突发事件检测： kleinberg 状态机模型

本文链接：https://blog.csdn.net/hero_fantao/article/details/69681239

范涛
发表于 2017-04-08

1 背景

现实中，我们接触到各种文本信息，大多是以相应的事件来组织的。针对每个特定事件，涉及的相关文档都会有相应的时间信息，我们称这种时间信息为文档的到达时间。 那针对某个特定事件，涉及的相关文档的按到达时间顺序形成文档数据流。这种文档数据流天然的包含有序的时序信息，通过这种时序信息，我们能观察到事件是何时发生的，何时突然爆发，又何时衰退的，比如“天津爆炸案”。在 TDT（topic detection and tracking）领域，如何检测和追踪突发事件是一个重要的研究方向。这里重点想分享下这篇文章《Bursty and Hierarchical Structure in Streams》，Kleinberg 在 2003 左右发表的。这篇文章主体思路是根据时间发生的时间序列来建立一种突发检测模型。

在说这篇文章之前先列举几个问题：

（1）不同媒体，文档数据流速度不一样，如媒体，email，学术论文期刊。媒体文档数据流速度快，学术论文期刊文档流速度慢；

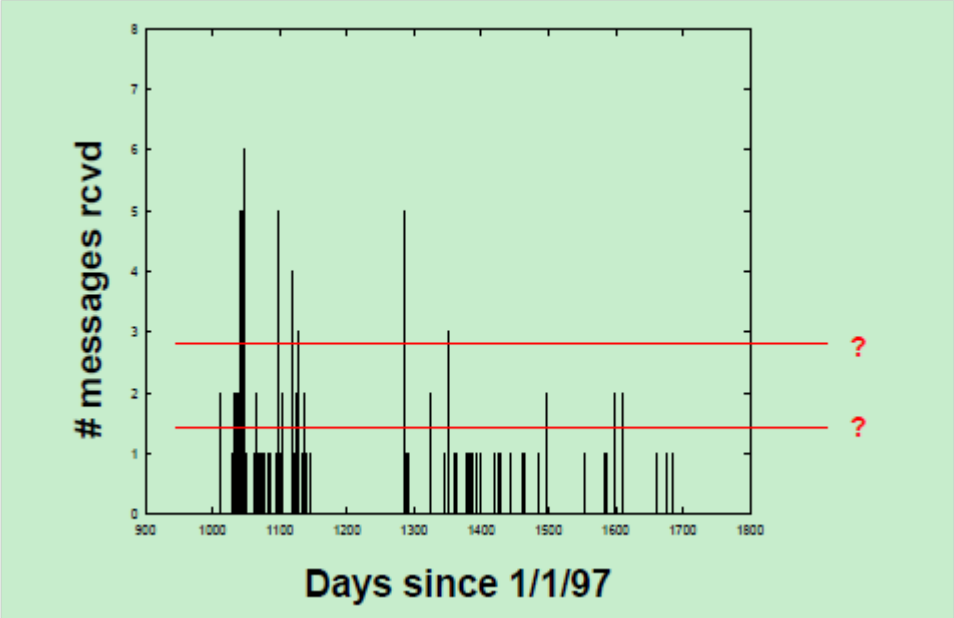
（2）如何通过模型来检测不同媒体，不同数据源下，不同演化速率主题突发行为以及持续周期？

（3）特定事件在突发周期里面是否包含多个嵌套的突发行为？

2 模型

2.1 Threshold-based Method

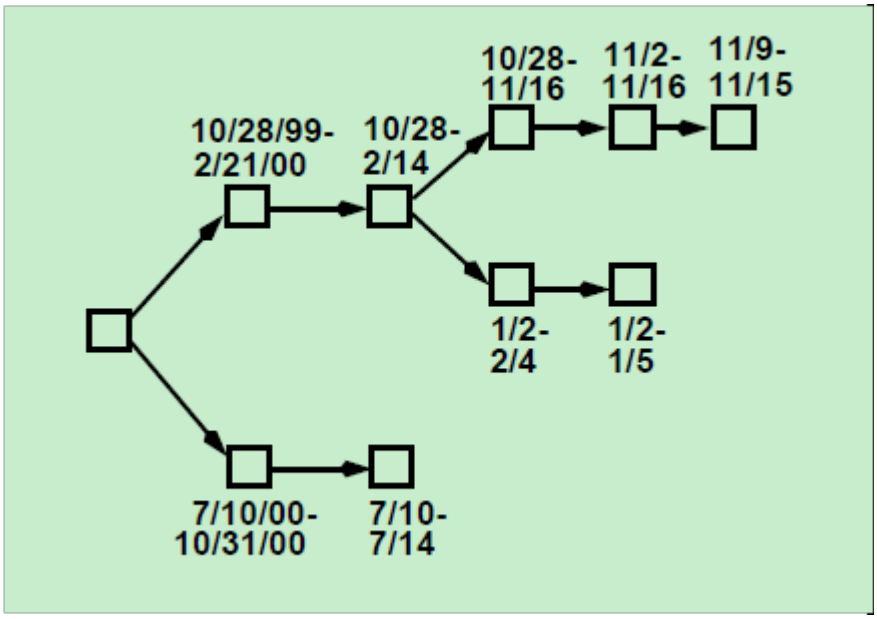
对同一事件的相关文档的到达时间，按天/小时 进行分箱，每个分箱包含一定文档数。设定文档数阈值，过滤出那些满足阈值条件的时间箱，连续的时间箱组成可以看做一个事件波峰。阈值的设定可以参考: x2 分布或者相似的分布检验方式。



