The Text Network Analysis: What Does Strategic Documentation Tell Us About Regional Integration?

A.Murashov[1], O.Shmelev[2]

[1] Yaroslav-the-Wise Novgorod State University
(murashov.andrey@mail.ru)
[2] Kalashnikov Izhevsk State Technical University

Abstract. Values and attitudes towards the regional integration process of the Russian political elites are considered as an indication of what regional integration (RI) tends to be and how it evolves over time. This paper suggests how to systematically grasp and integrate elite's attitude into the analysis of RI by means of text network analysis. The text network analysis allows one to visualize the meanings and agendas present within political manifests which are supposed to reflect values and attitudes towards RI of the local political elite.

Keywords. igraph, political elite attitude, R, regional integration, regional strategy, system of indicators, text mining, text network analysis (TNA)

1 Introduction

This paper is a part of PhD thesis aimed at constructing a so-called System of Indicators of Regional Integration in Russia. Values and attitudes towards the regional integration process of the Russian political elites are considered as an indication of what regional integration tends to be and how it evolves over time. One of the shortcomings of conventional approaches is insufficient and unsystematic consideration of political elite's attitude towards regional integration and decision-making in this field. The question how to systematically grasp and integrate elite's attitude into analysis remains open.

In comparative politics measuring the attitudes of the political elite is often undertaken either by expert surveys or by the analysis of political manifests [De Lombaerde P. et al., 2011]. Our interest lies in researching the political manifests – a regional strategy reflects opinion of a local authority, like a party manifest directly reflects opinion of a political party.

Research questions are those related to monitoring and analyzing regional integration process. How do regions cooperate? What forms of integration emerge within the selected regions? What problems / challenges does integration impose? Which industries are mainly affected by integration process? How do regions choose their region-partners? What is beyond the choice?

2 Background

From theoretical perspective this paper is supposed to contribute to the investigation of values and attitudes towards the regional integration process that are represented in political manifests. This topic is covered in particular by:

(1) Comparative Manifesto Project (CMP) maintained by Manifesto Research Group. Their purpose is to discover party stances by quantifying their statements and messages to their electorate, method used is quantitative content-analysis [CMP, 2014];

(2) Leontief Center's Study of Russian Regions' Strategies aimed at, among other things, building a classification of regional strategies based on their content, method used is expert review and content-analysis [Zhikharevich B. et al., 2013];

(3) Philippe De Lombaerde from United Nations University, Institute on Comparative Regional Integration Studies (UNU-CRIS) and his team who employing multidisciplinary approach in developing quantitative and qualitative tools to monitor regional integration process [De Lombaerde P. et al., 2011].

3 Method

From methodological perspective this paper applies an approach which combines two methods - comparative text-mining and graph analysis – "text network analysis". The text network analysis allows one to visualize the meanings and agendas present within political manifests. This approach outputs a graph of relations between key terms where each node represents a term and edges express logical associations between terms.

Putting it in a general scenario of social networks, the terms are taken as people and the segments of text as groups on LinkedIn or Facebook, and the term-document matrix can then be taken as the group membership of people. Several notions of co-occurrence have been used in the literature to group words [Saeedeh M. et al., 2010]: document-wise/sentence-wise /window-wise/syntax-wise co-occurrence. We build a network of terms based on their co-occurrence in the same text segments (paragraphs) extracted from the documents in the course of expert review. There is an edge between two terms if they appear in the same text segments (paragraphs). The weight of an edge is its frequency [Batagelj V. et al. 2002, Polanco X., 2006]. Such a network (or conceptual map [Chernyak E. et al., 2014]) visualizes logical associations between concepts presented in the political manifests.

1. Establish text corpus and transform it

Data to analyze is regional strategies of socio-economic development as a central and most capacious source of information about political elite's views on regional integration process. We are interested in 6 Russian regions situated alongside the Moscow – St Petersburg transport corridor: Moscow, Moscow region, Tver' region, Novgorod region, Leningrad region, St Petersburg. Their strategies are studied. There may exist a wide range of other official documents on regional integration but unfortunately we are not able to cover all of them, so we decided to limit our sample by regional strategies only.

Using Atlas.ti (qualitative data analysis software) we establish text corpus and retrieve those text segments (paragraphs) from the regions' strategies which refer to regional integration process, refine it in a specific way then (lemmatization, filter stopwords, punctuation and numbers removing, etc.).

