

Discovering Text Reuse in Large Collections of Documents: a Study of Theses in History Sciences

Anton S. Khritankov, Pavel V. Botov, Nikolay S. Surovenko, Sergey V. Tsarkov,
Dmitriy V. Viuchnov, Yuri V. Chekhovich

Anti-Plagiat JSC
Moscow, Russia
khritankov@antiplagiat.ru

Abstract—In this paper we investigate graphs of text reuse cases in scientific degree theses in history sciences (07.xx.xx of Russian Higher Attestation Committee topic codes). Using algorithmic and statistical methods we discovered groups of highly connected theses with large amount of text reuse between them. In addition we located works compiled from several other theses and point out sources of reuse.

I. INTRODUCTION

In this paper we extend and detail the preliminary results presented earlier in [23] on our research conducted for Russian State Library (RSL). We analyzed doctoral degree theses from the Digital Library of RSL (RSL DL) using Antiplagiat software as well as custom in-house machine learning software. RSL DL contains bibliographic records and full texts of thesis annotations and doctoral theses for different topics. The texts were previously scanned from paper originals and processed using OCR software.

Antiplagiat text reuse detection engine can be used to perform a comparative analysis between a given text document and a text corpus. Result of the analysis is a list of text blocks found both in the document and texts in the corpus. These common text blocks are usually referred to as text reuse blocks. This reuse can have different interpretation ranging from citing a source text in the corpus to common citing of a third document and fixed phrases and even pure coincidence.

Therefore, the result is usually checked by an expert who qualifies discovered text reuse cases and gives overall verdict on the overall academic quality of the work [21].

Experts' job requires substantial effort and professionalism for large texts as are theses – from several hours to days per thesis. Considering that about 25 000 doctoral degrees are awarded annually in Russia, it is a tough task to check all theses.

The main aim of current research was to determine if it is possible to perform deep automatic text reuse analysis in large corpora and develop a preliminary filter for further expert analysis. This filter would allow selecting only theses that are most likely to contain substantial text reuse worth of expert analysis.

RSL was the initiator and primary consumer of the research. Several research questions have been put forward, which are provided in section 2. In order to quantify text reuse

correctly we excluded blocks with legal citations (section 3), fixed phrases usually found in doctoral theses and bibliography (section 4). After these preprocessing steps a detailed analysis is performed as described in section 5.

II. RESEARCH QUESTIONS AND PROBLEM STATEMENT

The following research questions and goals have been stated:

- evaluate feasibility of finding pairwise text reuse cases between documents in a large corpus;
- estimate portion of texts with substantial reuse from other theses;
- find out if these theses are produced as part of a systematic “compilation” from other sources or constitute just separate unrelated cases.

III. EXTRACTION OF QUOTATIONS

Authors often cite other works in their theses and include text fragments as quotes. These quotes shall be formatted according to grammar rules [15] and citing standards [16]. The text reuse detection engine would identify these quotations as reuse blocks, therefore we shall detect them and exclude from further analysis.

We employed a machine learning approach to detect quotations what consist of the following steps:

- 1) Extract candidate fragments from text using heuristics.
- 2) Calculate feature values for the candidate fragment.
- 3) Apply a binary classifier to discover correct quotations.

At the first step candidate fragments are extracted according to the grammar of Russian language [15]. Quotations are almost always put in angled quotes. Exceptions are poetry what can be cited without quotes. Poetry is rare in theses on history topics therefore we decided to use quoted text as candidate fragments, taking into account that quotes can be nested.

Next, features presented in Table I are evaluated for each candidate fragment. These features have been derived from the standard [16] and citation formats found in theses manually. There are more than sixty features developed; with further selection only 23 of them have been retained. They are shown in Table I.

Features can be split into two broad categories: citation formatting features and positioning and other properties of text. The first category includes fifteen features that capture different quotation formatting heuristics: occurrence and position of specific punctuation, occurrence of footnotes and references. These features mostly influence recall of the classification.

The second category consists of eight features describing text properties: recognition errors, misspelling, and length of the candidate block in words, symbols, average word length and others. These features mainly affect precision.