但是这种方法存在如下问题：现实中很多文本数据流是稀疏和噪音的，图中存在一个没有连续 7 天非零时间箱，这会导致没法识别突发事件。另外，阈值方法没法检测不同尺度的突发事件，以及检测嵌套的突发事件结构。

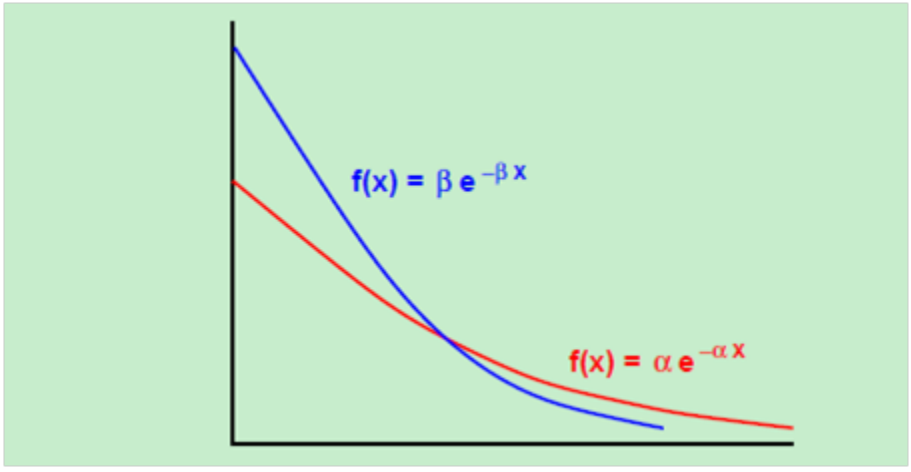
2.2 Kleinberg 状态机模型

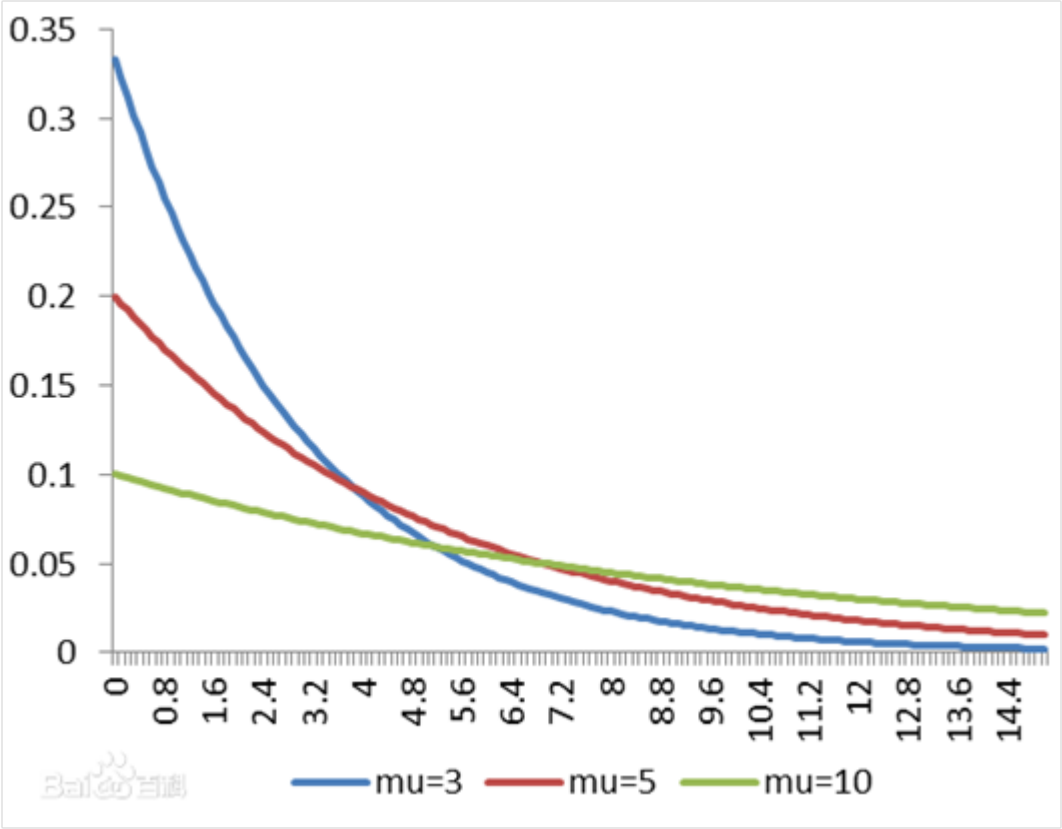
看一个例子： 一个 Email 数据流，文本包含“ITR“这个词，实际发现的突发事件结果。我们发现这个突发事件呈现不同持续周期以及嵌套层次结构。



