Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 60 (2015) 141 – 149

Procedia Computer Science

19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

User intent estimation from access logs with topic model

Keisuke Uetsuji[a], Hidekazu Yanagimoto[a,*], Michifumi Yoshioka[a]

[a]*Osaka Prefecture University, 1-1, Gakuen-cho, Naka-ku, Sakai, Osaka, 599-8531, Japan*

Abstract

As the Internet is widespread and there are many online shops in the Internet, many persons buy products in the online shops. Customer's behavior in the online shops is a sequence of customer driven activities intrinsically because his/her movement in an online shop occurs according to only his/her decision. Hence, to achieve satisfactory purchase experiments it is important how the shop supports them. Online shops have to predict visitors' intents correctly to support them effectively. One of information resources the shops can use is an access log including information on customer's movement in the online shop. If they are histories of customer's behaviors in online shops and the behaviors depend on customer's intents, we can extract new knowledge on them from the access logs. Speaking concretely, we can predict customers' intents from the access logs since their internal intents affect their activities. We can realized more appropriate recommendation service by changing recommendation strategy depending on customer's intents. In this paper, we propose a method to predict customer's intents from access logs in a real online shop. We adopt a Topic Tracking Model (TTM) to analyze the access logs.

Peer-review under responsibility of KES International

Keywords: Access log, Topic model

1. Introduction

As the Internet is widespread and there are many online shops in the Internet, many persons buy products in the online shops. Customer's behavior in the online shops is a sequence of customer driven activities intrinsically because his/her movement in an online shop occurs according to only his/her decision. Hence, to achieve satisfactory purchase experiments it is important how the shop supports them. Online shops have to predict visitors' intents correctly to support them effectively. One of information resources the shops can use is an access log including information on customer's movement in the online shop. These logs are originally intended to be used for only online shop service maintenance. If they are histories of customer's behaviors in online shops and the behaviors depend on customer's intents, we can extract new knowledge on them from the access logs. Speaking concretely, we can predict customers' intents from the access logs since their internal intents affect their activities. We assume that users visiting an online shop have some intents; for example, just a window shopping, comparing similar items, looking for a newer item than the one the customer has already bought. Detecting these intents are important to support them. We can realized

* Corresponding author. Tel.: +81-72-254-9279 ; fax: +81-72-254-9909.
E-mail address: hidekazu@kis.osakafu-u.ac.jp

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2015.08.113

more appropriate recommendation service by changing recommendation strategy depending on customer's intents. For example, if an online shop can predict a category a customer will buy from his/her intents, we can recommend more specific items related to the category. Moreover, if an online shop can predict a item a customer will buy, we give him/her extra information on a sale price.

In this paper, we propose a method to predict customer's intents from access logs in a real online shop. We adopt a Topic Tracking Model (TTM)[1] to analyze the access logs. A topic model is one of generative models which include latent properties called topics. The topic is an element to determine customer's behavior and customer's activities are generated according to the topic. In generative model, customer's behavior is generated with two-step probabilistic process; in the first step a topic is selected according to a topic probability distribution representing topic selection tendency. In the next step user's activity is determined according to activity probability distribution linked to the selected topic in the first step. Especially TTM can capture topic dynamics because TTM can consider time dependency and analyzes access logs considering dependencies between neighboring actitiveis in access logs.

In section 2, related works are introduced. In section 3 we explain TTM and how to train it and in section 4 we explain how to apply TTM to access log analysis. Experiments using actual access logs are executed in section 5. Finally in section 6, we describe conclusions and future works.

2. Related Works

There are many researches to predict users' intents from their activity histories. In this section we describe such related works to emphasize our research aims. First, we describe researches on search engines. Carman et al.[2] proposed extended LDA models including latent topics and achieved search result personalization. One of their model is similar to ours because they assume that a user has topics. Lin et al.[3] analyzed query logs using n-gram model to predict user behaviors (for example, clicking the item on search result, requesting the next page and so on) and Sadagopan et al.[4] applied Markov model to session classification (for example, sessions made by bots) with logs. In their experiments, logs contains only specifics behaviors, such as searching, requesting next pages, clicking a promotion link, and so on. Guo et al.[5] predicted user intents; whether or not the user purchases item. They used SVM and various operations on the search result pages as feathers; cursor movements, scrolls, keyboard inputs, and so on. What features are useful for predicting user behaviors were examined by Van den Poel et al.[6] works.

