Chapter 19
Event Detection Based on Call Detail Records

Huiqi Zhang and Ram Dantu

Abstract In this paper we propose the model of the inhomogeneous Poisson for call frequency and inhomogeneous exponential distribution for call durations to detect events based on mobile phone call detail records. The maximum likelihood method is used to estimate the rate of frequency and call duration. This work is useful for enhancing homeland security, detecting unwanted calls (e.g., spam) and commercial purposes. For validation of our results, we used actual call logs of 100 users collected at MIT by the Reality Mining Project group for a period of 8 months. The experimental results show that our model achieves good performance with high accuracy.

19.1 Introduction

Analyzing patterns of human behavior [15] is an area of increasing interest in a number of different applications. The automatic detection of events by studying patterns of human behavior is one of them and has recently attracted attention. An event is something that happens at a given point in time and at a given place. We use *event* to refer to a large-scale activity that is unusual relative to normal patterns of behavior. To understand such data, we often care about both the patterns of typical behavior and detecting and extracting information from deviations from this behavior. Almost all previous approaches for event detection are based on text, website data and video data (see related work).

In this paper we propose and investigate the inhomogeneous Poisson and inhomogeneous exponential distribution model to detect events, and we illustrate how to learn such a model from data to both characterize normal behavior and detect anomalous events based on call detail records. There are no contents in call detail records that are the main difference from the text and website data and more difficult to detect events hidden in them. We can only use information such as the time of

H. Zhang (✉) · R. Dantu
Computer Science and Engineering, University of North Texas, Denton, TX 76201, USA
e-mail: hz0019@unt.edu

R. Dantu
e-mail: rdantu@unt.edu

L. Cao, P.S. Yu (eds.), *Behavior Computing*,
DOI 10.1007/978-1-4471-2969-1_19, © Springer-Verlag London 2012

initiation of calls, number of calls in a period of time, call duration, incoming calls, outgoing calls and location. The maximum likelihood estimation is used to estimate the rates and the thresholds of the number of calls and call duration. The experimental results show that our model achieves good performance with high accuracy.

In Sect. 19.2 we briefly review the related work. In Sect. 19.3 the model is described. We perform the experiments with the actual call logs and discuss the results in Sect. 19.4. In Sect. 19.5, we conduct the validation of our model using the actual call logs. Finally, we have the conclusions in Sect. 19.6.

19.2 Related Work

There are a large amount of previous work on event detection in text, data stream and video. In [1] the authors proposed a method based on an incremental TF-IDF model and the extensions include generation of source-specific models, similarity score normalization based on document-specific averages, source-pair specific averages, term re-weighting based on inverse event frequencies, and segmentation of the documents. In [2] the authors examine the effect of a number of techniques, such as part of speech tagging, similarity measures, and an expanded stop list on the performance. In [3] the authors use text classification techniques and named entities to improve the performance. In [4] a novelty detection approach based on the identification of sentence level patterns is proposed. In [5] the authors propose a probabilistic model to incorporate both content and time information in a unified framework, which gives new representations of both news articles and news events. They did explorations in two directions because the news articles are always aroused by events and similar articles reporting the same event often redundantly appear on many news sources. In [6] the authors propose to detect events by combining text-based clustering, temporal segmentation, and graph cuts of social networks in which each node represents a social actor and each edge represents a piece of text communication that connects two actors. In [7, 8] the authors propose the conceptual model-based approach by the use of domain knowledge and named entity type assignments and showed that classical cosine similarity method fails for the anticipatory event detection task. In [9] the authors proposed the online new event detection (ONED) framework, which includes a combination of indexing and compression methods to improve the document processing rate, a resource-adaptive computation method to maximize the benefit that can be gained from limited resources, new events to be further filtered and prioritized before they are presented to the consumer when the new event arrival rate is beyond the processing capability of the consumer and implicit citation relationships to be created among all the documents and used to compute the importance of document sources.

In [10] the authors propose an iterative algorithm and use likelihood criterion to segment a time-series into piecewise homogeneous regions to detect the change points, which are equivalent to events defined by them and evaluate them with the highway traffic data. In [11] the authors use an infinite automaton in which bursts are state transitions to detect burst events in text streams and conduct the experiments with emails and research papers. In [12] the authors use suffix tree to encode the frequency of all observed patterns and apply a Markov model to detect patterns in the symbol sequence. In [13] the authors find piecewise constant intensity functions to represent continuous intensity functions using a combination of Poisson models and Bayesian estimation methods and use dynamic programming method to find them. In [14] the authors use a time-varying Poisson process model and statistical estimation techniques for unsupervised learning in the context. They applied this model to freeway traffic data and building access data.