2.2.1 消息到达时间分布

采用指数分布来模拟消息到达时间。





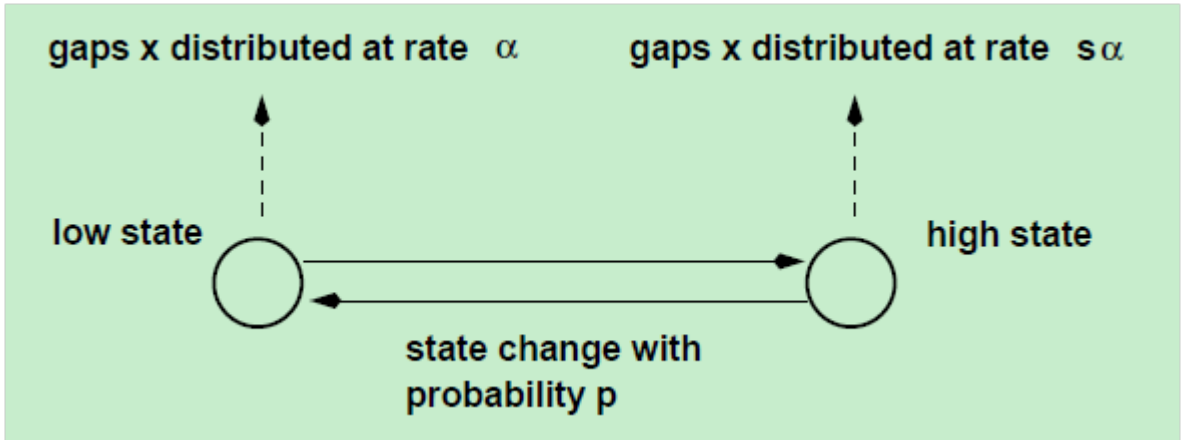
一个事件中下一个消息到达时间间隔服从指数分布（无记忆分布）：

$$f(x) = ae^{-ax}$$

其中时间间隔期望 $E(x) = a^{-1}$，其中 a 是消息到达速率。上图，展示了不同到达速率下指数分布。

2.2.2 Two state Model

在上节，提到用指数分布来模型消息到达间隔，那如何模型突发事件模型呢？ Kleinberg 提出的状态模型来模拟这种突发行为。先介绍下简单的 two state 模型。 设计两个状态，低状态（Low state）和高状态（High state），突发事件行为可以被模拟成一段周期内高低状态的转换。示意图如下：



低状态下服从指数分布，速度率为 a；高状态下，同样服从指数分布，速度率:s*a, 其中 s>1. 状态之间的转移概率为 p。

如果指定事件时间间隔序列 gaps 为 x = (x1,x2,... xn)， 存在状态序列 q = (qi1 ,.., qin),

在该状态序列下，事件时间间隔序列概率密度函数定义为：

$$f_{\mathbf{q}}(x_1, \ldots, x_n) = \prod_{t=1}^{n} f_{i_t}(x_t)$$

利用贝叶斯原理，得到后验概率:

$$\begin{aligned} \Pr[\mathbf{q} \mid \mathbf{x}] &= \frac{\Pr[\mathbf{q}]\, f_{\mathbf{q}}(\mathbf{x})}{\sum_{\mathbf{q}'} \Pr[\mathbf{q}']\, f_{\mathbf{q}'}(\mathbf{x})} \\ &= \frac{1}{Z} \left(\frac{p}{1-p} \right)^{b} (1-p)^{n} \prod_{t=1}^{n} f_{i_t}(x_t), \end{aligned}$$

其中 b 是状态转移次数。

最大化上述后验概率，等价于最小化下面公式：

$$-\ln \Pr[\mathbf{q} \mid \mathbf{x}] = b \ln \left(\frac{1-p}{p} \right) + \left(\sum_{t=1}^{n} -\ln f_{i_t}(x_t) \right) - n \ln(1-p) + \ln Z.$$