After all features have been computed we apply decision tree model that classifies candidate blocks into correct quotations and the rest.

Training data consists of manually labelled candidate blocks in real theses in history sciences. We developed labeling software with GUI to automate expert's work. Training data includes 16320 candidate blocks, 2848 from which are labelled as quotations. Test data consists of 8159 candidates with 1429 quotations.

We used Weka [17] toolkit for feature selection. Using only training data and Gain Ratio [18] criterion we selected 23 features for further analysis. The decision tree was built with

C4.5 algorithm [18] with maximum tree depth limited to 7. We chose decision tree over other models for its understandability for non-technical audience. Visual analysis of the tree produced by the algorithm shows that footnote occurrence features are most informative.

Classification quality was evaluated using precision and recall. On training precision was 96.8% and recall was 73.5%, on testing – 95.8% and 43.8% correspondingly.

These results have been obtained on texts from RSL DL with OCR errors.

IV. DATA PREPARATION

Antiplagiat text reuse detection engine processes texts, builds inverted index of word n-grams [19] and performs pairwise comparison between a given document and candidate text reuse sources found via the inverted index. We used a version of the algorithm that finds text reuse blocks that appear almost identical with respect to stop-words in both texts being compared. The algorithm is exact, meaning it finds all text exact reuse blocks and does not produce false positives by design.

During data preparation our goal was to find text reuse blocks between theses. We put Ph.D. and Sc.D. theses from RSL DL into the text reuse detection engine, totally over 14

TABLE I. FEATURES USED FOR CLASSIFICATION OF CANDIDATE TEXT BLOCKS

Feature name	Description
ExternalReference	External reference to a citation is found
FirstSymbolDots	Block starts with an ellipsis
FirstSymbolUpperCase	Block starts with a capital letter
Footnote	Footnote is present inside block
LastSymbolDots	Block ends with an ellipsis
InternalReference	Internal reference to a citation is found
ExternalExtReference	External reference is shown in parentheses
AuthorAfter	Text after block contains a last name from the dictionary
AuthorBefore	Text before block contains a last name from the dictionary
Length	Length of the block lies in a specific range (binary feature)
LinesInAfterText	Number of lines in the block
LinesInBeforeText	Number of lines before block lies in a specific range (binary feature)

Feature name	Description
NumericCount	Number of digits in the block lies in a specific range (binary feature)
List	Block has lines starting with a digit
SubStringCount	Total occurrences of given string in the block
Position	Relative position of the block in the document (in percent)
FootnoteInAfterText	A footnote immediately follows the block
FootnoteInAfterTextBeginning	A footnote immediately precedes the block
UpperCaseCharCount	Number of upper case character in the block
AvgWordLength	Average word length in the block
WordCount	Number of word in the block
SpecialCharsDensity	Number of non-alphanumeric symbols in the block
Dots	An ellipsis is present in the block

thousand theses on history topics written mostly between 1999 and 2012, see Fig. 1. Document metadata used was also obtained from RSL DL. From this data 51 empty documents and 114 documents with less than 15 thousand characters were excluded. Fig. 1 also illustrates availability of theses in RSL DL, many theses from 1985 to 1999 are not available in digital form.

When we look for text reuse in a collection of documents it is important to know the direction of each reuse in a pair of documents. In this research we used year of defense for each thesis to judge which of a pair was written first and set the direction correspondingly. We assumed that if a thesis had been defended a year before the other then it can be taken as the source of reuse. For each thesis we retained top 100 of source theses with the most reuse. Minimum reuse block size varied from three to seven words depending on context.

Data preparation was performed on a Xeon 1.6 Ghz server with eight virtual HT cores, 6 Gb RAM during four days. Total size of text reuse blocks in XML format was about 4 Gb. These blocks were further processed: nearby blocks were merged if they were less than 30 characters away, intersections with correct quotations excluded, then second merge with the same algorithm and filtering. Original and processed text reuse blocks distributions are shown at Fig. 2 and Fig. 3 correspondingly.