There are researches about user behaviors analysis on online shopping sites. Iwata et al. proposed Topic Tracking Model (TTM), applied it to purchase history analysis, and predicted items a user would purchase. In this paper we apply TTM to access log analysis and predict customer's intents. A work of Kumagae et al., predicted user's intent from access logs using a hidden Markov model.

Our proposed method is related to a topic model to predict user's intents. Latent Dirichlet Allocation (LDA)[7] is one of the most famous topic models and at first is proposed as a document generation model. Nowadays LDA is used more widely to capture hidden structure from observation. However, it cannot deal with time-series data well because LDA assumes document as bag-of-words that is a independent among word occurence. Therefore some LDA-based topic models which can deal with time-series data were proposed. Topic Tracking Model (TTM) is one of LDA variations and in our proposed method we use it.

Dynamic Mixture Model (DMM)[8] and Dynamic Topic Model (DTM)[9] are one of topic models dealing with time series. DMM supposes random variable θ, which denotes a topic, changes over time, in the left figure of Fig. 1. Since ϕ, which denotes user's behavior preference, does not depends on time, the dynamics of behavior probabilities cannot be captured. The graphical model of DTM is shown in the right figure of Fig. 1. DTM can deal with dynamics of topic probabilities and behavior probabilities. Speaking concretely, the model changes α, which denotes trends of topic selection, and β, which denotes behavior preference. However, when we apply it to user's intent analysis, we do not track each user appropriately because no parameters are shared over time. On the other hand, TTM is suitable for access log analysis because TTM define a topic probability and a behavior probability are different as time spends but common parameter, α and β are shared in each user.

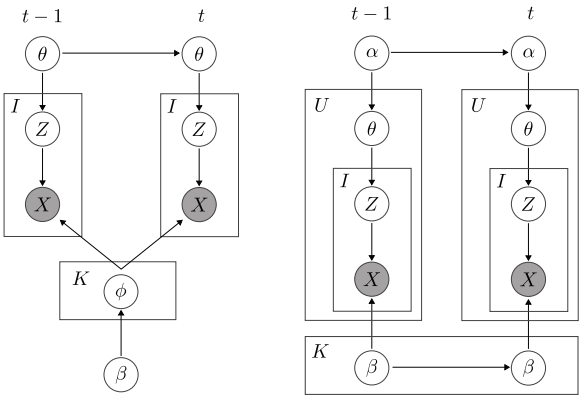

Fig. 1. Graphical model of DMM(left), DTM(right)

3. Topic Tracking Model

3.1. Generative Process

Table 1. Notation in this paper

symbol	description
t	discrete time
u	user
w	user's behavior
k	topic
U	the number of users
W	the number of behaviors users can take
K	the number of topics

In a generative process, observation, which denotes user's actual behavior, is generated from the following process. In the first step a topic is selected according to a topic probability distribution representing user's intent selection preference and then behavior is determined based on the selected topic. Considering $P(w|u,t)$ as a probability of behavior w user u takes in time t, it is expressed as Equation (1). We use notation described in Table 1.

$$P(w|u,t) = \Sigma_{k=1}^{K} \phi_{t,u,k}\theta_{t,k,w} \tag{1}$$

$$\phi_{t,u,k} = P(k|u,t) \tag{2}$$

$$\theta_{t,k,w} = P(w|k,t) \tag{3}$$

$\phi_{t,u,k}$ is a variable that denotes the k-th topic a user u selects at time t and $\theta_{t,k,w}$ is a variable that denotes the w-th behavior a user u take at time t. In this model a topic is selected for each user and affects user's behavior selection. Hence, $\phi_{t,u}$ is regarded as user's intents which select his/her behavior. Moreover, $\theta_{t,k}$ denotes preference of behavior when user's topic is determined. As explained above, the users intent and their behavior preference changes over time. In TTM $\phi_{t,u}$ and $\theta_{t,k}$ is calculated with a Dirichlet distribution conditioned by previous $\phi_{t,u}$ and $\theta_{t,k}$. Hence, we assume the model satisfies a Markov property.

$$\phi_{t,u} = \{\phi_{t,u,k}\}_{k=1}^{K} \tag{4}$$

$$P(\phi_{t,u}|\hat{\phi}_{t-1,u}, \alpha_{t,u}) \propto \prod_{k} \phi_{t,u,k}^{\alpha_{t,u}\hat{\phi}_{t-1,u,k}-1} \tag{5}$$