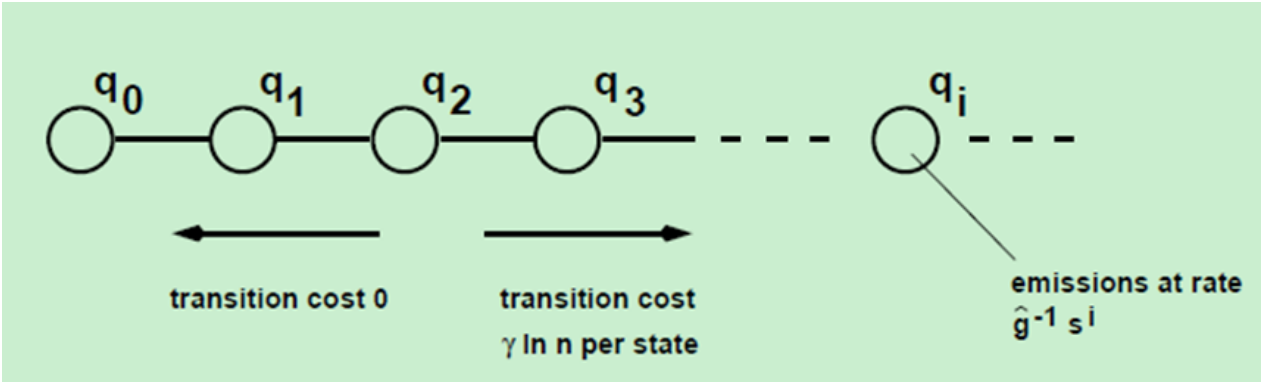
其中，第三项和第四项为常数项。

所以，设计代价函数 cost-function:

$$c(\mathbf{q} \mid \mathbf{x}) = b \ln \left(\frac{1-p}{p} \right) + \left(\sum_{t=1}^{n} -\ln f_{i_t}(x_t) \right)$$

2.2.3 Infinite-state model

2.2.2 描述了 two-state 模型，该模型只包含了 2 个状态，对于事件中出现的嵌套层次结构无法模拟，因此本节在 two-state 模型的基础上继续扩展，构造无限状态模型。示意图如下：



构造无限状态序列：q0, q1, q2...,qi,....，一段时间周期 T，时间间隔序列 gaps x = (x1,x2,... xn), 那么平均速率 a=n/T。设定 q0 的速率： a, qi (i > 0) 速率： (s^i) * a。

定义 transition cost:

- Low-> High cost: (j-i) * r * ln(n), where r>0 and j > i.
- High->Low cost: 0.

上面所说的无限状态机可以通过如下理论退化成有限状机, 存在个最高的状态 k：

Theorem 2.1 *Let* $\delta(\mathbf{x}) = \min_{i=1}^{n} x_i$ *and*

$$k = \lceil 1 + \log_s T + \log_s \delta(\mathbf{x})^{-1} \rceil.$$

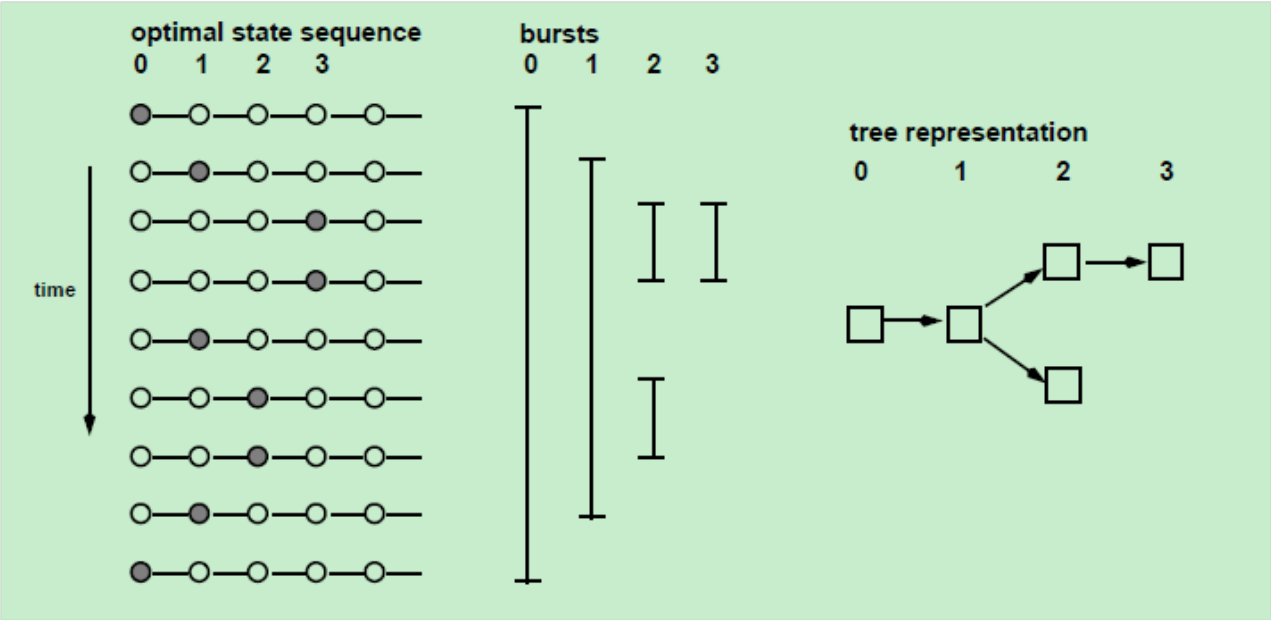
(Note that $\delta(\mathbf{x}) > 0$*, since all gaps are positive.) If* $\mathbf{q}^*$ *is an optimal state sequence in* $\mathcal{A}_{s,\gamma}^{k}$*, then it is also an optimal state sequence in* $\mathcal{A}_{s,\gamma}^{*}$.

