A probabilistic-entropy approach of finding thematically similar documents with creating context-semantic graph for investigating evolution of society opinion

Moloshnikov I.A., Sboev A.G., Rybka R.B., Gydovskikh D.V.
NRC “Kurchatov Institute” (Academic Kurachatov sq. 1, Moscow), Department of mathematical modelling and information technologies, Russia
E-mail: ivan-rus@yandex.ru, sag111@mail.ru, rybkarb@gmail.com, dmitrygagus@gmail.com

Abstract. The composite algorithm integrating, on one hand, the algorithm of finding documents on a given topic, and, on the other hand, the method of emotiveness evaluation of topical texts is presented. This method is convenient for analysis of people opinions expressed in social media and, as a result, for automated analysis of event evolutions in social media. Some examples of such analysing are demonstrated and discussed.

1. Introduction
The analysis of people opinions expressed in social media is the very important tool to investigate the level of social tension and indicate emerging critical situations. This analysis is complex one, it may be divided into the following four steps:

(1) selecting topical texts,
(2) selecting subtopics,
(3) emotiveness analysis of the subtopics,
(4) analysis of the temporary evolution of the subtopics and the whole theme.

Currently it is difficult to point out some systems to solve this problem in whole. Concerning points 1 and 2, the search engines, based on Apache Lucene, with implementations of the “bag of words” method may be mentioned. But despite such their advantages, as fast performance and universality, it can be difficult to use them for analysis of key word evolution in time and, as a result, for analysis of temporally changing topics (point 4). As for the use of statistical LDA, PLSA [1] and Doc2vec [3] based on neural network, their shortcomings are the following: the great training corpus must be used, not high precision of results and difficulty to determine necessary level of proximity. In this work, the method based on a combination of several probabilistic and entropic algorithms for extracting keywords and collocations with including additional information sources, such as the Russian National Corpus, is used for topic analyzes.

The emotiveness evaluation of topics which has been implemented in this work, is based on the set of special markers and the context-independent approach, proposed in [8] and briefly...
described in chapter 5. It allows to determine the emotional background of topics according to point 4 and demonstrate the emotional strengths of subtopic connections of documents in collection and also connections of keywords in document collection.

2. Search thematically similar documents
To find thematically similar documents the full-text search in Solar system, with setting weights of keywords and phrases in the boost query were used [7]. For this purpose, 15 words and 15 bigrams from the reference collection, with the highest weights calculated by the ranking function, were selected. This approach had shown the best results in comparing of different approaches on the corpus SCTM-ru [2], which is a set of labelled news topics from the free news source Wikinews.

2.1. Theme modelling
To simulate the theme for searching we compute relevant weights for keywords and key phrases (combinations of two or three words in one sentence), using the following indicators for terms:

- Kullback-Leibler divergence for the comparing of the term distribution in documents with theoretical distribution;
- Information entropy represents uniformity distribution in documents;
- The Bernoulli Model of Randomness distribution of terms;
- The Ginzburg semantic algorithm to determine the thematic proximity of the words.

The word “term” means a “word” in case of using the algorithm to extract keywords, and in case of using the algorithm to extract key phrases the word “term” means a “phrase”.

We combine this indicators to one term weight using ranking functions. The algorithm of ranking function:

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 term do
    term weight = summary of normalized parameters;
end for

This approach allows one to combine multiple parameters, smoothing out differences in the scales. The word weight shows how the word is important in the theme. And for bigrams this weight shows the strength of relationship between two words in a theme.

3. Extracting of subtopics
A subtopic is an automatically allocated set of weighted words and bigrams defining subtopics in the main theme of analysed documents collection. Furthermore, this topic must be significantly represented in the collection of analysed documents. Also, our proposed method allows to compute the strength of the relationship between nested topics.

The algorithm of extracting of subtopics is based on the construction of a graph of the relationship between the keywords of theme (bigrams). It consists of several steps:

1) allocation of 100 words with the highest weight of relevance to topic,
(2) formation bigrams based on previously selected keywords,
(3) filtering bigrams with weight below the arithmetic mean
(4) generating the affinity matrix, based on bigrams weights,
(5) applying the Affinity Propagation [5] clustering method for extracting subtopics

This algorithm does not require to set a number of clusters or centres of clusters. This is a big advantage in the analysis of the various thematic collections of documents when the user does not know the number of subtopics. The centres of cluster are the words which are most specific for subtopics, i.e. strongly associated with the other words and clusters through a common context.

