Evolution of Information Needs based on Life Event Experiences with Topic Transition

Naoto Takeda
University of Tsukuba
Tsukuba, Japan
s1621623@u.tsukuba.ac.jp

Yohei Seki
University of Tsukuba
Tsukuba, Japan
yohei@slis.tsukuba.ac.jp

Mimpei Morishita
kizasi Company, Inc.
Tokyo, Japan
mimpei@kizasi.jp

Yoichi Inagaki
kizasi Company, Inc.
Tokyo, Japan
inagaki@kizasi.jp

ABSTRACT

We propose a method to clarify the evolution of users' information needs related to a user's interests and actions based upon life events such as "childbirth." First, we extract topic transitions using dynamic topic models from blogs posted by users who have experienced life events. Next, we select the topics by computing the differences in topic probabilities before and after the life event. We evaluated our method based on three life events: "childbirth," "finding employment," and "marriage." Our method selected life event-relevant topics such as "child development," "working life," and "wedding ceremony." We found mothers' information needs such as "how to introduce baby food," employees' information needs such as "preparing an induction programme," and couples' information needs such as "wedding reception planning" in each topic.

CCS CONCEPTS

- **Information systems → Web and social media search**;

KEYWORDS

Life event; DTMs (Dynamic Topic Models); Information needs

1 INTRODUCTION

Users' interests and actions evolve as they experience life events such as "childbirth," "finding employment," or "marriage." We propose a method to clarify the evolution of the users' information needs related to a user's interests and actions based on the life events that they experience.

Many researchers have analyzed the changes in users' interests or actions in life events such as "engagement" [4] or "losing a job" [3] using the users' social media posts. In these studies, however, the analysts must specify the interests or actions about the life events in advance. In this paper, we propose a method for extracting topics relevant to life events and their transitions.

Our method uses dynamic topic models (DTMs) [1] and blog posts from users who have experienced a life event to extract the topics and their transitions. We select the topics by computing the difference in topic probabilities before and after the life event. Accordingly, the analysts[1] can provide information relevant to the life event at appropriate times. For instance, many new mothers become interested in child development after the "childbirth" life event. With our method, analysts can find trends of information needs such as "how to introduce baby food" or "usage of early development toys" in "child development" topic to provide supportive information timely for new mothers. We demonstrate the generality of our method by applying it to multiple life events. We also verify that our method can clarify the evolution of information needs by analyzing the chronological evolution of words in topics.

2 RELATED WORK

2.1 Users' Actions in Life Events

Choudhury et al. [4] analyzed the changes in user's actions of "marriage engagement" with Twitter posts. They clarified that the number of postings about "wedding planning" or "enjoyable activities (couple-centric)" increased after "engagement". Burke et al. [3] focused on the "losing a job" experience and analyzed the changes in users' activities on Facebook. They clarified that Facebook communications were effective in helping users find a new job and relieving their stresses.

In these studies, however, the analysts must specify his or her interests or actions about the life event in advance. In this paper, we propose an extraction method for topic transitions that reflect the evolution of the users' interests or actions in life events by computing the difference in topic probabilities before and after the life event.

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan
© 2017 ACM. 978-1-4503-5022-8/17/08...$15.00
DOI: http://dx.doi.org/10.1145/3077136.3080703

[1] The market analysts who investigate trends of users' interests.

2.2 Evolution of Information Needs

DTM is a topic model that extends latent Dirichlet allocation (LDA)[2], which enables us to obtain time-series changes of topic probability distributions and word probability distributions for a topic. Zhang et al. [6] proposed a different type of DTM for monitoring the temporal evolution of market competition using tweets and their associated images. Kanhabua et al. [5] proposed a method to detect event-related queries from Web search logs. These studies clarified that users' information needs, which were relevant to "market of interest" or "natural disasters," evolved over time. Many users shared similar information needs simultaneously. In this paper, we focus on the information needs in life events, which were shared by users at different times.

