Comment-based communities in the Russian Livejournal and their topical coherence

Olessia Koltsova, Sergei Koltcov (with special thanks to Sergei Nikolenko)

http://linis.hse.ru

RuSSIR 2013

Online discussions and networking have proved vitally important in the political life of contemporary societies, and sometimes have been crucial for political regime change. The structure of those discussions and communities that presumably arise around them presents an important and relatively new research question for social scientists. For Russia, most such discussions have been until recently housed by Livejournal blogging service. Most network analysis of Livejournal has been devoted to the networks of friendship (Zakharov 2007; Lescovec 2008), while real discussions develop in threads of comments that may also be represented as comment-based networks. Our goal in this study has been to determine if such networks form discussion communities based on a shared topic, or around an author (opinion leader), or neither. For our goals, we have constructed a graph with posts as vertices, where two posts were considered connected if they had comments written by the same blogger.

Data

The data were retrieved from Livejournal website based on its social capital rating list and its API into an MS SQL database with authors' Koltran BlogMiner downloading software. The data include all posts by top-2000 bloggers for 1 week between April 1 and 7, 2013, as well as relational structure of commenting (who commented which post and how many times). After clearing and excluding non-commented posts the resulting graph contains 17386 vertices (i.e. posts written by 1667 authors) and around 4.5 mln edges.

Community detection

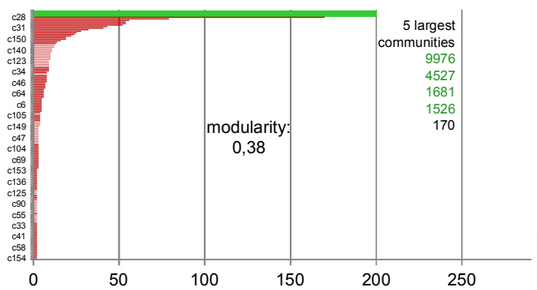
Community detection with Louvain algorithm (Blondel et al 2008) has revealed a moderately manifest, still evident community structure with modularity =0.38 and a highly skewed distribution of community sizes, the largest community comprising more than a half of vertices. A large number of small communities are isolated pairs and triads of little interest. Analysis of dependence of posts' belonging to a community on their authorship has revealed strong positive correlation (Pic. 1.)

Authorship

		Value	Asympt.Std.Error	Approx. T	Approx.sig.
Lambda	Symmetric	,209	,003	59,644	,000
	Dependent blogger	,057	,002	26,346	,000
	Dependent community	**,522**	,007	56,832	,000
Goodman & Kruskal Tau	Dependent blogger	,041	,001		,000
	Dependent community	**,510**	,004		,000
Cramer's V		**,466**			,000
Contingency Coefficient		**,985**			,000

Pic.1. Analysis of dependence of posts' belonging to a community on their authorship. Belonging of a post to a community strongly depends on the post's authorship. I.e. communities tend to form around authors.

Community structure



Pic.2. Analysis of dependence of posts' belonging to a community on their authorship. Belonging of a post to a community strongly depends on the post's authorship. I.e. communities tend to form around authors.

Topic similarity

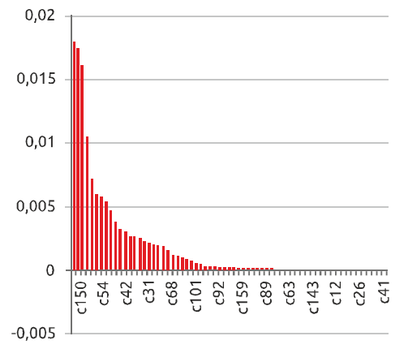
To detect topical similarity of texts within and outside communities a classical bag-of-words approach was used: texts were considered thematically similar if they shared a large amount of words, word sequence being not taken into consideration. Prior to calculating word frequencies (TF-IDF), each text was cleared of html tags, other impediment symbols and lemmatized with Yandex Mystem lemmatizer. After that two alternative methodologies were used: cosine similarity calculation and topic modeling with LDA algorithm.

Cosine similarity

Once all cosine distances between each pair of texts were calculated, it became possible to obtain: average distance within each community, average intra-community distance (=0,04916513) and global average distance (=0,00015924). As can be seen, distance between texts assigned to the same community is on average three orders of magnitude smaller than global average. At the same time, distribution of intra-community cosine similarity means is again highly skewed, with a minority of communities being highly above the average and a vast majority only slightly above or even slightly below the global average. The middle part of this distribution is shown on Pic. 2, where 0 on Y axis is global cosine similarity average, X axis contains communities sorted by their average cosine similarities.

