Scalable and Near Real-Time Burst Detection from eCommerce Queries

Nish Parikh, Neel Sundaesan
ACM SIGKDD '08

Outline
- Context in which the problem is bored
- Infinite-state automaton
  Bursty and Hierarchical Structure in Streams—ACM SIGKDD'02
- Main contribution of this work
- Former related work

Main Idea—Bursty and Hierarchical Structure in Streams
- Extract meaningful structure from document stream
- Burst of activity: certain features rising sharply in frequency as the topic emerges
- A formal approach for modeling such “bursts”
  - An infinite-state automaton
  - Bursts appear as state transitions
  - A nested representation of the set of bursts that imposes a hierarchical structure on the overall stream.

A Weighted Automation Model: One State Model
- Generating model:
  \[ f(x) = \alpha e^{-\alpha x} \]
  - \( x \): the gap in time of two consecutive messages
  - \( \alpha \): expectation:
    - : rate of message arrivals
- Why this model?

A Weighted Automaton Model: Two State Model
- Two states automaton \( A: q_0, q_1 \)
  \[ f_0(x) = \alpha_0 e^{-\alpha_0 x} \quad f_1(x) = \alpha_1 e^{-\alpha_1 x} \]
  - A changes state with probability \( p \), remaining in its current state with probability \( 1-p \), independently of previous emissions and state changes.
  - A begins in state \( q_0 \). Before each message is emitted, A changes state with probability \( p \). A message is then emitted, and the gap in time until the next message is determined by the distribution associated with \( A \)'s current state.

A Weighted Automaton Model: Two State Model
- Based on a set of messages to estimate a state sequence
  - Maximum likelihood
    \[ x = (x_1, x_2, \ldots, x_n) \quad q = (q_0, q_1, \ldots, q_b) \]
  - \( n \) inter-arrival gaps:
    - A state sequence:
    - \( b \) denotes the number of state transitions in the sequence, \( q \):
      \[
      L[q|x] = \frac{p^b}{\left(1-p\right)^n} \prod_{i=1}^{n} f_q(x_i)
      \]
      \[
      = \frac{1}{Z} (1-p)^b (1-p)^n \prod_{i=1}^{n} f_q(x_i)
      \]
A Weighted Automaton Model: Two State Model

- Finding a state sequence \( q \) maximizing previous probability is equivalent to finding one that minimizes

\[
\ln \Pr(q \mid x) = h \ln \left( \frac{p}{1-p} \right) + \left( \sum_{i} - \ln f_i(x_i) \right) = n \ln (1-p) + \ln Z
\]

- Equivalent to minimize the following cost function:

\[
c(q \mid x) = h \ln \left( \frac{p}{1-p} \right) + \left( \sum_{i} - \ln f_i(x_i) \right)
\]

Experiment related

- Dataset: 5 months of queries from eBay.com in 2007 (75+ TB of data).
- Assumption and pre-definition:
  i) The number of queries uniform distribute over time of one day;
  ii) Max number of segments of query arrivals per day is scaled to 48;
  iii) Each arrival is represented by a UNIX timestamp.

Incremental Burst Detection

- Based on the rate of change of percentage volume for a query
- Vs. change of absolute volume—Noiseless;
- Object to batched arrival of new queries– avoid re-calculate the entire state sequence when new batch arrives.
Burst Classification

- Method based: Wavelet transforms
- 4 classes:
  i) Matterhorns;
  ii) Cuestas;
  iii) Dogtooths;
  iv) Hogback.

Sorting and Ranking

- Concentration based ranking
  - Duration of burst (D);
  - Mass (Popularity) of Burst (M);
  - Arrival Rate for Burst (A);
  - Span Ratio (SR);
  - Momentum of Burst (Mo): \( Mo = (M \cdot A) \);
  - Concentration of Burst (Xc): \( Xc = SR^{-1} \)

Table 4 Table indicating calculated values of \( \alpha \) and \( \beta \) for five different days using method discussed above

<table>
<thead>
<tr>
<th>Sample Number</th>
<th>( \alpha = \frac{m(Ga - R)}{n(Ga)} )</th>
<th>( \beta = \frac{m(Ga - R)}{n(R)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.31</td>
<td>0.27</td>
</tr>
<tr>
<td>2</td>
<td>0.27</td>
<td>0.33</td>
</tr>
<tr>
<td>3</td>
<td>0.40</td>
<td>0.37</td>
</tr>
<tr>
<td>4</td>
<td>0.38</td>
<td>0.33</td>
</tr>
<tr>
<td>5</td>
<td>0.16</td>
<td>0.26</td>
</tr>
<tr>
<td>Average</td>
<td>0.3</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Implementation

- Method based: Wavelet transforms
- 4 classes:
  i) Matterhorns;
  ii) Cuestas;
  iii) Dogtooths;
  iv) Hogback.

Figure 10 Screenshot of an application using top burst of day to create a mash-up and attract online users based on diversity

Thank You!