Detecting Changes in Content and Posting Time Distributions in Social Media

Kazumi Saito*, Kouzou Ohara†, Masahiro Kimura‡ and Hiroshi Motoda§
*University of Shizuoka, Shizuoka 422-8526, Japan, Email: k-saito@u-shizuoka-ken.ac.jp
†Aoyama Gakuin University, Kanagawa 252-5258, Japan, Email: ohara@it.aoyama.ac.jp
‡Ryukoku University, Otsu 520-2194, Japan, Email: kimura@rins.ryukoku.ac.jp
§Osaka University, Osaka 567-0047, Japan, Email: motoda@ar.sanken.osaka-u.ac.jp

***Abstract*—We address a problem of detecting changes in information posted to social media taking both content and posting time distributions into account. To this end, we introduce a generative model consisting of two components, one for a content distribution and the other for a timing distribution, approximating the shape of the parameter change by a series of step functions. We then propose an efficient algorithm to detect change points by maximizing the likelihood of generating the observed sequence data, which has time complexity almost proportional to the length of observed sequence (possible change points). We experimentally evaluate the method on synthetic data streams and demonstrate the importance of considering both distributions to improve the accuracy. We, further, apply our method to real scoring stream data extracted from a Japanese word-of-mouth communication site for cosmetics and show that it can detect change points and the detected parameter change patterns are interpretable through an in-depth investigation of actual reviews.**

I. Introduction

It has become a part of our daily life that we post diverse information, *e.g.*, news, ideas, opinions, reviews, etc. directly to so-called social media on the Web, *e.g.*, weblogs, social blogs, wikis, etc. Once the information is posted on social media, it can be rapidly and widely spread through social networks on the Web and can be shared by a large number of people. Thus, it has a large influence on our decision making, and it is becoming pressingly important that we are able to efficiently analyze this huge amount of information or opinions.

There has been a large number of studies on social media from various angles. One typical direction is modeling how information propagates through a social network [1]–[5]. This would be useful for viral marketing to solve such a problem of influence maximization [6]–[9] in which the task is to identify a limited number of nodes which together maximize the information spread. Another direction is analyzing opinion formation, where the focus is to model how people are influenced by their neighbors in making decision [10]–[14]. At a more fundamental level, sentiment analysis tries to classify contents on social media for a certain topic [15]–[17], which could allow companies to know how their products are evaluated by consumers.

On one hand we are interested in knowing what is happening now and how it develops in the future, but on the other we are also interested in knowing what happened in the past and how this caused the change in the distribution of the information. For example, if the rank of an item dropped on a certain review site, the manufacturer that produces it would analyze the site to know when and how the consumers' evaluation to the item changed. Such changes may involve changes in the number of reviews posted in a certain period, *i.e.,* changes in the posting interval and frequency, in which case existing burst detection techniques [18]–[20] would be applicable. However, if no change in the time interval is involved, we would not be able to use these techniques because they do not focus on the change in the content. In other words, they intend to detect a burst for a single topic, and do not directly deal with multiple topics and the change of their distribution.

We note that this kind of change detection is substantially different from the typical anomaly detection or change point detection widely studied in machine learning, whose techniques are closer to those used in novelty detection or outlier detection [21]. For instance, statistical techniques used in anomaly detection fit a statistical model (usually for normal behavior) to the given data. Unseen data that have a low probability to be generated from the learned model are judged as anomalies. On the other hand, we are interested in identifying a model with time-varying parameters. A conventional approach for this direction includes studies of regime-switching models in economics (e.g., [22]), but they heavily rely on the Gaussian assumption. In decision support, a number of methods have been developed for detecting and excluding unfair ratings in online reputation systems [23], but they are clearly categorized as anomaly detection.

The problem we tackle in this paper is detecting changes not only in time intervals between posts, but also in their content distribution. As one typical example, we consider scoring streams observed on a review site where items are evaluated by multiple categorical scores. To handle this problem, we introduce a generative model that has parameters for such stream data. In reality the change in the distribution of content may involve the change in the time interval, or may not. There is no reason to believe that both changes must occur at exactly the same time. This leads us to model these changes as a combination of two distinct models, one for content distribution (the content model) and the other for time intervals (the timing model). This type of generative model can cover a wide range of phenomena such as information diffusion, opinion formation, and document generation by adopting an appropriate distribution for each component. In our case, we adopt an M-categorical distribution for the content model

since items are evaluated by M categorical scores, and an exponential distribution for the timing model as a typical time delay distribution.

We devise an efficient algorithm that accurately detects the changes in the parameter values from the observed stream data. More precisely, we approximate parameter changes in each model by a series of step functions and propose an optimization algorithm that maximizes the likelihood ratio (the ratio of the likelihood of observing the data assuming the parameter changes to the likelihood of observing the data assuming that there is no change in any parameters). The algorithm is a combination of a greedy search that recursively splits stream data and a local search that starts from the the greedy search results and seeks for a better solution. The time complexity is almost proportional to the length of observed data points (candidates of possible change points). We apply the proposed method to synthetic scoring data streams, and show that our method outperforms, in terms of accuracy, the methods that consider only either one of the content and the timing distributions. In addition, we apply it to the real scoring stream data from a Japanese word-of-mouth communication site for cosmetics and show that the detected change points are interpretable through an in-depth investigation of actual review content.