Analysis of the distribution showed that most of the blocks are located in the title page and, apparently, in bibliography. Assuming that these blocks are due to title page formatting and layout standards and due to common references in bibliography we excluded blocks located in the first 1000 characters and last 10% of thesis text.

Further distribution analysis had shown that there are a lot of minute blocks of size less than 250 characters that include common phrases related to many theses and usually found in introduction and conclusion. Note that later for reuse graph analysis we excluded blocks smaller than 750 characters and this leads to uniform distribution of blocks in text, see Fig. 3.

Filtering also substantially decreased the number of sources as shown at Fig. 4 and lowered text reuse in theses as shown at Fig. 5. Block length distribution became unimodal as well, see Fig. 6.

Some abnormal texts were spotted during analysis: about 50 documents contained two or more theses put together what also could be found separately.

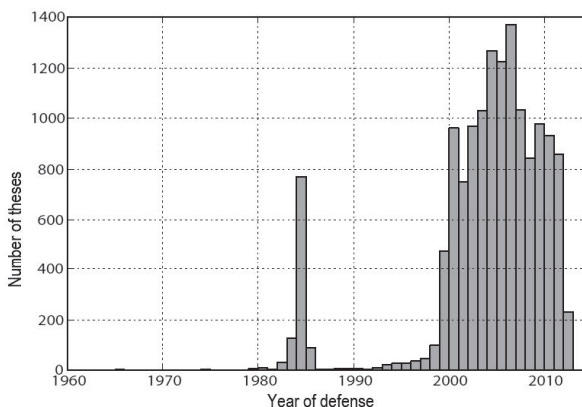


Fig. 1. Number of theses per year of defense

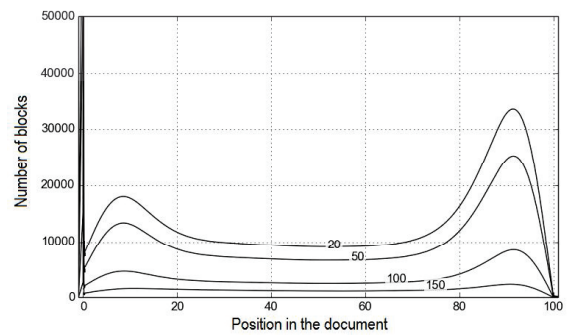


Fig. 2. Size and position of text blocks before preprocessing. Isodenses denote size of blocks

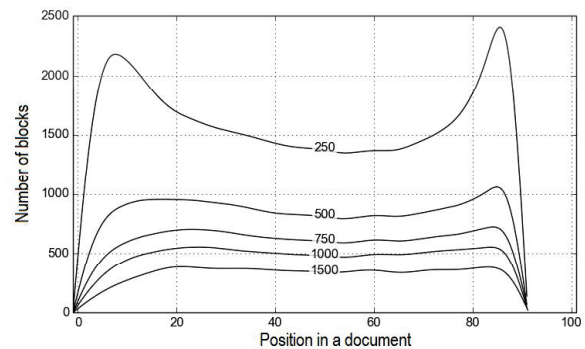


Fig. 3. Size and position of text blocks after preprocessing. Isodenses denote size of blocks

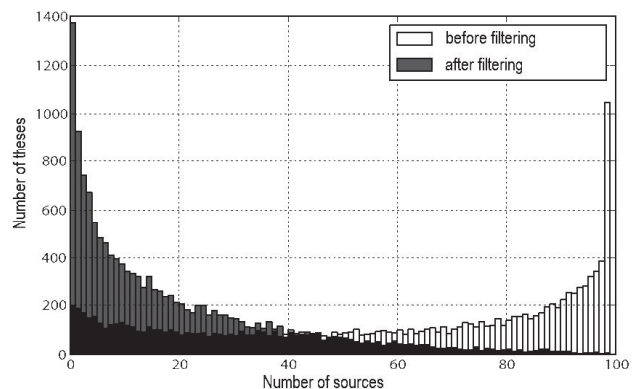


Fig. 4. Histogram of number of text reuse sources for theses before and after filtering