2. Explore text corpus (igraph & tm packages)

Text network analysis is performed with R [Yanchang Zhao, 2012], specifically, with packages {igraph} and {tm} (provides functions for text mining). We build a document-term matrix, after that, it is transformed into a term-term adjacency matrix, where the rows and columns represent terms, and every entry is the number of concurrences of two terms, after that, frequent words and their associations (fastgreedy.community) are found from the matrix.

3. Visualization

Finally, we visualize the result by means of {igraph} package in R environment: (1) plot the graph to show the relationship between frequent terms (graph.adjacency, layout = layout.fruchterman.reingold, delete.edges), (2) dendrogram (dendPlot).

4 Results

First we review general graph statistics. Snapshot of the network metrics is in the table following (tab.1). Volume of the strategies varies from 396 vertices for Moscow region to 800 vertices for Tver' region. Function assortativity.degree uses vertex degree (minus one) as vertex values. The coefficient throughout the corpora is negative suggesting that the connected vertices tend to have different degrees. Centralization is a general method for calculating graph-level centrality score based on node-level centrality measure. Novgorod region's strategy is that one having most centralized structure (centralization degree of 68% of its theoretical maximum). We also arrive at the conclusion that there is a substantial amount of centralization in the Moscow region's strategy. In general, the power of individual terms varies rather substantially, and this means that, overall, positional advantages are rather unequally distributed in each strategy. The global version of clustering coefficient (function transitivity) indicates that the degree to which nodes in a graph tend to cluster together is relatively low. This makes sense since we removed from the graphs singular edges for the sake of simplicity (here we refer to a parameter n which is discussed below). Fastgreedy algorithm identifies from 6 to 10 communities in the graphs with moderate modularity. As we can see from the table 1 the graphs are quite similar in terms of their mathematical conception. Much more insightful and interesting results come from analysis of the networks' content.

Table 1. Graphs' key metrics[1]

Parameter	{igraph} function	Strategy					
		SP	M	LO	NO	TO	MO
Number of vertices	vcount	737	463	490	491	800	396
Number of edges	ecount	5039	2311	2517	2913	7296	1897
Assortativity	assortativity.degree	-0.25	-0.34	-0.32	-0.24	-0.31	-0.29
Transitivity	transitivity	0.19	0.22	0.16	0.21	0.22	0.16
Average path length	average.path.length	2.54	2.64	2.48	2.43	2.52	2.58
Graph density	graph.density	0.019	0.022	0.021	0.024	0.023	0.024
Centralization Degree	centralization.degree	0.49	0.41	0.52	0.68	0.48	0.68
Centralization Closeness	centralization.closeness	0.54	0.48	0.53	0.67	0.50	0.61
Centralization Betweenness	centralization.betweenness	0.30	0.30	0.27	0.55	0.19	0.39
Eigenvector Centrality Scores	centralization.evcent	0.92	0.91	0.92	0.92	0.91	0.92
Diameter	diameter	13	10	13	13	10	14
Number of communities (best split)	fastgreedy.community	6	6	10	8	8	8
Modularity (best split)	fastgreedy.community	0.40	0.49	0.35	0.38	0.32	0.38

To demonstrate some examples for applying the strategies to study regional integration the graphs following are built (fig.1). They are based on the strategy of St Petersburg. The graph (fig.1,a) is crowded with many vertices and edges, it represents most of the ideas we can find in the strategy. To simplify the graph we remove insignificant terms. With function delete.edges, we remove edges which have weight less than a certain value. To do it in our experiment we introduce a parameter n referring to a number of text segments (paragraphs) where a certain term appears. After removing edges, some vertices become isolated and are also removed. The produced graph is on fig.1,b. The interpretation is that we exclude from the scope of analysis most rare and random concepts.

Let us set n equal to 8. The resulting graph on fig.2,a is crowded with many vertices and edges, we can interpret it at some extent but we need to get more precise picture. We identify vertices whose removal increases the number of connected components in the graph. They are: city, petersburg, development, etc. To simplify the graph and find relationship between terms beyond the selected keywords, we remove major articulation points (or alternatively those terms whose removal, we expect, will lead to a result we are looking for) so that the layout is rearrange and new concepts and links between them are revealed. We see that some of the articulation points are not necessarily meaningful but just the highly frequent words carrying less meaning than those with a moderate or low frequency and are thus not very valuable to explore.