II. Generative Model Settings

We consider a generative model $p(\mathrm{x}, \Delta t \mid t; \mathbf{\Phi})$ for stream data for a given time t. Here $p(\mathrm{x}, \Delta t \mid t; \mathbf{\Phi})$ stands for the probability that an event of content x occurs at time t and is updated at $t + \Delta t$, where $\mathbf{\Phi}$ is a parameter vector of our model. Here, we assume that our generative model is factorized as follows:

$$p(\mathrm{x}, \Delta t \mid t; \mathbf{\Phi}) = p_{\mathrm{X}}(\mathrm{x} \mid t; \boldsymbol{\theta}) p_{\Delta t}(\Delta t \mid t; \boldsymbol{\rho}), \;\; \mathbf{\Phi} = (\boldsymbol{\theta}, \boldsymbol{\rho}), \quad (1)$$

i,e, the probabilities, $p_{\mathrm{X}}(\mathrm{x} \mid t; \boldsymbol{\theta})$ and $p_{\Delta t}(\Delta t \mid t; \boldsymbol{\rho})$ are conditionally independent under the fixed time t. Note that Equation (1) does not mean that the probabilities, $p_{\mathrm{X}}(\mathrm{x}; \boldsymbol{\theta})$ and $p_{\Delta t}(\Delta t; \boldsymbol{\rho})$ are mutually independent. This type of generative model covers a wide range of phenomena such as information diffusion, opinion formation, and document generation. More specifically, as our generative model for content x, we can typically consider a Bernoulli distribution for information diffusion, a categorical distribution for opinion formation, and a multinomial distribution for document generation.

In this paper, we focus on a problem of scoring stream data over a review site where items are evaluated by scores with M-category. Thus, we consider an M-categorical distribution as our content model which is defined by

$$p_{\mathrm{X}}(\mathrm{x}|t; \boldsymbol{\theta}) = \prod_{m \in \mathcal{M}} \theta_m^{x_m}, \mathrm{x} = (x_1, \cdots, x_M), \boldsymbol{\theta} = (\theta_1, \cdots, \theta_M),$$

where $x_m \in \{0, 1\}$, $\theta_m \in (0, 1)$, $x_1 + \cdots + x_M = \theta_1 + \cdots + \theta_M = 1$, and $\mathcal{M} = \{1, \cdots, M\}$. Note that the functional form of this distribution can easily be replaced by other type of distribution. As for our timing model, we employ an exponential distribution with parameter r, which is defined by

$$p_{\Delta t}(\Delta t \mid t; \boldsymbol{\rho}) \;=\; p_{\Delta t}(\Delta t \mid t; r) = r \exp(-r \Delta t).$$

Again note that we can use other distributions such as power-law and Weibull according to the nature of data streams.

III. Problem Settings

We formally define the change point detection problem. As mentioned in Section I, we assume that some unknown change took place in the course of the data streaming process and what we observe is a sequence of stream data in which the change is encapsulated. Change can be in content or in speed of stream or in both. Thus, our goal is to detect each change point and how long the change persisted from there. Let's assume that we observe a set of pairs of content vectors and their time stamps, *i.e.,* $\mathcal{D} = \{(\mathrm{x}_0, t_0), (\mathrm{x}_1, t_1), \cdots, (\mathrm{x}_N, t_N)\}$. Let the time of the j-th change point be T_j ($t_0 < T_j < t_N$). The parameter vector that the distribution follows switches from $(r_j, \boldsymbol{\theta}_j)$ to $(r_{j+1}, \boldsymbol{\theta}_{j+1})$ at the j-th change point T_j. Namely, we are assuming a series of step functions as a shape of parameter vector changes. Let the set comprising J change points be $\mathcal{S}_J = \{T_1, \cdots, T_J\}$, and we set $T_0 = t_0$ and $T_{J+1} = t_N$ for the sake of convenience ($T_{j-1} < T_j$). Let the division of $\mathcal{D}$ by $\mathcal{S}_J$ be $\mathcal{D}_j = \{n; T_{j-1} < t_n \leq T_j\}$, *i.e.,* $\mathcal{N} = \{0, 1, \cdots, N\} = \{0\} \cup \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_{J+1}$, and $|\mathcal{D}_j|$ represents the number of observed points in $(T_{j-1}, T_j]$. Here, we request that $|\mathcal{D}_j| \neq 0$ for any $j \in \mathcal{J} = \{1, \cdots, J+1\}$ and there exists at least one t_n and $t_n \in \mathcal{D}_j$ is satisfied. The problem of detecting change points is equivalent to a problem of finding a subset $\mathcal{S}_J \subset \mathcal{T}$ where $\mathcal{T}$ is a set of the observed time points, *i.e.,* $\mathcal{T} = \{t_0, t_1, \cdots, t_N\}$.

The log-likelihood for the $\mathcal{D}$, given a set of change points $\mathcal{S}_J$, is calculated, by defining the parameter vectors, $\mathbf{\Theta}_{J+1} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{J+1})$ and $\mathrm{r}_{J+1} = (r_1, \cdots, r_{J+1})$, as follows.

$$\begin{aligned} &L(\mathcal{D}; \mathbf{\Theta}_{J+1}, \mathrm{r}_{J+1}, \mathcal{S}_J) \\ &\quad = \sum_{j \in \mathcal{J}} \Big(\sum_{m \in \mathcal{M}} X_{j,m} \log \theta_{j,m} + |\mathcal{D}_j| \log r_j - r_j \Delta T_j \Big), \end{aligned} \quad (2)$$

where $X_{j,m} = \sum_{n \in \mathcal{D}_j} x_{n,m}$ and $\Delta T_j = T_j - T_{j-1}$. Thus, the maximum likelihood estimators of Equation (2) is given by $\hat{\theta}_{j,m} = X_{j,m}/|\mathcal{D}_j|$ and $\hat{r}_j = |\mathcal{D}_j|/\Delta T_j$, for $j = 1, \cdots, J+1$, and $m = 1, \cdots, M$. Substituting these estimators to Equation (2) leads to

$$\begin{aligned} &L(\mathcal{D}; \hat{\mathbf{\Theta}}_{J+1}, \hat{\mathbf{r}}_{J+1}, \mathcal{S}_J) \\ &\quad = \sum_{j \in \mathcal{J}} \Big(\sum_{m \in \mathcal{M}} X_{j,m} \log X_{j,m} - |\mathcal{D}_j| \log \Delta T_j \Big) - N. \end{aligned} \quad (3)$$