A parameter of the Dirichlet distribution is $\alpha_{t,u}\hat{\phi}_{t-1,u}$. $\hat{\phi}_{t-1,u}$ is a estimates of $\phi_{t-1,u}$ and $\alpha_{t,u}$ is a persistency parameter. Then, expectation of the distribution is $\hat{\phi}_{t-1,u}$, a variance is $1/\alpha_{t,u}$. Therefore $\alpha_{t,u}$ means how $\phi_{t,u}$ tends to be close to

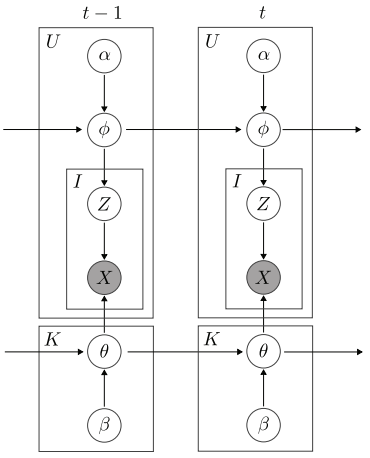

Fig. 2. Graphical model of TTM

$\hat{\phi}_{t-1,u}$. In other words, $\alpha_{t,u}$ is a parameter which represents fluidity of users intents when the time changes from $t-1$ to t. Similarly, the prior distribution of $\theta_{t,k}$ is Equation (7).

$$\theta_{t,k} = \{\theta_{t,k,w}\}_{w=1}^{W} \tag{6}$$

$$P(\theta_{t,k}|\hat{\theta}_{t-1,k}, \beta_{t,k}) \propto \prod_{w} \theta_{t,k,w}^{\beta_{t,k}\hat{\theta}_{t-1,k,w}-1} \tag{7}$$

$\beta_{t,k}$ is a persistency parameter and $\hat{\theta}_{t-1,k}$ is an estimated value of $\theta_{t-1,k}$.

The behavior generation in TTM is summarized as follows. Here, $I_{t,u}$ is the number of behaviors user took on at time t, $x_{t,u,i}$ is the ith behavior of the user u at time t, and $z_{t,u,i}$ is a topic and affects a selection of $x_{t,u,i}$.

1. for each $k = 1, \cdots, K$
   (a) sampling behavior probabilities
   $\theta_{t,k} \sim \text{Dirichlet}(\beta_{t,k}\hat{\theta}_{t-1,k})$
2. for each $u = 1, \cdots, U$
   (a) sampling topic probabilities of user
   $\phi_{t,u} \sim \text{Dirichlet}(\alpha_{t,u}\hat{\phi}_{t-1,u})$
   (b) for each behavior $i = 1, \cdots, I_{t,u}$
      i. sampling a topic
      $z_{t,u,i} \sim \text{Multinomial}(\hat{\phi}_{t,u})$
      ii. sampling a behavior
      $x_{t,u,i} \sim \text{Multinomial}(\hat{\theta}_{t,z_{t,u,i}})$

Fig. 2 is a graphical model which describes dependencies of variables in TTM.

3.2. TTM training

To train TTM from observations, we use a probabilistic EM algorithm[10]. Specifically, we apply the following two step repeatedly.

1. Topics estimation: estimate Z_t using gibbs sampling[11].
2. Persistency parameters update: apply update rules of fixed-point iteration to α_t and β_t.

Repeating the probabilistic EM algorithm until convergence, then we estimate $\hat{\phi}_{t,u}$, $\hat{\theta}_{t,k}$ using samples of Z_t, α_t, β_t.

$$\hat{\phi}_{t,u,k} = \frac{n_{t,u,k} + \alpha_{t,u}\hat{\phi}_{t-1,u,k}}{n_{t,u} + \alpha_{t,u}} \tag{8}$$

$$\hat{\theta}_{t,k,w} = \frac{n_{t,k,w} + \beta_{t,k}\hat{\theta}_{t-1,k,w}}{n_{t,k} + \beta_{t,k}} \tag{9}$$

$n_{t,u}$ is the number of behaviors generated by the user u at time t. $n_{t,k}$ is the number of behaviors that were assigned to topic k at time t. $n_{t,u,k}$ is the number of behaviors generated by the user u at time t and assigned to topic k.