定义 Cost Function:

$$c(\mathbf{q} \mid \mathbf{x}) = \left(\sum_{t=0}^{n-1} \tau(i_t, i_{t+1}) \right) + \left(\sum_{t=1}^{n} -\ln f_{i_t}(x_t) \right).$$

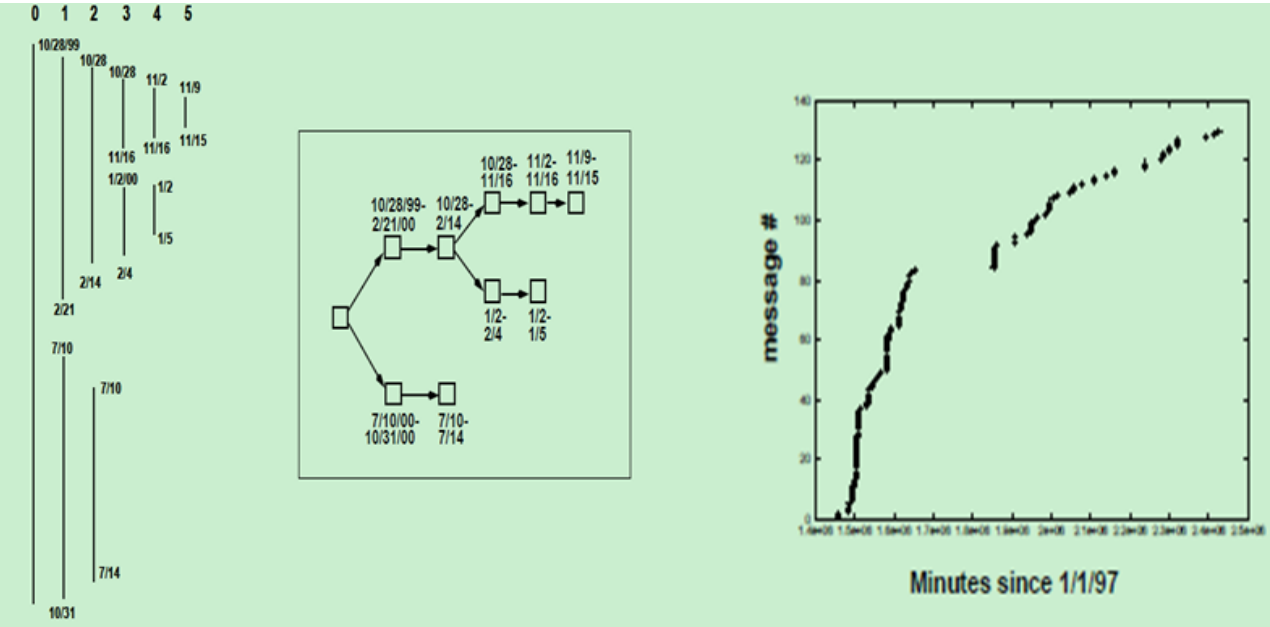
其中 $\tau(i_t, i_{t+1})$ 为状态转移代价。可以采用动态规划去寻找最优状态集合。

采用 infinite-state 模型我们可以检测和追踪突发事件，挖掘类似结构如下:



定义突发强度 j 为保持状态 qj 以及更高状态最大时间间隔。其中呈现的嵌套层次结构为：突发强度 j 为突发强度 j-1 的子节点。

在 email 数据流中挖掘出如下突发事件结构：



3 应用

其中以“二胎政策“事件以例。下面的例子中，该算法能够很好地检测出事件爆发的时刻以及持续时间。

挖掘出来的突发时间段：

State begin end

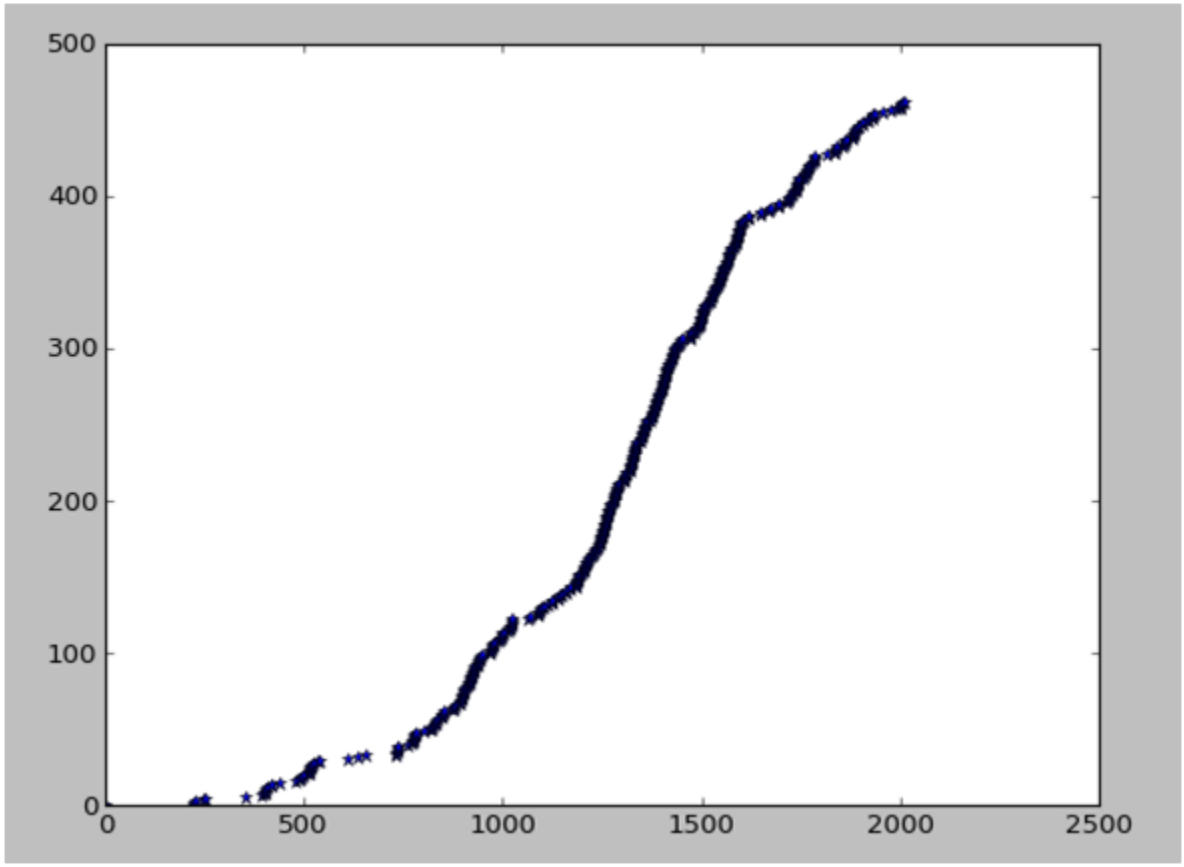
（状态， 开始日期， 结束日期）

0 2015-09-08 2015-11-30

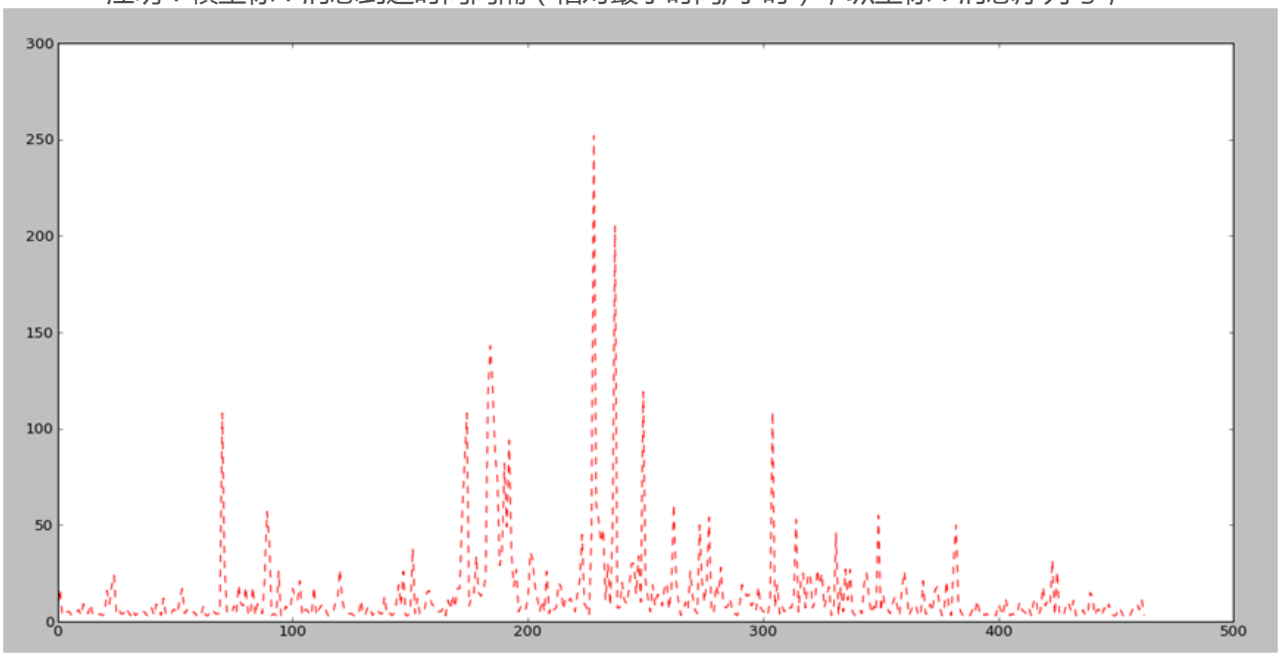
1 2015-10-15 2015-10-17

1 2015-10-27 2015-11-13

注明：10-29 日国家发布二胎政策



注明：横坐标：消息到达时间间隔（相对最小时间/小时），纵坐标：消息序列号；



注明：横坐标：消息对应的时刻序（从小到大），纵坐标：对应时刻的消息数；

