2015 International Conference on Computational Science and Computational Intelligence

Visualization of Subtopics of the Thematic Document Collection Using the Context-Semantic Graph

Alexander Sboev[1,2], Ivan Moloshnikov[1], Dmitry Gudovskikh[1], Roman Rybka[1]
[1] NRC "Kurchatov Institute", Moscow, Russia
[2] NRNU "MEPhI", Moscow, Russia
sag111@mail.ru, ivan-rus@yandex.ru, dvgudovskikh, rybkarb@gmail.com

***Abstract*—An approach for visualization of nested topics within large collections of documents is proposed. The approach is based on set of parameters: information entropy, Kullback-Leibler divergence, Ginzburg algorithm, similarity the distributions of keywords and key phrases in the documents with Bernoulli's theoretical distributions. The results of comparisons of our approach with implementations based on TF-IDF approaches are presented.**

***Keywords*-contex-semantic graph, sub topic graph, topic visualization.**

I. Introduction

The analysis of large volumes of unstructured information contains a number of topical tasks, as examples:

1) annotation of collections of documents,
2) visual query building,
3) visualization of the relationship between nested topics,
4) the allocation of sub-topics.

To meet these challenges requires:

1) Allocation of nested topics. A nested topic is an automatically allocated set of weighted words and phrases defining domain specific terminology. Furthermore, this topic must be significantly represented in the collection of analysed documents .
2) Visual representation by a concisely identified visual summary of the collection. To allocate nested topics a set of methods, both statistical LDA, PLSA [3] and based on neural network Doc2vec [5] is used.

There are a number of works using above mentioned methods: Visualization of correlation of latent topics in the documents highlighted by LDA which is performed in several steps [10]:

1) highlighting themes by the LDA method,
2) calculating the correlation by the method of Correlated topic model.
3) visualization using the TopicPanorama for a Full Picture of Relevant Topics

In the TopicNets [11] the authors propose an approach based on statistical topic models that is similar to the LDA.

As shortcomings of these approaches the following may be pointed: a great training corpus must be used, not high precision of results and difficulty to determine necessary level of proximity. In this paper we propose an approach to extract the nested topics based on complex of probabilistic and entropic methods and semantic algorithm of Ginzburg. Our approach allows to display the strength of connection of subtopics in document collection and their keywords. Unlike other approaches it is based on probability and entropy characteristics and it works better for thematically similar documents.

II. Topic modelling

A. Probability-entropic characteristics

The indication based on **Kullback-Leibler divergence** is computed for words or phrases. It compares how the term w is represented in the collection D with the random representation of that term in a document of the collection according to the count of different terms in the document in terms. The word "term" means a "word" if the indicators are used to extract keywords or if the algorithm is used to extract key phrases, the word "term" means a "bigramm" (2 words in sentence regardless of the order of the words).

$$D(w) = \sum_{d \in D} p_{doc}(w,d) \cdot ln\left(\frac{p_{doc}(w,d)}{p_n(d)}\right) \quad (1)$$

$$p_n(d) = \frac{N(d)}{\sum_{x \in D} N(x)} \quad (2)$$

$$p_{doc}(w,d) = \frac{TF(w,d)}{F(w)} \quad (3)$$

Where $TF(w,d)$ is the frequency of the term w in the document d, $F(w)$ is the frequency of the term w in the collection, and $N(d)$ is the count of all terms in the document d, $\sum_{x \in D} N(x)$ is the count of all terms in the collection D

The value of $D(w)$ characterizes the actual distribution of the term w in documents of the collection relatively to the random theoretical distribution of the term w in documents of the collection in the proportion with the number of all terms in documents of the collection. At small values of $D(w)$ the word occurs in the document according with the length of the document and apparently it is the word of common using.

Information entropy is the distribution of terms in the documents in the collection.

978-1-4673-9795-7/15 $31.00 © 2015 IEEE
DOI 10.1109/CSCI.2015.124

$$H(w) = \sum_{d \in D} p_{doc}(w,d) \cdot ln\left(\frac{1}{p_{doc}(w,d)}\right) \quad (4)$$

If $H(w)$ is large, the term is uniformly presented in all the documents of the collection, if $H(w)$ is 0 it means that all terms w are concentrated in a single document.