In a topic estimation step, we use gibbs sampling to get topics Z_t. Defining X_t as behaviors of all users, rules are following.

$$\hat{\Phi}_{t-1} = \{\hat{\phi}_{t-1,u}\}_{u=1}^{U} \tag{10}$$

$$\hat{\Theta}_{t-1} = \{\hat{\theta}_{t-1,k}\}_{k=1}^{K} \tag{11}$$

$$P(z_j = k|X_t, Z_{t\backslash j}, \hat{\Phi}_{t-1}, \hat{\Theta}_{t-1}, \alpha, \beta) \propto \frac{n_{t,u,k\backslash j} + \alpha_{t,u}\hat{\phi}_{t-1,u,k}}{n_{t,u\backslash j} + \alpha_{t,u}} \frac{n_{t,k,x_j\backslash j} + \beta_{t,k}\hat{\theta}_{t-1,k,x_j}}{n_{t,k\backslash j} + \beta_{t,k}} \tag{12}$$

In a persistency parameters update step, we use update rules of fixed-point iteration. Those iterations are to maximize likelihood of persistency parameters; $\alpha_{t,u}, \beta_{t,k}$. Update rules are:

$$\alpha_{t,u} \leftarrow \alpha_{t,u} \frac{\Sigma_k \hat{\phi}_{t-1,u,k} A_{t,u,k}}{\Psi(n_{t,u} + \alpha_{t,u}) - \Psi(\alpha_{t,u})} \tag{13}$$

$$\beta_{t,k} \leftarrow \beta_{t,k} \frac{\Sigma_w \hat{\theta}_{t-1,k,w} B_{t,k,w}}{\Psi(n_{t,k} + \beta_{t,k}) - \Psi(\beta_{t,k})} \tag{14}$$

where

$$A_{t,u,k} = \Psi(n_{t,u,k} + \alpha_{t,u}\hat{\phi}_{t-1,u,k}) - \Psi(\alpha_{t,u}\hat{\phi}_{t-1,u,k}) \tag{15}$$

$$B_{t,k,w} = \Psi(n_{t,k,w} + \beta_{t,k}\hat{\theta}_{t-1,k,w}) - \Psi(\beta_{t,k}\hat{\theta}_{t-1,k,w}) \tag{16}$$

Ψ is the digamma function, $\Psi = \frac{\partial \log \Gamma(x)}{\partial x}$.

4. Application of TTM to access log analysis

TTM is usually used to analyze purchase history [1]. In this paper we apply TTM to access log analysis and extract a topic, which denotes user's intent to select his/her actual behavior, from the access logs. Hence, purchase history is different from access logs since generally some access logs are not related to purchase. For example, customers often visit online shop to know new product.

1. Remove user's access log without item purchase because we focus on user's intent related to item purchase.
2. Divide user's access logs into some sequences including only a purchase event because we clarify relationship between purchase and user's behavior.
3. Adjust the number of sessions in the sequences. The sequence consists of some sessions separated with predefined duration, for example half an hour. We assume the session is constructed based on single intent and regard the sessions as features of the sequence. Since the length of the sequence is different, sequences include the different number of sessions. Hence we insert dummy sessions in head of the sequence to adjust the number of sessions in the sequence. The dummy session is considered as a access to nonexistent pages. After that, the sequence to analyze user's intent includes purchase in the last session of the sequence.

Since the session is a set of web pages (URLs) which a user visited, we represent a session using these web pages. Generally URL description denotes the structure of online shop and user's operation, for example search, purchase, and so on. For example, a URL includes item categories, item brand, search conditions. Hence, we separate URLs

with symbols, "/", "&", and "=" and so on, and each separated parts, which we call an activity, is used as features to describe the session. For example, the separated parts include item name, brand name and so on.

We confirm relation between our study and TTM in Section 3. We sort sessions of each sequence in chronological order and describe the session order as time t. And we expressed user's behavior w as the activity extracted from a session, such as item name, brand name. We set the sequence as u.

After TTM training, we discuss change of topics and predict user's intent. We determine a topic in a session which has the highest probability in all topics. Hence, since we assume a session include only one topic, we can capture a topic sequence easily.

5. Experiments

5.1. Dataset and evaluation

We used access logs in an actual online shop related to golf. These logs collected for a month and 24,050,086 accesses of 779,059 users were included. We executed above preprocess for the access logs. The number of sessions in a sequence is 10. After that, the dataset which contains 43,516 sequences, 31,935,160 activities (82,517 unique activities) were obtained. Finally, we set the number of topics to 50 and estimated topics by TTM. We assume the customer's intent as topic which has highest probability on the user.

Since customer's intents are not observed directly, we do not know correct intent label for each session. Hence, it is difficult to evaluate how correctly our proposed method assigned a topic to a session. We discuss coocurrence frequency of a topic and item purchase. When a topic frequently occurs with the same item purchase, we can judge strong relation between the topic and the item purchase. Moreover, we discuss when a topic related to item purchase occurs in a sequence. If the topic appears near the final session, we can predict customer's purchase using the topic occurrence.

