Olessia Koltsova
Kirill Maslinsky
Sergei Koltcov
National Research Univeristy Higher School of Economics, Russia

Protests, Elections and their contribution to the topical structure of the Russian blogosphere: a "Big Data approach"

Paper to be submitted at "Internet, Politics, Policy 2012: Big Data, Big Challenges?" conference
Oxford Internet Institute, September 20-21, 2012.

Blogs and social networks have proved important in recent years in various senses, but particularly for political mobilization in a whole range of societies. They include not only European countries such as Spain (May 2011) or Russia (December 2011), but also countries with relatively low Internet penetration, such as Egypt. This paper is based on the study of the Russian blogosphere aimed at mapping its thematic, discursive and network structure.
The long-term methodological goal of the study is to borrow and/or develop and introduce into social science a range of new methods of collection and analysis of big data in order to be able to derive sociological conclusions from them. The immediate goal of the study presented is to describe topical structure of the Russian language blogosphere as some reflection of public opinion in order to see what social issues are important, what cleavages exist and how it changes over time. Russian elections and protests in December 2011 is the key test issue explored in the project.

1. Blogs and the Russian blogosphere

A blog, as an online "genre" that has a form of a diary, is most often expected to be maintained by an individual beyond his or her professional activity – that is, to be a kind of user generated content. Therefore, this content may be viewed as a mode of existence and formation of public opinion, reflecting some cleavages and solidarities present in the society. Russian-language blogosphere contains about 58 million blogs, of which stand-alone blogs are about 7 million, while others are hosted by about a hundred of blog-platforms (www.blogs.yandex.ru, accessed 31.07.2012). Blog platforms differ technically which may have social consequences: for instance, some platforms, including LiveJournal maintain dendroid structure of comments (that is, users may comment not only on posts, but on each other's comments), while other platforms do not provide this possibility, and this has a direct effect on the structure of discussions.

A feature that until recently has been specific to Russia is merge of blog platforms and social networks. Classical American blog service offers a function of a blog-roll – a list of hyperlinks to favorite blogs independent of their location. That is why connectedness of American blogs is platform-independent and may be explained by social factors. Most Russian blog services offer a function of friending instead that can not be applied to bloggers from other platforms. This makes access to blog content outside the platform more difficult than inside it, which leads to closure of discussions and other hyperlinking within separate platforms. Since no single blog platform clearly dominates the Russian blogosphere, the latter has got clearly clustered.

The situation may change with social networks developing blogging features. Russian audience of blogs is considered to be close to saturation both in terms of number of accounts and unique daily users, while social networks are growing rapidly. The leading network Vkontakte unites three times more users than there are blog accounts in the entire Russian blogosphere, where leadership is shared between four blog platforms that house only one fifth of all blog accounts.

However, so far as social networks have not become “proper blogs”, the real blogs play an important role in a society where Internet is not technically filtered, while leading traditional media are tightly controlled by the political elite. Most political and social discussion is housed by Livejournal (Etling et al, 2010; Alexanyan & Koltsova, 2009; Gorny, 2004) which is perceived as a place for public-oriented accounts. It is mostly LJ blogs that collect thousands of friends and thus function as alternative media: top 30 LJ users have 20000+ friends each, which is a good circulation for an average Russian newspaper. LJ is one of the four leaders in terms of the number of accounts and is the leader in the number of daily posts and of unique daily users.

2. Defining and operationalizing the concept of topic

Although intuitively clear, the concept of topic has been fraught with theoretical and practical problems. Most generally, it may be defined as the main subject matter of a text. But on a more detailed level it has been extensively debated in linguistics where it found very little consensus or clarity; and it has been never problematized in topic modeling and similar statistical approaches where it has always been defined algorithmically, in order to find a mathematical model that fits best some human coding that is seldom questioned. At the same time, in our experience determining the topic of a text has always aroused a lot of difficulties and discrepancies among coders. Everyday topical classifications of discourse have neither clear grounds, nor shared ad hoc reasons, and they are very difficult to operationalize for research unambiguously and comprehensively. In people’s minds topics may be centered around events, social problems, unproblematic issues, constant or transient aspects of life, discourse types or value systems. With varying “power of a microscope” through which a text (document) collection (corpus, sample) is viewed, topics of different scales may be found, some of them being organized in hierarchies, and others standing alone. Often topics meet or overlap within the same text, but some texts are monotopical. In some discursive spaces, such as blogosphere, topics change fast, not only emerging and disappearing, but also breaking, uniting and mutating. Finally, some portion of texts has no topic or legible meaning and can not be understood or classified at all. Such texts can be considered noise. Despite all this, media, search engines and blog platforms constantly generate topical classifications of their content, while millions of users successfully apply them on a daily basis.