There are some indicators to compare the distribution of keywords and key phrases in the documents of the collection with **Bernoulli's theoretical distribution** [1]. The probability of term occurrence in a document is given by $Prob_1(w,d)$.

$$Prob_1(w,d) = B(N,F,X) = \binom{F}{TF} p^{TF} q^{F-TF} \quad (5)$$

Where $p = 1/N$, $q = (N-1)/N$ and N is the number of documents in the collection.

The Ginzburg algorithm is designed for extraction of words [2] related by their meanings.

$$ind(a|c) = \frac{N_{ac} \cdot N_t}{N_{tc} \cdot N_a} \quad (6)$$

According to this algorithm, the index of significance of the word A in the context of the word C is calculated by the formula 6, where N_{ac} is the number of occurrences of the word A together with the word C, N_{tc} is the total amount of words in the sentences, where the word C appeared, N_t is the amount of words in the collection of documents, N_a is the number of the occurrences of the word A in the collection.

The Ginzburg connection index is calculated using the significance indexes. The Ginzburg index determines the intensity of the connection between the two words.

$$ginz(a,c) = 1 - \frac{sum(a) + sum\_razn + sum(c)}{sum\_all} \quad (7)$$

Where $sum(a)$ is the sum of the indexes of the significance for the word A, which are more than 1 and do not belong to C. The $sum(c)$ is the sum of the indexes of the significance for the word C, which are more than 1 and do not belong to A. The sum_razn is the sum of the absolute values of differences of the indexes of the significance for A and C. And the sum_all is the sum of all indexes of the significance, which are more than 1.

B. Ranking functions

Using ranking functions different probability-entropic and semantic parameters are combined into a bigramm weight. This weight showed the strength of relationship between two words. The algorithm of ranking function:

This approach allows you to combine multiple parameters, smoothing out the differences in the scales. It has shown good results for the the task of searching thematically similar documents

Algorithm 1 Algorithm of ranking function

```
for each parameters do
    Number the unique value of the parameter in the sort order, from 0 to the number of unique values;
    Normalize the obtained number in the range from 0 to 1;
end for
for each bigramm do
    bigramm weight = summery of normalized parameters;
end for
```

C. Allocation of the subtopic

proposed approach is based on the method of Affinity Propagation Clustering [8] (from scikit-learn library [6]). This method generates clusters on base of relation nodes through neighbours. Application to the text information is based on the relations of words in their context. This algorithm is not requires to set a finite number of clusters and also centres of selected clusters. The centres are the most specific words for subtopics which is strongly associated with the other words and clusters through a common context.

III. Constructing the graph of word connections

We use the previously described probabilistic and entropy metric to allocate the bigramms and key words of the collection. The nodes of the graph are the keywords and the edges are weighted bigramms. The graph reflect the nested themes and the relationship between subtopics for a large collection of documents.

To lay the graph we use the algorithm "Force Atlas 2" [9]. It has worked well in the analysis of the social graph, and clearly reflects the group is strongly connected nodes.

The size of the nodes, the distance and the thickness of the lines reflect how words and phrases characterize the collection of documents. Such a graph is difficult to read, even with the highlight color clusters and strengthening relations between the words of one cluster. The example of a part of the full graph, for the theme "rocket Bulava" based on the collection of 9000 documents of news and blogs, is presented on the Fig. 1.

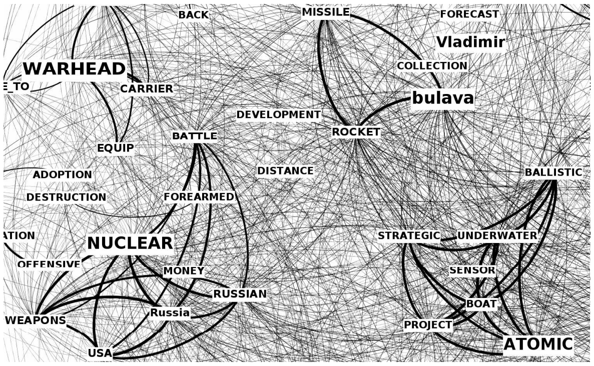

Fig. 1. The part of the full graph, for the theme "rocket Bulava"

The main problem is the lack of clear boundaries of clusters and a large number of connections between words. To solve these problems, we leave only relation between the nodes within the cluster and between the centres of the clusters. This allows you to clear a visual representation. To visualize the relationships between the two clusters of weights of the edges for the nodes of different clusters are combined into one bond that reflects the relationship between subtopics. The example of the part of the revision graph, for fourteen thousand documents containing the word “Arctic” is presented on Fig. 8.

