Estimation for the closeness to a semantic pattern without paraphrasing, and a hierarchy of topical texts

Mikhaylov D., Emelyanov G.

Yaroslav-the-Wise Novgorod State University

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

December 8–11, 2020

Moscow, Russian Federation

Text of scientific paper and rational (i. e. standard) sense transfer

Requirements for the solution

1. Hierarchization of information sources by degree of reflection of the most significant concepts of the studied subject area at maximal compactness and non-redundancy of narration.
2. The expert should not rephrase the text to search the semantically equivalent natural-language forms of description of knowledge unit.
3. Revelation of a set of text units and their relations necessary and enough to represent a knowledge unit and satisfies the sense standard.
4. The standard of a higher-level document must redefine the standard of a directly related lower-level document in the formed hierarchy.

A set of text units and their relationships, which is *necessary and enough* to represent a unit of knowledge, ensures the standard sense transfer.

Abstract and title of scientific paper

1. Reflect *the main content* and *the most important results* obtained by authors *without unnecessary methodological details*.
2. The title reflects *the name of method, model, algorithm* presented by paper, as well as the *theoretical basis* of the *proposed solutions*.

Closest researches and projects

- Probabilistic topic modeling and exploratory search [Vorontsov K., 2019].
- Hierarchical thematic modeling of major conference proceedings [Strijov V., 2014].
- Quantile-base approach to measuring cognitive complexity of text [Eremeev M., 2019].
- Representation of the ontology of an image analysis domain through a thesaurus [CC RAS, The Black Square System].
- Preparation of tagged text corpora for training the system of automatic paraphrasing [the ParaPhraser project].

Main problems:

- the proper qualitative analysis of linguistic expressional means, meaningful for choosing the best variants among possible paraphrases, is not provided;
- it is necessary (de facto) to reveal and analyze the relationship of standards for separate text documents for estimating their mutual complexity.

TF-IDF metrics and key word combinations

According to classic definition, TF-IDF is the product of two statistics:
term frequency (TF) and *inverse document frequency (IDF)*.

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.

The value of IDF is unique for each unique word in corpus D and can be determined as follows:

$$\mathrm{idf}\,(t_i, D) = \log\left(\frac{|D|}{|D_i|}\right), \tag{2}$$

where numerator represents the total number of documents in corpus,
and $|D_i \subset D|$ is a number of documents where the word t_i appears.

Interpreting the TF-IDF for word combinations, let's identify the numerator value in (1) with the number of co-occurrences of all combination words in the phrases of given $d \in D$; when calculating the value in denominator of (1) we'll separately take into account the cases of co-occurrence of combination words and occurrence without simultaneous presence in a phrase.

Clustreing the vocabulary of initial phrase by TF-IDF: basic assumptions

1. The words, which are the most unique in document and have the largest values of TF*IDF, must be related to terms of document's topical area.
2. The fact that the term has synonyms at the same document means the decrease of TF metrics for this word relatively to given document.
3. For words of general vocabulary and for those terms which are prevail in corpus the value of IDF tends to zero.
4. Synonyms, unique for some documents of corpus, will have a higher values of IDF.

   For example: general-vocabulary words which are define the conversive replacements, like *«приводить ⇔ являться следствием»* (in Russian).

Statement 1

The value of TF-IDF metrics for key word combination should not be less than the minimum of values of the mentioned measure for its separate words.

Let

D be an initial text set considered as a topical corpus.

X be an ordered descending sequence of $\mathrm{tf}\,(t_i, d) \cdot \mathrm{idf}\,(t_i, D)$ values for all words t_i of initial phrase relatively to document $d \in D$.

F be the sequence of clusters $H_1, \ldots H_r$ as a result of splitting the initial X by means of algorithm close to FOREL class taxonomy algorithms.

As the mass center of cluster H_i the arithmetic mean of all $x_j \in H_i$ is taken.

Herewith elements of X *can be assigned to one cluster* if

$$\begin{cases} \left|\,\mathrm{mc}(X) - \mathrm{first}(X)\,\right| < \dfrac{\mathrm{mc}(X)}{4} \\ \left|\,\mathrm{mc}(X) - \mathrm{last}(X)\,\right| < \dfrac{\mathrm{mc}(X)}{4} \end{cases}. \tag{3}$$

For estimating the phrase affinity to the sense standard
the most significant words are related to the clusters:

$H_1\,(X)$ — the *terms* from initial phrase which are the *most unique* for d;

$H_{r/2}\,(X)$ — *general vocabulary* as a basis of *synonymic paraphrases*, and those *terms* which have *synonyms*;

$H_r\,(X)$ — those *terms* which are *prevail* in corpus.