This success may be explained by that in everyday life approximate classifications are sufficient or at least acceptable. For research purposes this may not be enough; however, non-academic benchmark text collections (such as Reuters corpora http://trec.nist.gov/data/reuters/reuters.html, accessed 1 August 2012) are often the ones against which topic models are tested, and statistical approaches to topics are the only ones that try to define and investigate topics empirically on large text data. Therefore, scholars doing social research, including ourselves, have to “consume” these models acknowledging the listed above limitations. Conventionally, all statistical approaches – i.e. those that define topics through analysis of word frequencies and their co-occurrence in texts – may be divided into classical cluster analysis and topic modeling, currently dominated by Latent Dirichlet Allocation algorithm (LDA) and its derivatives. In text clustering, topics are not defined explicitly; texts are directly compared to each other and divided into groups based on their lexical similarity. Since texts on the same topic may be expected to share a lot of words, the resulting clusters may be interpreted as topical aggregates. In topic modeling, topics are viewed as latent sets of words that co-occur in texts of the corpus most frequently. Texts, then, are compared not to each other, but to these word sets / topics and are assigned to these topics with varying probabilities.

3. Data collection

In sociology, data collection is a well-developed area of knowledge and practice that deserves special attention in methodological research, textbooks and separate courses in university curricula. However, sociological literature is surprisingly scarce when it comes to collection of internet data. Sociological works on methodology are dominated with discussion of online surveys (see e.g. Lauret et al, 2003); internet content analysis works rely on manual collection of small samples for hand coding, which is usually not problematized (Papacharissi, 2007; Herring et al, 2005). On the data mining side, the main subject of study is optimization of data delivery to users of commercially generated content, and – to our knowledge – no attention is paid to data collection for research goals, at least in the sphere of text analysis (in network online data collection the situation seems more optimistic, but this is beyond the scope of this article). For computer scientists, technical aspects of such data collection are too trivial to be of research interest, while for sociologists this is an unsolvable problem. To bridge this gap, a whole set of issues – sample design, data search, downloading, storage, navigation and export to analytical tools – has to be addressed by interdisciplinary teams. The results of our attempt to do so are presented below.

To form samples, one has to have an idea about general population; while exhaustive lists of blogs are not publicly available or even non-existent in most countries / language domains, Russia is a notable exception in this respect. Its leading search engine company, Yandex, provides access to a relatively full list ranked according to its criteria. However, topic-related descriptive statistics of blogs is largely unknown – at least in public domain, partially because of the ambiguity of the concept of topic, but also because of many other technical and methodological obstacles. Blogs are usually manifestly polythematic. What can have one or at least limited number of topics is a post. Therefore, it is posts that should be sampled, and this is a much more difficult task than sampling from blogs. Russian-language blogosphere produces 10^5 posts a day (with microblogs - 10^6) and several times more comments (www.blogs.yandex.ru, accessed 31 July 2012). Full public lists of posts do not exist, so it is impossible to sample from them, unless one has a Google database in her pocket. Furthermore, topic modeling studies suggest that topics are usually numerous and may amount to dozens and hundreds in collections of 10^5 texts (Blei, 2003: 4). What follows from this is the necessity to form large samples that can not be either analyzed or even downloaded manually. This in turn demands developing special software for each task, which is definitely a challenge for social scientists.

One seemingly easy way to get around technical difficulties is to try to benefit from data aggregated by search engines. Automatic downloading of search results has the following drawbacks: (a) search engines can not provide all posts in a given period, while limitations of keyword search are even more severe; (b) they usually do not give away more than one thousand pages; (c) they rank results with algorithms protected as commercial secrets and thus methodologically opaque. On the other hand, developing a software that could aggregate data from all blog platforms in a structured searchable form would actually mean repeating the work of a large search engine. This is unfeasible for most academics. One alternative is to download the first available pages of blogs whose URLs are obtained from blog lists, such as the one maintained by Yandex, without parsing – i.e. without separating titles, tags, ads, html codes etc. from main texts. A simple word count applied to these noisy data gives a snapshot of the blogosphere's short-term agenda across platforms, however, possibilities for more detailed, precise and long-term analysis are very limited. This approach is used in MediaCloud developed by Berkman Center for Internet and Society at Harvard (www.mediacloud.org, accessed 2 August 2012). Another approach is to concentrate effort on one platform that is found most relevant to the research goals.