The above approaches are to detect events for text, novel and unusual data points or segments in time-series that have either contents or are traffic data. However, none of the previous work focuses on the specific problem we study here, using the inhomogeneous Poisson and inhomogeneous exponential distribution model by studying the calling pattern based on call detail records to detect events that reflect the human activity.

19.3 Model

19.3.1 Formulation

In event detection we need to analyze and classify categorical data, either in an exploratory or in a confirmatory context. Exploratory analysis of such data often has to do with extracting relevant hidden knowledge from a large dataset. We need to develop and use robust and flexible classification methods. In this paper we use probabilistic models for the classification of variables, based on inhomogeneous Poisson process for number of calls and exponential distribution for call durations.

The observed data can be represented in a bi-dimensional matrix, where rows describe data units and columns describe categorical variables. Empirical clustering models are usually used to analyze such data.

We assume that number of calls follows inhomogeneous Poisson process and call duration follow inhomogeneous exponential distribution.

Let $N_i = \{n_{i1}, \ldots, n_{ik}\}$ be random variable for number of calls of a given day i, $D_i = \{d_{i1}, \ldots, d_{ik}\}$ be random variable for call duration for day i, $i = 1, 2, \ldots, 7$ be a day of week, 1 for Sunday, …, 7 for Saturday. Then

$$N = \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1k} \\ n_{21} & n_{22} & \cdots & n_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ n_{71} & n_{72} & \cdots & n_{7k} \end{bmatrix}$$

is the matrix of number of calls on 7 days of week and

$$D = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1k} \\ d_{21} & d_{22} & \dots & d_{2k} \\ \dots & \dots & \dots & \dots \\ d_{71} & d_{72} & \dots & d_{7k} \end{bmatrix}$$

is the matrix of call duration on 7 days of week.

Then Poisson density function for day i is given by

$$P_{N_i}(N_i = n_{ij}) = \frac{e^{-\lambda_i}\lambda_i^{n_{ij}}}{n_{ij}!} \tag{19.1}$$

where λ_i is the rate (average) of number of calls for day i.

By the properties of Poisson distribution, the mean $= \lambda_i$, the variance $var = \lambda_i$ and the standard error $\sigma = \pm\sqrt{\lambda_i}$.

The exponential distribution density function of call duration for day i is given by

$$P_{D_i}(D_i = d_{ij}) = \frac{1}{\mu_i} e^{-\frac{d_{ij}}{\mu_i}} \tag{19.2}$$

where μ_i is the mean of call duration for day i.

By the properties of exponential distribution, the variance $var = \mu_i^2$ and the standard error $\delta = \pm\sqrt{\mu_i^2} = \pm\mu_i$.

Now using maximum likelihood estimates [16] to estimate the λ_i for day i. The cumulated probability distribution function is

$$\begin{aligned} P_{N_i}(N_i = n_{i1}, n_{i2}, \dots, n_{ik}|\lambda_i) &= \prod_{j=1}^{k} \frac{e^{-\lambda_i}\lambda_i^{n_{ij}}}{n_{ij}!} = \frac{e^{-k\lambda_i}\lambda_i^{\sum_{j=1}^{k} n_{ij}}}{\prod_{j=1}^{k} n_{ij}!} \\ \ln P_{N_i} &= -k\lambda_i + (\ln\lambda_i)\sum_{j=1}^{k} n_{ij} - \ln\left(\prod_{j=1}^{k} n_{ij}\right) \\ \frac{d(\ln P_{N_i})}{d\lambda_i} &= -k + \frac{\sum_{j=1}^{k} n_{ij}}{\lambda_i} = 0 \\ \hat{\lambda}_i &= \frac{\sum_{j=1}^{k} n_{ij}}{k} \end{aligned} \tag{19.3}$$

For μ_i The cumulated probability distribution function of call duration is

$$\begin{aligned} P_{D_i}(D_i = d_{i1}, d_{i2}, \ldots, d_{ik} | \mu_i) &= \prod_{j=1}^{k} \tfrac{1}{\mu_i} e^{-\frac{d_{ij}}{\mu_i}} = \tfrac{1}{\mu_i^k} e^{-\frac{1}{\mu_i} \sum_{j=1}^{k} d_{ij}} \\ \ln P_{D_i} &= -k \ln \mu_i - \tfrac{1}{\mu_i} \sum_{j=1}^{k} d_{ij} \\ \frac{d(\ln P_{D_i})}{d\mu_i} &= -\frac{k}{\mu_i} + \frac{1}{\mu_i^2} \sum_{j=1}^{k} d_{ij} = 0 \\ \hat{\mu}_i &= \frac{\sum_{j=1}^{k} d_{ij}}{k} \end{aligned} \tag{19.4}$$

