<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Event detection using blog tags</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Chen, Wenda</td>
</tr>
<tr>
<td><strong>Citation</strong></td>
<td>Chen, W. (2008, March). Event detection using blog tags. Presented at Discover URECA @ NTU poster exhibition and competition, Nanyang Technological University, Singapore.</td>
</tr>
<tr>
<td><strong>Date</strong></td>
<td>2008</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://hdl.handle.net/10220/9062">http://hdl.handle.net/10220/9062</a></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>© 2008 The Author(s).</td>
</tr>
</tbody>
</table>
Event Detection Using Blog Tags

Background

Bloggers often add tags (or keywords) chosen from an uncontrolled vocabulary to blog posts to enable browsing and searching on posts. Many of the tags are related to the events described in the blog posts. The purpose of this project is to detect events based on the tags and also the blog posts content. The evolution of the tags used to describe the same event at different time points reflects the change of views among bloggers.

Event Definition

An event is represented by a subset of blog posts describing what happened in a particular place and location. With the tags attached to blog posts, we aim to firstly obtain those tags related to the same event and then select blog posts according to their tags and publication time.

Event Detection Algorithm

Step 1: obtain all the occurrences of a given tag in ascending order of publish time. Calculate the time gaps $x$ to form a sequence $q$:

\[
\begin{array}{cccc}
\text{states}(q) & 0 & 1 & 2 & 3 \\
\text{time(day)} & 6 & 8 & 20 & 21
\end{array}
\]

Step 2: apply 2-state automaton algorithm\[1\]. State for each occurrence is determined by a genetic algorithm to find the state sequence $q'$ that can maximize the function

\[
Pr(q|x) = \frac{Pr[q]|q(x)}{\sum_{q'} Pr[q']|q(x)} = \frac{1}{Z} p^x (1 - p)^{1-x} \prod_i f_i(x)
\]

where $p$ is consecutive state transition probability and $f_i(x)$ is the probability of time gap $x$ in state $i$.

Step 3: cluster the time series $a$ of the tags in set $B$ into events $E$ by maximizing

\[
P(E_k | D) = \frac{P(D|E_k)P(E_k)}{P(D)} = \prod_{j=1}^{M} D_j
\]

\[
P(D|E_k) = \prod_{j=1}^{M} \left[ \frac{D_j}{M} \right]^x \left[ 1 - \frac{D_j}{M} \right]^{1-x}
\]

which considers the co-occurrence of the tags belonging to the event $E_k$ (indicated by $\alpha_k = 1$) and the percentage of posts $D$ carrying the tag over the period of the time when the event is bursty \[2\].

Future Work

In our future work, we plan to study event detection based on the named entities in blog posts together with the tags and also event evaluation methods.

References: