Bursty and Hierarchical Structure in Streams*  
Jon Kleinberg

Presented By: Amir Sadovnik

*Or “What happens when Prof. Jon Kleinberg wants to organize his email”

Motivation

- Many documents can be viewed as streams that arrive continuously over time. (e.g. email, news articles, conference papers).
- An appearance of a topic in a document stream is signaled by a burst of activity.
- The goal of this paper is to model such bursts in a formal way which will provide a framework for analyzing the underlying content.

Motivation

Search/Veronica

Two State Model

- Bursts correspond to points at which the intensity of message arrival increases
- Rate of arrival does not rise smoothly and then fall but exhibits frequent alternations
- Analyzing gaps in a too simplistic way can lead to wrong results

Poisson arrival of messages:

![Diagram of Two State Model]

Gaps PDF: 
\[
f_0(x) = \lambda e^{-\lambda x} \quad \lambda < \lambda_1 \]
\[
f_1(x) = \lambda_1 e^{-\lambda_1 x}
\]

Sequence of Gaps: 
\[
\tilde{x} = (x_1, x_2, \ldots, x_n)
\]

Sequence of States: 
\[
\tilde{q} = (q_1, \ldots, q_n)
\]

Probability of gaps given states: 
\[
f_{\tilde{q}}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i)
\]
Two State Model

Cost Function:
\[ c(q \mid x) = b \ln \left( \frac{1-p}{p} \right) + \sum_{i=1}^{n} \ln f_i(x_i) \]

- First term tries to minimize state transitions
- Second term tries to maximize probability of \( x \).
- Can be solved using dynamic programming.

This gives us an optimum which tracks the global structure of bursts in the gap sequence while holding to a single state through non-uniformity.

An Infinite-State Model

- A burst of intensity is a maximal interval over which \( q \) is in a state of index \( j \) or higher.

Hierarchical Structure

- For modeling papers gap time cannot be used (appear in batches once a year).
- Instead we can use the portion of documents that are relevant in a batch (e.g. contain a specific word)
- Parameters:
  - \( r_i \) - # of relevant doc. In batch \( t \)
  - \( d_i \) - total # of doc. In batch \( t \)

\[ R = \sum_{i=1}^{n} r_i, \quad D = \sum_{i=1}^{n} d_i \]
Analogous Model

Number of states: \( P_0 \leq 1 \)

Transition Cost: \( \tau(i,j) = \begin{cases} (j - i)^n \ln n & \text{if } j > i \\ 0 & \text{if } j < i \end{cases} \)

State Sequence Cost: \( \sigma(i, r_i, d_i) = -\ln \left( \frac{d_i}{r_i} \right) _0 ^\infty (1 - p_0)^s \)

Cost Function: \( \psi(q, r_q, d_q) = \sum_{i=1}^{p_0} \tau(i, r_i, d_i) + \sum_{i=1}^{p_0} \ln \left( \frac{d_i}{r_i} \right) _0 ^\infty (1 - p_0)^s \)

Weight of Burst

- If we consider just 2 states in the automaton we can define the weight of a burst as:

\[ \sum_{i=0}^{p_0} (\sigma(0, r_i, d_i) - \sigma(1, r_i, d_i)) \]

- Using this the following experiment was conducted:
  - Analysis to the titles STOC and FOCS papers 1969-2001
  - All words were tracked in experiment

Results

<table>
<thead>
<tr>
<th>Words</th>
<th>Interval of Bursts</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;data base&quot;</td>
<td>1790 - 1800</td>
</tr>
<tr>
<td>&quot;database&quot;</td>
<td>1801 - 1816</td>
</tr>
<tr>
<td>&quot;gentlemen&quot;</td>
<td>1857 - 1860</td>
</tr>
<tr>
<td>&quot;militia&quot;</td>
<td>1859 - 1863</td>
</tr>
<tr>
<td>&quot;whilst&quot;</td>
<td>1861 - 1871</td>
</tr>
<tr>
<td>&quot;slaves&quot;</td>
<td>1859 - 1863</td>
</tr>
<tr>
<td>&quot;rebellion&quot;</td>
<td>1861 - 1871</td>
</tr>
<tr>
<td>&quot;depression&quot;</td>
<td>1930 - 1937</td>
</tr>
<tr>
<td>&quot;recovery&quot;</td>
<td>1930 - 1937</td>
</tr>
<tr>
<td>&quot;banks&quot;</td>
<td>1931 - 1934</td>
</tr>
<tr>
<td>&quot;democracy&quot;</td>
<td>1937 - 1941</td>
</tr>
<tr>
<td>&quot;wartime&quot;</td>
<td>1937 - 1941</td>
</tr>
<tr>
<td>&quot;that's&quot;</td>
<td>1982 - 1982</td>
</tr>
<tr>
<td>&quot;we're&quot;</td>
<td>1982 - 1982</td>
</tr>
<tr>
<td>&quot;we've&quot;</td>
<td>1982 - 1982</td>
</tr>
<tr>
<td>&quot;schools&quot;</td>
<td>1996 - 1996</td>
</tr>
<tr>
<td>&quot;teachers&quot;</td>
<td>1996 - 1996</td>
</tr>
<tr>
<td>&quot;21st century&quot;</td>
<td>1997 - 1997</td>
</tr>
</tbody>
</table>

Discussion

- The interplay between time and content is crucial.
- This model can be applied in other areas (e.g. web usage data)
- Bursts have sharp boundaries, therefore can be mapped to specific documents/events.