ВИЗУАЛИАЗИЯ ТЕКСТОВ В ВИДЕ РЕФЕРЕНТНЫХ ГРАФОВ

Черняк Е. Л. (echernyak@hse.ru),
Дубов М. С. (mdubov@hse.ru),
Миркин Б. Г. (bmirkin@hse.ru),
Шишкова А. С. (ashishkova@hse.ru)

Национальный исследовательский университет
«Высшая школа экономики», Москва, Россия

Ключевые слова: анализ текстов, визуализация текстов, веб-корпус, аннотированное суффиксное дерево

REFERENCE GRAPH AS A TOOL FOR TEXT VISUALISATION

Chernyak E. L. (echernyak@hse.ru),
Dubov M. S. (mdubov@hse.ru),
Mirkin B. G. (bmirkin@hse.ru),
Shishkova A. S. (ashishkova@hse.ru)

National Research University Higher School of Economics,
Moscow, Russia

This paper presents the ongoing work on developing a tool for text collection visualization. We suggest visualizing a collection of texts as a so-called reference graph. The nodes of this graph stand for key words and key phrases extracted from the texts. There is a directed edge between two nodes A and B if node A refers to node B. The reference relation between two key words or phrases is defined in a way similar to the association rule technique. The resulting visualization of the text collection shows some hidden relations between the key words and phrases. The reference graph can be investigated by graph-theoretic algorithms for further analysis of the text collection. To test the visualization technique we collect our own Web-based collection of Russian-language newspapers. Several examples of reference graphs are provided. The annotated suffix tree measure is used throughout the paper to measure the relevance of a key word/phrase to a text.

Keywords: text visualization, web-corpus, annotated suffix tree, text analysis

1. Introduction

The concept of text visualization is rather ambiguous. The most straightforward approach to text visualization is generation of a scene, described in a text. For example, given the text "There is a room with a chair and a computer" the visualization tool should first infer that the desk should support the computer and the chair should stand in front of the desk, and next draw the scene [5]. However another approach to text visualization achieves currently more attention. The main idea of this approach is to plot important elements of the text (such as key word or phrases, named entities, terms). Such pictures can be seen as a tool for text summarization and information extraction / presentation [19]. The most known text visualization technique is tag clouds [7]. The tag cloud shows the key words or phrases (i.e. tags) extracted from a text on a plain. The size of the tag depends on its frequency or any other statistical feature. There are dozens of tag cloud services on the Web, such as Wordle, TagCrowd, TagCloud, etc. Sometimes the tags may form a cloud, a flower or a heart, whatever the user prefers. Tag clouds help to achieve a very general understanding of the text and are nowadays usually used as navigation tool on a web-site [9]. The majority of text visualization techniques extend the idea of tag cloud. In [20] the tags extracted from tweets were color-coded according to the politics of the user. Vennclouds, introduced in [6] are an extension of the tag cloud idea. Instead of one tag cloud, a Venncloud presents three tag clouds, which are used to contrast two texts. One tag cloud presents the key words and phrases of the first text, the second presents the key words and phrases of the second text and the third tag cloud presents the words and phrases two texts have in common. In [19] the tag clouds are placed inside the nodes of the graph and the nodes are connected by an edge if they have a lot in common. Furthermore, the nodes are sorted according to the time axis. This way a metro map of a temporal text collection is constructed. Another extension of the tag cloud idea is the tag graph. To achieve the tag graph one needs to introduce some sort of relation between the tags. For example, in [11] the tags stand for named entities and the edges between them show whether they co-occur. The layout of such graph is another research question. For example, in [1] the tag graph is divided in several compounds. The numbers of compounds coincided with the number of sources and every compound presents source-specific tag. Latent topics might be also a subject of visualization [1, 18].

Our project of text collection visualization belongs to the tag cloud direction. We construct so-called reference graphs, where nodes stand for key words and phrases, which are extracted from the whole collection. There is a directed edge between two nodes A and B if node refers to node B: $A \Rightarrow B$. The referral relation between two nodes $A \Rightarrow B$ shows that the key word or phrase B occurs with a higher probability if the key word or phrase A occurs. Hence the reference graph is a directed graph, which is a very well studied mathematical structure. This gives us plenty of opportunities for further analysis.

The paper is organized as follows. Section 2 presents the text collection, its construction strategy and evaluation. Section 3 is devoted to key words and phrases extraction. The method for reference graph construction is described in Section 4. The resulting visualization is presented in Section 5. Section 6 lists future directions for our project. Section 7 concludes.

2. Data

Choosing the test text collection we decided to follow the trend of natural language processing, that is newspaper analysis [2, 10, 11, 19].

We have limited the range of text sources to a relatively small set of Russian newspaper and newsportals ("Izvestia", "Nezavisimaya gazeta", "Moscow Komsomoltes", "Kommersant"), to be more precise—to their sections devoted to economics topics. The texts that form the basis of our collection are the materials published on Web-sites and in RSS feeds of those newspapers. All the text sources used for collection construction are distributing their content with a free license, which lifts any restrictions regarding the usage of their texts. Following [17], the usage of a limited set of sources allows us to overcome technical difficulties intrinsic to text collection fetched using Web-crawlers: there is no need of automated text language detection, html markup removal, text deduplication, dealing with occasional advertisement messages and other types of noise in the collected texts. Consequently, one of the features of our collection is both high accuracy of text preprocessing and high quality of the collected data.

We process and aggregate all the articles published in 2014. This gave us a total of 4,061 articles (1,109 from "Kommersant", 1,061 from "Izvestia", 1,284 from "Nezavisimaya gazeta", and 613 from "Moscow Komsomoltes"). Using our tokenizer, based on regular expressions, the texts were split into 10,032,509 tokens, among them 130,138 unique tokens. There are 578.35 tokens per text on average. Other collection properties are the following. The most frequent part of speech is noun (57,832 unique tokens were annotated as nouns by pymoprhy2 [15] which is an open source Russian morphological parser trained on Open Corpus [14]), followed by full adjectives (30,852 unique tokens) and verbs (16,013 unique tokens). Prepositions and conjuctions are significantly less frequent: only 108 and 109 unique were annotated as prepositions and conjunctions correspondingly. These numbers are quite natural for the Russian newspaper style of writing [16].

3. Key word and phrase extraction

We follow the key word and phrase extraction strategy proposed in [8], which consist of two main steps. On the first step the candidate words and phrases are extracted from the texts. The candidate words and phrases should satisfy certain part of speech patterns. Then on the second step the candidate word and phrases are sorted according to their frequency. The most frequent words and phrases form the resulting set of key words and phrases.

To apply this procedure, we need to define the part of speech patterns. Let us define a key word as a single word noun and a key phrase a phrase of two or more words that satisfy a certain part of speech patterns, such as NOUN + NOUN or ADJECTIVE + NOUN or NOUN + PREPOSITION + NOUN, etc. The whole list of patterns was adopted from [13]. We set a threshold for frequency of candidate phrases and select only frequent phrases. We calculate frequency of the candidate phrase in the whole corpus, not in individual texts. Finally, we achieve a list of phrases that satisfy grammar patterns and are frequent enough. We chose the threshold for frequency empirically so that we get top 250 candidate phrases and top 100 candidate words. We remove phrases, that are newspaper-specific but not semantically important ("Izvestia reporter" ["korrespondent Izvestij"], for example) manually and consider the remaining key phrases. Using only the part of speech patterns we can not filter such phrases, but their removal is necessary because the frequently co-occur with other key phrases and hence refer to them. Note, that there are not many such phrases and their automatical removal can be easily conducted if a special list of such ley phrases is constructed.

This approach to key phrase extraction has several advantages. First of all, it is easy. Since all the texts in the collection belong to the same domain and are written using specific vocabulary, there is no need for more complex extraction procedure. Secondly, it allows us to get coherent key phrases of different length. Replacement of manual key phrase processing with some computational techniques, which take newspaper specific vocabulary into account, is an important part of future work.