We chose the second approach and, given our goals, Livejournal became a natural choice for us. An additional argument was that LJ, among its other ratings, maintains a rating which sorts users solely by the number of friends and is therefore absolutely transparent in the methodological terms. We have developed a software that downloads bloggers' nicknames, texts and URLs of their posts with dates, texts of related comments with dates and commentators' nicknames. This is a relational SQL text-only database with full text search that does not rate or cut off the search results. The database maintains different kinds of sampling and exports data to a number of analytical tools. For the research presented here we have formed three datasets containing all posts from top n bloggers from three different periods:

- August 15 – September 15 2011 (politically "calm" period), top 1400, no more than 50 posts per blogger, 24 thousand posts;
- November 27 – December 27 2011 (period around Russian parliamentary election of December 4, which covers both pre-election discussion and all protest actions before the Christmas break), top 1400, no more than 50 posts per blogger, 27 thousand posts;
- March 4 – April 4 2012 (period after the Presidential elections with some protest actions), top 2000, 67 thousand posts.

The dataset for the period before the presidential elections have been also formed, but not yet analysed. The length of periods is taken on the basis of previous studies of news lifecycles (Koltsova, 2011; Wu, 2011).

4. Data analysis approaches, algorithms and software

4.1. Related work and its unresolved problems

The demand for social analysis of large text data is currently ahead of the development not only of data collection instruments, but also of analytical instruments that social scientists could use to their ends. Both clustering and topic modeling algorithms are multiplying rapidly, while their comparative testing is lagging back. Since samples should be large, a first problem is that of computational complexity of algorithms, and its relation to quality. Some works in this field contain no information on computational complexity at all (Mizral, 2009), others (e.g. Zhong, 2005) derive it from algorithm's mathematical properties solely, without reference to real experiments, that could clearly tell what time is needed to process a given number of texts on a given computer, and what is the maximum number of texts/megabites the algorithm can handle at all. Quality is often assessed from the experiments, however, first, different measures and different datasets are applied in different articles, each time for a narrow set of algorithms, so cross-article comparison is difficult. Second, again, no reference is usually made to experimental conditions such as concrete software tested and computer used, so boundaries of applicability of tests' results are unclear.

All this deprives social scientists and other potential algorithm "consumers" of criteria that could clearly help choose between existing software. Some review articles (Andrews & Fox, 2007) contain some comparative data on computational complexity, but they are obtained from the developers, not from the independent experiments. So far, we have not found any comprehensive comparative study that would test a wide range of algorithms on the same (large) datasets with the same quality measures, with clear reference to software and computers involved. Neither we have seen works comparing cluster solutions obtained from classical cluster analysis and from topic modeling approaches, where they are comparable.

The problem of quality is related to the problem of determining the number of clusters or topics, because quality measures may be used to assess which solution is better. We have not found any software that would automatically optimize any quality function, at best some offer entire dendrograms or lists of solutions with respective values of some quality functions. However, most quality functions, both internal and external, change monotonously, and it is still very hard to choose between solutions. And, of course, the number of topics/clusters in the blogosphere is not known a priori – furthermore, there may be no single "right" answer to this question. As is known from Milligan and Cooper's (1985) comprehensive study of different stopping rules for hierarchical clusterization, two of the best functions they have tested are Calínski & Harabasz pseudo-*F* index and Duda & Hart Je(2)/Je(1) index. They are present in several standard statistical packages, however, to our knowledge, not in software capable of working with large text corpora. We thus have developed a piece of software based on one of recent approaches that allows finding jumps in the function measuring internal cluster similarity (Sugar & James, 2003), which is present in the clustering software we have chosen. However, we have not yet found an approach that allows measuring jumps in perplexity function used to assess quality of topic modeling approaches, in particular LDA.