In the graph we can see the two ”poles”, the first relates to the Department of Defence (Fig. 2), and the second relates to develop resources in the northern latitudes (Fig. 3).

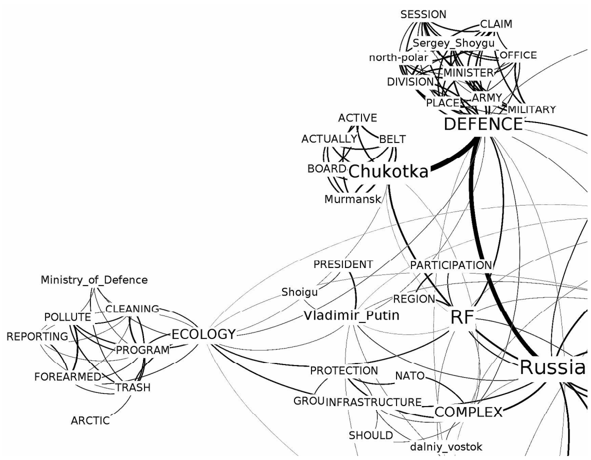

Fig. 2. The part for the Department of Defence of the full graph, in the theme “Arctic”

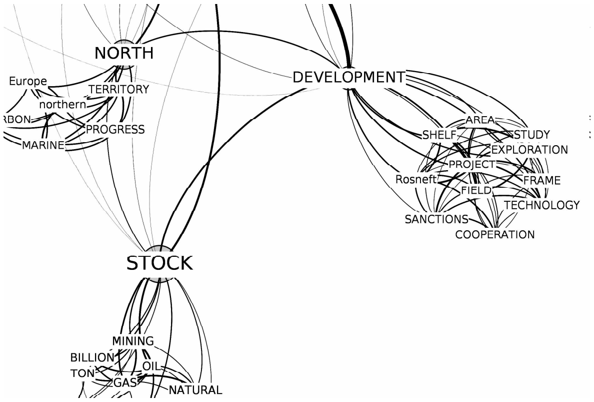

Fig. 3. The part related to the develop resources of the full graph, in the theme “Arctic”

Count Centre is embedded Russian theme, since It is connected with all themes. The shape of the graph indicates the presence of several large so loosely connected with the other one. You can see a strong connection between Chukotka and the Ministry of Defence, this relationship is expressed in the news and blogs related to the conduct of the exercises in this area. On the other hand the news related to the development of resources, little overlap with the topic of Defense, which is clearly seen on the graph.

For visualizing we use the open source software Gephi [7].

IV. Testing

Also we present the results of testing on the corpus SCTM-ru [4], which is a set of labelled news topics from the free news source Wikinews (https://ru.wikinews.org).

The collection for testing was selected from news from the corpus about 1000 documents. The documents with themes labeled were selected on the following topics: ”Space”, ”Russia”, ”USA”, ”Science and technology”, ”Culture”, ”human Rights”.

The distribution of analyzed documents in the collection are presented on the histogram in Fig. 4.

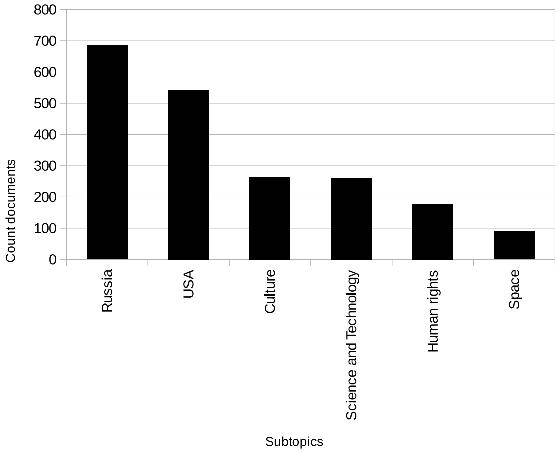

Fig. 4. Distribution of documents on topics in the analysed collections

Our approach identified a cluster corresponding in meaning to the following topics: Russia, USA, Space exploration, Right. We are rated how subtopics are presented in a document collection, by formula 8.

$$Cw(c) = \sum_{d \in D} \sum_{w \in W(c)} TF(w,d) * R(w) \quad (8)$$

Where $Cw(c)$ is weight of subtopic c in collection and d is a text of collection of documents D, and w is word of words in subtopic c ($W(c)$), and $R(w)$ is the summery weight of term w. The calculation results can be seen in Fig. 5.

