Trend analysis and anomaly detection in time series of language usage

Krisztian Buza[1], Gabor I. Nagy[1], Alexandros Nanopoulos[2]

[1] Budapest University of Technology and Economics, Hungary, chrisbuza@yahoo.com, nagy.gabor.i@gmail.com
[2] University of Eichstätt-Ingolstadt, Germany, alexandros.nanopoulos@ku.de

1. Background

- Growing importance of social media
- Some authors suggest that sudden (unexpected) changes in the usage of language may be indicative of psychological disorders and/or diseases [1]
- **Goal**: examine how the usage of language changes

language usage
trend tweets anomaly detection
analysis change social time series
historical blogs textual data media recognition
data LARGE predictions for the future?
collection

2. Data and Methods

- BlogData: 37,279 blog posts from 1200 sources (users) spanning several years [2].
- Financial Tweets: 174,826 financial tweets from 82,653 users within a period of two weeks

For each user (source) we extracted time-series of
- total length of his/her texts (number of the words),
- number of different words used by that user,
- entropy of the text

in monthly (BlogData) and daily (Financial Tweets) resolution.

Entropy: $H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$

- in our case: each x_i is a word of the text

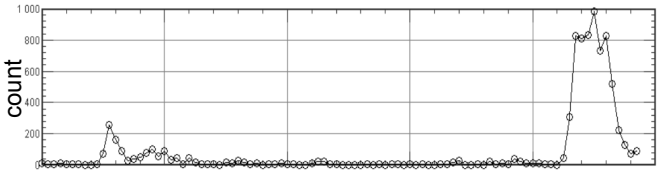
Anomaly score – distance from the *k*-th nearest neighbour with *k*=10 [3].

Topic trends – time series of the frequencies of keywords being characteristic to a topic

Acknowledgements – We thank DAAD and Magyar Ösztöndíj Bizottság (grant no. 39859) and the Hungarian Scientific Research Fund (grant no. OTKA 108947). This work is supported by the grant: FUTURICT, TÁMOP-4.2.2.C-11/1/KONV, „Financial Systems" subproject.

3. Results

Daily trend for the keyword "schmitt" in blog feedbacks



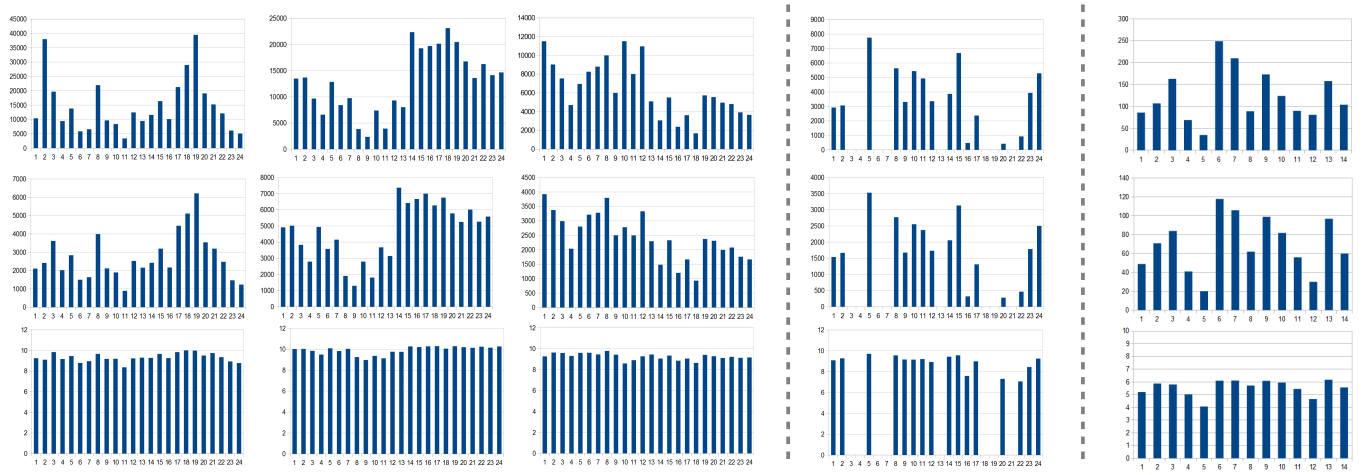
1 Jan 2012 8 April 2012

- Usage of language may substantially vary over time and across users
- Entropy seems to be more stabil than the number of (different) words of a user

References

[1] Deborah Estrin (2013). Small, n=me, Data. Invited talk at Neural Information Processing Systems Conference (NIPS).
[2] Buza, K. (2014). Feedback Prediction for Blogs. In Data Analysis, Machine Learning and Knowledge Discovery. Springer.
[3] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys (CSUR), 41(3), 15.

Drivers' blog | Consumers' blog | Political blog | Blog with the highest anomaly score | Financial tweets (a characteristic user)



From the top to the bottom: total length of the blogs in words, number of different words, and entropy for selected sources (users) as function of time (May 2010 – April 2012).