The distribution of logarithms of cosine similarity (Pic. 3) shows that while globally they are distributed as some kind of a perfect bell (black line), some communities that stand high above the global cosine similarity average gain additional peaks shifted closer to the higher values of cosine similarity (X axis).

The preliminary selective coding of communities shows that those with cosine similarity above the average tend to be dominated by a set of posts covering a roughly similar set of issues and written by the same author or by a very limited number of authors, while a relatively large number of disconnected posts by a large number of authors "sticks" to this relatively coherent core. Presumably, it is this core that produces additional peaks in Pic. 4.



Pic.3. Distribution of intra-community cosine similarities in comparison with global average.

Topic modeling

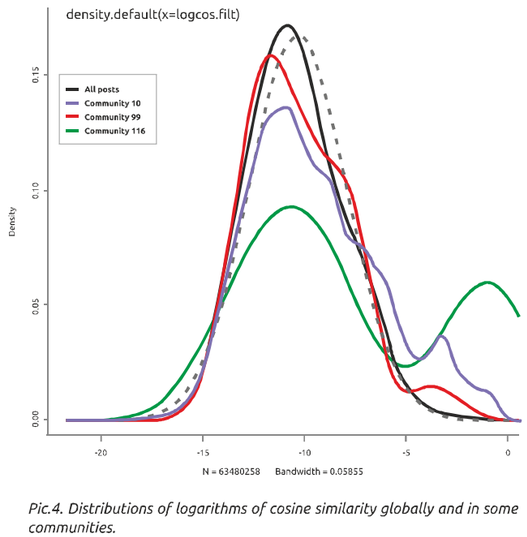
Topic modeling (N of topics =100), LDA with Gibbs sampling algorithm (authors' LINIS TopicMiner software) was then performed on the dataset. Total weight of each topic for each comment-based community was calculated, being a sum of probabilities of belonging of each text in a given community to a given topic. After normalization, variances of topics' weights for each community were calculated, ranging from 0 to 100. If texts in a community were evenly distributed across all topics, the variance was low; if one or a small number of topics dominated, the variance was high and the graph of distribution had sharp peaks at certain topics. The largest community containing more than the half vertices naturally had the lowest variance, while among other communities different types could be observed.

Conclusions

People commenting top LJ bloggers tend to unite into moderately manifest (modularity =0.38) communities by unintentionally commenting roughly the same sets of posts. The graph of co-commenting is sparse and connected by a minority of active commentators that tend to be non-top bloggers themselves (fandom commenting). Communities strongly tend to emerge around authors of posts and to a less visible degree – around topics of posts. Topical coherence of some communities is presumably connected with the topical coherence of the author (or a small number of authors) dominating these communities, while a large number of communities are not topically coherent at all. This research has still to be replicated on other datasets and supplemented with analysis of commentators composition.

Acknowledgements

This research is supported by the Basic Research Program of the National Research University Higher School of Economics, in 2013.



Pic.4. Distributions of logarithms of cosine similarity globally and in some communities.

References

1. Adamic L.A., Zhang, J., Bakshy, E., Ackerman M.S. Knowledge sharing and yahoo answers: everyone knows something. // WWW '08: Proceeding of the 17th international conference on World Wide Web. – NY: ACM, 2008. P. 665–674

2. Ali-Hasan N., Adamic L.A. Expressing social relationships on the blog through links and comments. // International conference on weblogs and social media. – San Jose, CA, USA, 2009.

3. Blondel V.D., Guillaume J.-L., Lambiotte R., Lefebvre E., Fast unfolding of communities in large networks, J. Journal of Statistical Mechanics, P10008 (2008)

4. Gomez V., Kaltenbrunner A., Lopez A. Statistical analysis of the social network and discussion threads in slashdot. // WWW '08: Proceeding of the 17th international conference on World Wide Web. – NY: ACM, 2008. P. 645–654.3

5. Jamali S., Rangwala H. Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis. // International Conference on Web Information Systems and Mining. – Shanghai, China, 2009. P. 32-38.

6. Leskovec J., Lang K.J., Dasgupta, A., Mahoney M.W. Statistical properties of community structure in large social and information networks. // WWW '08 Proceedings of the 17th international conference on World Wide Web. – Beijing, China: ACM, 2008. P. 695-704

7. Zakharov P. Diffusion approach for community discovering within the complex networks: LiveJournal study. // Physica A: Statistical Mechanics and its Applications, vol. 378, Issue 2. – Switzerland: Department of Physics, University of Fribourg, 2007. P. 550-560.