4 扩展 kleinberg 状态机模型

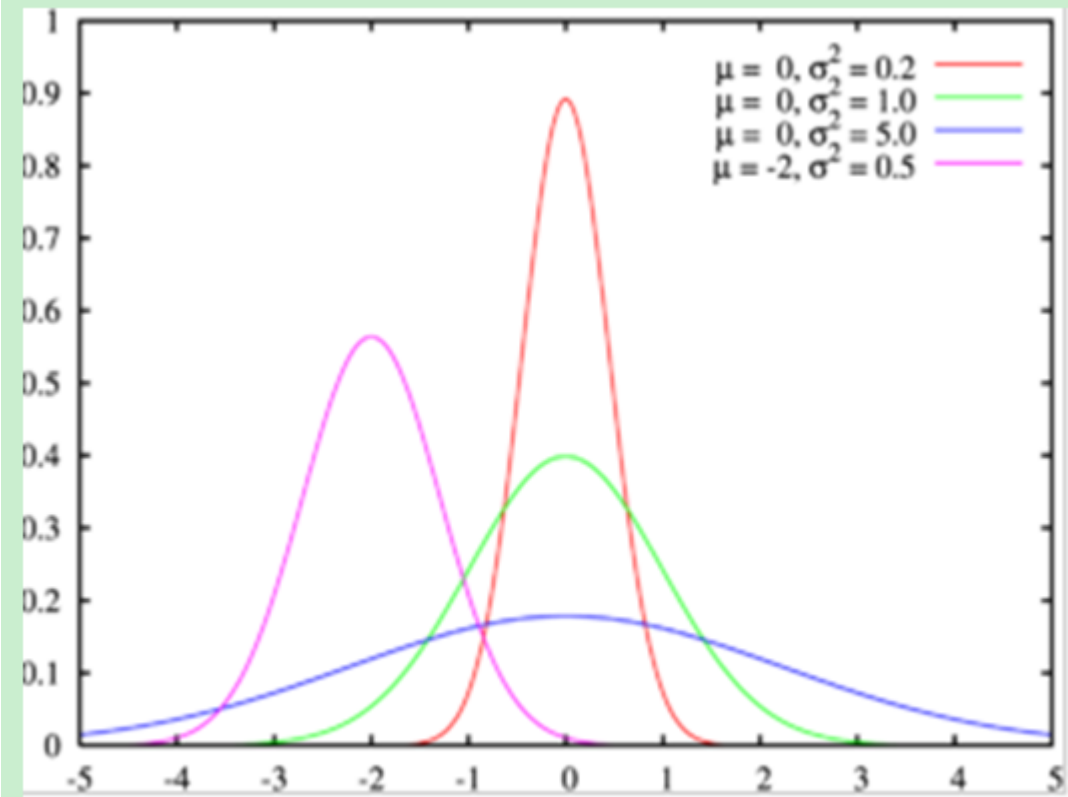
2 节详细讨论了基于文档时间序列构建的状态机模型，里面只考虑文档的到达时间。现实中，考虑到实际计算成本，还是会限定最小的时间粒度，比如分钟，小时等。这样，每个最小时间粒度下，可能会有多篇文档信息，也就是

说单位时间文档发生频次。同样，针对点击日志数据，经常会有单位时间点击数等。那如何在 kleinberg 状态机模型的基础上进行扩展，来模拟这种数据行为？

总体来说，kleinberg 状态机模型框架可以很容易被修正，来支持这种数据行为。重要改动的的部分是 cost-function。Cost Function 部分分为两个分布：状态转移代价和当前状态概率密度。状态转移代价这块可以和 2 节完全一样。需要改动概率密度这块。所以，重点是如何采用什么样的分布来拟合频次，点击数这样的数据分布？本次实验中主要采用正态分布（x2 分布应该也可以）来模拟消息单位时间频次分布。