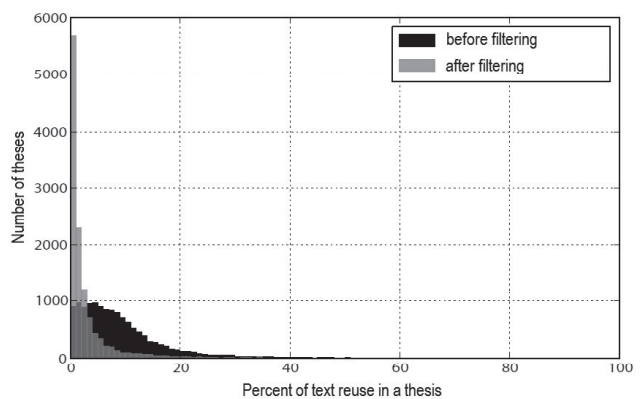


Fig. 5. Histogram of text reuse amount in theses before and after filtering.

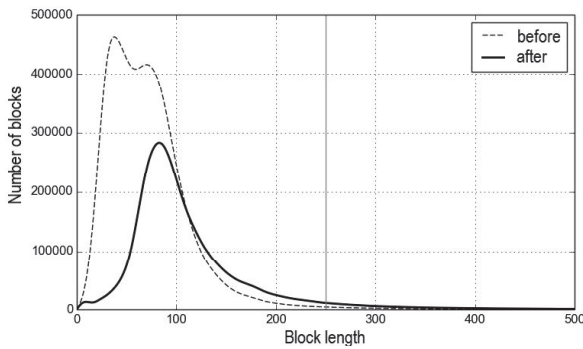


Fig. 6. Number of text reuse blocks before and after filtering

V. COMMUNITY ANALYSIS

A. Research basis

Aiming at discovering systematic reuse we take a hypothesis that systematic reuse should leave traces such as common reused blocks of texts found in a series of theses produced together.

Exploring this further we build a text reuse graph as a tool to analyze relations between theses based on the reused text. Text reuse graph contains theses as nodes and reuse cases as directed edges with weight corresponding to the total characters reused between a pair of theses.

Community analysis applied to text reuse graph helps to deduce context of the reuse between texts and discover hidden systematic reuse.

We applied a fast community detection algorithm based on maximization of an internal quality criterion – modularity [2]:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

where A_{ij} – edge weight between nodes i and j , $k_i = \sum_j A_{ij}$ – total weight of all edges from node i , c_i – community, which node i belongs to, δ -function $\delta(u, v)$ equals 1, when $u = v$, and 0 otherwise, and $m = \sum_{i,j} A_{ij} / 2$.

Community detection algorithm [2] consists of two iteratively repeated steps.

At the first step each node is assigned its own community. Then for each node i and node j if there is an edge between i and j and modularity increases then assign i and j to the same community. The procedure repeats until modularity reaches local maximum.

At the second step the communities discovered earlier form a new graph for its nodes and edges summarize edges between nodes of the communities. Then a new iteration continues with the new graph.

Iterations continue until communities stabilize.

B. Discovering graph structure

Original graph contains about 13 000 nodes and 164 000 edges. Before filtering, the graph contained a giant component

with over 12 000 nodes, what suggests a large number of “noise” edges. Assuming that these noise edges have small weight we could pick a threshold that filters out most of the noise. At the other hand, filtering out edges may lead to loss of meaningful relations between nodes that would form communities otherwise. Therefore we aimed at setting a minimum edge weight threshold sufficient to extract communities. For convenience when a node loses all of its edges due to filtering it also gets removed from the graph.

We analyzed how the following parameters depend on the threshold: number of communities, number of loosely coupled components, and size of a giant component (see Fig. 9 and Fig. 10).

As threshold increases the number of communities grows due to destruction of the giant component (Fig. 9), then it reached maximum, and then decreases. We chose this extreme point as a filtering threshold because it minimized loss of communities.

As a result the threshold was set to 0.05 that accounts to about 7500 reused characters between theses. A total of 748 communities have been discovered. These communities have more reuse between theses inside them than reuse with theses from other communities.

C. Comparison with random graphs

In order to demonstrate that structures we could find in a graph are not accidental we repeated the same structure finding procedure for two random graphs build using nodes and edges of the original text reuse graph.