The test results shows that the proposed approach was able to allocate 4 of the 6 clusters: Russia, USA, Right and Space. This is set occurrence of them in the same order as that noted by people.

Also, an experiment was conducted on the same data, but using only measures TF-IDF for weighing words and bigramms. The result is shown in Fig. 6.

As you can see, the test results shows that the TF-IDF approach was able to allocate 2 of the 6 clusters: Russia and USA.

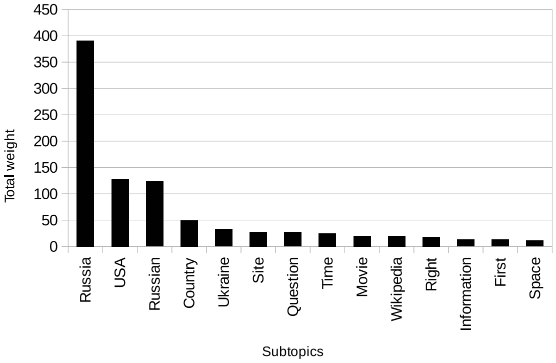

Fig. 5. The distribution of topic weights in the collection

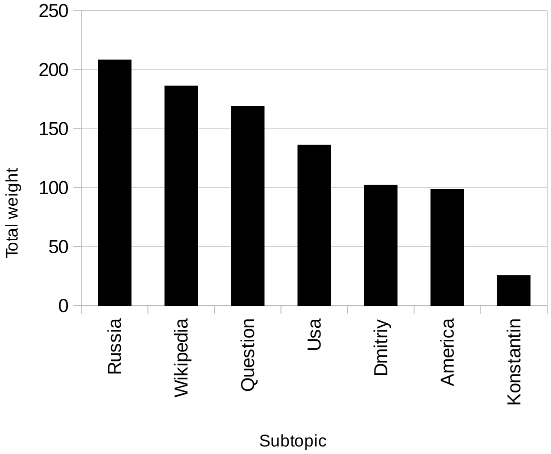

Fig. 6. The distribution of topic weights in the collection

A. Relevant documents for subtopics

In this visualisation there is no rigid connection between the subtopic and documents relevant to this subtopic. But we can weigh each document belonging to a selected attachment topic by the formula 9, similar to the previously proposed:

$$Dwc(d,c) = \sum_{w \in W(c)} TF(w,d) * R(w) \tag{9}$$

Using the $Dwc(d,c)$ weight of the document we can evaluate the quality of the clusters for the task of searching for documents, the most relevant for selected subtopic. For this purpose, we construct the ROC curves and calculate the area under the curve, the result is shown in Fig. 7

In three selected topics AUC was significantly higher than 0.5, indicating that the suitability of this method to search for documents relevant to subtopic.

B. Emotiveness of subtopics

We combined the proposed approach for searching documents by cluster with approach for analysis of emotiveness of texts referred in article ”A quantitative method of text emotiveness evaluation on base of the psycholinguistic markers founded on morphological features” [12]. This paper shows that the value of the of emotiveness text reflect the degree of emotional exhilaration author of the text at the time of writing. Values are calculated based on psycholinguistic markers of text that are allocated using morphological characters of words. For this purpose, we have analysed the corpus of news and blogs (about 7,000 documents) on the topic ”Armata tank” for days near the Victory Parade date. The results of summery emotivnes rank for subtopics is presented in table I.

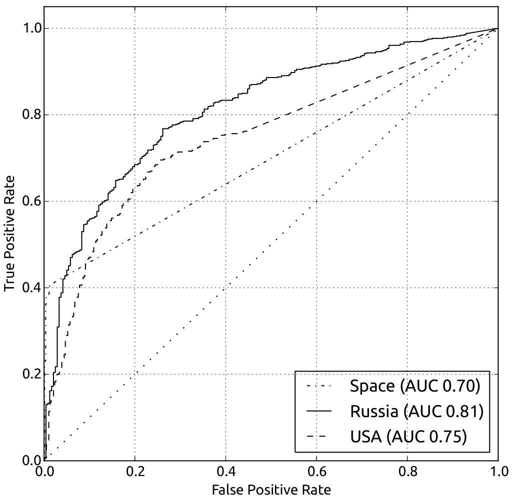

Fig. 7. ROC for subtopics relevant documents

TABLE I
EMOTIVENESS OF SUBTOPICS IN THEM ”ARMATA TANK”

Cluster name	Emotiveness
REPETITION	321
VICTORY	318
SAMPLE	317
ALPHA BANK	316
TANK	315
IMMORTAL	306
TOWER	304
PUTIN	301
PRESIDENT	297
GREAT	292
BOOMERANG	291
Armed with	287
ARMENIA	286
WORLD	271

As you can see, the themes related to the parade and the new tanks have a greater of emotiveness value that is a hotly debated topic with a large number of excited users reviews.