正态分布：

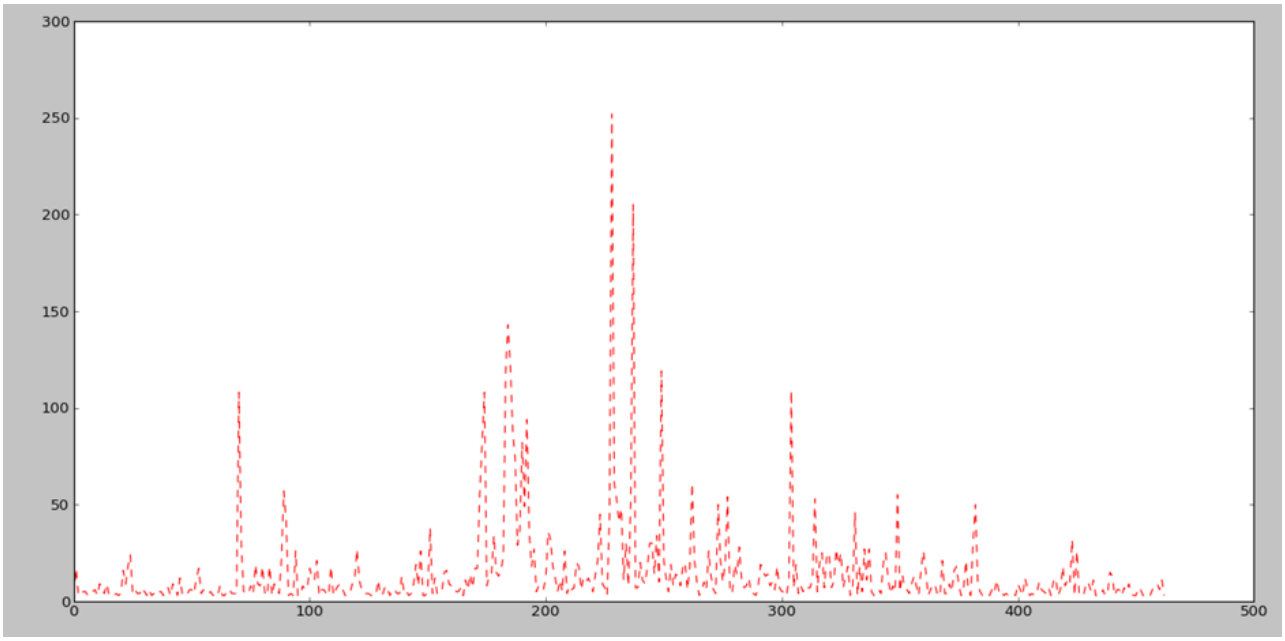
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



四个不同参数集的概率密度函数（绿色线代表标准正态分布）

优化方案中，设定$\mu = 0$，不同状态只是设定σ不同。$p0 = \sigma 0$，$pi = s^i * \sigma 0$。

还是以“二胎政策”为例，下面是频次分布（横坐标是时间序列，纵坐标是频次）：

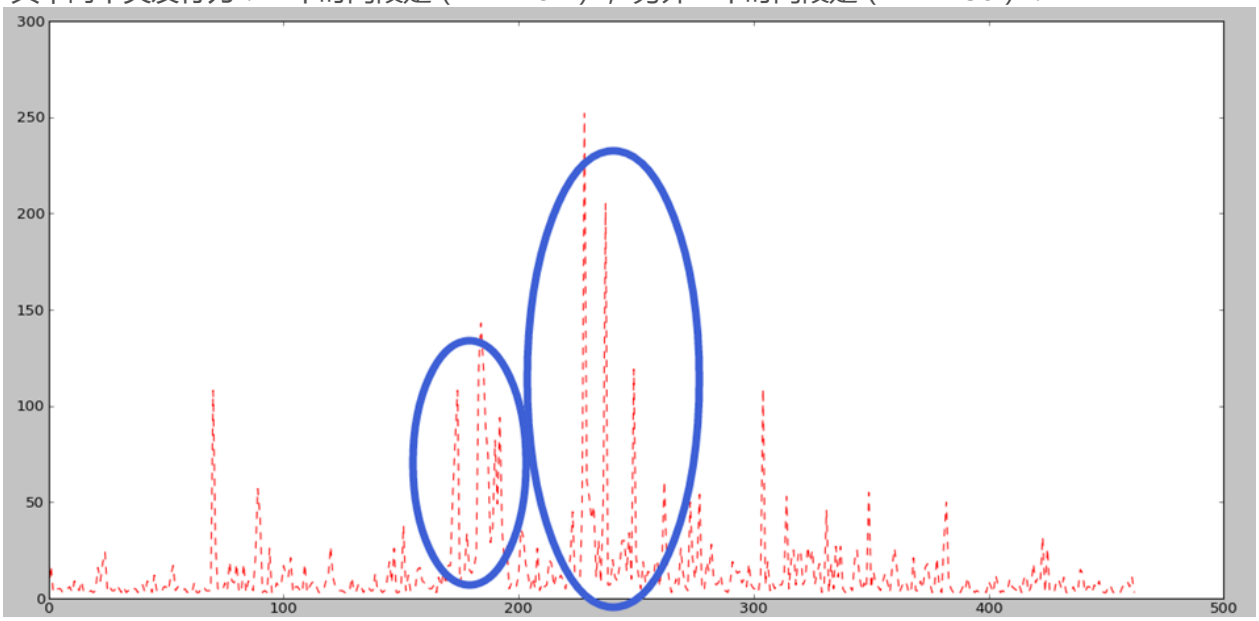


我们模型挖掘出两个下面突发行为：

[1 172 194]

[1227 250]

其中两个突发行为：一个时间段是（172-194），另外一个时间段是（227-250）：



5 附录

这篇文章有开源的代码，我这里做了些修改和改进。

代码 github 路径： https://github.com/dylan-fan/kleinberg_bursts

6 参考文献

Kleinberg J. Bursty and hierarchical structure in streams[C]// Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2003:91-101.