4. Reference graph construction

The reference graph construction method is based on the procedure of scoring key phrase to text relevance. Because the key phrases are extracted from the whole collection, we do not know how relevant they are to individual texts. We use annotated suffix tree (AST) scoring to compute key phrase to text relevance in the same fashion as it is presented in [12]. This scoring takes all fuzzy matches between the key phrase and the text into account. It helps to cope with some typos and replaces stemming in a sense (see [12]). For example, the Porter stemmer will stem the words "Ukraina" and "ukrainskij" as "ukrain" and "ukrainsk", but the AST procedure allows us to detect the matching fragment "ukrain". Then while scoring the word "Ukraina" both "Ukraina" and "ukrainskij" would be taken into account.

Using AST scoring we estimate the relevance of every key phrase to every text. If the relevance value is lower than the given threshold, we suppose the text is not about this particular key phrase. Usually we set up the relevance threshold at the level of 0.2, which makes up around a third of the maximum experimental AST relevance value. Given the relevance threshold we define the set of texts, which are relevant for every key phrase. Let us denote key phrases as k_i, $i = 1:n$, and let $F(k_i)$ be the set of texts, relevant to key phrase k_i. Let us consider that key phrase k_i refers to key phrase k_j ($k_i \Rightarrow k_j$), if the number of texts which belong both to $F(k_j)$ and $F(k_i)$ makes out a significant part of $F(k_j)$:

$$\frac{|F(k_j) \cap F(k_i)|}{|F(k_j)|} > r$$

where r is the confidence threshold and belongs to the (0.5, 1) interval. This gives us the structure of the referrals between key phrases and can be represented as a graph, where nodes are key phrases and edges are referral. We also introduce the support threshold in a way similar to associative rule framework [3]:

$Support(F(k_i)) = |F(k_i)|$ and use for further analysis only those key phrases, whose support value are higher than the given threshold. From the associative rule framework we inherit the problem of the confidence and support thresholds selection. Both are very important, but there is no technique to set them automatically. The association rules are found with user defined minimum support and confidence values. So do we. We set the relevance threshold at 0.2, the confidence threshold at 0.7 and the support threshold at 5.

5. Reference graph visualisation

As soon as we get the set of referrals $k_i \Rightarrow k_j$, their confidence and support values we can plot the reference graphs. For the sake of space we replace key words and phrases with their index numbers. The size of the node depends on the support value. The nodes are color-coded in the following way: the green nodes only refer to other nodes, the violet nodes are referees and are only referred by other nodes, the rest of the nodes are of intermediate type and are blue.