3 TOPIC TRANSITION EXTRACTION AND TOPIC SELECTION

First, to extract topic transitions, we use DTM, which enables us to obtain time-series changes of topic probability distributions and word probability distributions in a topic. In DTM, time slices (TS) are used to separate the data. We divided the blogs per month that were posted by users who experienced a life event and extracted the topic transitions. In this paper, we set $TS = 25$ to analyze the 12 months before the life event, 12 months after, and the month in which the life event occurs.

Next, to select the life event-relevant topics, which evolve with the life event experience, we select the topics by computing the difference in topic probabilities before and after the life event. In our method, we use the difference between the means (variances) of topic probabilities before and after the life event as the clue to find the life event-relevant topics. With the difference of the means, we can extract the topics for which the probabilities increase or decrease drastically after the life event. With the difference of variance, we can extract the topics for which the probabilities stay flat or are unstable before and after the life event.

In Figure 1, we illustrate the topic selection process with topic probabilities. The topic transition with a red line in Figure 1 represents topics for which the probabilities drastically increase. For instance, this line corresponds to the "working life" topic in the "finding employment" life event because this topic was newly introduced after the life event experience. However, the topic transition with a green line in Figure 1 corresponds to the topics for which the probabilities drop gradually before the life event and remain constant after the life event. For instance, this line corresponds to the "job hunting" topic for the "finding employment" life event. In this transition, the variance of topic probabilities after the life event decreased compared with that before the life event.

We defined $score_t$ for topic selection as the sum of the difference score of the means of topic probabilities and the difference score of the variances of topic probabilities. The difference score of the means ($m\text{-}score_t$) and the difference score of the variances ($v\text{-}score_t$) are normalized as a z score for each. These scores are defined as follows:

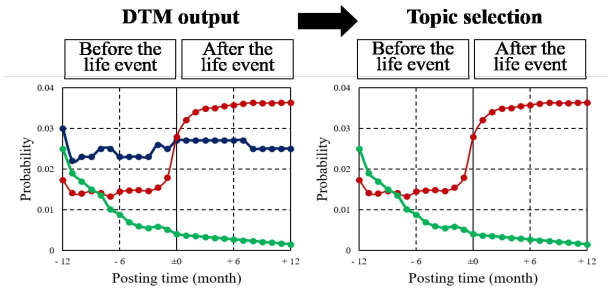

Figure 1: Topic selection with topic probabilities

$$score_t = m\text{-}score_t + v\text{-}score_t \tag{1}$$

$$m\text{-}score_t = \frac{|E(\boldsymbol{x}_t) - E(\boldsymbol{y}_t)| - \mu_m}{\sigma_m} \tag{2}$$

$$v\text{-}score_t = \frac{|V(\boldsymbol{x}_t) - V(\boldsymbol{y}_t)| - \mu_v}{\sigma_v} \tag{3}$$

$$\mu_m = \frac{1}{K}\sum_k |E(\boldsymbol{x}_t) - E(\boldsymbol{y}_t)| \tag{4}$$

$$\mu_v = \frac{1}{K}\sum_k |V(\boldsymbol{x}_t) - V(\boldsymbol{y}_t)| \tag{5}$$

$$\sigma_m = \sqrt{\frac{1}{K}\sum_k (|E(\boldsymbol{x}_t) - E(\boldsymbol{y}_t)| - \mu_m)^2} \tag{6}$$

$$\sigma_v = \sqrt{\frac{1}{K}\sum_k (|V(\boldsymbol{x}_t) - V(\boldsymbol{y}_t)| - \mu_v)^2} \tag{7}$$

In these equations, $\boldsymbol{x}_t$ is the topic probability distribution before the life event, $\boldsymbol{y}_t$ is the topic probability distribution after the life event, $E(\boldsymbol{x})$ is the mean of $\boldsymbol{x}$, $V(\boldsymbol{x})$ is the variance of $\boldsymbol{x}$, K is the number of topics, and μ_m (μ_v) is the mean of the differences of the mean (variance) topic probabilities before and after the life event. σ_m (σ_v) is the standard variation of the difference of mean (variance) topic probabilities before and after the life event.