Still another important problem is automatic cluster labeling: if determining what a topic or a cluster is about, demands reading all corpus by a human, algorithms for automatic text grouping are of little use. Automatically generated labels can take form of: lists of most frequent / most probable words, information about cluster's centroid and the list of texts with shortest distances from it, lists of texts with highest probabilities of belonging to a cluster or to a topic in fuzzy clustering and topic modeling. Carpineto et al (2009) point out that academic developers usually choose to concentrate their effort on quality without much attention to labeling (though they are not directly competing parameters), while commercial developers pay much more attention to speed and labeling than to quality and large data. Our experience with text analysis software confirms this statement.

Finally, blog corpora have their specific problems outlined e.g. in Perez-Tellez et al (2010). First, texts are short which leads to low term frequency and to low separability of groups. Li et al (2007) have offered to extend bodies of posts by bodies of their comments and have demonstrated in experiments that this increases the accuracy of clustering, especially if terms from comments are weighted higher than those from posts. This, however, makes large collections even larger. Big size of weblog data has, on the contrary, lead some researchers to refuse from full-text clustering and to find more greedy solutions (e.g. Agarwal et al, 2010). Another cause of low discriminability of blog corpora is general character of topics and overlapping vocabulary – that is, topics in blogs have much less topic-specific words than in research papers from different domains of science. At the same time, writing styles in blogs differ much more than in scientific texts, so some topics get formed on the basis of style. Thus, texts with extensive use of abusive vocabulary form separate topics in different datasets, independently of the subjects to which this vocabulary is applied. Most existing algorithms have been tested on non-blog collections (either science or media texts, such as Reuters corpus) that presumably have better separated clusters. Research on algorithms of community detection in graphs shows that algorithms which demonstrate similarly good quality on graphs with strong community structure, deteriorate with varying speed on graphs with overlapping communities. Unfortunately, we have not found such tests for text clustering or topic modeling algorithms.

4.2. Algorithms and software used

Having acknowledged the described above limitations, we have decided to choose both from classic cluster analysis and topic modeling. Among about 30 clustering packages studied, gCLUTO (http://glaros.dtc.umn.edu/gkhome/cluto/gcluto/overview (accessed 19 April 2012) turned to be one of few methodologically transparent and able to handle large data. Unfortunately, it does not do any text preprocessing, demanding a series of scripts to be developed, and it has been not easy to make it display Cyrillic in its outputs. At the same time it contains four algorithms (direct, agglomerative, repeated bisection and graph), several similarity measures, several criterion functions, several quality functions: internal – intercluster and intracluster similarity – and external – entropy and purity. The latter can be used for algorithm selection if a training collection is available. The former can be used to determine the optimal number of clusters for an unknown collection, through calculation of the function's jumps. Criterion functions in gCLUTO are the methods to calculate distances between clusters, and the sets of all distances for a particular solution are maximized by an optimization algorithm in direct clustering and at each stage of hierarchical clustering. All algorithms and other features are published and tested on various data, including multidimentional (text) data (e.g. Rasmussen, 2004; Zhao & Karypis, 2004, 2005).

Our coders have created a benchmark sample of three hundred Russian language posts belonging to three different topics (Islam, football and earthquake in Japan). Basing on gCLUTO developers' tests, we have selected two algorithms – agglomerative and repeated bisection, and two criterion functions called *I2* and *H2* that performed best in their tests. We have set the cosine similarity as constant to all four combinations because it this measure that allows to compensate for the different length of texts and thus to compare them correctly. For our data, the leader is the combination of repeated bisection with *H2* (entropy 0,14 compared to 0.47-0.6 for other combinations; purity 0,94 compared to 0.62-0.75 for others).

Using this combination, we got six cluster solutions both for August-September and December datasets with step k=50, and then five more around the cluster solution k=100, with step 10, obtaining the total of eleven solutions for each sample. Using our software, we calculated the jump of the internal quality function, ISim (average pairwise distance within clusters of a given solution). The optimal solution for September turned to be k=120, for December k=130. We labeled clusters manually basing on four top-words and on random samples of texts from each cluster, which did not leave an impression of sufficiency.