V. Conclusion

The proposed approach allows to allocate much present themes (Russia, USA) or poorly represented (Kosmonavtika, Right). General topics such as science, culture, "fall apart" into several clusters (Wikipedia, channel, Movies, etc.). In the case of using the standard TF-IDF measure to compute the connection between words, the results become much worse and rougher. It highlighted major themes, Russia and the USA, and the rest of the words go astray in an uncertain cluster.

The proposed approach are well fit for the analysis of documents, united by a common theme, such as the Arctic. It can be used to reflect in succinct form the results of a search and visually annotate collections of documents. A further development consists in to building a dynamic graph that reflecting the dynamics of change theme. This can be useful for the analysis of the events with long histories, for example, to search for the crucial event that dramatically changed the composition of nodes and links in the graph.

Use of syntax parser will allow to build the bigramms founded on syntactic tree. This will improve the formation of relationships of the graph as compared to an approach based on a bag of words. Using the services for allocation of named entities gives us possibility to reflect the main objects of the subtopics, such as a persons, organizations, geographical places, etc. Application of morphological analyser enables us to remove homonymy and in the future, to build grammar rules and then to filter out insignificant words in the graph. Using service for definition of emotiveness and tone of the texts we will allow to add new dimensions for the subtopics, reflecting the mood of the users in the analysis of social media.

Further development of the visualization is to add a brief annotation for subtopics. This can be done on the basis of the selection of sentences which relevant to words and phrases in the cluster. It will give to users a possibility of easily navigation in graph, as opposed to if the graph consists only of keywords.

The proposed approach can be applied for visual modelling topics for the search. It can be used in various systems aimed on analyzing the various topics on the internet, for example, market research on new products. Also, this approach can be applied when setting news streams. The user can selects interesting topics and deselects uninteresting.

VI. Acknowledgment

This study was supported by the Russian Foundation for Basic Research (project 15-29-01173 ofi_m).

References

[1] Amati, Gianni and Van Rijsbergen, Cornelis Joost, "Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness", *ACM Transactions on Information Systems (TOIS)*, vol. 20, 2002, pp. 357–389

[2] I.E. Voronina, A.A. Kretov and I.V. Popova, "ALGORITMY OPREDELENIYA SEMANTICHESKOJ BLIZOSTI KLYUCHEVYX SLOV PO IX OKRUZHENIYU V TEKSTE", *Vestn. Voronezh. gos. un-ta. Seriya Sistemnyj analiz i informacionnye texnologii*, 2010, pp. 148–153

[3] Blei, David M and Ng, Andrew Y and Jordan, Michael I, "Latent dirichlet allocation", *the Journal of machine Learning research*, 2003, pp. 993–1022

[4] Karpovich S.N., "The Russian language text corpus for testing algorithms of topic model", *SPIIRAS Proceedings*, 2015, pp. 123–142

[5] Le, Quoc and Mikolov, Tomas, "Distributed Representations of Sentences and Documents", *Proceedings of The 31st International Conference on Machine Learning*, 2014, pp. 1188–1196

[6] Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E., "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, 2011, pp. 2825–2830

[7] Mathieu Bastian and Sebastien Heymann and Mathieu Jacomy, "Gephi: An Open Source Software for Exploring and Manipulating Networks", *International AAAI Conference on Weblogs and Social Media*, 2009.

[8] Frey B. J., Dueck D. "Clustering by passing messages between data points" *science*, 315.5814, 2007, pp. 972-976.

[9] Jacomy M. et al. "Forceatlas2, a continuous graph layout algorithm for handy network visualization", *Medialab center of research.*, 2011

[10] Liu S. et al. "Topicpanorama: a full picture of relevant topics" *Visual Analytics Science and Technology (VAST)*, 2014 IEEE Conference on. IEEE, 2014, pp. 183-192.