4 EXPERIMENT: TIME EVOLUTION OF LIFE EVENT TOPICS

4.1 Method

In our experiment, to verify whether the selected topics reflect users' interests or actions, we checked typical words and transitions of the selected topics in multiple life events. We also checked the chronological evolution of words in the selected topics to verify whether they evolved according to users' information needs.

Experimental Data: We collected blogs based on three life events: "childbirth," "finding employment," and "marriage." We extracted blogs published from January 11, 2008 to March 13, 2011 from blogram.jp[2], which is a popular blog ranking site in Japan.

(1) **Life Event User Selection**
We labeled the users who had experienced the life event. First, we extracted the blog posts with the life event queries that describe users' experience. Table 1 shows a list of queries for each life event. We extracted the users who posted blogs that contained the queries for each life event (3,357 users

[2] http://blogram.jp/

Table 1: Query list for each life event

Life event	Query
Childbirth	Gave birth to [syussan shimashita]
Finding employment	New profession [shin-syakaijin], Initiation ceremony[3] [nyuusya-shiki]
Marriage	Getting married [kekkon shimashita], Marriage registered [nyuuseki shimashita]

Table 2: Results of labeling for "childbirth"

label	Blog posts
1	I gave birth to a bouncing baby boy at dawn.
1	I gave birth to a baby 10 days before my due date. Although I had a difficult childbirth, we left the hospital in good health.
0	My close friend gave birth to a baby boy.

in "childbirth", 3,719 users in "finding employment", and 3,619 users in "marriage" life events). To filter out advertisements or posts with other users' life events, we labeled the users who experienced their own life event up to 100 users by alphabetical order of blog accounts in each life event (300 users in total). To extract enough users' interests and actions, we restricted the selection to users who posted over 30 blogs. In total, we extracted 34,753 posts for "childbirth," 40,238 posts for "finding employment," and 30,605 posts for "marriage," which were posted by the users described above. Table 2 shows examples of blog posts with life event labels (the original texts were in Japanese). In this Table, label 1 means "experienced their own life event" and label 0 means "did not experience their own life event."

(2) **Verification of Labeling Reliability**
To verify the labeling reliability, we extracted 50 users who were labeled as "experienced their own life event" and 50 users who were labeled as "did not experience their own life event" for each life event. We selected 50 blogs that described the life event experience from the 50 labeled users who experienced their own life event. We also selected 50 blogs that contained the queries in Table 1 from the 50 labeled users who did not experience their own life event. Two annotators[4] labeled 100 blog posts in total. Interannotator agreement was 1.00 in Cohen's kappa for each life event. Accordingly, we concluded that our data were reliable.

4.2 Experimental Environment

To implement DTM, we used the Python library code Gensim[5]. To determine the number of topics, we estimated the topic independence. Specifically, we computed the dissimilarity of topics using Jensen–Shannon (JS) divergence, which was applied to all combinations of topics. We determined the optimal number of topics to maximize the mean of JS divergence by changing the number of topics from 10 to 100 with step size of 10. The JS divergence between the probability distributions P and Q is defined as follows, where M is the mean of P and Q ($M(i) = \frac{P(i)+Q(i)}{2}$).

$$JSD(P \parallel Q) = \frac{1}{2}\Big(\sum_i P(i) \log \frac{P(i)}{M(i)} + \sum_i Q(i) \log \frac{Q(i)}{M(i)}\Big) \qquad (8)$$

Using this equation, we determined the number of topics as follows: $K = 100$ in "childbirth," $K = 60$ in "finding employment," and $K = 80$ in "marriage." We set the hyperparameter $\alpha = 0.01$ for each life event and set the TS as $TS = 25$, as mentioned previously. We restricted the chosen words to the grammatical categories noun, verb, and adjective only. We performed a morphological analysis of the data using MeCab[6], which is an open-source morphological analysis tool for word segmentation in Japanese. We also used mecab-ipadic-Neologd[7] as the morphological dictionary to expand the informal expressions.