[1] SP = St Petersburg, M = Moscow, LO = Leningrad region, NO = Novgorod region, TO = Tver' region, MO = Moscow region

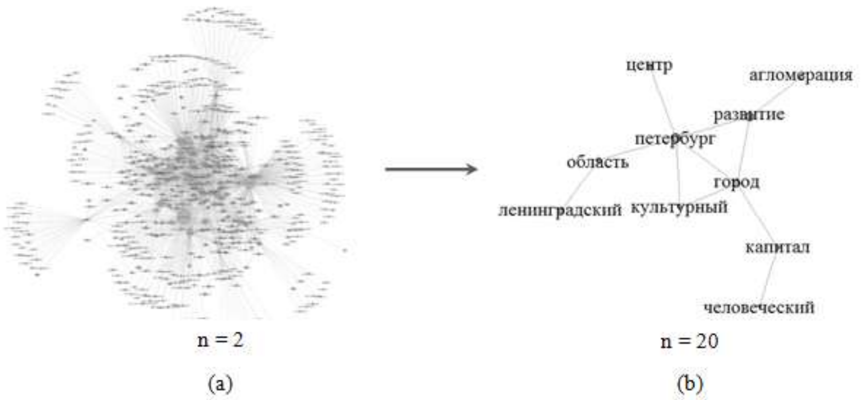

Fig. 1. Example of graph evolution (a – initial graph; b – truncated graph)

Next, we try to detect communities in the graph. Graph community structure is calculated with the fastgreedy algorithm [Kincharova A., 2013]. The nodes that cluster together (communities) are shown with the same color on fig. 2, indicating contextual proximity of the terms used. The communities tell us that the local authorities focus quite heavily on patterns of spatial development , unique role of St Petersburg and its attractiveness for migrants, close association between the City and Leningrad region, etc.

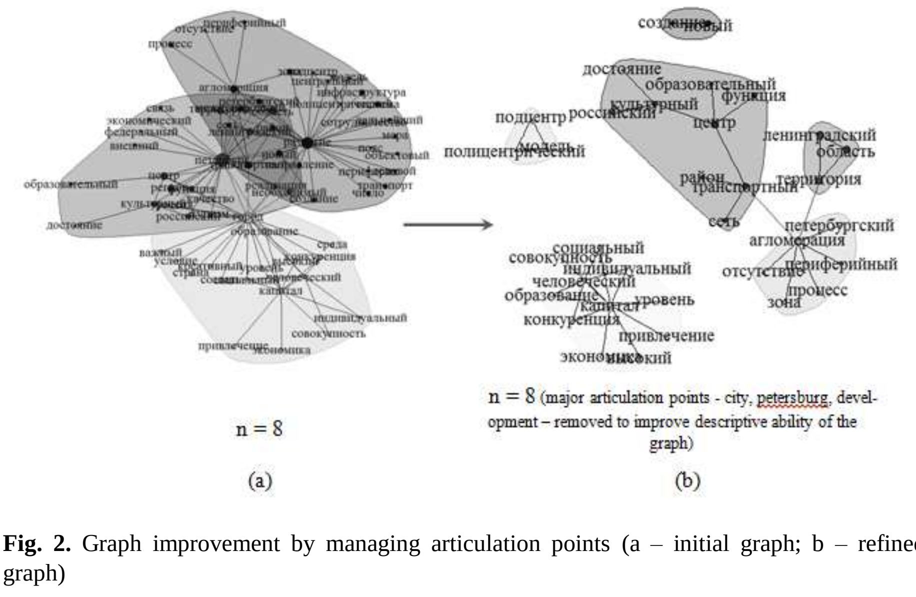

Fig. 2. Graph improvement by managing articulation points (a – initial graph; b – refined graph)

We can also have a further look at which terms colocations are most frequent in each strategy (fig.3). Parameter n tells us how many times the plotted collocations appear. Parameter n is a lower bound of the frequency, that is, collocation «Moscow –

St Petersburg» appears not less than 33 times in different text segments within Tver' region's strategy. Each strategy mentions those regions more frequently which are supposed to be their main partners.