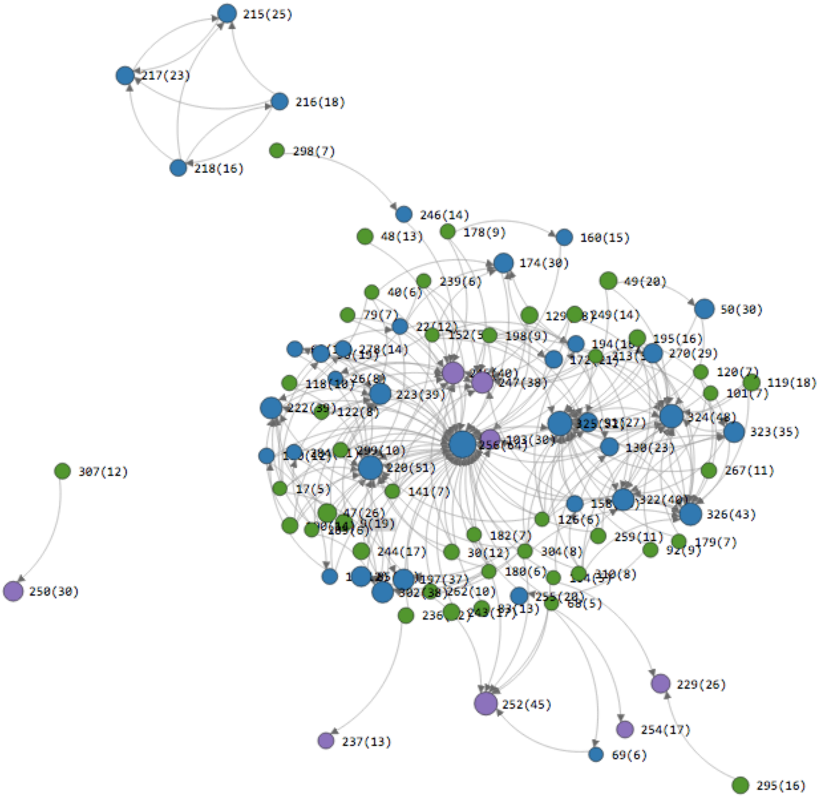

Fig. 1. Reference graph for "Nezavisimaya gazeta", December 2014

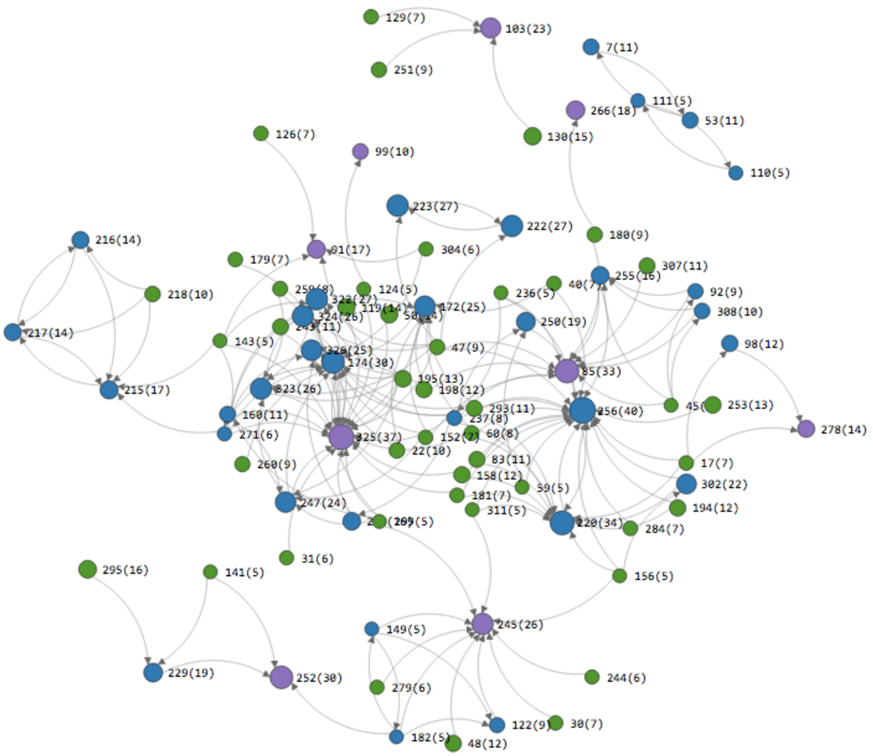

Fig. 2. Reference graph for “Moscow Komsomolets”, December 2014

Let us consider two reference graphs constructed for two different newspapers (“Nezavisimaya gazeta” (Fig. 1) and “Moscow Komsomolets” (Fig. 2)) based on articles published in December 2014. First of all, the graph are of similar size: there are 88 nodes in the first graph and 85 nodes nodes in the second graph. The graphs are of different shapes: the first one in centered around the node 256 (“Russian Government” [“Rossijskoe Pravitel’stvo”]), that has the highest support. The second graph is sparse and there is no clear center. The highest support get the nodes 256 (“Russian Government” [“Rossijskoe Pravitel’stvo”]) and 325 (“Economic growth” [“Ekonomicheskij rost”]). The both graph share in common a strongly connected component of four nodes 215, 216, 217, 218, that describes consumer behavior (“Consumer price” [“Potrebitel’skaja tsena”], “Consumer credit” [“Potrebitel’skij kredit”], “Consumer demand” [“Potrebitel’skij spros”], “Consumer lending” [“Kreditovanie potrebitelej”]), which is no surprise. However there are little intersections in content of the graphs. The nodes “Vladimir Putin” and “Dmitry Medvev” are absent in the second graph. There are four nodes that deal with Ukraine in the first graph, and only two of them appear in the second. At the same time, there is a node “Saudi Arabia” [“Saudovskaja Aravia”] in the second graph. The majority of nodes in the second graph are Russian Government and Ministry of Finance related, while in the first graph the majority of nodes relate to ruble devaluation and business in Russia.