Out of about 20 topic modeling packages, for none of which we have found sufficient description and sound criteria for choice, we have chosen Stanford Topic Modeling Toolbox, for Linux (http://nlp.stanford.edu/software/tmt/tmt-04/, accessed 19 April 2012): unlike most other tools, it has been developed for social scientists with modest scripting skills and has been used for large data, including web-data (Ramage et al, 2010). It applies basic LDA (Blei, 2003) and Labeled LDA variant offered by its developer (Ramage et al, 2009), as well as a possibility to trace a change in topic structure between two datasets. An in-built quality function is perplexity. TMT does a lot of text preprocessing, except lemmatization, and works with Cyrillic. TMT output includes full matrices of texts' weights of belonging to topics, and of words' probabilities of belonging to topics. Topics, thus, can be easily labeled with top-words and characterized by top-texts. In our experience, top 20 words usually allow defining a broad topic class (political, recreational, noise), while reading top 20 texts most of the time allows specific topic labeling. Each topic is assigned a weight, or a score, which is a sum of all probabilities with which texts are assigned to this topic normalized in certain way. One of the problems of TMT is scarce description, which is partially compensated with the open code.

We have used the same datasets with TMT as with gCLUTO: August-September and December. For both periods, multiple solutions with varying number of topics were obtained, and the graph of perplexity dependence on this number was plotted, that – not unexpectedly – has demonstrated a rather monotonous decline. Since we have not so far found a sound proof of applicability of the jump approach to perplexity function, we had to make a choice between different solutions based on visual analysis of the graph and on hand-coding of topics' interpretability. The solution with 100 topics was selected for both datasets.

Unlike with clustering, we could not apply external quality measures to LDA. It produces highly overlapping clusters, so it was difficult to test it on our single-label benchmark set and even harder to compare results directly with those for clustering algorithms. In general, evaluation of topic models is a new and still underdeveloped area of inquiry (Wallach et al, 2009). One of the problems of external evaluation is that topic models are aimed at discovering latent cluster structures, and not only at detecting those known beforehand. Therefore, comparisons with hand-coded collections may be not always the best way to assess quality of topic models. All this left us little choice except subsequent interpretation of the obtained topics.

Results

Hand-labeling of TMT and gCLUTO clusters has shown that approximately two thirds of clusters are easily interpretable by a human and can be called domain clusters, that is containing texts on a specific subject area or event. The remaining one third is shared between language, "style" and noise clusters. Language clusters contain texts in languages other than Russian (Ukrainian and English); style clusters are centered around writing styles – e.g. offensive vocabulary – or excessive use of specific terms: proper names, digits, measures or computer terms. Noise contains uniterpretable texts or uninterpretable combinations of meaningful texts. The presence of noise might indicate the excessive number of topics / clusters in the given solutions, however, some domain clusters obtained from these solutions contain important political events that are not visible in solutions with fewer clusters. Solutions with fewer groupings draw a better overall picture and detect larger topics that change less with time (politics & state, recreational activity, private life etc). Solutions with more subdivisions detect smaller topics, while larger split into sub-topics, so their structure may be better studied. Thus, Islam as one of the test topics did not appear unless the number of subdivisions reached one hundred, and even then it appeared as a component of neighboring topics (terrorism, Lybian conflict, Israeli-Palestinian relations etc). Elections & protests topic, on the contrary, at the level of 100 clusters is well divided into 6-15 meaningful subtopics, the concrete number depending on the period and the method applied.

In addition to modeling two periods separately, we have also used the TMT function of diachronic analysis. It first merges the data from the two given periods, though keeping the labels of a period to which each text belongs, and divides the merged collection into a given number of topics as a single dataset. Then using the period-labels it calculates how the weight of a given topic in each period differs from that in the merged collection, thus detecting growth and decline of topics. It thus assumes that all topics exist in both periods, which is a simplification, but on the whole the approach is convenient for quick automatic comparison of two datasets – whether belonging to different periods, different authors or editions. This approach has also shown a pronounced growth of elections & protests related topics in December, with decline of a number of other topics (Syrian conflict, new academic year, travel etc).

One of the lines of the analysis, namely separate modeling of each period, was extended to March 2012 – a period following the presidential elections whose results have aroused as much doubt in the opposition as those of the parliamentary elections. General topic structure is very stable across all three periods, although its proportions change moderately (see table 1). Two large groups may be discerned within the cluster of domain topics. Social & political issues group includes international relations, Russian politics, social issues and some economics. Would there be no elections & protests, social issues would be the most volatile group. Culture & private issues group demonstrates manifest stability, for the composition of its four sub-groups – culture, recreational activity, consumption and private life – virtually does not change, but for seasonal holidays. The private life group includes personal relations, family and everyday issues.