The maximum likelihood estimates are used to estimate average number of calls and call duration. Next we consider the maximum average number of calls and call duration obtained for all weekday/weekend and week by week. Suppose that the m week data is used to compute the rates of number of calls and call duration for user p. Let $\hat{\lambda}_{d1}^{p}, \hat{\lambda}_{d2}^{p}, \ldots, \hat{\lambda}_{d7}^{p}$ be the rate of number of calls obtained for all weekday/weekend and $\hat{\lambda}_{w1}^{p}, \hat{\lambda}_{w2}^{p}, \ldots, \hat{\lambda}_{wm}^{p}$ be the rate of call duration obtained week by week for m weeks of user p respectively. Let $\hat{\mu}_{d1}^{p}, \hat{\mu}_{d2}^{p}, \ldots, \hat{\mu}_{d7}^{p}$ be the mean of call duration obtained for all weekday/weekend and $\hat{\mu}_{w1}^{p}, \hat{\mu}_{w2}^{p}, \ldots, \hat{\mu}_{wm}^{p}$ be the mean of call duration obtained week by week for m weeks of user p respectively.

Then the maximum means of number of calls and call duration are respectively computed by:

$$\hat{\lambda}_{\max}^{p} = \max\left(\hat{\lambda}_{d1}^{p}, \hat{\lambda}_{d2}^{p}, \ldots, \hat{\lambda}_{d7}^{p}, \hat{\lambda}_{w1}^{p}, \hat{\lambda}_{w2}^{p}, \ldots, \hat{\lambda}_{wm}^{p}\right) \tag{19.5}$$

$$\hat{\mu}_{\max}^{p} = \max\left(\hat{\mu}_{d1}^{p}, \hat{\mu}_{d2}^{p}, \ldots, \hat{\mu}_{d7}^{p}, \hat{\mu}_{w1}^{p}, \hat{\mu}_{w2}^{p}, \ldots, \hat{\mu}_{wm}^{p}\right) \tag{19.6}$$

where $\hat{\lambda}_{\max}^{p}$ and $\hat{\mu}_{\max}^{p}$ are the maximum likelihood estimates of number of calls and call duration for user p over the number of days specified respectively. The thresholds define the limits for all weekday/weekend and week by week. The assumption is that the calling pattern could be different. Each person has his/her own thresholds, and if the number of calls or call duration are greater than the thresholds of their own for some day, we define that there is some event in that day.

To calculate the threshold of number of calls for user p, N_{thres}^{p}, we define

$$N_{\text{thres}}^{p} = \hat{\lambda}_{\max}^{p} + \hat{\sigma}_{\max}^{p} \tag{19.7}$$

where $\hat{\lambda}_{\max}^{p}$ and $\hat{\sigma}_{\max}^{p}$ are the maximum rate of number of calls and correspondent standard error with positive $\hat{\sigma}_{\max}^{p}$.

To calculate the threshold of call duration for user p, D_{thres}^{p}, we define

$$D_{\text{thres}}^{p} = \hat{\mu}_{\max}^{p} + \hat{\delta}_{\max}^{p} \tag{19.8}$$

where $\hat{\mu}^p_{\max}$ and $\hat{\delta}^p_{\max}$ are the maximum mean of call duration and correspondent standard error with positive $\hat{\delta}^p_{\max}$.

Definition of an Event A collection of call log data can be represented as

$$C = \left\langle (t_1, a_1, d_1, l_1), (t_2, a_2, d_2, l_2), \ldots, (t_n, a_n, d_n, l_n) \right\rangle,$$

where t_i is a time point, d_i is a call duration, l_i is a location and a_i is a pair of actors, caller-callee $\langle s_i, r_i \rangle$ where s_i is an actor who initiates a call at time t_i and r_i is an actor who receive a call. An event is defined as a subset $E \subset C$ of a tuple

$$E = \left\{ (t_1, a_1, d_1, l_1), (t_2, a_2, d_2, l_2), \ldots, (t_m, a_m, d_m, l_m) \right\}$$

such that either $\sum_{i=1}^{m} d_i > D_{\text{thres}}$ or $count(d_i) > N_{\text{thres}}$ defined as the above in the time period $\Delta t = t_m - t_1$.