Therefore, the change point detection problem is reduced to the problem of finding the change point set $\mathcal{S}_J$ that maximizes Equation (3). However, Equation (3) alone does not allow us to directly evaluate the effect of introducing $\mathcal{S}_j$. It is important to evaluate how the log-likelihood improves over the one obtained without considering the parameter changes. We, thus, reformulate the problem as the maximization problem of log-likelihood ratio. If we do not assume any changes, *i.e.,* $\mathcal{S}_0 = \emptyset$, Equation (3) is reduced to

$$L(\mathcal{D}; \hat{\mathbf{\Theta}}_1, \hat{r}_1, \mathcal{S}_0) \;\; = \;\; \sum_{m \in \mathcal{M}} X_m \log X_m - N \log \Delta T - N,$$

where $X_m = \sum_{n \in \mathcal{N}} x_{n,m}$ and $\Delta T = T_{J+1} - T_0 = t_N - t_0$. Thus, the log-likelihood ratio of the two cases, one with J change points and the other with no change points is given by

$$\begin{aligned} LR(\mathcal{S}_J) &= L(\mathcal{D}; \hat{\mathbf{\Theta}}_{J+1}, \hat{\mathbf{r}}_{J+1}, \mathcal{S}_J) - L(\mathcal{D}; \hat{\mathbf{\Theta}}_1, \hat{r}_1, \mathcal{S}_0) \\ &= \sum_{j \in \mathcal{J}} \Big(\sum_{m \in \mathcal{M}} X_{j,m} \log \frac{X_{j,m}}{X_m} - |\mathcal{D}_j| \log \frac{\Delta T_j}{\Delta T} \Big). \end{aligned} \quad (4)$$

In summary we consider the problem of finding the set of change points $\mathcal{S}_J$ that maximizes $LR(\mathcal{S}_J)$ defined above.

On the other hand, we can derive two specialized change detection problems from the above arguments, *i.e.*, detection only from the content distribution and detection only from the timing distribution, defined by the following objective functions, $LR_{\mathrm{X}}(\mathcal{S}_J)$ and $LR_{\Delta t}(\mathcal{S}_J)$.

$$LR_{\mathrm{X}}(\mathcal{S}_J) = \sum_{j\in\mathcal{J}}\Big(\sum_{m\in\mathcal{M}} X_{j,m}\log\frac{X_{j,m}}{X_m} - |\mathcal{D}_j|\log\frac{|\mathcal{D}_j|}{N}\Big), \quad (5)$$

$$LR_{\Delta t}(\mathcal{S}_J) = -\sum_{j\in\mathcal{J}}\Big(|\mathcal{D}_j|\log\frac{\Delta T_j}{\Delta T} + |\mathcal{D}_j|\log\frac{|\mathcal{D}_j|}{N}\Big). \quad (6)$$

In our experiments, we empirically evaluate how detection performance improves by considering both the content and the timing distributions simultaneously, in comparison to those by considering only either one of these distributions.

IV. Change Point Detection Method

For a given number of change points J, we search for J time points that are most likely to be the change points from a sequence of N observation points. In what follows, we explain our detection method that is a combination of a greedy search (A1) and a local search (A2). The algorithm is given below.

A1. Produce $\mathcal{S}_J$ from input data, J and $\mathcal{D}$, by a greedy search.
A2. Improve $\mathcal{S}_J$ and output the final result by a local search.

We first describe the procedure of A1. This is a progressive binary splitting without backtracking. We fix the already selected set of $(j-1)$ change points $\mathcal{S}_{j-1}$ and search for the optimal j-th change point T_j and add it to $\mathcal{S}_{j-1}$. We repeat this procedure from $j=1$ to J. The algorithm is given below.

A1-1. Initialize $j=1$, $\mathcal{S}_0=\emptyset$.
A1-2. Search for $T_j = \arg\max_{t_n\in\mathcal{T}}\{LR(\mathcal{S}_{j-1}\cup\{t_n\})\}$.
A1-3. Update $\mathcal{S}_j = \mathcal{S}_{j-1}\cup\{T_j\}$.
A1-4. If $j=J$, output $\mathcal{S}_J$ and stop.
A1-5. $j=j+1$, and return to A1-2.

Here note that in A1-3 elements of the change point set $\mathcal{S}_j$ are reindexed to satisfy $T_{i-1}<T_i$ for $i=2,\cdots,j$. Clearly, the time complexity of this simple method is $O(NJ)$ which is fast. Thus, it is possible to obtain the result within an allowable computation time for a large N. However, since this is a greedy algorithm, it can be trapped easily to a poor local optimal.

Next, we describe the procedure of A2. We start with the solution $\mathcal{S}_J$ obtained by A1, pick up a change point T_j from the list, fix the rest $\mathcal{S}_J\setminus\{T_j\}$ and search for the better value T'_j of T_j, where $\cdot\setminus\cdot$ represents set difference. We repeat this from $j=1$ to J. If no replacement is possible for all j ($j=1,\cdots,J$), *i.e.*, $T'_j=T_j$ for all j, no better solution is expected and the iteration stops. The algorithm is given below.