Selecting the estimation for the phrase affinity to the sense standard

Basic empirical considerations

- the division into general vocabulary and terms should be expressed as much as possible;
- the words in clusters $H_1, \dots H_r$, formed by the TF-IDF of words of the source phrase relative to a certain $d \in D$, should be distributed more or less evenly;
- the number of resulted clusters on the sequence X must be close to three as much as possible at maximum of TF-IDF values for words related to the cluster H_1.

Documents of corpus D are sorted descending the product of estimations:

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

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

and, correspondingly,

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

where Σ_{H_1} is the sum of TF-IDF values for words related to the cluster H_1 concerning to $d \in D$;
$\sigma\left(\left|H_i,\, i = \left\{1,\, r/2,\, r\right\}\right|\right)$ is the RMSD of number of elements in $H_i \in \left\{H_1,\, H_{r/2},\, H_r\right\}$;
$\mathrm{len}\left(X\right)$ is the length of X.

Remarks

- in a case of $\Sigma_{H_1} = 0$ the value of val_1 is assumed to be zero;
- if the number of clusters TF-IDF-obtained is smaller than two, the values of $\left|H_{r/2}\right|$ and $\left|H_r\right|$ are assumed to be zero;
- in a case of only two TF-IDF-obtained clusters the value of $\left|H_r\right|$ is assumed to be zero.

Let

Ts be a *group of phrases*, first of which is *the title* of scientific article and others represent its *abstract*.

The first variant of estimation:

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

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 $Ts_i \in \mathrm{Ts}$.

Remarks

- the estimation (7) depends on the selection of corpus D by expert;
- the offered estimation *does not imply sorting* of phrases $Ts_i \in \mathrm{Ts}$ by *affinity to the sense standard* and corresponds essentially to the order of selection of articles with *analysis of the title at first*;
- the apriori assumption of maximal closeness to the standard exactly of the title of the article is not always performed in practice.

The second variant of estimation:

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

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

Statement 2

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

- the correctly application of *Statement 2* 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 2*.

Ranking of texts in the initial collection

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 2*

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

Here:

good is the function that returns true/false depending on the fulfillment of the condition (3);

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

Let's enter the following denotations:

$\mathbf{H}_1(Ts_i)$, $\mathbf{H}_{r/2}(Ts_i)$, and $\mathbf{H}_r(Ts_i)$ are sets of words of clusters H_1, $H_{r/2}$ and H_r, respectively, for the phrase $Ts_i \in \mathbf{Ts}$ relative to the document $d \in D$, concerning which the maximum of affinity to the standard has been achieved, $\mathbf{Ts} \in S_{res}$;

$\mathbf{H_1}(\mathbf{Ts}) = \bigcup_{Ts_i \in \mathbf{Ts}} \mathbf{H}_1(Ts_i)$;

$\mathbf{H}_{\overline{\mathbf{Z}}}(Ts_i)$ is a set of words of the phrase Ts_i with nonzero values of TF-IDF relative to the same document d;

$\mathbf{H}_{\overline{\mathbf{Z}}}(\mathbf{Ts}) = \bigcup_{Ts_i \in \mathbf{Ts}} \left(\mathbf{H}_{\overline{\mathbf{Z}}}(Ts_i) \setminus \mathbf{H}_1(Ts_i)\right)$.

Let $\mathbf{Ts}_i$ and $\mathbf{Ts}_j$ be texts from S_{res} and $i > j$, i. e., the rank of the article that matches a phrase group $\mathbf{Ts}_i$ is higher than for $\mathbf{Ts}_j$.

The main hypothesis

The measure of how the text $\mathbf{Ts}_j$ is complemented by a sense of the text $\mathbf{Ts}_i$ has corresponded to the value $\left|\left(\mathbf{H}_{\overline{\mathbf{Z}}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\right) \cap \mathbf{H_1}(\mathbf{Ts}_i)\right|$.

The sense complementarity of text $\mathbf{Ts}_j$ by text $\mathbf{Ts}_i$ can be defined as

$$K_1(\mathbf{Ts}_j, \mathbf{Ts}_i) = \frac{\left|\left(\mathbf{H}_{\overline{\mathbf{Z}}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\right) \cap \mathbf{H_1}(\mathbf{Ts}_i)\right|}{\left|\mathbf{H_1}(\mathbf{Ts}_i)\right|}. \quad (9)$$

Let

$\mathbf{Kw}\,(\mathbf{Ts}_i)$ be a set of key word combinations satisfying the condition of *Statement 1* and found for $\mathbf{Ts}_i$;

$\mathbf{H_{Kw}}\,(\mathbf{Ts}_i)$ be the set of words within the mentioned combinations.