19.3.2 Real-Life Data Sets and Parameters

Real-Life Traffic Profile In this paper, the actual call logs are used for analysis. These actual call logs are collected at MIT [17] by the Reality Mining Project group for a period of 8 months. This group collected mobile phone usage of 100 users, including their user IDs (unique number representing a mobile phone user), time of calls, call direction (incoming and outgoing), incoming call description (missed, accepted), talk time, and tower IDs (location of phone users). These 100 phone users are students, professors and staff members. The collection of the call logs is followed by a survey of feedback from participating phone users for behavior patterns such as favorite hangout places; service provider; talk time minutes and phone users' friends, relatives and parents. We used this extensive dataset for our social group analysis and validation of 20 sample users in this paper. More information about the Reality Mining Project can be found in [17].

We use the formulas (19.3) and (19.4) to detect events based on the day of week, call frequencies and call duration.

Day of week: Everyone has his/her own schedule for working, studying, entertainment, traveling and so on. The schedule is mainly based on the day of the week.

Call frequencies: The call frequency is the number of incoming or outgoing calls in a period of time. The greater the number of incoming or outgoing calls in a period of time, the more socially close the caller and callee relationship.

Call duration: The call duration is how long both caller and callee want to talk to each other. The longer the call duration is in a period of time, the more socially close the caller and callee relationship.

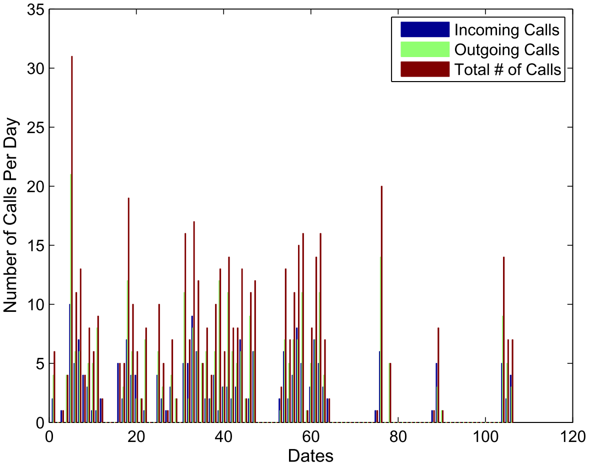


Fig. 19.1 The number of incoming, outgoing and total calls per day for user3

19.3.3 Computing the Thresholds of Frequency and Duration

The thresholds of frequency and duration are computed by formulas (19.7) and (19.8) based on day of week (Sunday, Monday, . . . , Saturday) and week sequence (1st week, 2nd week, . . .). Then the thresholds of frequency and duration are chosen to compare with the frequency and duration for each day. If the frequency or duration of some day is greater than the thresholds of frequency or duration, we define that there is an event in that day.

We used the data from the data set of four months, a semester since the communication members were relatively less changed in a semester for students.

19.4 Experiment Results and Discussion

Figures 19.1 and 19.2 show the number of calls and call duration for user3, where the x-axis indicates the days and y-axis indicates the number of calls and call duration (incoming, outgoing and total of them) respectively.

In Fig. 19.3 the x-axis is days and y-axis indicates the number of calls and call duration (incoming, outgoing and total of them) respectively, which show that there are events in these dates. From the Fig. 19.3 we may see that there are 7 event days which are the 5th, 18th, 31st , 33rd, 58th, 62nd, 76th, days during 106 days.

The experiment results of user3 and user74 as examples are listed in Table 19.1. In Table 19.1 the thresholds of number of calls and duration are calculated by maximum likelihood estimates. There are two types of events: one has location change and the other has no location change. For example, in Table 19.1, there are 17 calls, which is greater than 15 (the threshold of the number of calls) for user3 on the 33rd day and the location is on campus. We define that there is some event in that day. For