A2-1. Initialize $j=1$, $k=0$.
A2-2. Search for $T'_j = \arg\max_{t_n\in\mathcal{T}}\{LR(\mathcal{S}_J\setminus\{T_j\}\cup\{t_n\})\}$.
A2-3. If $T'_j=T_j$, set $k=k+1$, otherwise set $k=0$, and update $\mathcal{S}_J=\mathcal{S}_J\setminus\{T_j\}\cup\{T'_j\}$.
A2-4. If $k=J$, output $\mathcal{S}_J$ and stop.
A2-5. If $j=J$, set $j=1$, otherwise set $j=j+1$, and return to A2-2.

It is evident that the proposed method requires computation time several times larger than that of the greedy method. In our previous studies using a similar strategy, the increase of the computation time was not that large, but the solution quality was substantially improved from the greedy solutions.

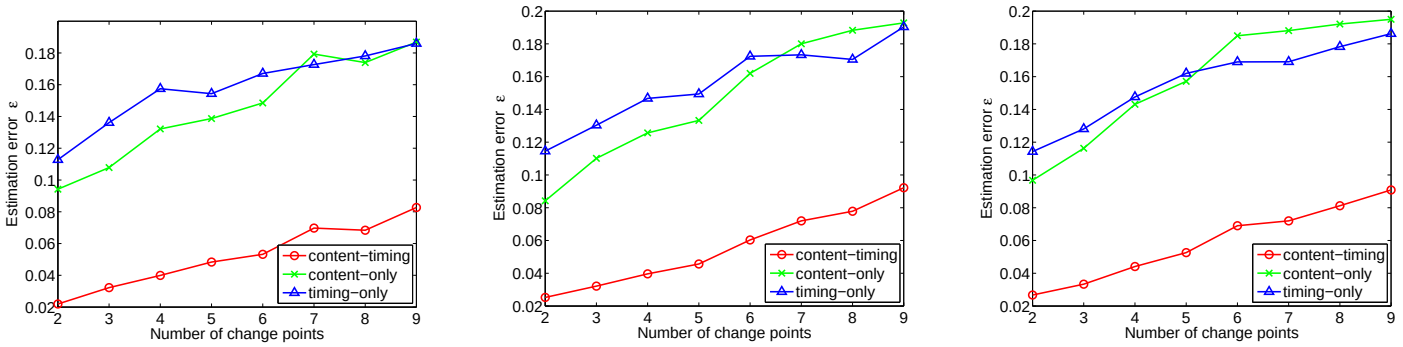
V. Experimental Evaluation

We experimentally evaluated the accuracy of the change point detection of our method using synthetic scoring stream data and confirmed that the algorithm works satisfactorily. We further applied our method to the real scoring data and show that the change points detected by the method are indeed interpretable through the analysis of actual review content.

A. Evaluation by Synthetic Data

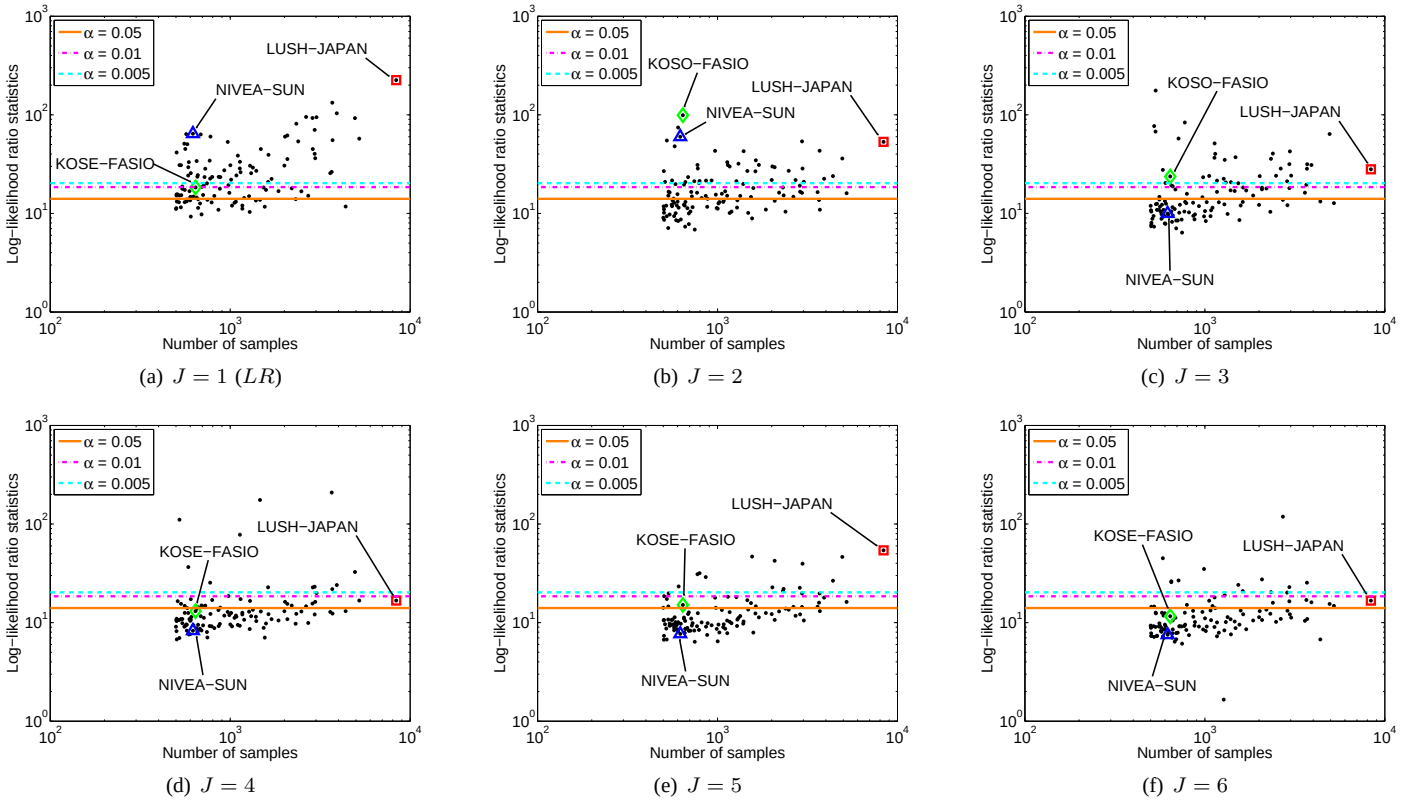
Using synthetic scoring stream data, we examined the accuracy of the proposed method by comparing it with two other methods that adopt objective functions given by Equations (5) and (6), respectively. Namely, the former detects change points only from the content distribution, while the latter only from the timing distribution. According to the generative models defined in Section II, we generated a scoring stream $\mathcal{D}$ for an item, consisting of pairs (x,t) of content vector x and its time stamp $t\in[0,1000]$ such that $|\mathcal{D}|=N$. Since we assume items are evaluated by M-categorical scores, x is an M-dimensional vector in which only the m-th element is 1 and the remaining ones are 0 at time t if the corresponding item is given the m-th score at time t. We embedded J change points in each stream such that each interval $(T_{j-1},T_j]$ becomes almost the same, where $j\in\{1,\cdots,J+1\}$, $T_0=t_0$ (the first observation time), and $T_{J+1}=t_N$ (the final observation time). Here, we assumed both content and timing distributions change at these change points in a synchronized manner so that we can fairly compare our proposed method with the other two methods. Thus, for each of $J+1$ periods, we determined parameters of both distributions according to their typical prior distributions, and then generated samples based on them. We adopted Dirichlet distribution with all concentration parameters set to 2 as the prior of the categorical distribution, and Gamma distribution with both shape and scale parameters set to 1 as the prior of the exponential distribution. The latter makes the expected number of N about 1,000. Varying the number of categorical scores M in the range of $M=5,7,10$ by reference to typical review sites on the Web, and the number of embedded change points J in the range of $J=2,\cdots,9$, we generated $K=1,000$ streams for each combination of M and J.