1) A graph *GR1* in which for each node and each edge the source of the edge remains and target is chosen at random uniformly among all nodes of the graph, edge weight preserved.

2) A graph *GR2* in which for each edge the source and target nodes are chosen at random uniformly from all nodes of the original graph, edge weight also preserved.

Comparing size of the giant component and number of communities for original graph (Fig. 9), *GR1* (Fig. 11) and *GR2* (Fig. 13), we can see that general picture is roughly the same – there is a local maximum at about 10 000 symbols threshold, but still different for each graph.

Notice that *GR2* has much more communities and loses its giant component earlier than original graph and *GR1*.

Another important point is that number of communities and connected component coincide that *GR1* and *GR2* at or just right after the maximum, but remain separate for the original graph and its giant component deteriorates much slower (see Fig. 10, Fig. 12, Fig. 14). That means that inside connected components in the original graph there are communities what are not caused by noise as they are in random graphs.

C. Exploring communities

Let us continue and explore some examples of communities. An example text reuse community graph is shown at Fig. 7. In a community a thesis can play two roles: a

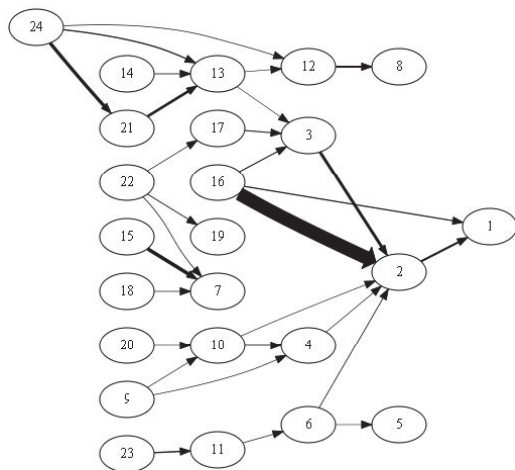


Fig. 7. An example of a community found in text reuse graph. Numbered nodes correspond to theses, edges show text reuse, edge weight is proportional to the volume of reuse

source of reuse and a recipient of reuse. At Fig. 7 we can see theses 24, 16, 22 as popular recipients of reuse in the community, while theses 2, 3, 7, 13 are the sources of reuse. Note that 2, 3 and 13 are also recipients of further reuse. Bold edge between 2 and 16 shows large text reuse between these theses.

Distinctive sources and recipients of reuse could be found in most of the communities. In these communities overall amount of reuse is large that can be interpreted as existence of some organizations that produce new theses by compilation from old ones. Assigning a source thesis of reuse to a community is therefore does not show the author of the thesis as part of the community but is useful for analysis of the community itself.

Let us put all communities to a diagram with average reuse and total reuse within community as axes (Fig. 8). Communities can be divided into three clusters. We would call small communities, apparently, compiled separately for a few theses “individual reuse”. Large communities with moderate average reuse volume between theses are “factories of reuse”. Third theses cluster is “strange communities”, they cannot be placed in the first two clusters. Theses that do not belong to any of these clusters are considered as being produced without systematic reuse.

As we analyze text reuse only among theses in RSL DL we did not consider possible reuse from third-party sources such as scientific papers and books. In the reuse graph begin considered, such cases may constitute themselves as reuse between theses when there is a third text they both reuse.

VI. RELATED RESEARCH

The D.Sc. and Ph.D. theses being defended in some area of research generally reflect structure and state-of-the-art in the area. Therefore they are an interesting research matter. In the past, scientific theses and relations between them have been considered for other research areas [8-13]. In [8, 9] authors study theses and their abstract in an attempt to discover science schools and research collaborations, relations between