5.2. Results

We found some estimated topics were related to a purchased items. Table 2 shows probabilities of item purchased on topic 29, 39 or 40 in the 10th session. 44% of sessions assigned Topic 29 to long pant was bought. Hence, we can regard Topic 29 as customer's intent showing customers look for long pants.

Table 2. Item purchase probability in Topic 29, 39, or 40 the final session

Topic 29 item	probability	Topic 39 item	probability	Topic 40 item	probability
long pants	0.44	driver	0.37	driver	0.25
cap	0.06	putter	0.06	fairway wood	0.15
half-sleeved shirt	0.06	fairway wood	0.05	iron	0.12
ball	0.05	ball	0.04	utility	0.10
half pants	0.03	cap	0.03	wedge	0.08
shoes	0.03	long pants	0.03	putter	0.05
glove	0.03	glove	0.03	glove	0.02
other goods	0.02	half-sleeved shirt	0.03	ball	0.01
tee	0.02	wedge	0.03	long paths	0.01
underwear	0.02	utility	0.03	cap	0.01

Topic 38 and Topic 39 is related to driver purchase strongly. However, they include different activity patterns. In Table 3 we discuss activity patterns of Topic 38 and Topic 39 focusing on search operations. One of operations is searches related to flex, which is a attribute of driver, and the another is searches related to brand name of driver. In Topic 39 searches specifying flex occurs more frequently than in Topic 38. On the other hand, in Topic 38 searches with a brand name are executed more frequently than in Topic 39. Therefore, our propose method can classify topics

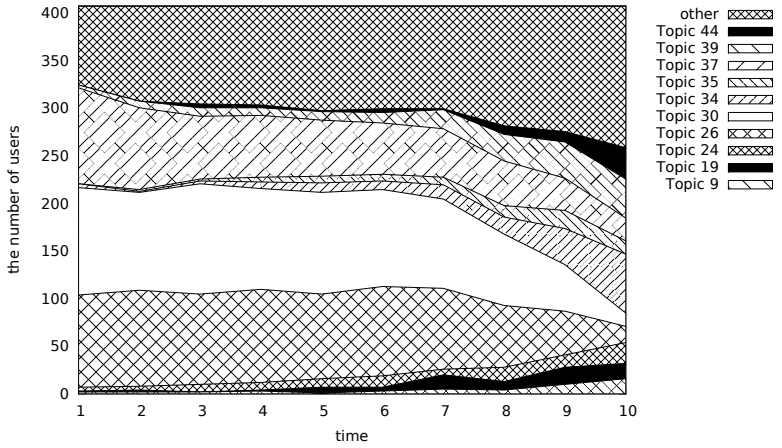

Fig. 3. Topic changes for sessions including iron purchase

(customer's intents) not only according to purchased items but also according to how to search item. Hence, we could confirm our proposed method extract customer's intent from access logs appropriately.

Table 3. Flex search, brand search probability of users who have Topic 38 or Topic 39

	flex search		brand search	
t	Topic 38	Topic 39	Topic 38	Topic 39
1	0.0002	0.0107	0.0170	0.0090
2	0.0007	0.0071	0.0179	0.0079
3	0.0011	0.0061	0.0159	0.0066
4	0.0002	0.0074	0.0168	0.0051
5	0.0014	0.0080	0.0157	0.0055
6	0.0024	0.0089	0.0161	0.0063
7	0.0022	0.0088	0.0175	0.0065
8	0.0015	0.0080	0.0153	0.0089
9	0.0021	0.0080	0.0165	0.0073
10	0.0016	0.0087	0.0128	0.0091

Fig. 3 shows topic changes for sessions including iron (a category of golf club) purchase. The top 10 frequently occurred topics are shown individually in Fig. 3 and the other 40 intents are contained in "others" because such topics occurred less frequently than the top 10 topics. At $t = 1$, Topic 26, 30, 37 occupied in almost all sessions. Since the Topic 26, 30, 37 includes many activities related to dummy sessions, we regarded the topics as meaningless topics. Since almost all sequences includes the smaller number of sessions than 10 sessions, they have dummy sessions at $t = 1$. Sessions including Topic 26, 30, 37 decrease gradually as t increases. This shows dummy sessions decreases and topics constructed with actual customer's activities emerge.