[11] Gretarsson B. et al., "Topicnets: Visual analysis of large text corpora with topic modeling", *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, 2012, pp. 23.

[12] "A quantitative method of text emotiveness evaluation on base of the psycholinguistic markers founded on morphological features", A. Sboev, D. Gudovskikh, R. Rybka1, and I. Moloshnikov, in press.

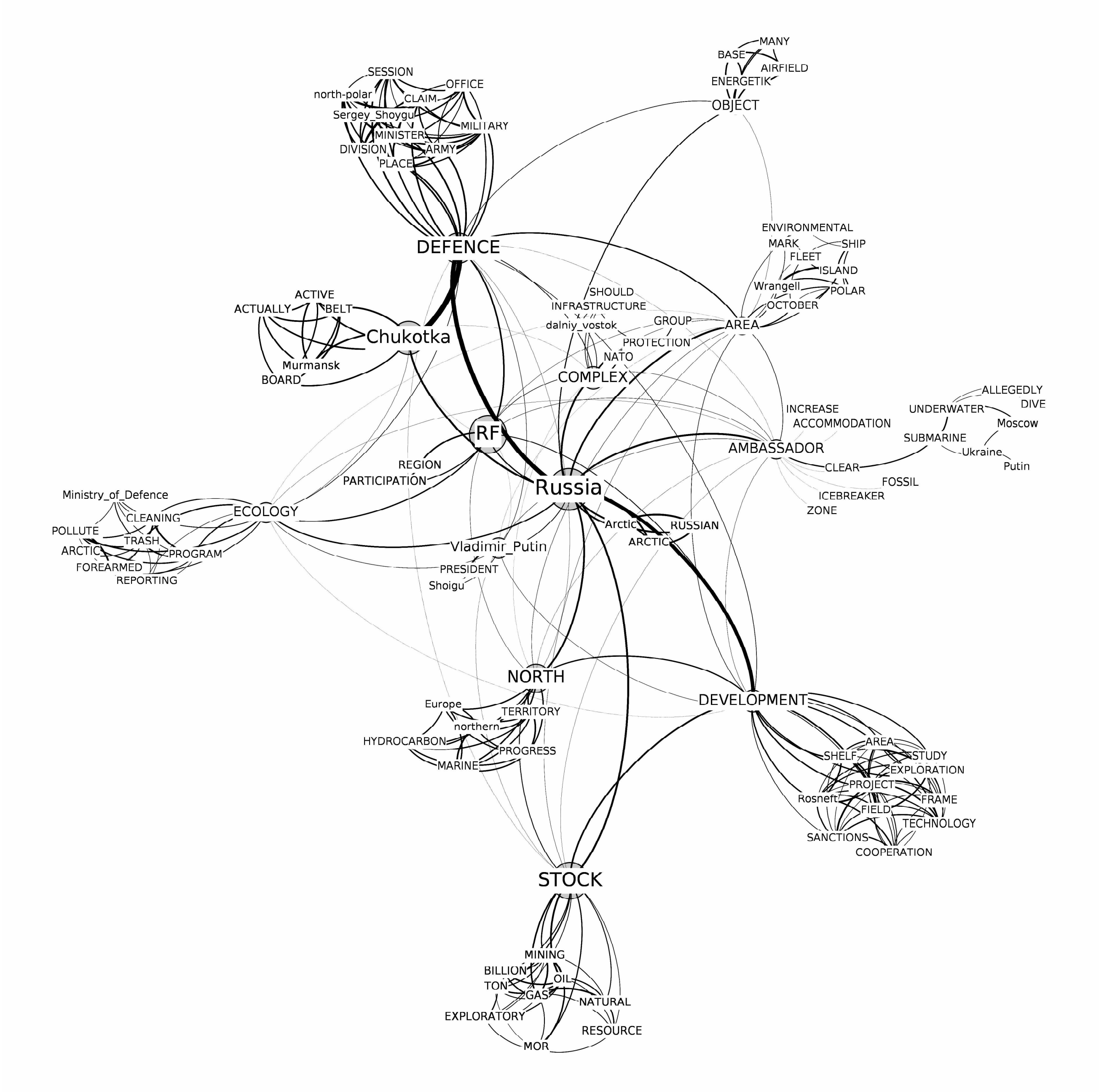

Fig. 8. The full graph, for the theme “Arctic”