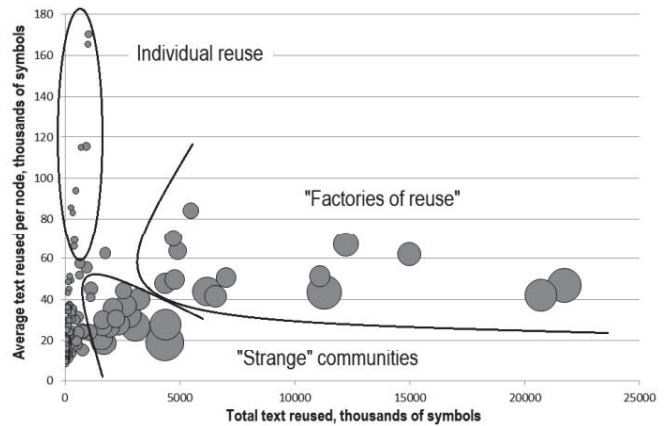


Fig. 8. Theses communities with respect to average reuse volume (bottom-up) and total reuse volume (left to right) in a community with hypothetical classification. Area of a circle corresponds to the size of the community from 4 to 189 theses

scientific advisers and students using text analysis methods. In recent research [10] quality of theses defended in 2008-2011 is studied, data published at Russian HAC website is used (Higher Attestation Committee (HAC), a government educational agency that assigns Sc.D. and reviews Ph.D. degree nominations [22]).

Current research is different in employing data from RSL DL [7], analyzing full texts of theses in only history sciences, in using text reuse to establish relations between theses and applying graph-based methods to study these relations. We believe that text reuse cases between theses indicate commonalities in their preparation.

Evaluation commonalities between scientific works using text reuse is a common method [1], [5], but there are other approaches, that use text analysis [13] and study co-citations between documents [14].

VII. CONCLUSION

To the best of our knowledge the research performed to study text reuse communities are novel and was not performed before. Research questions studied and answered have not been asked previously. Therefore it is also important that research methods were envisioned and elaborated.

Results show technological feasibility to perform text reuse analysis in large full-text corpus using Antiplagiat engine together with machine learning and data mining methods to extract suspicious documents that require expert attention.

We have discovered that most of theses processed have low text reuse. Nevertheless, more than 500 theses have as much as 33% of text common with other theses that may indicate common third-party sources or just direct text reuse.

In the reuse graph we could identify “communities” of theses, existence of which is, apparently, connected with their preparation process. Communities with large amount of reuse have been attributed to groups that have arranged for the text writing process by compilation from available sources.

Research results have been reported to the experts of RSL and received positive response. In the future works we plan to further our research for other areas of study.

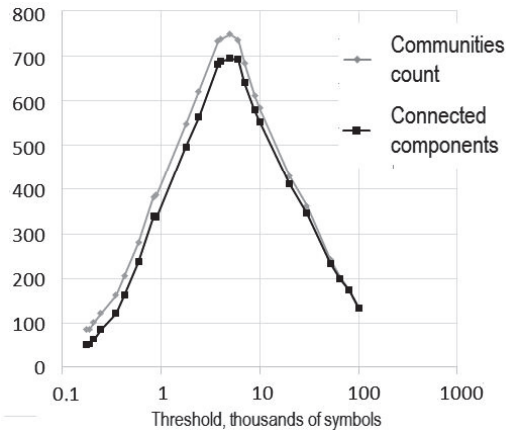


Fig. 9. Number of connected components in text reuse graph and number of communities depending on the edge weight cut-off threshold

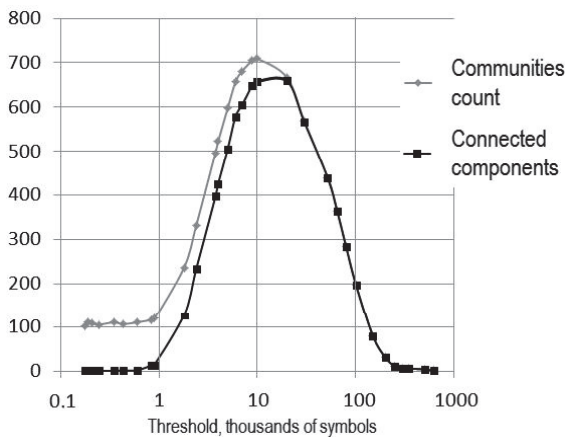


Fig. 11. Number of connected components in graph *GRI* and number of communities depending on the edge weight cut-off threshold