Fig. 19.2 The duration of incoming, outgoing and total calls per day for user3

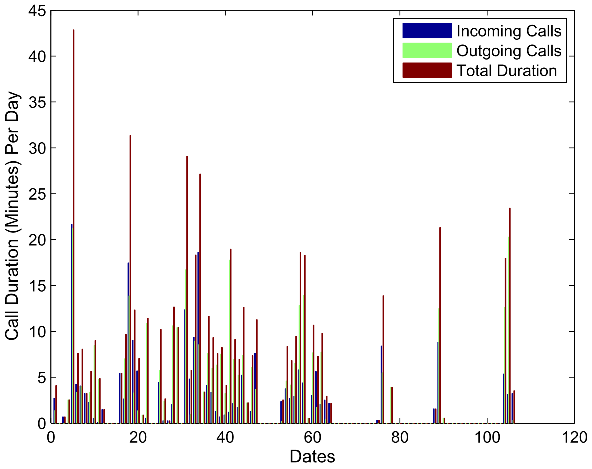
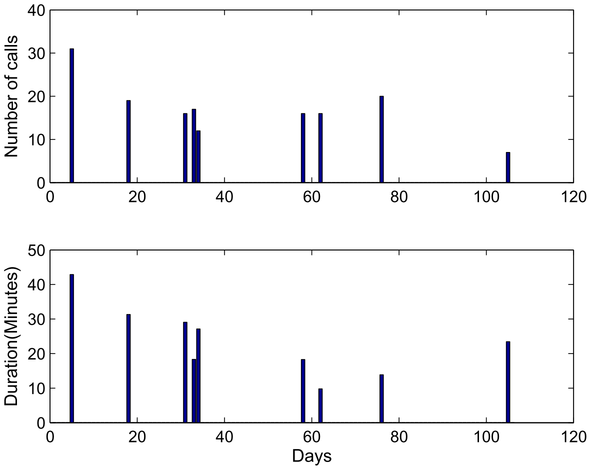


Fig. 19.3 There are events in these days for user3



user74 on the 24th day, although there are only 2 calls, the call duration is 76.8 minutes, which is much greater than 18 minutes (the maximum rate of the call duration) and there is some event in that day.

Note: the thresholds of number of calls and duration are computed by formulas (19.7) and (19.8).

19.5 Validation

To evaluate the accuracy of our model, we used actual call logs of 100 phone users and randomly choose 20 phone users. These users include students, professors and staff members. The best way to validate the results is to contact the phone users to

Table 19.1 Event dates and locations

Users	Event	Days	# of contacts	# of calls the day	Duration (minutes)	Location	Note
User3	1	5	11	31	42.9	Visit world trade center	Both large # of calls and duration
	2	18	3	19	31.3	Visit Harvard Univ.	Both large # of calls and duration
	3	31	8	16	29.1	Visit Harvard Univ.	Large # of calls
	4	33	6	17	18.3	Campus	Large # of calls
	5	58	11	16	18	Campus	Large # of calls
	6	62	10	16	9.8	Campus	Large # of calls
	7	76	7	20	13.9	Campus	Large # of calls
User74	1	24th	1	2	76.8	Campus	Large duration
	2	26th	4	13	20.3	Visit central square	Large # of calls
	3	27th	4	8	15.6	Campus	Large # of calls
	4	28th	2	11	64.2	At home	Large # of calls
	5	46th	1	6	40.2	Visit Stateplace-New York	Large duration
	6	78th	1	2	54.4	At home	Large duration
	7	86th	2	6	82.8	At home	Large duration
	9	93rd	3	10	24.5	Visit Stateplace-New York	Large # of calls
	10	122nd	4	9	7.9	Visit Stateplace-New York	Large # of calls
	11	130th	2	7	46.8	At home	Large duration
	12	131st	2	9	53.7	Home (next day visit Cityplace-Providence StateRI)	Both large # of calls and duration
	13	153rd	2	4	60	Visit Stateplace-New York	Large duration

get feedback, but because of the privacy issues it is almost impossible to use this way. Thus we use hand labeling method to validate our model. We used the data of the four months to detect events. Note that we cannot use the data of the next four months to validate our model since the events may happen in a different time period. In order to validate our model, we hand labeled the events based on the number of calls, duration of calls in the day, history of call logs, location, time of arrivals, and other humanly intelligible factors.

Table 19.2 shows the validation results. We achieve 92% accuracy.