Figures 1(a) to 1(c) illustrate the results for $M=5,7,9$, respectively, in which the horizontal axis means the number of change points embedded, J, while the vertical axis means the average relative erorr with respect to each change point for the corresponding M and J, which is defined as $\mathcal{E}=K^{-1}\sum_{k=1}^{K}J^{-1}\sum_{j=1}^{J}|t_{k,j}-\hat{t}_{k,j}|/t_{k,j}$, where $t_{k,j}$ and $\hat{t}_{k,j}$ denote the true time stamp of the j-th change point of the k-th stream and its estimation, respectively. It is clearly found that considering both the content and the timing distributions significantly improves the accuracy compared with considering only one of them, even if distribution changes are made in a synchronized manner. Especially, note that the estimation error of the proposed method is much less than the half of the estimation error of the one considering only one distribution



(a) $M = 5$ (b) $M = 7$ (c) $M = 9$

Fig. 1. Estimation errors of the three methods for $M = 5$, 7, and 9.



(a) $J = 1$ (LR) (b) $J = 2$ (c) $J = 3$

(d) $J = 4$ (e) $J = 5$ (f) $J = 6$

Fig. 2. Relation between log-likelihood ratio statistic and the number of samples. Note that only the results for $J = 1$ plot the log-likelihood ratio defined in Equation (4). The results for $J \geq 2$ plot the log-likelihood ratio between J and $J - 1$ to see the diminishing effect of J.

in every case, which implies that the effect of considering both distributions is more than additive of the two, each considering only one distribution. In addition, we can say that the accuracy of every method is insensitive to the number of categorical scores M, while the estimation errors linearly increase as the number of change points J increases although we evaluated the average relative error with respect to each change point. This means that this detection problem becomes substantially more difficult as the number of change points increases.

B. Change Point Detection Results for Real Dataset

1) Dataset: We collected real scoring stream data from "@cosme" [1], which is a Japanese word-of-mouth communication website for cosmetics. In @cosme, a user can post a review and give a score of each brand (one from 1 to 7). When one user registers another user as his/her favorite user, a "fan-link" is created between them. We traced up to ten steps in the fan-link network from a randomly chosen user in December 2009, and collected a set of (b, m, t, v)'s, where (b, m, t, v) means that user v gave m points to brand b at time t. In the collected data, the number of brands was 7,139, the number of users was 45,024, and the number of reviews posted was 331,084. For each brand b, we constructed a scoring data stream $\mathcal{D}$ consisting of pairs (m, t). In particular, we focused on these brands such that the number of observations $N = |\mathcal{D}|$ was greater than 500. Then, the number of brands was 120. We refer to this dataset extracted through the fan-link network as the @cosme dataset.

[1] http://www.cosme.net/

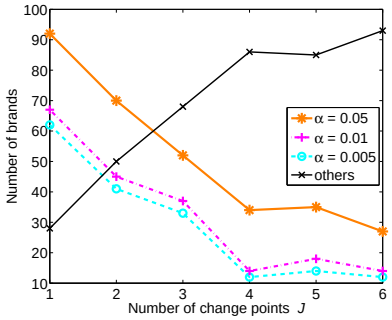


Fig. 3. Relation between the number of brands that pass the significant test and J for three different significant levels

2) Results: First, we show plots of the log-likelihood ratio statistic for each brand as a function of sample size N for the number of change points $J = 1$ to 6 in Figures 2(a) to 2(f). The objective function used was Equation (4) throughout these experiments, but the way the log-likelihood is plotted in these figures are slightly changed except for Figure 2(a). We plotted the log-likelihood ratio between J and $J-1$ for $J = 2$ to 6. Since we do not know the most appropriate value for J, we thought that it is best to evaluate the statistical significance in going from $J-1$ to J. Thus, in these figures three horizontal lines, each corresponding to a different significance level $\alpha = 0.05, 0.01, 0.005$ are also shown. We used χ^2 test with 7 degrees of freedom (6 score parameters and 1 time decay constant). From these figures, we can see the diminishing effect of J over the log-likelihood ratio in going from $J-1$ to J.[2] This means that from the statistical point of view, it does not make much sense to use a very large J to discuss the global feature of the parameter change. Figure 3 shows the relation of the number of the brands that pass the significant test with the number of change points J for three different significance levels. The black line is the complement of the orange line, *i.e.*, the number of brands that do not pass the statistical test for $\alpha = 0.05$. It is clear that for most of the brands, $J = 4$ is enough to capture the change patterns.