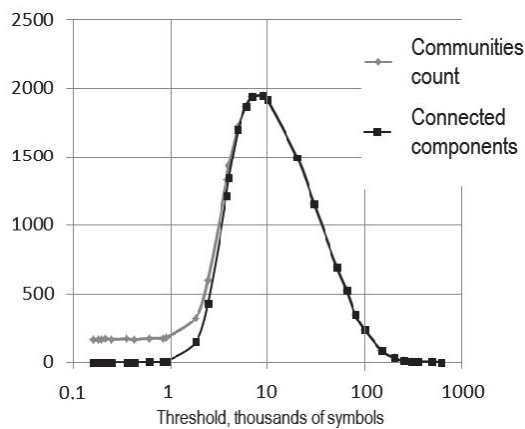


Fig. 13. Number of connected components inn graph *GR2* and number of communities depending on the edge weight cut-off threshold

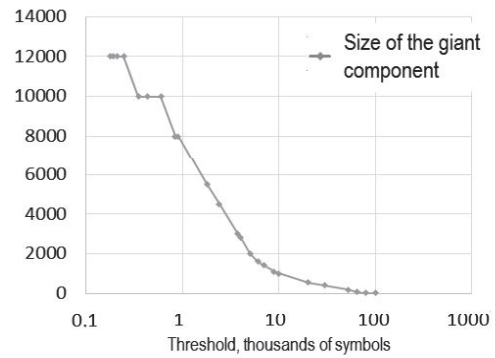


Fig. 10. Size of the largest connected component with respect to the edge weight cut-off threshold

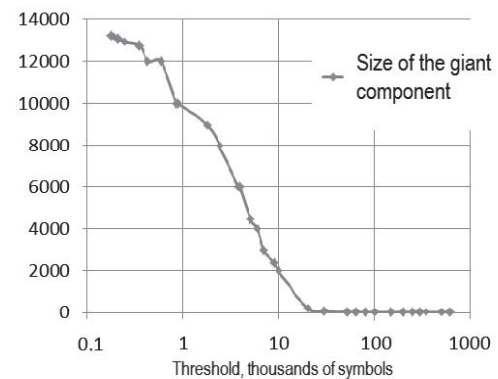


Fig. 12. Size of the largest connected component in graph *GRI* with respect to edge weight cut-off threshold

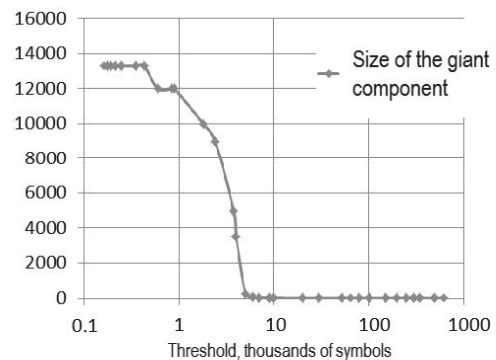


Fig. 14. Size of the largest connected component in graph *GR2* with respect to edge weight cut-off threshold

REFERENCES

- [1] N.V. Avdeeva, P.V. Botov, A.S. Bukaev, A.I. Visliy, I.A. Gruzdev, D.A. Zhitlukhin, M.Yu. Romanov, Yu.V. Chehovich. Deployment of Antiplagiat software in Russian state library, in Proc. of *Intelligent Information Processing: 8-th international conference*, Oct. 2012, pp 499-503. (in Russian)
- [2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre. Fast unfolding of communities in large networks, *J. of Statistical Mechanics: Theory and Experiment*, vol. 10, 2008.
- [3] R. Lambiotte, J.C. Delvenne, M. Barahona. *Laplacian dynamics and multiscale modular structure in networks*, arxiv.org preprint arXiv:0812.1770 (2008)
- [4] Anti-Plagiat JSC, Antiplagiat. A text reuse detection engine, Web: <http://www.antiplagiat.ru>