To compute relationships between the clusters, the weights of the edges for the nodes of different clusters are combined into one bond that reflects the relationship between subtopics.

4. Visualizing
We construct a context-semantic graph for visual annotations of large thematic collection of documents. The nodes of the graph are keywords, and the edges are key bigramms obtained during the analysis of search results using presented earlier methods. The size of the nodes, the distance and the thickness of the lines reflect how words and phrases characterize the collection of documents.

5. Methods of emotiveness evaluation
The approach presented in this work is based on the psycholinguistic diagnosis of stylistic features of the text with application of natural language processing methods. Initially we chose the following types of psycholinguistic markers from the article “A quantitative method of text emotiveness evaluation on base of the psycholinguistic markers founded on morphological features” [8]. This paper shows that the emotiveness value of a text reflects the degree of emotional exhilaration of the 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. There are indicators that reflect the psychological state of the author at the time of writing the text. In this article we use:

- the ratio of verbs to adjectives,
- the ratio of verbs to nouns,
the ratio of the number of verbs and their forms (participles and gerunds) to the number of all words,
• the ratio of prepositions to the total number of words.

We combine this markers to calculate emotive weight of document using ranking function similar to 1.

6. An analysis of the dynamics of topics

We propose a method for visualising of the dynamics of topics. It is based on sequential analysis of the flow of messages on a sliding time window algorithm 2.

Algorithm 2 Algorithm of dynamic graph

| DynGraph = dynamic graph |
|--------------------------|
| list_of_time_windows = overlapping time windows with size of 6 hours, and step of 1 hour |
| for time window in list_of_time_windows do |
| wgraph = graph for one window; |
| extract 10 keywords from time window documents; |
| add keywords with weights as nodes into wgraph; |
| extract key bigrams on base of 10 keywords; |
| add bigrams with weights as edges into wgraph; |
| for each node in wgraph do |
| node visual size = node weight * sum(of adjoining edges weight)/count(of adjoining edges); |
| if node in list of nodes of DynGraph then |
| append to list of node visual size with time of start and end for window; |
| else |
| add new node to DynGraph with visual size and time of start and end for window; |
| end if |
| end for |
| for each edge in wgraph do |
| if edge in list of edges of DynGraph then |
| append to list of edges new weight with time of start and end for window; |
| else |
| add new node to DynGraph with weight and time of start and end for window; |
| end if |
| end for |
| end for |

The graph is generated in a GEXF format (http://gexf.net/format/), which allows one to specify different values of weights for nodes and edges in the corresponding periods of time. For visualization we use the program Gephi [4] and online laying algorithm Force Atlas 2 [6]. An example of such a graph is shown in figure 2. It is seen as a term of “Greenland” is growing. This is associated with a gradual increase in number of messages containing a discussion of the ice cover in the Arctic and Greenland.

7. The emotiveness evaluation of subtopics

To find documents of the subtopic we form a boost query to Solr. It consists of the key words and phrases of the cluster with given weights. We analyze only the documents which contain at least one phrase from a subtopic. Solr calculates a score for each document from the search results. This reflects the weight of whether the document belongs to the cluster.
7.1. Emotiveness of subtopics

We combined the proposed approach for searching documents by cluster with an approach for analysis of emotiveness of texts. For each document of cluster overall rank of emotiveness is calculated, based on the selected psycholinguistic markers and ranking algorithm, similar to 1. Total emotive weight of the subtopic is calculated by the formula 1.

\[ E(c) = \sum_{d \in D(c)} S(d) \cdot M(d) \]

\[ \frac{N(c)}{E(c)} \]  \hspace{1cm} (1)

Where \( E(c) \) is the total weight of the cluster \( c \). And \( d \) is the document of subtopic, \( D(c) \) is a collection of documents belonging to the cluster, \( M(d) \) is the total weight of emotiveness of document, calculated on the basis of psycholinguistic markers, \( N(c) \) is the number of documents in the cluster.