We assume that many users mention the topics which reflect users' interests and actions. To evaluate the topics that were posted by a sufficient number of users in the time periods (25 months), the topics were restricted as follows:

(1) more than u users posted the topic;
(2) "user who posted topic" was determined by whether the number of blog posts by the user relevant to the topic was more than d;
(3) "blog post relevant to the topic" was determined by whether the topic probability in the blog post as a percentage was more than r.

In this experiment, we set $u = 30$, $d = 3$, and $r = 30$.

4.3 Results

Users' Interests and Actions in the Selected Topics: We checked topic transitions and the words in two selected topics for each life event to verify that they reflect users' interests and actions. Tables 3 and 4 show the five words which appeared many times chronologically for two topics that were extracted for the life events "childbirth" and "finding employment"[8]. The topic name is defined by the author to refer to the words in the topics. Figures 2 and 3 each show two topic transitions. In these figures, ±0 indicates the month when the life event occurred[9].

For this experiment, we selected the topics "status of unborn baby" and "child development" for the "childbirth" life event. The "status of unborn baby" topic contained words about the unborn baby's weight or prenatal care. We also selected the topics "working life" and "job hunting and study" for the "finding employment" life event. The "working life" topic contained words about the user's instruction works or relationships. The "job hunting and study" topic contained words about job interviews or visiting a company. The "status of unborn baby" and "job hunting and study" topics were selected by both scores of the difference score of the mean and of the variance. The "child development" topic was selected by the difference of the variance score. The "working life" topic was selected by the difference of the mean score.

Consequently, we have confirmed that our method was effective for selecting the topics that reflect users' interests

[3] In Japan, initiation ceremonies for all new employees are held through the recruitment process.
[4] The annotators were male college students.
[5] https://radimrehurek.com/gensim/
[6] http://taku910.github.io/mecab/
[7] https://github.com/neologd/mecab-ipadic-neologd
[8] In the "marriage" life event, we selected topics such as "wedding ceremony" or "cooking"; however, for want of space, the results for "marriage" are omitted.
[9] To clarify the probability change, the vertical scales are different for each topic

Table 3: Five typical words from the "status of unborn baby" and "child development" topics

"Status of unborn baby"	"Child development"
weight [taijuu]	month [kagetsu]
baby belly [onaka]	development [seityou]
prenatal care [kenshin]	in recent days [saikin]
week [syuukan]	fast [hayai]
pregnant [ninshin]	gain [hue]

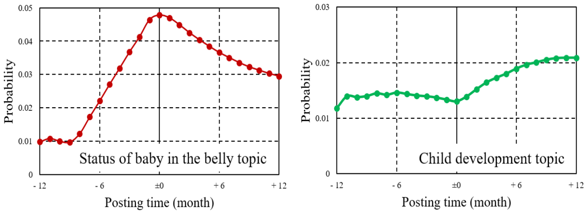

Figure 2: Topic transitions for "status of unborn baby" and "child development" topics

Table 4: Five typical words from the "working life" and "job hunting and study" topics

"Working life"	"Job hunting and study"
job [shigoto]	job hunting [syuukatsu]
office [kaisya]	study [benkyou]
induction course [kensyuu]	interview [mensetsu]
senior colleague [senpai]	exam [shiken]
colleagues [douki]	company [kigyou]

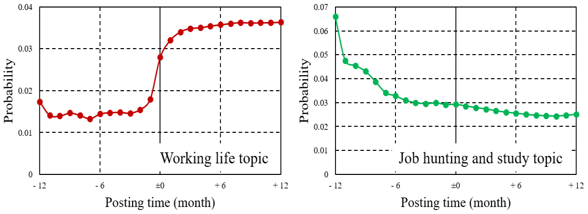

Figure 3: Topic transitions in "working life" and "job hunting and study" topics

or actions in multiple life events: Many new mothers become interested in child development after the "childbirth" life event and senior-year college students become interested in job hunting before "finding employment." Next, we clarify their information needs by investigating the evolution of words in the topics with relevant blog posts.