- [5] iParadigms, LLC. Turnitin. Plagiarism prevention engine. Web: <http://www.turnitin.com>
- [6] R.V. Sharapov, E.V. Sharapova. System of Duplicate Texts Detection, in Proc. of the 13-th Russian scientific conference "Digital Libraries: future methods and technology, digital collections – RCDL-2011", 2011, pp. 121-126. (in Russian)
- [7] O.A. Lavrenova, Development of a Digital Library of Ph.D. Theses and Abstracts for public access, in Proc. of *Educational Technology & Society*, vol. 9, no. 3, 2006, pp. 335-341. (in Russian)
- [8] Yu.V. Leonova, A.M. Fedotov, Extraction of knowledge and facts from texts of theses and abstracts for studying of communications of scientific communities, in Proc. of the 15-th Russian scientific conference "Digital Libraries: future methods and technology, digital collections – RCDL-2013", Oct. 2013. (in Russian)
- [9] Yu.V. Leonova, A.A. Dobrynin, A.Yu. Vesnin, Build a Graph of Dissertations, in Proc. of the 15-th Conference on Distributed Informational and Computing Resources (DICR-2012), Nov. 2012, p. 17. (in Russian)
- [10] S.S. Donetskaya, Statistical investigation of structure and quality of doctoral dissertations in Russia, *Voprosy Statistiki*, no. 12, 2012, pp. 71-76. (in Russian)
- [11] E.V. Deskaravainaya, I.A. Mitroshin, Analysis of Dissertations Database PSC RAS, *Informatsionnoe obespechenie nauki. Noveye tekhnologii*, 2011, pp 124-133. (in Russian)
- [12] Yu.N. Klimov, Kolichestvenno-informatsionnyy analiz potoka publikatsiy po bibliotekam i bibliotekovedeniyu na osnove poiska po kluchevym slovam v base dannyh Scienet-Direct, *Mezhotraslevaya informatsionnaya sluzhba*, issue 3, 2011. (in Russian)
- [13] V.N. Zakharov, A.A. Khoroshilov, Automatic assessment of similarity of the texts' thematic content on the base of their formalized semantic descriptions comparison in Proc. of the 14-th Russian scientific conference "Digital Libraries: future methods and technology, digital collections – RCDL-2012", Oct. 2012. (in Russian)
- [14] B. Gipp, J. Beel, Citation Proximity Analysis (CPA) – A new approach for identifying related work based on Co-Citation Analysis", in Birger Larsen and Jacqueline Leta, editors, *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*, vol. 2, Jul. 2009, pp. 571–575.
- [15] D.E. Rosental, D.E. Dzhavanzhakova, N.P. Kabanova, *Reference on Russian grammar, pronunciation, literature editorship, Second Ed.* CheRo, 1998, 400 pp. (in Russian)
- [16] GOST R 7.0.5-2008. Bibliographic reference, General requirements and rules of making, *System of standards on information, librarianship and publishing*, 2009.
- [17] University of Waikato. Weka Toolkit, Web: <http://www.cs.waikato.ac.nz/~ml/weka/>
- [18] J. R. Quinlan. *C4.5: Programs for Machine learning*. Morgan Kaufmann Publishers 1993.
- [19] C.D. Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press; 1 edition, 2008, 506 pp.
- [20] N.V. Nikulina, A.M. Sologubov, Antiplagiat-RSL software and dishonest authors: who will win?, *J. Scientific Proceedings: Problems and Solutions*, no. 5, issue 11, Sept-Oct 2012, pp. 11-16. (in Russian)
- [21] N.V. Avdeeva, G.A. Lobanova, Classification of text fragments in expert evaluation of dissertations for text reuse (plagiarism), *J. Russian Information Resources: research and practice*, no. 11, 2014, pp. 2-6. (in Russian)
- [22] Russian Higher Attestation Committee (HAC), Web: <http://vak.ed.gov.ru>
- [23] P.V. Botov, Y.V. Chehovich, A.S. Khritankov, N.S. Surovenko, S.V. Tsarkov, D.V. Viuchnov. Structures of Text Paraphrasing and Plagiarism in Dissertations on Historical Sciences, in Proc. of the 16-th Russian scientific conference "Digital Libraries: future methods and technology, digital collections – RCDL-2014", Oct. 2014, pp. 174-180. (in Russian)