We analysed the corpus of news and blogs (about 7,000 documents) on the topic “Armata tank” for days near the Victory Parade date. Graph of subtopics is shown in figure 1(a). The thickness of the lines shows the context strength of the connection between embedded themes. The size of the nodes show the summary of emotiveness rank for subtopics, that is presented in table 1.

“Immortal Regiment” is a public event, held in Russia on Victory Day, during which the participants carry banners with photographic portraits of their relatives who participated in the Great Patriotic War. Subtopics “Repetition” and “Armata” refer to the description of Victory Parade rehearsal and ascendent with the breakdown of the new Armata tank. As you can see, the themes related to the parade and the new tanks have a greater emotiveness value, therefore it is a hotly debated topic with a large number of excited users reviews.

7.2. An example of dynamic of subtopics

To reflect the dynamics of the embedded threads, we use 3 characteristics with a window at 6 hours:

- the number of documents in the window
- the relative score of documents in the cluster for window
- relative emotiveness weight \( M(d) \) of documents for windows

For the subtopic of “Alfa-Bank” these graphs are shown in figure 1(b). This topic relates to the discussion of financial claims of Alfa Bank to the manufacturer of the new Armata tank. From these graphs it can be concluded that the main peak of messages came in the dates from
### Table 1. Emotiveness of subtopics

| Subtopics     | Emotiveness | Subtopics     | Emotiveness |
|---------------|-------------|---------------|-------------|
| IMMORTAL      | 0.60        | TOWER         | 0.27        |
| REPETITION    | 0.58        | ARMENIA       | 0.24        |
| ARMATA        | 0.47        | SAMPLE        | 0.16        |
| ALPHA-BANK    | 0.33        | PRESIDENT     | 0.15        |
| VICTORY       | 0.31        | WORLD         | 0.11        |
| GREAT         | 0.31        | ARMED         | 0.09        |
| BOOMERANG     | 0.30        | PUTIN         | 0.06        |

8 to 9, but basically, the topic was discussed in passing, it is seen by the average values in the second chart. On the 3rd graph of relative emotiveness, there is a clearly visible decline of emotional stress in the subject with the beginning of the working week, which means rather small interest in the topic.

### 8. Conclusion

The system which is suitable for analysis of people opinions expressed in social media based on selection of thematic documents of given topic and emotiveness evaluations has been designed and tested. The presented algorithms have been applied for emotiveness evaluation of subtopics in frame of whole topic for some thematic examples. The results demonstrate that hotly debated subtopics with large number of reviews of excited users have higher values of emotiveness. Thus, the proposed composite approach is good to reflect a given topic in the form of a weighted set of keywords and phrases along with evaluation of the emotiveness of subtopics. It is convenient for the analysis of themes developing in time and social processes. The further development of this algorithm will be directed to adding a visualization in form of thematic-emotion graph with brief annotation for subtopics.

### Acknowledgment

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

### References

1. Blei, David M and Ng, Andrew Y and Jordan, Michael I, 2003, “Latent dirichlet allocation”, *the Journal of machine Learning research*, 993–1022
2. Karpovich S.N., 2015, “The Russian language text corpus for testing algorithms of topic model”, *SPIIRAS Proceedings*, 123–142
3. Le, Quoc and Mikolov, Tomas, 2014, “Distributed Representations of Sentences and Documents”, *Proceedings of The 31st International Conference on Machine Learning*, 1188–96
4. Mathieu Bastian and Sebastien Heymann and Mathieu Jacomy, 2009, “Gephi: An Open Source Software for Exploring and Manipulating Networks”, *International AAAI Conference on Weblogs and Social Media*
5. Frey B. J., Dueck D., 2007, “Clustering by passing messages between data points” *science*, 315.5814, 972-976.
6. Jacomy M. et al., 2011, “Forceatlas2, a continuous graph layout algorithm for handy network visualization”, *Medialab center of research*.
7. I. Moloshnikov, A. Sboev, D. Gudovskikh, R. Rybka, 2015, “An algorithm of finding thematically similar documents with creating context-semantic graph based on probabilistic-entropy approach”, *Procedia Computer Science*, in press
8. A. Sboev, D. Gudovskikh, R. Rybka, and I. Moloshnikov, 2015, “A quantitative method of text emotiveness evaluation on base of the psycholinguistic markers founded on morphological features”, *Procedia Computer Science*, in press.