Evolution of users' information needs: We selected the two words of which probabilities changed drastically after "childbirth" ("baby food" [rinyuusyoku] and "child crawling" [haihai]) in the "child development" topic to verify their evolution. Figure 4 shows the time evolution of word probabilities for the words "baby food" [rinyuusyoku] and "child crawling" [haihai]. In Figure 4, "baby food" increased and first appeared as top 20 words in the "child development" topic two months after "childbirth." "Child crawling" increased and first appeared as top 20 words in the topic eight months after "childbirth." Table 5 shows blog posts relevant to the "child development" topic that contained the word "baby food" [rinyuusyoku] (the original texts were in Japanese). In Table 5, we identify mothers' information needs such as "the relationship between an allergy and the elapsed months to give baby food" or "what types of foods should mothers give their babies."

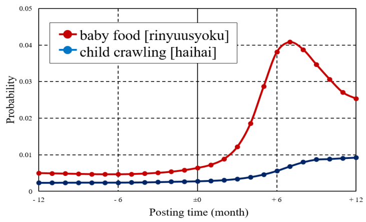

Figure 4: Time evolution of word probabilities for "baby food" [rinyuusyoku] and "child crawling" [haihai]

Table 5: Blog posts about "baby food"

Posting time	Blog post
5 months after	I worry about when to start giving **baby food**. I heard that offering **baby food** too early could lead to allergies for babies.
5 months after	My baby tried to start eating rice porridge as **baby food**.

5 CONCLUSION

In this paper, we have proposed a method to clarify the evolution of users' information needs based on when they experienced life events. In previous research, the analysts must specify the interests or actions about the life event in advance. In this paper, we proposed a method for extracting topics relevant to the life event by computing the difference in topic probabilities before and after the life event. Specifically, we used two difference scores to select the life event topics: the difference score of the mean (or variance) of topic probabilities. We confirmed that the two scores were effective in finding the topics relevant to the life event in multiple life events. In the "child development" topic, we also found chronological evolution of words such as "baby food" could characterize mothers' information needs.

ACKNOWLEDGMENTS

This work was partially supported by a JSPS Grant-in-Aid for Scientific Research (B) (#16H02913) and ACM SIGIR Student Travel Grant.

REFERENCES

[1] David M. Blei and John D. Lafferty. 2006. Dynamic Topic Models. In *Proc. of the 23rd Int'l Conf. on Machine Learning (ICML 2006)*. Pittsburgh, PA, USA, 113–120.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

[3] Moira Burke and Robert Kraut. 2013. Using Facebook after Losing a Job: Differential Benefits of Strong and Weak Ties. In *Proc. of the 2013 Conf. on Computer Supported Cooperative Work (CSCW 2013)*. San Antonio, TX, USA, 1419–1430.

[4] Munmun De Choudhury and Micheal Massimi. 2015. "She said yes!" Liminality and Engagement Announcements on Twitter. In *Proc. of iConference 2015*. Newport Beach, CA, USA, 1–13.

[5] Nattiya Kanhabua, Tu Ngoc Nguyen, and Wolfgang Nejdl. 2015. Learning to Detect Event-Related Queries for Web Search. In *Proc. of the 24th Int'l Conf. on World Wide Web (WWW 2015)*. Florence, Italy, 1339–1344.

[6] Hao Zhang, Gunhee Kim, and Eric P. Xing. 2015. Dynamic Topic Modeling for Monitoring Market Competition from Online Text and Image Data. In *Proc. of the 21th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD 2015)*. Sydney, Australia, 1425–1434.