

Tracking User Generated Content

Krisztián Balog
Information and Language Processing Systems
University of Amsterdam

User generated content

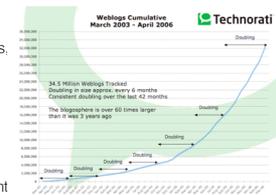
- ▶ In recent years the content of the web is changing
- ▶ User-generated content
 - Web pages and content created and published by users of websites, rather than website owners
 - Everyone is a contributor
 - People use the web to publish and share photos, videos, audio, and various forms of written material
 - A particular type of user-generated content is the *blog*

Blogs

- ▶ Blogs are frequently updated web pages, with entries shown in reverse chronological order
- ▶ Many blogs function as online diaries
 - personal-oriented content written in an informal manner, inviting interaction with readers
 - offers a unique window into people's areas of interest, thoughts, feelings

The blogspace as a corpus (Differences from other web corpora)

- ▶ Content
 - personal nature, thoughts, emotions, commentary
- ▶ Structure
 - social network of people
- ▶ Timeline
 - every blog post, comment has a timestamp attached
- ▶ Growth



Our interest in weblogs

- ▶ Address the *information overload* problem
 - we are no longer able to digest all material we have access to
- ▶ Need mechanism for
 - searching
 - categorizing
 - summarizing
 - ... simply understanding this data

Blog research @ ILPS

- ▶ Moodviews - blog mood analysis
- ▶ Advertisement placement, product recommendation
- ▶ Opinion search in blogs
- ▶ People search in blogs, recognizing social roles
- ▶ Blog community discovery
- ▶ Blog genre classification
- ▶ Faceted search in blogs
- ▶ Impact of cultural events ("a concert in Paradiso")
- ▶ Tracking the Dutch elections in news and blogs (verkiezingskijker.nl)
- ▶ Correlations between financial news, financial blogs, and the stockmarket

MoodViews.com

Gilad Mishne, Krisztian Balog, Maarten de Rijke

Collection of tools for tracking the stream of mood annotated text made available by Livejournal

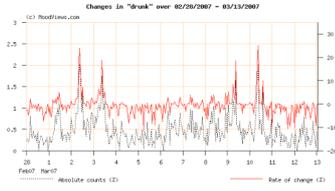
Livejournal

- ▶ Popular blogging platform
- ▶ Users can annotate their posts with their mood at the time of writing
 - choose from a list of 132 moods or enter free-text
- ▶ ~1 million posts/month



Aggregating moods

- ▶ The totality of mood reports gives an "internet global mood"
 - Over time: "internet mood swings"

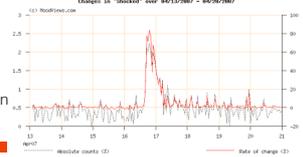


Moodgrapher

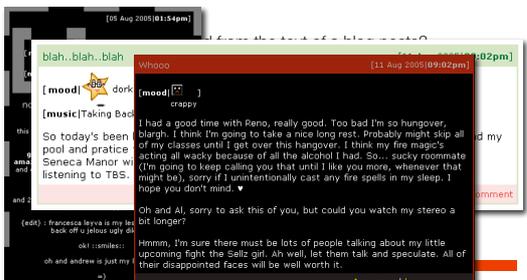
- ▶ Simply tracks mood levels as they develop
- ▶ Some moods display a cyclic behavior
 - Awake, sleepy, hungry, drunk, stressed, ...
- ▶ Observe strong responses to global world events

Questions

- Can we predict the mood levels?
- Can we identify and explain mood peaks?



Moodteller



Moodteller

- ▶ Can we predict the mood from the text of a blog posts?
- ▶ A text classification problem
 - Individual post classification is hard
 - ▶ Even with many, diverse features
 - ▶ Main setback: short text → meaningless statistics
- ▶ Aggregation solves sparseness
 - Can mood levels be derived from the language used by bloggers?

Mood level prediction

- ▶ Estimate intensity level (for each mood)
- ▶ Mood "recipe"
 - temporal metadata (hour, day of week)
 - "most indicative terms"
- ▶ Build linear regression model
 - Models must be constantly updated

```

Happy = 23.345 +
  0.0318 * total-posts +
  -2.4026 * count(always) +
  -114.9877 * count(day) +
  16.2727 * count(excited) +
  55.3942 * count(finally) +
  129.2576 * count(happy) +
  223.8079 * count(home) +
  -246.8737 * count(know) +
  506.9564 * count(lol) +
  5.7815 * count(thoughtful) +
  -88.1313 * count(will be)
    
```

Evaluation

- ▶ Corpus
 - 39 days of blog posts from LiveJournal
 - 8.1M posts, of these 3.5M indicate moods
 - 2.2GB of text
- ▶ Experiments
 - 10-fold cross validation
 - Measure: Pearson's correlation, Relative Error
- ▶ Baseline: use only temporal data

Results

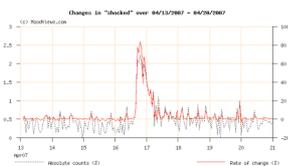
- ▶ Baseline alone gets 0.71 correlation and 64% relative error
 - Strong prior bias
- ▶ Regression improves both by ~20%, achieving 0.83 correlation and 52% relative error
 - Some moods have correlations of >0.95
 - ▶ "bored", "happy"

moodviews.com/Moodteller



Moodsignals

- ▶ Moods often reflect people's responses to global world events
- ▶ Detect unusual behavior (swings in mood levels)
- ▶ Explain peaks



Moodsignals

- ▶ Detecting peaks
 - need to deal with cyclic events
 - calculate *expected mood level* based on historical data
 - if *divergence* exceeds a threshold a spike has occurred
 - ▶ divergence = actual / expected mood level

www.moodviews.com

Questions?