Among the 120 brands we examined three brands which are shown in the above figures using $J = 2$, the minimum value to detect an abnormal period in which either or both of the content and the timing distributions are meaningfully different from those in the other periods if such period exists. The three brands chosen are "LUSH-JAPAN", "KOSE-FASIO" and "NIVEA-SUN". "LUSH-JAPAN" has the largest log-likelihood ratio (as measured by Equation (4) for $J = 2$), which implies that the detected change pattern is more distinctive than change patterns for the others. "KOSE-FASIO" and "NIVEA-SUN" have relatively high log-likelihood ratios although their sample sizes are rather small, which is against the standard tendency that there is a positive correlation between the sample size and the log-likelihood ratio.

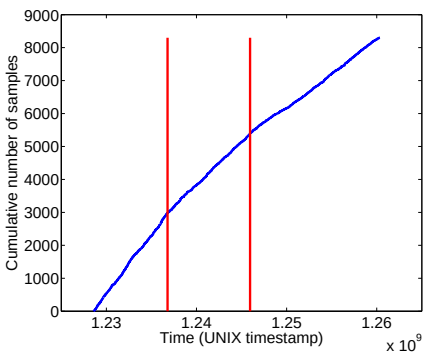
First, Figure 4 presents the results for the brand "LUSH-JAPAN" from three distinct angles. Figure 4(a) shows the cumulative number of reviews (scores) as a function of observed time, in which red vertical lines indicate the detected change points, while Figure 4(b) illustrates the detected change pattern of the timing parameter r of the exponential distribution, in which the value of r is normalized such that $r = 1$ corresponds to a delay of one day, meaning $r = 0.1$ corresponds to delay of 10 days. In fact, the differences between before and after the detected change points are negligible in Figure 4(a) and very small in Figure 4(b). It is read that the number of reviews for this brand is gradually decreasing, but it is unlikely that we find clear change points only from this timing distribution. On the other hand, in Figure 4(c) that depicts the change in the distribution of each score in the three periods made by partitioning the whole period using the two change points detected as the dividing points, we can clearly find distinctive changes in the distributions for some scores. The number of scores larger than 4 decreases in later period, while the number of score 3 increases. This result shows the importance of considering both the timing and the content distributions. We should emphasize that the algorithm does not assume the synchronous changes in both distributions and it can handle both changes separately. The results indicate that the two change patterns in the timing distribution and the content distribution coincide with each other because the decay of the attractiveness of the brand manifested itself in both distributions in the same way.

Second, we illustrate the results for the brand "KOSE-FASIO" in Figure 5 in the same way as in Figure 4. We can find that, during the second period, the number of reviews sharply increased in Figures 5(a) and 5(b), and similarly the number of relatively low scores also intensively increased in Figure 5(c). By examining the actual reviews more in depth, we found that this is due to bursty posts by a specific user who tends to give relatively low scores to this brand. Actually, she never gave this brand scores larger than 4, which seems to be an intentional posting. Thus, we can say that our method could be applicable to detect such intentional posts aiming at affecting the customer evaluation to a specific item.

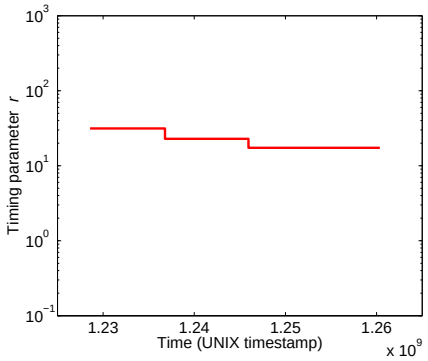
Third, we show the results for the brand "NIVEA-SUN" in Figure 6 as well. From Figures 6(a) and 6(b), the relatively large number of reviews can be observed during the second period. We found that this is a brand for sunscreen cosmetics. The first change point is "2009/4/7 15:46:49" and the second one is "2009/9/14 18:24:47". Namely, this period is the season when the ultraviolet rays are strong in Japan. Thus, during this period, the number of users who buy products of this brand increases. We note that in Figure 6(c) the number of high scores larger than 4 drops in the second period. This is mainly attributed to the two cosmetics both of which are sunscreen and whose average scores in the period are relatively low compared to those in the other periods. The number of negative reviews to them increased in the second period in response to the increase in the number of users who tried these cosmetics. The major claims are that both dehydrate the skin and one of them is hard to wash off. This case is a typical example that two distributions change simultaneously and that this kind of analysis contributes to improving the quality of products.

The above results are all for $J = 2$. As Figure 3 indicates that $J = 4$ is enough (we may be able to say $J = 3$ is enough), we believe that the above results can indeed capture the major characteristics of change patterns for these three brands. Just to make sure that this is true, we did an additional experiment for "LUSH-JAPAN" by increasing J to 6. Figure 7 shows the

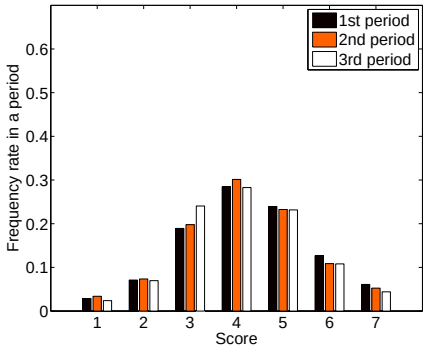
[2]This does not mean that the submodularity holds for Equation (4).



(a) Detected change points

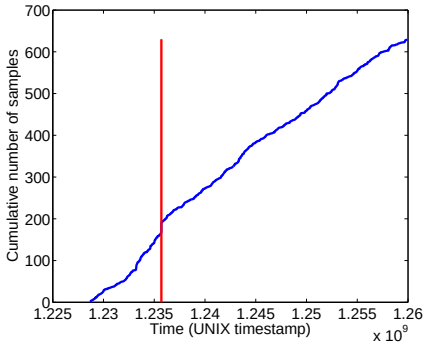


(b) Change of timing parameter r

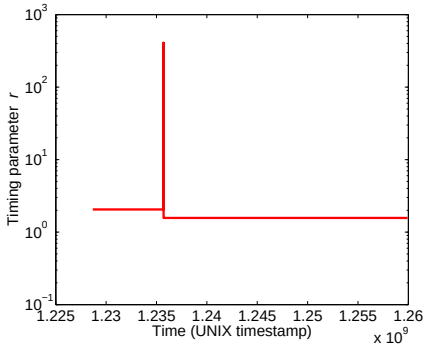


(c) Change of score distribution

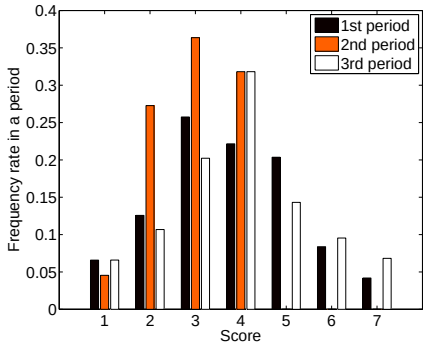
Fig. 4. Results for the brand "LUSH-JAPAN".



(a) Detected change points

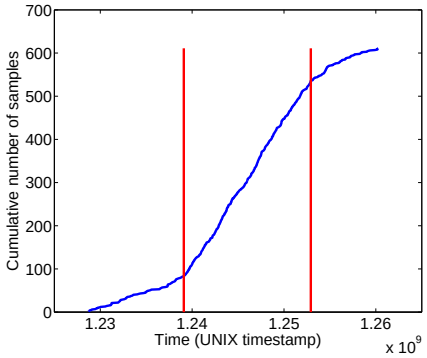


(b) Change of timing parameter r

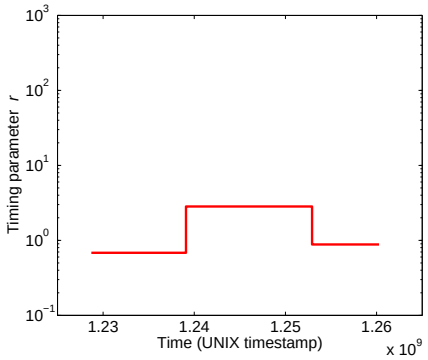


(c) Change of score distribution

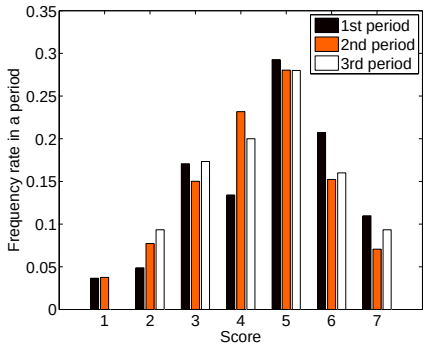
Fig. 5. Results for the brand "KOSE-FASIO".



(a) Detected change points



(b) Change of timing parameter r



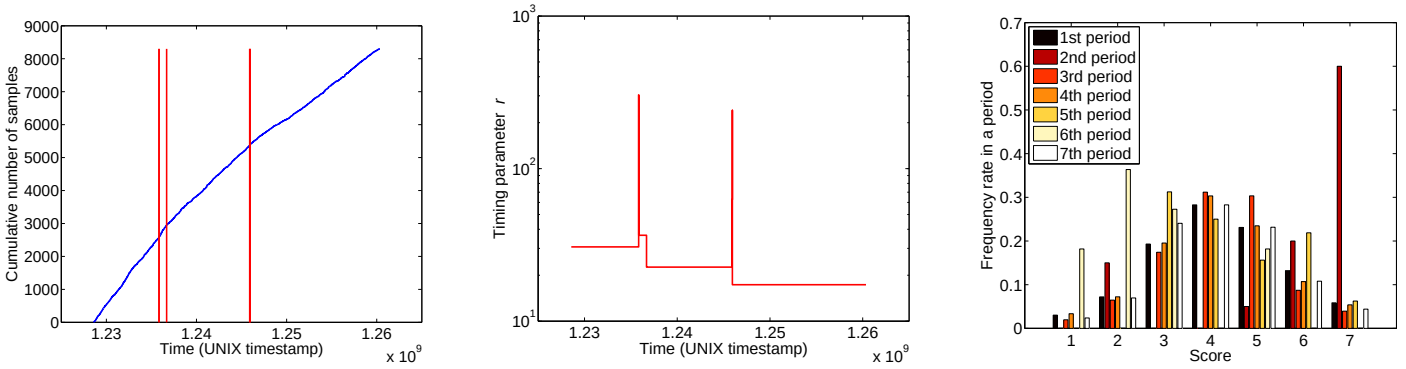
(c) Change of score distribution

Fig. 6. Results for the brand "NIVEA-SUN".