These two graphs clearly show the difference between two newspapers. “Nezavisimaja gazeta” being more politics and business oriented presents the year end situation in Russia as crisis, while the “Moscow Komsomolets” is more oriented towards international relations of Russia and consumer needs. This can be proved by diversion of key words and phrases that are included in the graphs, their support values and their relations to other nodes.

It might not be very clear at this stage of the project how can we use the edge direction. We will provide several ideas in the next section.

6. Future work

Let us list future directions for our project.

1. **Analysis of reference graphs.** It is necessary to test some methods for graph analysis such as clustering nodes, measuring centrality, finding cycles of minimal length, bridges, connected components, which can provide us with some insights. For example, calculating centrality with HITS algorithm will allow us to achieve two types of important nodes: hubs, that only refer to other nodes, and authorities, that are referees. In our preliminary experiments “Russian Government” [“Rossijskoe Pravitel’stvo”] is usually the main authority. This fact confirms the dependence of Russian market and economy on Russian Government decisions. Another way to measure centrality in reference graphs would be using the PageRank algorithm. Clustering nodes of reference graph may serve as latent topic detection technique.
2. **Temporal analysis**. Since our text collection is obviously a sort of temporal textual data, we might extend reference graphs to time-depended case. This will allow us to detect trends and/or events in newspapers by finding temporal references between key word or phrases that occurred yesterday and today’s key word or phrase.
3. **Coloring the nodes of the reference graph.** For this purpose we plan to use the LDA [4] or LDA-like methods to group key phrases into latent topics and color the corresponding nodes of the graph according to the topic mixture.
4. **Text preprocessing improvement.** Further directions of text processing module development include implementation of such functions as syntactic annotation of input publications, word sense disambiguation and disambiguation for morphological analysis. As it was said above, we need to develop some filters that distinguish between newspaper-specific vocabulary and general vocabulary.
5. **Open access.** The final step of this project will be providing public access to the web service for the newspaper collection and visualisation. This requires some technical efforts like developing security and stability modules.

7. Conclusion

In this paper we had presented an ongoing project on developing a tool for text visualization. Our approach is based on the idea of tag cloud: drawing key words and phrases (tags) on the plain and provide, so that some of their features such as frequency or negativity/positivity are represented by the size or the color of the tag. Despite the fact we do not actually plot the tags, but only their indexes, we extend the idea of the tag cloud to the directed graph of the tags. Every node of so-called reference graph is a key word or phrase. The edges of the graph represent reference relation, which means that if node A refers to node B, B is more likely to co-occur with A. The reference graphs can serve not only as a visualization tool, but also as a tool for further text analysis. We have tested the method for reference graph construction and visualization on newspaper collection and plan to continue our research in several directions, including time-depended analysis by means of graph-theoretic and latent topic detection. Preliminary experiments nevertheless show that reference graphs represent some deep characteristics of the texts.

Acknowledgements

For collaboration: Artiom Shal, Egor Morenko, Nikita Levitsky. The project is being developed by the “Methods of web corpus collection, analysis and visualization” research and study group under guidance of prof. B. Mirkin (grant 15-05-0041 of Academic Fund Program).