Gaining general topic structure allows further fast selection of texts belonging to topics of researcher's interest and further in-depth manual analysis. Having cast a closer look at elections

& protests topics, we are able to describe the composition of this group of messages and their communicative roles. Basing on separate clustering of the three periods by TMT, this area counts to the number of 13 topics in December, when it reaches almost 15% of topic weight. Four groups of topics may be discerned within this subject area. A first group of six subtopics is re-transmission of news from on-line media which is, however, different from them by its selective agenda shifted in favor of conventionally oppositional viewpoint (firing the *Kommersant* newspaper top-management, arrest of political activist Udaltsov, registrations and refusals to register presidential candidates etc). A second group of four subtopics contains opinion messages: short emotional utterances about political characters, namely about protest actions participants and presidential candidates, and long sophisticated speculations on forthcoming presidential elections, as well as for and against street actions. A third group of two subtopics are alternative news – personal reports from visiting street actions and from being observers at the parliamentary elections. The last very specific topic contains results of voting, turnout, exit polls and regular polls and conclusions about the legitimacy of the parliamentary elections based on juxtaposing these data. What is worth mentioning, LJ posts are NOT used for direct mobilization and coordination of protest activity, at least massively; rather they promote specific agenda and discussion, while mobilization activity may presumably be found in Twitter and social networks.

In March, the proportional weight of elections & protests subject area declines to about 9% - protests may be considered a lost game by the moment Putin has won the elections. However, the proportion of all social & political topics becomes bigger than both in August-September and in December. That is, while in December elections & protests topic expanded at the expense of other social and political issues, in March the social and political topic expanded at the expense of protests & elections AND noise. It brought about widening of the spectrum of issues addressed within socio-political domain (32 non-elections social & political topics in March compared to 20 in December). The enrichment of this sector of the LJ agenda may be seen as one of the consequences of the protests that have been fruitless in the sense of direct influence on the elections outcome, however did not pass without other social consequences.

Acknowledgements

The research is carried out by the Internet Studies Lab of National Research University Higher School of Economics (NRU HSE), St.Petersburg branch, and is supported by the NRU HSE Research Foundation, Moscow, Russia, grant No 11040006, 2011-2012.

References

1. Agarwal, N.; Galan, M.; Liu, H. and Subramanya, S.(2010) “Clustering of Blog Sites Using Collective Wisdom”. In Abraham, A.; Hassanien, A.E. and Snášel, V. (eds) *Computational Social Network Analysis: Trends, Tools and Research Advances.* Springer, p. 107-134. *Computer communications and networks* book series. DOI: 10.1007/978-1-84882-229-0.
2. Alexanyan, K. and Koltsova, O. “Blogging in Russia is not Russian blogging”. In Russel, A. and Echchaibi, N. (eds) *International Blogging: Identity, Politics and Networked Publics*. Peter Lang, 2009.
3. Andrews, N.O and Fox, E.A. (2007). “Recent Developments in Document Clustering”, October 16, 2007. http://eprints.cs.vt.edu/archive/00001000/01/docclust.pdf. (Accessed April 17 2012).
4. Blei, David M.; Ng, Andrew Y.; Jordan, Michael I; Lafferty, John. (2003). “Latent Dirichlet allocation”. *Journal of Machine Learning Research* 3: pp. 993–1022. doi:10.1162.

5. Carpineto, C.; Osiński, S.; Romano, G. and Weiss, D. (2009). “A Survey of Web Clustering Engines”. *ACM Computing Surveys (CSUR)*, Volume 41, Issue 3, Article No. 17.
6. Etling B., Alexanyan, K., Kelly, J., Faris, R., Palfrey, J. and Gasser, U. “Public Discourse in the Russian Blogosphere: Mapping RuNet Politics and Mobilization”. Berkman Center Research Publication No. 2010-11. October 19, 2010. http://cyber.law.harvard.edu/publications/2010/Public_Discourse_Russian_Blogosphere (accessed 31 July 2012).
7. Gorny, E. (2004) “Russian LiveJournal: National specifics in the Development of a Virtual Community”. Available at: *Russian-cyberspace.org* http://www.ruhr-uni-bochum.de/russ-cyb/library/texts/en/gorny_rlj.pdf. (accessed 05 April 2012).
8. Herring, S. C., Scheidt, L.A., Bonus, S. and Wright, E. (2005). “Weblogs as a bridging genre”. *Information, Technology & People, 18(2)*, 142-171.
9. Koltsova, O. (2011) “Coverage of Social Problems in St.Petersburg Press”. In: Feilitzen, C. von and Petrov, P. (eds). *Use and Views of Media in Sweden & Russia*. Södertörn Academic Studies, no. 44, Mediestudier vid Södertörns högskola, no. 2011:1, Huddinge.
10. Lauret, D.; Lewson, C.; Yule, P. and Vogel, C. *Internet Research Methods: A Practical Guide for the Social and Behavioral Sciences*. Sage.
11. Li, B.; Xu, S. and Zhang, J. (2007). “Enhancing Clustering Blog Documents by Utilizing Author/Reader Comments”. In *Proceedings of the 45th ACM Southeast Conference (ACMSE 2007)*, pp.94-99, March 23-24, 2007, Winston-Sale, North Carolina http://portal.acm.org/citation.cfm?id=1233359 (accessed 16 February 2011).
12. Milligan, G.W. and Cooper, M.C. (1985). “An Examination of Procedures of Determining the Number of Clusters in Data Set”. *Psychometrika* Vol. 50, No. 2, 159-179.
13. Mizral A. (2009). “Weblog Clustering in Multilinear Algebra Perspective”. *International Journal of Information Technology*, Vol. 15 No. 1.
14. Papacharissi, Z. (2007). “Audiences as Media Producers: Content Analysis of 260 blogs”. In Tremayne, Mark (ed). *Blogging, Citizenship and the Future of Media*. NY&London: Routledge.
15. Perez-Tellez, F.; Pinto, D.; Cardiff J. and Rosso, P. (2010). “Characterizing Weblog Corpora”. H. Horacek et al. (Eds.): *NLDB 2009, LNCS 5723*, pp. 299–300.
16. Ramage, D., Dumais, S. & Liebling, D. (2010). “Characterising Microblogs with Topic Models”. *ICWSM 2010*. http://research.microsoft.com/pubs/131777/twitter-icwsm10.pdf (accessed 4 August 2012). Wallach, H., Murray, I., Salakhutdinov, R. & Mimno, D. (2009). “Evaluation methods for topic models”. Proceedings of the 26 th International Conference on Machine Learning, Montreal, Canada, 2009.
17. Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). “Labeled LDA: A supervised topic model for credit attribution in multi-label corpora”. *EMNLP 2009*.
18. Rasmussen, M. and Karypis, G. (2004). “gCLUTO: An Interactive Clustering, Visualization, and Analysis System”. *UMN-CS TR-04-021*.
19. Sugar, C. and James, G. (2003). “Finding the Number of Clusters in a Data Set: An Information Theoretic Approach”. *Journal of the American Statistical Association*, 98:750–763.
20. Wu, S.; Hofman, J.M.; Mason, W. and Watts, D.J. (2011) “Who Says What to Whom on Twitter”. *International WWW Conference 2011*, March 28–April 1, 2011, Hyderabad, India. ACM 978-1-4503-0637-9/11/03.
21. Zhao, Y. and Karypis, G. (2005) Hierarchical Clustering Algorithms for Document Clustering. Data Mining and Knowledge Discovery, Vol. 10, No. 2, pp. 141-168.
22. Zhao, Y. and Karypis, G. (2004). Emprical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering. Machine Learning, 55, pp.311-331.
23. Zhong, S. (2005). Efficient online spherical k-means clustering. In IEEE *International Joint Conference on Neural Networks*, volume 5, pages 3180-3185.

TABLE 1. Topic structure of three periods modeled separately, 100-topic solution

	Topics’ weight shares & quantity, by period		
	August-September 2011	December 2011	March 2012
Social and political issues	32% (37 topics) *International relations:* US – Middle East relations; Russian-Ukrainian-Belorussian relations; Russia – Caucasus – Central Asia relations; Arab-Israeli relations & Islam; Asia; Caucasus; Moldova; Syria; 9/11. *Russian politics & history:* general politics, governmental projects & decision making; parties & elections; protest actions; Putin & Medvedev; nazi crime; military; aircrafts & tanks; World War 2; Soviet history; general history. *Economics*: oil & Russian economy; industry; world economic crisis. *Social issues*: law & statehood; Motherland – friends & foes; courts & trials; crime & police; church & orthodoxy; faith, God, sin; transport; space; population problems; higher education; new academic year; air crush in Yaroslavl; Perm events; Dagestan university incident.	33,7% (33 topics) *International relations*: US – Middle East relations; Russian-Ukrainian-Belorussian relations; Russia – Israel & Russian-Jewish relations; history of Russia – Caucasus relations; Kim Chen Ir’s death. *Russia politics & history*: governmental projects & decision making; regional administration; political slogans; Word War 2 & aviation; military; *Elections & protests* (14,9%): opinions for & against protests; protests at Sakharova & Bolotnaya – personal accounts; protests – media reports; protest participants – media reports; protest participants – opinions; authorities’ reaction on protests – media reports; authorities’ reaction on meetings – opinions; protesters’ arrests; firing *Kommersant* journalists; parliamentary elections results; observers’ reports on parliamentary elections; anticipating presidential elections; registration of candidates for presidency. *Economics*: industry; world finance; China & world economics. *Social issues*: church & orthodoxy; education; corruption; crime & courts; automobiles in the streets, space.	41% (40 topics) *International relations*: US – Middle East relations & UN; Belorussian politics, Baltics & Poland; Ukrainian politics; US politics. *Russian politics & history*: power, politics & economics; political ideologies; governmental policies; Moscow & regional administration; military transport; military & Ulianovsk; World War 2; Soviet history; mixed opinionated texts. Elections & protests (8,8%): Putin’s victory; voting & falsifications; elections, including mayoral; March protests & Navalny; protesters’ arrests; protests, TV & Sobchak; Shein’s hunger strike; law, society & freedom. *Economics*: financial markets & crisis; large business & policy on it; state budget & wages; employment, companies & their clients. *Social issues*: space & nuclear energy; public transport; healthcare & medicine; terrorism & related crime; nationalism & migration; Jews, Arabs, Judaism & Islam; orthodoxy; Pussy riot arrest; scandal with patriarch’s flat; misconduct & tortures in police; criminal proceedings; accusations against Shuvalov;
Cultural & private issues	30,4% (39 topics) *Culture*: architecture; music & concerts; literature & writing; book publishing; visual art; movies; photography; television; computers; Internet & blogs; LJ; Twitter; art auction in England, philosophy. *Recreational activity:* cars & races; football; pets; cooking receipts; competitions; fortune-telling; tourism; eco-travel; nature& fishing; weather. *Consumption & goods*: fashion & consumption; wine consumption; on-line games; phones; internet-trade. *Private*: love & personal relations; sex; sex, smoking & drinking; wedding; family; children; work & money gaining; personal health; doctors & hospitals.	32,6 % (39 topics) *Culture*: architecture & churches; museums & art; theaters & concerts; book publishing; movie about Vysotsky; television; photography; online photo & video; blogs (2 topics) *Recreational activity*: New Year & Christmas; transporting; football; animals; general cooking receipts; sweet cooking receipts; competitions; tourism to European cities; eco-travel & India; nature; weather; communication (letters, phones, internet). *Consumption & goods*: restaurants; clothes & body; mobile devices & internet; housing – buying; housing – repair; goods through internet (2 topics); announcements *Private*: poems & romantic photos; love & interpersonal relations; sex & beauty; family & children; home, family & work; work, salary & credits.	31% (40 topics) *Culture*: memorials & buildings; museums & exhibitions ; literature & books; ideas, knowledge & culture; science & research; concerts, albums & movies; movies & *Men in Black*; photography; internet, blogs, social networks & infographics (5 similar topics). *Recreational activity:* Easter, cars, football, pets, animals & zoo; cooking receipts; seasons & weather; nature; travel & rest; tourism – Asia; miss Belarus competition. *Consumption & goods*: restaurants & food; fashion & clothes; drinks; cars& traffic jams; banks; photos & toys; computers & mobile devices; internet & Yandex; goods through internet. *Private*: sex; love, coupling & wedding; family & relatives; family, children & parenting; home & everyday life.
Other	37,6% (24 topics) *Language*: Ukrainian & English language texts (3 clusters) *“Style”*: offensive vocabulary; calendar; personal names; English-language Internet terms (2 clusters) *Noise*: uninterpretable & mixed (14 topics).	33,7% (28 topics) *Language*: Ukrainian & English language texts, English-Russian translation (4 topics) *“Style”*: offensive vocabulary; calendar; measures; English-language scripts. *Noise*: uninterpretable & mixed (18 topics).	28% (20 topics) Language: Ukrainian texts, English-Russian translations *“Style”*: offensive vocabulary, personal names, calendar, digits & measures (8 topics). Noise: uninterpretable (10 topics)