results. We see that in Figure 7(a) the first and the second change points and the fourth, the fifth and the sixth change points are indistinguishable, and there seem to be only three change points. From Figure 7(b) we can see that there should be three peaks but the last two peaks are indistinguishable. Actually the second peak is much smaller than the third one. We can say that the general trend of the timing parameters is gradually decreasing except for the three peaks and the third short period. This confirms our belief. Further, by increasing J to a larger value, we are able to detect abnormal peaks. The first peak corresponds to the second period in Figure 7(c) and is found to be caused by the explosive increase of the number of reviews with the highest score 7 from Figure 7(c), while the third one which corresponds to the sixth period in Figure 7(c) is due to a large number of reviews with low scores posted during this period. By checking the raw reviews, we confirmed that such biased reviews were posted by a specific user in each case.

VI. Conclusion

This paper addressed the problem of detecting the change points from observed time series data of information posted to social media, taking into consideration not only the change in the distribution of time interval between posts but also the change in the distribution of their content. We introduced a generative model that consists of two primitive components



(a) Detected change points (b) Change of timing parameter r (c) Change of score distribution

Fig. 7. Results for the brand "LUSH-JAPAN" in the case of $J = 6$.

that model the distributions of the content and the timing, respectively, and formally defined the problem of 1) detecting the change points and 2) finding the model parameter values such that the likelihood of generating the observed data stream is maximized. We then devised an efficient iterative algorithm to search for the change points, whose time complexity is almost linear to the number of data points. We empirically tested the proposed algorithm with the synthetic scoring data streams and demonstrated that considering both distributions is essential for accurate detection. Further, we applied the proposed method to the real scoring stream data from a Japanese word-of-mouth communication site for cosmetics and experimentally confirmed the versatility of the method through three typical cases: 1) the case that only the distribution of the content clearly changed, but the change in the distribution of the timing was not clear; 2) the case that there occurred an intentional bursty posts by a certain user; 3) the case where the distributions of both the content and the timing changed concurrently. In all of these cases, our method successfully detected the change patterns, which we were able to interpret through in-depth analysis of actual review articles. Our immediate future work is to experimentally compare the proposed method with existing burst detection methods.

ACKNOWLEDGMENT

This work was partly supported by Asian Office of Aerospace Research and Development, Air Force Office of Scientific Research under Grant No. AOARD-13-4042, and JSPS Grant-in-Aid for Young Scientists (B) (No. 23700181).

REFERENCES

[1] J. Yang and S. Counts, "Predicting the speed, scale, and range of information diffusion in twitter," in *Proc. of ICWSM 2010*, 2010.

[2] J. Yang and J. Leskovec, "Modeling information diffusion in implicit networks," in *Proc. of ICDM'10*, 2010, pp. 599–608.

[3] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: Quantifying influence on twitter," in *Proc. of WSDM'11*, 2011, pp. 65–74.

[4] P. Cui, F. Wang, S. Yang, and L. Sun, "Item-level social influence prediction with probabilistic hybrid factor matrix factorization," in *Proc. of AAAI 2011*, 2011, pp. 331–336.

[5] A. Guille and H. Hacid, "A predictive model for the temporal dynamics of information diffusion in online social networks," in *Proc. of WWW'12*, 2012, pp. 1145–1152.

[6] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proc. of KDD 2003*, 2003, pp. 137–146.

[7] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proc. of KDD 2007*, 2007, pp. 420–429.

[8] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proc. of KDD 2009*, 2009, pp. 199–208.

[9] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *Proc. of ICDM'10*, 2010, pp. 88–97.

[10] V. Sood and S. Redner, "Voter model on heterogeneous graphs," *Physical Review Letters*, vol. 94, p. 178701, 2005.

[11] P. Holme and M. E. J. Newman, "Nonequilibrium phase transition in the coevolution of networks and opinions," *Physical Review E*, vol. 74, p. 056108, 2006.

[12] E. Even-Dar and A. Shapria, "A note on maximizing the spread of influence in social networks," in *Proc. of WINE 2007*, 2007, pp. 281–286.

[13] C. Castellano, M. A. Munoz, and R. Pastor-Satorras, "Nonlinear q-voter model," *Physical Review E*, vol. 80, p. 041129, 2009.

[14] H. Yang, Z. Wu, C. Zhou, T. Zhou, and B. Wang, "Effects of social diversity on the emergence of global consensus in opinion dynamics," *Physical Review E*, vol. 80, p. 046108, 2009.

[15] P. Melville, W. Gryc, and R. D. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification," in *Proc. of KDD 2009*, 2009, pp. 1275–1284.

[16] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proc. of LREC'10*, 2010, pp. 1320–1326.

[17] K. Glass and R. Colbaugh, "Estimating sentiment orientation in social media for business informatics," in *AAAI Spring Symposium: AI for Business Agility*, 2011.

[18] J. Kleinberg, "Bursty and hierarchical structure in streams," in *Proc. of KDD 2002*, 2002, pp. 91–101.

[19] Y. Zhu and D. Shasha, "Efficient elastic burst detection in data streams," in *Proc. of KDD 2003*, 2003, pp. 336–345.

[20] A. Sun, D. Zeng, and H. Chen, "Burst detection from multiple data streams: A network-based approach," *IEEE Transactions on Systems, Man, & Cybernetics Society, Part C*, pp. 258–267, 2010.

[21] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, 2009.

[22] C. Kim, J. Piger, and R. Startz, "Estimation of markov regime-switching regression models with endogenous switching," *Journal of Econometrics*, vol. 143, pp. 263–273, 2008.

[23] A. Josang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," *Decision support systems*, vol. 43, pp. 618–644, 2007.