References

1. *Alexander, E., Kohlmann, J., Valenza, R., Witmore, M., Gleicher, M.* (2014). Serendip: Topic model-driven visual exploration of text corpora. Proceedings of IEEE Conference on Visual Analytics Science and Technology. IEEE. pp. 173–182.
2. *Alekseev A. A., Lukashevich N. V.* (2012), Combining Attributes for Topic Chain Extraction in News Cluster [Kombinirovanie priznakov dlya izvlecheniya tematicheskih tsepochek v novostnom klastere], Works of RAS Instutute for Systems Analysis [Trudy Instituta sistemnogo programmirovaniya RAN], Vol. 23, pp. 257–276, available at: www.ispras.ru/ru/proceedings/archives/isp_23_2012/isp.
3. *Agrawal R., Imieliński T., Swami A.* (1993) Mining association rules between sets of items in large databases. ACM SIGMOD Record. Vol. 22. No. 2.
4. *Blei D. M., Ng, A. Y., Jordan M. I.* (2003), Latent Dirichlet allocation, Journal of Machine Learning Research, Vol. 3 (4–5): pp. 993–1022.
5. *Chang A. X., Savva M., Manning C. D.* (2014), Semantic parsing for text to 3d scene generation. In Proceedings of the ACL 2014 Workshop on Semantic Parsing, Baltimore, pp. 17–22.
6. *Coppersmith G., Erin K.* (2014), Dynamic Wordclouds and Vennclouds for Exploratory Data Analysis. Association for Computational Linguistics, pp. 22–29.

7. *Coupland D.* (1996), Microserfs, Flamingo
8. *Hulth A.* (2003), Improved automatic keyword extraction given more linguistic knowledge, The conference on Empirical methods in natural language processing, pp. 216–223.
9. *Owen K., Lemire D.* (2007), Tag-cloud drawing: Algorithms for cloud visualization, Cornell University Library, available at: arxiv.org/abs/cs/0703109
10. *Liu, S., Wang, X., Chen, J., Zhu, J., Guo, B.* (2014,). TopicPanorama: A full picture of relevant topics. Proceedings of IEEE Conference on Visual Analytics Science and Technology. IEEE. pp. 183–192.
11. *Lloyd L., Kechagias D., Skiena S.* (2005), Lydia: A system for large-scale news analysis, String Processing and Information Retrieval, Springer Berlin Heidelberg.
12. *Mirkin B. G., Chernyak E. L., Chugunova O. N.* (2012), Scoring String to Text Similarity by Means of Annotated Suffix Trees [Metod annotirovannogo suffiksnogo dereva dlya otsenki stepeni vhozhdeniya strok v tekstovye dokumenty], Business Informatics [Biznes-Informatika], Vol. 3, № 21, pp. 31–41.
13. *Mitrofanova O. A., Zaharov V. P.* (2009), Automatic Analysis of Terminology in the Russian Text Corpus on Corpus Linguistics [Avtomatizirovannyy analiz terminologii v russkoyazyichnom korpuse tekstov po korpusnoy lingvistike], Dialog, available at: http://www.dialog-21.ru/digests/dialog2009/materials/.
14. *Open Corpus*. Russian-language open corpus, available at: http://opencorpora.org/
15. *Pymoprhy2*. Russian morphological parser, available at: https://pymorphy2.readthedocs.org
16. *Savchuk S.* (2011), A Corpus-based Study of Morphological Variability: Variation in Gender Forms of Russian Nouns. In Conference Proceedings of «Computational Linguistics and Intellectual Technologies», pp. 562–580.
17. *Schäfer R., Barbaresi A., Bildhauer F.* (2014), Focused Web Corpus Crawling, 9th Web as Corpus Workshop (WaC-9)
18. *Smith, A., Chuang, J., Hu, Y., Boyd-Graber, J., Findlater, L.* (2014). Concurrent Visualization of Relationships between Words and Topics in Topic Models Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics.
19. *Shahaf, D., Yang, J., Suen, C., Jacobs, J., Wang, H., Leskovec, J.* (2013, August). Information cartography: creating zoomable, large-scale maps of information. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM. pp. 1097–1105.
20. *Wang H., Can D., Kazemzadeh A., Bar F., Narayanan S.* (2012), A system for real-time twitter sentiment analysis of 2012 US presidential election cycle. Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics.