The Blogosphere at a Glance — Content-Based Structures Made Simple

Olof Görnerup and Magnus Boman
Swedish Institute of Computer Science (SICS)
olofg@sics.se
The blogosphere is goldmine of information
Dynamic repository of thoughts and opinions
Raw data semantically unstructured
Provide an overview of contents
Back to basics
Back to basics

We have blogs
Back to basics

We have blogs
We have blogs

We have implicit (note) semantic relations
Back to basics

- We have blogs
- We have implicit (note) semantic relations
Back to basics

- We have blogs
- We have implicit (note) semantic relations
- We have a network
Define “blog”

- A set (bag) of words
- That’s it — word order discarded
- Collected during a certain time period
Define “semantic relation”
Define “semantic relation”

Assumption: Blogs that share many words are semantically similar
Define “semantic relation”

Assumption: Blogs that share many words are semantically similar

Similarity quantified as the Jaccard index
Jaccard index
Jaccard index

Two blogs, \( A = \{v_1, v_2, \ldots, v_m\} \) and \( B = \{w_1, w_2, \ldots, w_n\} \)
Two blogs, $A=\{v_1, v_2, \ldots, v_m\}$ and $B=\{w_1, w_2, \ldots, w_n\}$

Similarity as “the fraction of all of $A$ and $B$’s words that are shared by $A$ and $B$”
Two blogs, $A=\{v_1, v_2, \ldots, v_m\}$ and $B=\{w_1, w_2, \ldots, w_n\}$

Similarity as “the fraction of all of A and B’s words that are shared by A and B”

Put differently, $S(A,B)=\frac{|A \cap B|}{|A \cup B|}$
Blog similarity network
Blog similarity network

Nodes are blogs
Blog similarity network

Nodes are blogs

Nodes are linked with edges with weights $S$

$S = 0 \iff$ no edge
Case study

Swedish blog data from 5-month period
Collected through Twingly* blog search engine
About 20k blogs considered

*www.twingly.com
Minimal pre-processing

- No stemming etc
- Keep words that are uncommon, but not too uncommon
- Info content per word likely to increase
- Computationally more tractable
The Swedish blogosphere
First observation
First observation

Blogs are clustered
First observation

- Blogs are clustered
- Identify blog groups automatically
First observation

- Blogs are clustered
- Identify blog groups automatically
- Agglomerative clustering technique [Clauset, 2005]
First observation

- Blogs are clustered
- Identify blog groups automatically
- Agglomerative clustering technique [Clauset, 2005]
- Partition that maximizes modularity
Second observation
Second observation

Clusters are clustered
Second observation

Clusters are clustered

Infer hierarchical model [Clauset et al., 2008]
Second observation

- Clusters are clustered
- Infer hierarchical model [Clauset et al., 2008]
- Monte Carlo sampling
Second observation

- Clusters are clustered
- Infer hierarchical model [Clauset et al., 2008]
- Monte Carlo sampling
- Maximum likelihood
Third observation
Third observation

Spam blogs (splogs) are revealed as outliers
Conclusions
Conclusions

Simple methods reveal valuable information
Conclusions

Simple methods reveal valuable information

The signal in raw blog data is strong!
Conclusions

- Simple methods reveal valuable information
- The signal in raw blog data is strong!
- Blogosphere highly structured
Conclusions

- Simple methods reveal valuable information
- The signal in raw blog data is strong!
- Blogosphere highly structured
- Applications on multiple granularity levels
Conclusions

- Simple methods reveal valuable information
- The signal in raw blog data is strong!
- Blogosphere highly structured
- Applications on multiple granularity levels
- Navigation and monitoring
Conclusions

- Simple methods reveal valuable information
- The signal in raw blog data is strong!
- Blogosphere highly structured
- Applications on multiple granularity levels
- Navigation and monitoring
- Splog detection
Thank you