Term frequency, reference text corpus and estimating the closeness of short texts to the semantic standard

Mikhaylov D., Emelyanov G.

Yaroslav-the-Wise Novgorod State University

14^{th} International Conference
on Intelligent Data Processing: Theory and Applications,

December 6–9, 2022

Moscow, Russian Federation

Selection of publications and rational (i. e. standard) sense transfer

Basic requirements to the representative (i. e. reference) text collection

1. Maximal disclosure of topic interesting for the end user in each text of the forming collection.
2. Texts in a collection should be relevant as much as possible to the given topical area from the point of expert view both in vocabulary, and in internal text relations (syntactic, semantic, etc.).
3. Maximum of an average number of the most significant terms per a one simple spread sentence (i. e. phrase) at minimum of its length measured in words, that satisfy to the *standard variant* of sense transfer.

How to minimize the handwork of expert: basic considerations

1. It is advisable for an expert to use short texts which are comparable in vocabulary and (possible) in relationships between words with documents being added to the reference collection.
2. In a role of such texts abstracts of scientific articles or other texts that are resume significant facts of a given topical area can entirely be.
3. The specified selection to the reference collection is the task inverse to the *abstractive summarization*: to find a text, in which general ideas described in an abstract (or in a collection of abstracts) are reflected the most fully.

«Classic» problem statement [Eremeev M., 2019]:

1. For each linguistic level, its own *alphabet of tokens* is defined. For example, words or terms for the lexical level, types and lengths of syntactic links for syntactic one.
2. The occurrence frequency for token is considered as abnormally high if it is greater than 95^{th} *percentile* of its frequency in a reference corpus of texts which are not complicated for the implied readership.
3. The assumption was make about that 95% *of tokens* in a reference corpus on each of language level *not exceed* their *fixed occurrence frequency*, which is determined experimentally (by the results of neurophysiological and psychophysiological studies).

Percentile is a some value which the investigated random variable not exceed with a fixed probability *measured in percents*.

Sampling documents to the reference corpus basing on collection of abstracts

Since we are considering here the *minimum necessary* representation level of words (terms) from abstracts in an analyzed document, then it's reasonable to assume that the 5^{th} *percentile* of frequency characteristic of word relatively to the given document here we should consider.

Justification of frequency characteristic selection

The main requirement here is independence from the number of document words.

Let's calculate for each phrase in each abstract the share of non-zero values of *TF-measure* for phrase words relatively to the analyzed document.

TF-measure (term frequency) estimates the significance of word t_i within the document d and can be defined as

$$\mathrm{tf}\,(t_i, d) = \frac{n_i}{\sum_k n_k}, \tag{1}$$

where n_i is the number of times that t_i occurs in document d,
and denominator contains the total number of words for d.

Remarks

- one phrase here corresponds to the simple spread natural-language sentence (according to the terminology of «Meaning⇔Text» approach);
- it is admissible that the same phrase may appear in more than one abstract of the collection (for example, if these articles denoted to the same author);
- in any case each phrase is accepted to consideration only once;
- using exactly the share of non-zero values of TF-measure, and not the *term frequency* values themselves for words of a phrase, allows solving the problem of dependence of significance estimation value for a document from the number of words in it;
- only the presence of the maximum number of words from the abstracts in the analyzed document is important, while the frequency of individual words is not principled here.

Document significance estimation for adding to the reference collection

Let document d be a candidate for adding to the reference collection (or corpus). For each word w of each phrase Ts of each abstract from the collection formed by an expert the value of TF-measure relatively to d, $\mathrm{tf}(w,d)$, is calculated. Herewith the share of non-zero TF values for separate phrase Ts is defined as

$$c(Ts,d) = \frac{\left|w\colon\ (w \in Ts) \wedge (\mathrm{tf}(w,d) > 0)\right|}{\left|w\colon w \in Ts\right|}. \tag{2}$$

Let's designate as $C_5(\mathbb{T}\mathrm{s},d)$ the 5^{th} *percentile* of empirical distribution of estimation (2) value concerning the document d for the given collection of abstracts $\mathbb{T}\mathrm{s}$.

We'll associate $\mathbb{T}\mathrm{s}$ with the combining of sets of phrases for separate abstracts.

Let's sort documents that are candidates for adding to the reference corpus, by decreasing of $C_5(\mathbb{T}\mathrm{s},d)$. Let $d_{\max}$ be a document with the maximal value of C_5 among the documents $d \in D$ for phrases of abstracts represented in $\mathbb{T}\mathrm{s}$.

Let's enter into consideration for each $d \in D$ the vector of quantiles values

$$\overline{V}(\mathbb{T}\mathrm{s},d) = \Big(C_\gamma(\mathbb{T}\mathrm{s},d)\Big)_{\gamma\in[5,10,20,25,30,40,50,60,70,75,80,90,95]}, \tag{3}$$

into which deciles together with the first and third quartiles will be included in addition to above-mentioned 5^{th} and 95^{th} percentiles.

Document significance estimation for adding to the reference collection

Let $\overline{V}(\mathbb{T}\mathrm{s}, d_{\max})$ be the vector of the kind (3) for the document $d_{\max}$.

Let's designate the sequence of mentioned vectors like (3) obtained for the collection $\mathbb{T}\mathrm{s}$ relatively to documents $d_j \in D\colon d_j \neq d_{\max}$ and sorted by descending the Euclidean distance to $\overline{V}(\mathbb{T}\mathrm{s}, d_{\max})$, as $\mathbb{V}(\mathbb{T}\mathrm{s}, D)$.

Let us split the sequence $\mathbb{V}(\mathbb{T}\mathrm{s}, D)$ into clusters $H_1, \ldots, H_r$ using an algorithm close in meaning to the FOREL class of algorithms. Herewith the cluster H_r will correspond to documents with the shortest distance to the document $d_{\max}$.

Statement 1

The most significant for the target collection will be documents $d \in D$ related to the cluster H_r, and the document $d_{\max}$ itself.

Remarks

1. To improve the *recall of search* the significant documents for reference collection the above-mentioned classification of documents $d \in D$ should be implemented independently for several collections of abstracts of articles denoted to close scopes.
2. *The recall of search* is estimated here by the ratio of the number of documents that meet the condition of *Statement 1* and classified as significant by an expert, to the total number of documents $d \in D$ from recognized as significant by an expert.

Abstract significance at selection documents to the target collection

Let $\mathbb{T}\mathrm{s}_i \subset \mathbb{T}\mathrm{s}$ be the set of phrases of the i^{th} abstract from the collection $\mathbb{T}\mathrm{s}$, and $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ be the 5^{th} percentile of empirical distribution of estimation (2) value concerning the document $d_{\max}$ relatively to phrases of this abstract.

Let's designate the document with the maximal value of C_5 among the documents $d \in D$ for phrases of $\mathbb{T}\mathrm{s}_i$, as $d_{\max(i)}$.

Statement 2

According to significance for precise calculating the value of $C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$ among abstracts related to $\mathbb{T}\mathrm{s}$ the following five groups can be distinguished:

- *group 1:* abstracts, where $d_{\max} = d_{\max(i)}$ and $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right) > C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$;
- *group 2:* abstracts, where $d_{\max} \neq d_{\max(i)}$, but $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right) > C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$, at that $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max(i)}\right)$ and $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ related to one cluster;
- *group 3:* abstracts, where $d_{\max} \neq d_{\max(i)}$, but $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right) > C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$, at that $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max(i)}\right)$ and $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ cannot be assigned to one cluster;
- *group 4:* abstracts, where $d_{\max} \neq d_{\max(i)}$ and $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right) < C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$, but $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max(i)}\right)$ and $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ related to one cluster;
- *group 5:* abstracts, where $d_{\max} \neq d_{\max(i)}$ and $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right) < C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$, at that $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max(i)}\right)$ and $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ cannot be assigned to one cluster.

Herewith the highest precision for search of significant documents is reached with abstracts of *groups* from *1* to *3*. In meaning, *1st* group abstracts are closest to the sense standard.

Initial data for experiment: documents that are reference corpus candidates

- Taurida journal of computer science theory and mathematics (TJCSTM, 3 papers);
- Proceedings of International conferences «Intelligent Information Processing» IIP-8 and IIP-9 of the years 2010 and 2012 (2 papers);
- Proceedings of All-Russian Conference with International Participation on Mathematical Methods for Pattern Recognition (MMPR-15, 1 paper);
- Proceedings of the Conference MMPR-13 (2007, 2 papers);
- Proceedings of the Conference MMPR-16 (2013, 14 papers);
- Proceedings of the Conference IIP-10 (2014, 2 papers);
- the text of a scientific report prepared in 2003 by Dmitry Mikhaylov.

Remark

The number of words in documents of corpus varied here from 218 to 6298, and the number of phrases per document varied between 9 and 587.

Initial data for experiment: the scope of candidates for adding to the corpus

- mathematical methods for learning by precedents (K. Vorontsov, M. Khachay, E. Djukova, N. Zagoruiko, Yu. Dyulicheva, I. Genrikhov, A. Ivakhnenko);
- methods and models of pattern recognition and forecasting (V. Mottl, O. Seredin, A. Tatarchuk, P. Turkov, M. Suvorov, A. Maysuradze);
- intelligent processing of experimental information (S. Dvoenko, N. Borovykh);
- image processing, analysis, classification and recognition (A. Zhiznyakov, K. Zhukova, I. Reyer, D. Murashov, N. Fedotov, V. Martyanov, M. Kharinov).

Initial data for the formation of collections of abstracts

- proceedings of «Intelligent Information Processing» conference of the year 2012, section «Theory and Methods of Pattern Recognition and Classification» (14 articles);
- proceedings of the 14^{th} All-Russian conference «Mathematical Methods for Pattern Recognition», section «Methods and Models of Pattern Recognition and Forecasting» (2009, 35 articles);
- proceedings of the 15^{th} All-Russian conference «Mathematical Methods for Pattern Recognition» (2011), section «Theory and Methods of Pattern Recognition and Classification» (18 articles) and «Statistical Learning Theory» (10 articles).

Some technical details

- Values of *term frequency* are calculated disregard of prepositions and conjunctions.
- Text extraction from a PDF file was implemented using the functions of the *pdfinterp*, *converter*, *layout* and *pdfpage* classes as part of the *PDFMiner* package.
- In order to be correctly recognized, all formulas from the analyzed documents here were translated by an expert manually into a format close to used in LaTeX.
- To select the boundaries of sentences in the text by punctuation marks, the method *sent_tokenize()* of the *tokenize* class from the open-source library *NLTK* was used.
- Lemmatization of words was performed using the morphological analyzer *PyMorphy2*.
- If a word has more than one parsing variant when determining its initial form (lemma), the closest one issued by the n-gram tagger from the *nltk4russian* library is taken.

software implementation (in Python 2.7) and experimental results

Sampling documents to the reference corpus

Table 1. The most significant documents for target collection.

No.	Author(s), title, and imprint of the paper, $d \in D$	N_1	N_2	N_3
1	Vorontsov K. V. The review of contemporary investigations at the problem of quality of learning of algorithms // TJCSTM. 2004. No. 1. P. 5–24.	667	6299	4
2	Vorontsov K. V. The combinatorial theory of overfitting: results, applications, and open problems // Proceedings of the 15th All-Russian Conference on Mathematical Methods for Pattern Recognition (MMPR-15). Moscow: Russian Academy of Sciences, 2011. P. 40–43.	230	2345	4
3	Dyulicheva Yu. Yu. Pruning strategies of decision trees (review) // TJCSTM. 2002. No. 1. P. 10–16.	153	2360	1
4	Dyulicheva Yu. Yu. About software implementation and approbation of Decision Forest Building Sequencing Algorithm for empirical decision forest synthesis // TJCSTM. 2003. No. 2. P. 35–44.	174	2075	2
5	Martyanov V. Yu., Polovinkin A. N., Tuv E. V. Image classification with codebook based on decision tree ensembles // Proceedings of the 9th International Conference «Intelligent Information Processing» (IIP-9). Moscow: Russian Academy of Sciences, 2012. P. 480–482.	139	1602	1

Table 2. Documents not meeting the condition of *Statement 1* and affinity to the sense standard by phrases.

Author(s), title, and imprint of the paper	N_1	N_2
Djukova E. V., Peskov N. V. About classification algorithm based on complete decision tree // Proceedings of the MMPR-13 All-Russian Conference. Moscow, 2007. P. 125–126.	23	348
Ishkina Sh. Kh., Ivakhnenko A. A. Combinatorial estimations for the overfitting of threshold decision rules // Proc. of the MMPR-16 All-Russian Conference. Moscow, 2013. P. 23.	14	278

N_1 is the number of phrases for d; N_2 is the total number of words with the respect of all occurrences of each word; N_3 is the number of collections of abstracts, where d meets the *Statement 1* condition.

Example: MMPR-15, Statistical Learning Theory

Table 3. Abstracts in descending order of their rank according to conditions of *Statement 2*.

i	Author (s) and article heading	N_{gr}
1	Frei A. I. The method of generating and destroying sets for randomized minimization of empirical risk	1
2	Vorontsov K. V., Makhina G. A. The principle of gap maximization for nearest neighbor monotonic classifier	1
3	Botov P. V. Reducing the probability of overfitting for iterative methods of statistical learning	1
4	Ivakhnenko A. A., Vorontsov K. V. Informativity criteria for thresholded logical rules with the correction for overfitting of thresholds	1
5	Zhivotovskiy N. K. Combinatorial estimations for the probability of test error deviation from the cross-validation error	1
6	Nedelko V. M. Empirical confidence intervals for conditional risk in the classification problem	1
7	Guz I. S. Hybrid estimations of complete cross-validation for monotonic classifiers	2
8	Kanevskiy D. Yu. Overfitting and combinatorial Rademacher complexity in regression recovery tasks	2
9	Khachay M. Yu. The convergence of empirical random processes generated by procedures of learning	3
10	Senko O. V., Kuznetsova A. V. Systems of reliable empirical regularities in models of optimal partitionings and methods to analyze them	5

Here N_{gr} is the group number from defined by *Statement 2*; as an $d_{\max(i)}$ on following slides, the number of corresponding document according to *Table 1*, is indicated. For comparison, the document $d_{\max}$ here has the serial number 2 by *Table 1*, and $C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right) = 0{,}53409091$.

Example: MMPR-15, Statistical Learning Theory

Let's enter into consideration the sequence X consists of values of $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ and $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max(i)}\right)$ for abstracts $\mathbb{T}\mathrm{s}_i$ within $\mathbb{T}\mathrm{s}$.

Let's order X by descending with splitting into clusters $H_1^X, \ldots, H_{r(X)}^X$.

Table 4. Calculated estimations for abstracts.

Author (s)	$d_{\max(i)}$	$C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$	$C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max(i)}\right)$	related to H_1^X
Frei A. I.	1	0,93571429	0,93571429	true
Vorontsov K. V., Makhina G. A.	1	0,93461539	0,93461539	true
Botov P. V.	1	0,79114286	0,79114286	true
Ivakhnenko A. A., Vorontsov K. V.	1	0,75500000	0,75500000	true
Zhivotovskiy N. K.	1	0,74787879	0,74787879	true
Nedelko V. M.	1	0,61312500	0,61312500	true
Guz I. S.	2	0,79000000	0,90000000	true
Kanevskiy D. Yu.	2	0,62222222	0,78222222	true
Khachay M. Yu.	2	0,58214286	0,93214286	true
Senko O. V., Kuznetsova A. V.	2	0,50000000	0,62500000	false

Let's note, that:

- papers represented in *Tables 3* and *4* are related to one cluster according to the value of $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ except for the article whose serial number is 10;
- when splitting these papers into clusters according to $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ with adding the value of $C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$ into the split sequence we'll obtain two clusters: to the first will be related articles with serial numbers 1 and 2, all others will be related to the second.

Documents of set D are sorted by the descending product of estimations:

$$val_1 = -1/\log_{10}\left(\Sigma_{H_1}\right), \tag{4}$$

$$val_2 = 10^{-\sigma\left(\left|H_i,\, i=\{1,\, r/2,\, r\}\right|\right)}, \tag{5}$$

and, correspondingly,

$$val_3 = \left|H_1 \setminus H_{r/2} \setminus H_r\right| / \mathrm{len}\,(Ts), \tag{6}$$

where Σ_{H_1} is the sum of TF-IDF values for words related to the cluster H_1 concerning to $d \in D$;
$\sigma\left(\left|H_i,\, i = \{1,\, r/2,\, r\}\right|\right)$ is the RMSD of number of elements in $H_i \in \{H_1,\, H_{r/2},\, H_r\}$;
$\mathrm{len}\,(Ts)$ is the length of phrase Ts in a group $\mathbb{T}\mathrm{s}$ consists of article title and abstract.

The first variant of estimation:

$$\mathrm{N}_1\left(\mathbb{T}\mathrm{s}, D\right) = \frac{\max\limits_{d\in D}\left(val_1\left(Ts_1, d\right) \cdot val_2\left(Ts_1, d\right) \cdot val_3\left(Ts_1, d\right)\right)}{\sigma\Big(\max\left(val_1\left(Ts_i, d\right) \cdot val_2\left(Ts_i, d\right) \cdot val_3\left(Ts_i, d\right)\right), Ts_i \in \mathbb{T}\mathrm{s}\Big) + 1}. \tag{7}$$

Here:

the *numerator* is the estimation of *affinity to the standard* for the *article title* (Ts_1);
the first summand in *denominator* is the RMSD for affinity to standard for all $Ts_i \in \mathbb{T}\mathrm{s}$.

The second variant of estimation:

$$\mathrm{N}_2\left(\mathbb{T}\mathrm{s}, D\right) = \frac{\max\limits_{d\in D}\left(val_1\left(Ts_{\max}, d\right) \cdot val_2\left(Ts_{\max}, d\right) \cdot val_3\left(Ts_{\max}, d\right)\right)}{\sigma\Big(\max\left(val_1\left(Ts_i, d\right) \cdot val_2\left(Ts_i, d\right) \cdot val_3\left(Ts_i, d\right)\right), Ts_i \in \mathbb{T}\mathrm{s}\Big) + 1}, \tag{8}$$

where $Ts_{\max} \in \mathbb{T}\mathrm{s}$ is the phrase for which the affinity to the sense standard is maximal.

Statement 3

The *maximal final* rank in the collection will be designated to the article with a greatest value of estimation (7) related to the same cluster with the value of estimation (8) for the same paper.

Remarks

- elements of numerical sequence X sorted in descending order *can be assigned to one cluster* if

$$\begin{cases} \left| \operatorname{mc}(X) - \operatorname{first}(X) \right| < \dfrac{\operatorname{mc}(X)}{4} \\ \left| \operatorname{mc}(X) - \operatorname{last}(X) \right| < \dfrac{\operatorname{mc}(X)}{4} \end{cases}, \tag{9}$$

  where $\operatorname{mc}(X)$ is the mass center of this sequence considered as a single cluster. As the mass center the arithmetic mean of all $x_j \in X$ is taken here;
- the correctly application of *Statement 3* assumes the relating to the same cluster the value of estimation (7) for article with a maximal final rank, and a maximal value of estimation (7) in the collection for paper selection;
- in a case of absence of article meets this requirement, the *maximal final rank* will be designated to the article with a greatest value of estimation (7) in analyzed collection;
- since the title and phrases of the article abstract (by definition) represent a certain single semantic image, it is entirely acceptable to swap with each other the estimations (7) and (8) in *Statement 3*.

To compare: ranking a collection of texts using *Statement 3*

Input: S; // the sequence of texts in the initial collection
// sorted in descending order of estimation (7)

Output: S_{res}; // the result of ranking the initial collection using *Statement 3*

1: $S_{res} := \varnothing$;
2: **while** $S \neq \varnothing$
3: $Flag := \text{false}$;
4: **for all** $\mathbb{T}\text{s} \in S$
5: $Tmp := \{\text{N}_1(\text{first}(S), D), \text{N}_1(\mathbb{T}\text{s}, D), \text{N}_2(\text{first}(S), D)\}$;
6: sort Tmp in the descending order;
7: **if** $\text{good}(Tmp) = \text{true}$ **then**
8: $Flag := \text{true}$;
9: $S_{res} := S_{res} \odot \{\mathbb{T}\text{s}\}$; // "$\odot$" is the concatenation operation
10: $S := S \setminus \{\mathbb{T}\text{s}\}$;
11: **exit from the cycle** $\{for\}$
12: **end if**
13: **end for**
14: **if** $Flag = \text{false}$ **then**
15: $S_{res} := S_{res} \odot \{\text{first}(S)\}$;
16: $S := S \setminus \{\text{first}(S)\}$;
17: **end if**
18: **end while**

Here:

good is the function that returns true/false depending on the fulfillment of the condition (9);

first is the function that returns the first element of a given sequence.

To compare: ranking results

Table 5. Ranking of articles according to algorithm on *Slide 16* concerning estimation (7).

No.	Author (s) and article heading	Estimation (7)	Estimation (8)
1	Vorontsov K. V., Makhina G. A. The principle of gap maximization for nearest neighbor monotonic classifier	0,07112036	0,07112036
2	Guz I. S. Hybrid estimations of complete cross-validation for monotonic classifiers	0,05185727	0,05185727
3	Khachay M. Yu. The convergence of empirical random processes generated by procedures of learning	0,05169631	0,05169631
4	Frei A. I. The method of generating and destroying sets for randomized minimization of empirical risk	0,03992817	0,03992817
5	Zhivotovskiy N. K. Combinatorial estimations for the probability of test error deviation from the cross-validation error	0,02178213	0,02178213
6	Kanevskiy D. Yu. Overfitting and combinatorial Rademacher complexity in regression recovery tasks	0,01969541	0,01969541
7	Nedelko V. M. Empirical confidence intervals for conditional risk in the classification problem	0,01851287	0,01851287
8	Botov P. V. Reducing the probability of overfitting for iterative methods of statistical learning	0,01731464	0,01731464
9	Ivakhnenko A. A., Vorontsov K. V. Informativity criteria for thresholded logical rules with the correction for overfitting of thresholds	0,01591723	0,01591723
10	Senko O. V., Kuznetsova A. V. Systems of reliable empirical regularities in models of optimal partitionings and methods to analyze them	0,00285329	0,03573024

To compare: ranking results

Table 6. Ranking of articles according to algorithm on *Slide 16* concerning estimation (8).

No.	Author (s) and article heading	Estimation (8)	Estimation (7)
1	Vorontsov K. V., Makhina G. A. The principle of gap maximization for nearest neighbor monotonic classifier	0,07112036	0,07112036
2	Guz I. S. Hybrid estimations of complete cross-validation for monotonic classifiers	0,05185727	0,05185727
3	Khachay M. Yu. The convergence of empirical random processes generated by procedures of learning	0,05169631	0,05169631
4	Frei A. I. The method of generating and destroying sets for randomized minimization of empirical risk	0,03992817	0,03992817
5	Senko O. V., Kuznetsova A. V. Systems of reliable empirical regularities in models of optimal partitionings and methods to analyze them	0,03573024	0,00285329
6	Zhivotovskiy N. K. Combinatorial estimations for the probability of test error deviation from the cross-validation error	0,02178213	0,02178213
7	Kanevskiy D. Yu. Overfitting and combinatorial Rademacher complexity in regression recovery tasks	0,01969541	0,01969541
8	Nedelko V. M. Empirical confidence intervals for conditional risk in the classification problem	0,01851287	0,01851287
9	Botov P. V. Reducing the probability of overfitting for iterative methods of statistical learning	0,01731464	0,01731464
10	Ivakhnenko A. A., Vorontsov K. V. Informativity criteria for thresholded logical rules with the correction for overfitting of thresholds	0,01591723	0,01591723

Main conclusions and open problems

1. The main *result* of current work is the proposed *method* for formation of reference text collection for revelation of dependencies within texts of a given scope. *Dependencies* here *can be arbitrary* and are not restricted to the co-occurrence of lexical units and their relationships typical for the most rational (i. e. standard) sense transfer.
2. The proposed solution gives *at least fivefold* reduction in the number of documents of minimally relevant to a given topical area when implementing the selection to the reference collection.
3. *The higher estimation of significance* for reference collection will have those documents, which *at greater number of phrases a higher average number of the most significant terms* per a one phrase *at minimum of its length*, contain. Substantially, this corresponds to the most *brief, but succinct* narration, that satisfy to the «good manner» rule of publications in Physics, Mathematics and Technical Sciences.
4. To improve the search accuracy for significant documents, it is of interest *to adapt offered estimations* to other linguistic levels in addition to lexical. The comparison of classifications relatively to different levels allows making a conclusion about document significance in disputable cases, for example, at non-fulfillment of *Statement 1* condition on one of the levels.