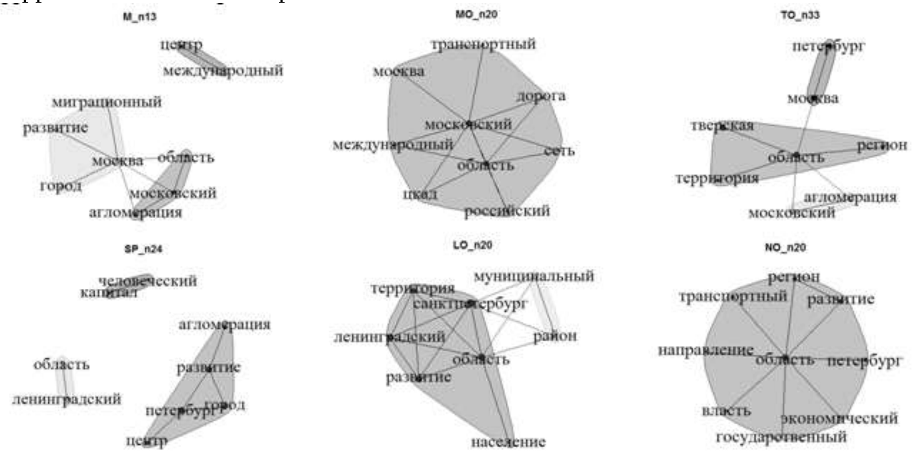

Fig. 3. Most frequent terms colocations

Plots for Moscow region and Leningrad region suggest that the development of the two regions is strongly connected to two largest Russian cities that they surround. For instance, the Leningrad region strategy suggests strong spatial planning on territories adjoined to St Petersburg. The Moscow region strategy repeatedly highlights importance of Moscow and its agglomeration.

5 Concluding remarks

Given the work-in-progress style of our study, it has a number of weaknesses; among them is a lack of the comparative perspective. A solution might be to build a network of text segments (paragraphs) based on the terms shared by them (two-mode network). Alternatively, we would probably arrive at some interesting insights if could put a network on another one to see joint term clusters and terms-outliers [Ermakov A. et al., 2002]. We appreciate the more sophisticated models use phrases or n-grams instead of words and this also could be a possibility to improve our analysis. These basic tools implemented here are only the beginning of possibilities for applications of TNA. It is worth noting that we were not meant to provide an extension of the base technology of computer-supported TNA but suggest an example of its practical implementation in social science.

The text-mining approach combined with graph theory appears to be a valuable method for extracting elite's attitude towards regional integration process from public strategic documentation (allows us to access a large amount of textual material, regional analysis may provide interesting input for text network analysis, etc.). Modeling the data using this method provided us with the specific insights on what local authorities really focus on and how the strategies differ from each other.

References

1. Batagelj V., Mrvar A., Zaversnik M. Network analysis of texts. URL: http://nl.ijs.si/isjt02/zbornik/sdjt02-24bbatagelj.pdf (25.03.2015).
2. Chernyak E., Morenko E., Mirkin B. Conceptual Maps: Construction Over a Text Collection and Analysis. Analysis of Images, Social Networks and Texts Communications in Computer and Information Science Volume №436, 2014, pp.163-168.
3. De Lombaerde P. et al. The Regional Integration Manual: Quantitative and Qualitative Methods. Routledge, London, 2011.
4. Ermakov A., Pleshko V. Informatization and information security of enforcement officials. XI International scientific conference. Conference proceedings, Moscow, 2002, pp. 343-347.
5. Kincharova A. Application of community detection algorithms for sociological investigation of blogs: results of a piloting study. URL: www.hse.ru/data/2013/06/10/1283702757/dzh.pdf (25.03.2015).
6. Manifesto Project Database. URL: https://manifestoproject.wzb.eu/ (25.03.2015).
7. Polanco X., San Juan E. Text data network analysis using graph approach. Vicente P. Guerrero-Bote. I International Conference on Multidisciplinary Information Sciences and Technology, Oct 2006, Merida, Spain. Open Institute of Knowledge, vol. 2, pp.586-592. URL: https://hal.archives-ouvertes.fr/hal-00165964 (25.03.2015).
8. Saeedeh M. et al. A Comparative Study of Word Co-occurrence for Term Clustering in Language Model-based Sentence Retrieval. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pp. 325–328, Los Angeles, California, June 2010. URL: http://www.aclweb.org/anthology/N10-1046 (25.03.2015).
9. Yanchang Zhao. R and Data Mining: Examples and Case Studies. Academic Press, Elsevier, 2012.
10. Zhikharevich B., Zhunda N., Rusetskaya O. Proclaimed and actual priorities of regional and local authorities: approaches to reveal and compare // The Region: Economics and Sociology, 2013, №2, pp. 108 – 132.