is related to 'Apple Inc.' using machine learning. With the exception of the different features, the classifier algorithm and all options are the same as in Step 1. For Step 3, we calculated the scores of each tweet using the linear interpolation between the probability $P_{baseline}$ estimated by the Step 1 classifier and the probability P_{lda} estimated by the Step 2 classifier. We then estimated the probability that each tweet is related to 'Apple Inc.' by each score. The interpolation coefficient λ was evaluated for all values between 0 and 1 in 0.5 increments. We used a λ value that yielded the highest accuracy.

4.2. Results and discussion

Table 3 shows the results identified by the three methods: Step 1 using the information in the tweet, Step 2 using only LDA and Step 3 using a combination of the other two using linear interpolation (lerp) as the baseline. As can be seen from the accuracy rate in Table 3, Step 2 was lower than the baseline. On the other hand, the accuracy rate of Step 3 improved by 8.5% compared with the baseline. In general, it is difficult to correct a misclassification of a tweet. The number of false negatives (FN) in the contingency table in Table 4 indicates this. From the FN rate shown in Table 3, it can be seen that the FN rate of Step 3 improved by 16.5% compared with the baseline. From the comparison of Step 1 and Step 2, the accuracy rate obtained using only the 'user interest model' of the user was lower than the accuracy obtained using only the surface information of the tweet. These results suggest that it is not possible to identify whether a tweet is related to 'Apple Inc.' or other apple-related topics using only latent user interest information. Furthermore, from the comparison of Step 1 and Step 3, it is confirmed that the accuracy rate using the surface information of the tweet and the user interest model was higher than that of the method using only the surface information. These results suggest that, for the case that a target topic cannot be identified using only the surface information, combining surface information with latent user interest information improves the ability to identify the target topic.

Table 3. Evaluation experiment results for the three methods.

	Step 1: Baseline	Step 2: Only LDA	Step 3: Lerp
# correct	704	664	776
% correct	83.3	78.6	91.8
# FN	115	64	17
% FN	19.4	10.8	2.9

We also examined the data that could not be correctly identified. With regard to the test data of all 847 tweets and the 141 tweets that were incorrectly identified using the baseline, Fig. 2 illustrates the results of the density estimation of the number of proper nouns and common nouns extracted from each tweet. The 141 tweets that were incorrectly

Table 4. Example contingency table.

	Classified as 'Apple Inc.'	Classified as Others
'Apple Inc.'	TP (True positive)	FN (False negative)
Others	FP (False positive)	TN (True negative)

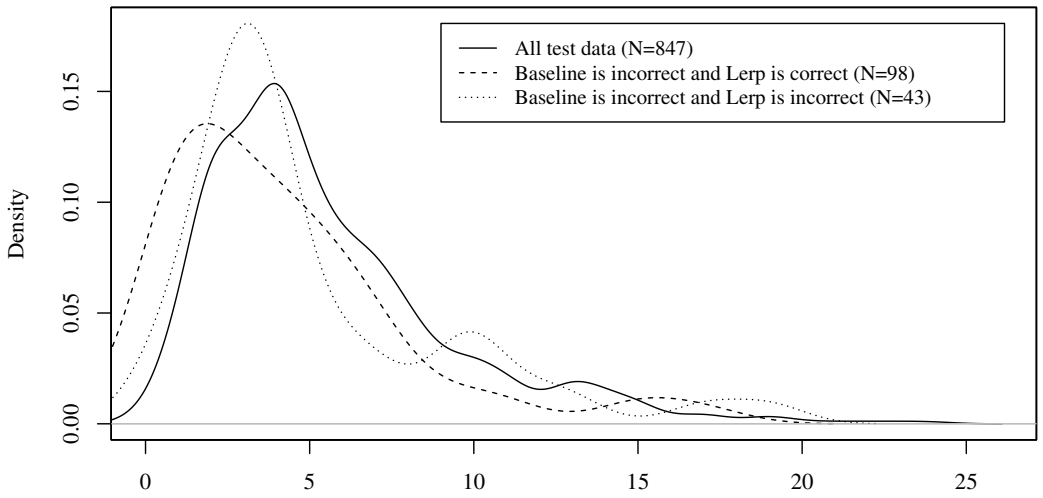

Fig. 2. Distributions of the number of words.

identified using the baseline are plotted separately; 98 tweets that yielded a correct answer by lerp and 43 tweets that yielded incorrect answers by lerp. The peaks of the two distributions of the 141 tweets that could not be identified correctly can be seen on the left side. This indicates that fewer extracted words are one of the misidentification causes for the baseline. In particular, the 43 tweets (i.e. incorrect answer by lerp) strongly demonstrate this trend. These results suggest that, for a case that cannot be identified using only the surface information as the baseline, it is effective to combine the latent user interest information, which increases the identification ability.

5. Conclusions

In this paper, we have focused on noisy data contained in tweets collected from a keyword search. We have proposed a method to classify tweets from search results, including noisy data, using topic models. In addition, we have reported the evaluation experiment results. To identify a target topic tweet in the evaluation experiment, the proposed method showed an accuracy improvement of 8.5% and a reduction of 16.5% in misidentification rate (% FN) compared with conventional methods.

We have confirmed that the features derived by the proposed method can be used to identify target topics in tweets. However, using the supervised classifier discussed in this paper requires training data for each target topic. Hence, the application of features derived for classification by unsupervised learning, such as clustering, and improvement of the accuracy of these features will be the focus of future work.

References

1. Hofmann, T.. Probabilistic latent semantic indexing. In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM; 1999, p. 50–57.
2. Blei, D.M., Ng, A.Y., Jordan, M.I.. Latent dirichlet allocation. *the Journal of machine Learning research* 2003;**3**:993–1022.
3. Griffiths, T.L., Steyvers, M.. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America* 2004;**101**(Suppl 1):5228–5235.
4. Okumura, M.. Microblog mining. *IEICE technical report Natural language understanding and models of communication* 2012;**111**(427):19–24.
5. Weng, J., Lim, E.P., Jiang, J., He, Q.. Twitterrank: finding topic-sensitive influential twitterers. In: *Proceedings of the third ACM international conference on Web search and data mining*. ACM; 2010, p. 261–270.
6. Pennacchiotti, M., Popescu, A.M.. A machine learning approach to twitter user classification. In: *ICWSM*. 2011, .
7. Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.. Probabilistic author-topic models for information discovery. In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM; 2004, p. 306–315.
8. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H., et al. Comparing twitter and traditional media using topic models. In: *Advances in Information Retrieval*. Springer; 2011, p. 338–349.
9. Sasaki, K., Yoshikawa, T., Furuhashi, T.. A proposal of online topic model for twitter considering temporal dynamics of user interests and topic trends. *Transactions on Mathematical Modeling and its Applications (TOM)* 2014;**7**(1):53–60.
10. Iwata, T., Watanabe, S., Yamada, T., Ueda, N.. Topic tracking model for analyzing consumer purchase behavior. In: *IJCAI*; vol. 9. 2009, p. 1427–1432.
11. Kudo, T., Yamamoto, K., Matsumoto, Y.. Applying conditional random fields to japanese morphological analysis. In: *EMNLP*; vol. 4. 2004, p. 230–237.
12. Iwata, T.. Data mining using latent topic model. In: *Proceedings of the Collection of Technical Reports of the First Workshop on Latent Dynamics (LD-1)*. 2010, .
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.. The weka data mining software: an update. *ACM SIGKDD explorations newsletter* 2009;**11**(1):10–18.