Using TTM we can capture how the occurrence frequency of a topic changes as time spends. Table 4 shows occurrence probability of Topic 47, which is related to browsing reviews pages as time spends. The probabilities increase over time, thus it shows customers tend to check review pages as a probability of purchase is high.

Finally, we discussed whether an estimated topic can predicted item category that purchase item belongs to. After topic estimation with TTM and LDA, we divide sequences into two data sets randomly. We assigned the item category to a topic based on cooccurrence frequency of a topic and purchase item category in a data set. In evaluation step we evaluate category prediction accuracy using the another data set, which is not used for category assignment. The results are shown in Table 5. There are three topic estimation method, TTM, LDA(all) and LDA(last). TTM is our proposed method, and LDA(all) is the method which applies LDA and uses all sessions before t-th session to estimate topic. LDA(last) uses only the t-th session to estimate topic. In evaluating an accuracy rate we used only the last 3

Table 4. Transition of review list probability of intent 47

t	probability
1	0.00008
2	0.00000
3	0.00000
4	0.00000
5	0.00000
6	0.00017
7	0.00019
8	0.00028
9	0.00032
10	0.00206

sessions since in our proposed method we inserted dummy sessions into sequences and topics in earlier sessions is not reliable.

The accuracy rate of TTM is exceed the rate of LDA(all) and LDA(last) on every session. Since TTM can track changes of topic transition, TTM can estimate a topic in a session more correctly. Hence, our proposed method improved category prediction considering change of topics over time.

Table 5. Accuracy rate of prediction of purchase item category

t	TTM	LDA(all)	LDA(last)
8	0.239	0.233	0.220
9	0.268	0.248	0.218
10	0.355	0.321	0.339

6. Conclusion

In this paper, we proposed a access log analysis method using Topic Tracking Model. And we confirmed our proposed method could discover useful topics related to item purchases.

In future works, we will find some topic transition patterns shared with many customer's activities. In this study we focus on individual topics but I think topic transition patterns is valuable. And those are useful for recommendation in earlier session.

We will make training faster. Since it takes too long to train TTM, it is difficult to process access logs in real time fashion. We will use the methods that achieve faster LDA estimation. For example, the efficient method to calculate Z_t, and the parallelized algorithms for LDA were proposed.[12,13]. These methods are also adoptable to the TTM estimation, because gibbs sampling in TTM is the same process as in LDA.

References

1. T. Iwata, S. Watanabe, T. Yamada, N. Ueda, Topic tracking model for analyzing consumer purchase behavior., in: IJCAI, Vol. 9, Citeseer, 2009, pp. 1427–1432.
2. M. J. Carman, F. Crestani, M. Harvey, M. Baillie, Towards query log based personalization using topic models, in: Proceedings of the 19th ACM international conference on Information and knowledge management, ACM, 2010, pp. 1849–1852.
3. J. Lin, W. J. Wilbur, Modeling actions of pubmed users with n-gram language models, Information retrieval 12 (4) (2009) 487–503.
4. N. Sadagopan, J. Li, Characterizing typical and atypical user sessions in clickstreams, in: Proceedings of the 17th international conference on World Wide Web, ACM, 2008, pp. 885–894.
5. Q. Guo, E. Agichtein, Ready to buy or just browsing?: detecting web searcher goals from interaction data, in: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, ACM, 2010, pp. 130–137.
6. D. Van den Poel, W. Buckinx, Predicting online-purchasing behaviour, European Journal of Operational Research 166 (2) (2005) 557–575.

7. D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, the Journal of machine Learning research 3 (2003) 993–1022.
8. X. Wei, J. Sun, X. Wang, Dynamic mixture models for multiple time-series., in: IJCAI, Vol. 7, 2007, pp. 2909–2914.
9. D. M. Blei, J. D. Lafferty, Dynamic topic models, in: Proceedings of the 23rd international conference on Machine learning, ACM, 2006, pp. 113–120.
10. H. M. Wallach, Topic modeling: beyond bag-of-words, in: Proceedings of the 23rd international conference on Machine learning, ACM, 2006, pp. 977–984.
11. T. L. Griffiths, M. Steyvers, Finding scientific topics, Proceedings of the National Academy of Sciences 101 (suppl 1) (2004) 5228–5235.
12. L. Yao, D. Mimno, A. McCallum, Efficient methods for topic model inference on streaming document collections, in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2009, pp. 937–946.
13. A. Ihler, D. Newman, Understanding errors in approximate distributed latent dirichlet allocation, Knowledge and Data Engineering, IEEE Transactions on 24 (5) (2012) 952–960.