Table 19.2 Validation results

Users	# of events	Threshold of # of calls per day	Threshold of duration per day (minutes)	Ave. # of calls per day	Ave. duration per day (minutes)	False positive	False negative
3	7	15	30	5	5.5	0%	11%
14	10	9	82	4	40	0%	9%
15	9	14	21	6	7	0%	10%
16	11	8	24	4	8	0%	8%
21	8	11	56	5	18	0%	0%
22	4	22	60	11	20	0%	9%
29	12	10	14	4	7	0%	7%
33	9	6	43	1	8	0%	10%
35	5	21	76	10	21	0%	8%
38	15	15	67	10	29	0%	6%
39	13	13	52	5	14	0%	9%
50	9	16	75	7	31	0%	10%
57	8	8	15	3	5	0%	11%
72	13	12	70	6	23	0%	6%
74	13	7	36	2	7	0%	7%
78	13	10	76	3	13	0%	6%
83	12	10	28	5	9	0%	7%
85	14	9	24	4	9	0%	6%
88	11	7	16	3	4	0%	8%
95	8	4	10	2	4	0%	10%

Note: the thresholds of number of calls and duration are computed by formulas (19.7) and (19.8).

19.6 Conclusion

In this paper we proposed the inhomogeneous Poisson process model for detecting events based on mobile phone call detail records. We used the data from the data set of four months, a semester since the communication members were relatively less changed in a semester for students.

The maximum likelihood estimates are used to estimate average number of calls and call duration to compare with the frequency and duration for each day. If the frequency or duration of some day is greater than the thresholds of frequency and duration, we define that there is some event in that day.

The best way to validate the results is to contact the phone users to get feedback, but because of the privacy issues it is almost impossible to use this way. Thus we use our defined conditions to validate our model.

This work is useful for enhancing homeland security, detecting unwanted calls (e.g., spam), communication presence, marketing etc. The experimental results show that our model achieves good performance with high accuracy.

In our future work we plan to detail the event classification and to analyze and use some criterion to optimize the detection process.

Acknowledgement We would like to thank Nathan Eagle and Massachusetts Institute of Technology for providing us the call logs of Reality Mining dataset.

References

1. Brants, T., Chen, F.: A system for new event detection. In: Proceedings of International ACM SIGIR Conference, pp. 330–337 (2003)
2. Chen, F., Farahat, A., Brants, T.: Story link detection and new event detection are asymmetric. In: Human Language Technology Conference (HLT-NAACL) (2003)
3. Kumaran, G., Allan, J.: Text classification and named entities for new event detection. In: Proceedings of international ACM SIGIR Conference, pp. 297–304 (2004)
4. Li, X., Croft, B.W.: Novelty detection based on sentence level patterns. In: Proceedings of ACM CIKM, pp. 744–751 (2005)
5. Li, Z., Wang, B., Li, M.: A probabilistic model for retrospective news event detection. In: Proceedings of International SIGIR Conference, pp. 106–113 (2005)
6. Zhao, Q., Mitra, P.: Event detection and visualization for social text streams. In: Proceedings of International Conference on Weblogs and Social Media (ICWSM) (2007)
7. He, Q., Chang, K., Lim, E.P.: A model for anticipatory event detection. In: Proceedings of the 25th International Conference on Conceptual Modeling (ER). LNCS, vol. 4215, pp. 168–181. Springer, Berlin (2006)
8. He, Q., Chang, K., Lim, E.P.: Anticipatory event detection via sentence classification. In: Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, pp. 1143–1148 (2006)
9. Luo, G., Tang, C., Yu, P.: Resource-adaptive real-time new event detection. In: Proceedings of International ACM SIGMOD Conference on Management of Data (2007)
10. Guralnik, V., Srivastava, J.: Event detection from time series data. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 33–42. ACM Press, New York (1999)
11. Kleinberg, J.: Bursty and hierarchical structure in streams. In: Proceedings of the Eigth ACM SIGKDD International Conference Knowledge Discovery and Data Mining, pp. 91–101. ACM Press, New York (2002)
12. Keogh, E., Lonardi, S., Chiu, B.Y.: Finding surprising patterns in a time series database in linear time and space. In: Proceedings of the Eigth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 550–556. ACM Press, New York (2002)
13. Salmenkivi, M., Mannila, H.: Using Markov chain Cityplace Monte Carlo and dynamic programming for event sequence data. Knowl. Inf. Syst. **7**(3), 267–288 (2005)
14. Ihler, A., Hutchins, J., Smyth, P.: Adaptive event detection with time-varying Poisson processes. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 207–216 (2006)

15. Cao, L.: In-depth behavior understanding and use: the behavior informatics approach. Inf. Sci. **180**, 3067–3085 (2010)
16. Harris, J.W., Stocker, H.: Maximum likelihood method. In: Handbook of Mathematics and Computational Science, p. 824. Springer, New York (1998), §21.10.4
17. Massachusetts Institute of Technology: Reality mining. http://reality.media.mit.edu/ (2008)