Let's enter into consideration $\mathbf{Kw'}\,(\mathbf{Ts}_j, \mathbf{Ts}_i) \subset \mathbf{Kw}\,(\mathbf{Ts}_i)$, which includes combinations of words of set $\mathbf{H}_{\overline{\mathbf{Z}}}\,(Ts_{jk}) \setminus \mathbf{H_1}\,(\mathbf{Ts}_j)$ for each phrase $Ts_{jk} \in \mathbf{Ts}_j$, and for each combination, at least one word must belong to $\mathbf{H_1}\,(\mathbf{Ts}_i)$.

Taking into account the sought word combinations, estimation (9) takes the following form:

$$K_2(\mathbf{Ts}_j, \mathbf{Ts}_i) = \\ = \frac{\left|\mathbf{Kw'}(\mathbf{Ts}_j, \mathbf{Ts}_i)\right| + \left|\left(\left(\mathbf{H}_{\overline{\mathbf{Z}}}\,(\mathbf{Ts}_j) \setminus \mathbf{H_1}\,(\mathbf{Ts}_j)\right) \cap \mathbf{H_1}\,(\mathbf{Ts}_i)\right) \setminus \mathbf{H_{Kw'}}\,(\mathbf{Ts}_j, \mathbf{Ts}_i)\right|}{\left|\mathbf{Kw}(\mathbf{Ts}_i)\right| + \left|\mathbf{H_1}\,(\mathbf{Ts}_i) \setminus \mathbf{H_{Kw}}\,(\mathbf{Ts}_i)\right|}, \quad (10)$$

where $\mathbf{H_{Kw'}}\,(\mathbf{Ts}_j, \mathbf{Ts}_i)$ is the set of words in the combination from $\mathbf{Kw'}(\mathbf{Ts}_j, \mathbf{Ts}_i)$.

Representation of words of phrase $Ts_{jk} \in \mathbf{Ts}_j$ in clusters $\left\{H_1, H_{r/2}, H_r\right\} := Cl$:

$$\mathrm{N}\Big(Ts_{jk}, Cl\left(Ts_{jk}\right)\Big) = \frac{\sqrt{\sum_{\mathrm{m}\in\{1,\mathrm{r}/2,\mathrm{r}\}}\Big(\left|\mathbf{H}_{\mathrm{m}}\left(Ts_{jk}\right)\right|/\mathrm{len}\left(Ts_{jk}\right)\Big)^2}}{\sigma\Big(\left|\mathbf{H}_{\mathrm{m}}\left(Ts_{jk}\right)\right|/\mathrm{len}\left(Ts_{jk}\right)\Big)+1}, \qquad (11)$$

where $\mathrm{len}\left(Ts_{jk}\right)$ is the number of words in phrase Ts_{jk}.

Remarks

- if the number of clusters according to the TF-IDF is less than two, then values $\left|\mathbf{H}_{\mathrm{r}/2}\left(Ts_{jk}\right)\right|$ and $\left|\mathbf{H}_{\mathrm{r}}\left(Ts_{jk}\right)\right|$ are taken equal to zero;
- in case of exactly two clusters, value $\left|\mathbf{H}_{\mathrm{r}}\left(Ts_{jk}\right)\right|$ is considered zero.

Let's define the complement of standard for text $\mathbf{Ts}_j$ with the standard of $\mathbf{Ts}_i$, by introducing sets

$$\mathbf{H}'_1\left(Ts_{jk}, \mathbf{Ts}_i\right) = \mathbf{H}_1\left(Ts_{jk}\right) \cup \left(\left(\mathbf{H}_{\overline{\mathrm{Z}}}\left(Ts_{jk}\right) \setminus \mathbf{H}_1\left(Ts_{jk}\right)\right) \cap \mathbf{H_1}\left(\mathbf{Ts}_i\right)\right);$$

$$\mathbf{H}'_{\mathrm{r}/2}\left(Ts_{jk}, \mathbf{Ts}_i\right) = \mathbf{H}_{\mathrm{r}/2}\left(Ts_{jk}\right) \setminus \left(\left(\mathbf{H}_{\overline{\mathrm{Z}}}\left(Ts_{jk}\right) \setminus \mathbf{H}_1\left(Ts_{jk}\right)\right) \cap \mathbf{H_1}\left(\mathbf{Ts}_i\right)\right);$$

$$\mathbf{H}'_{\mathrm{r}}\left(Ts_{jk}, \mathbf{Ts}_i\right) = \mathbf{H}_{\mathrm{r}}\left(Ts_{jk}\right) \setminus \left(\left(\mathbf{H}_{\overline{\mathrm{Z}}}\left(Ts_{jk}\right) \setminus \mathbf{H}_1\left(Ts_{jk}\right)\right) \cap \mathbf{H_1}\left(\mathbf{Ts}_i\right)\right).$$

into estimation (11), which itself takes the following form:

$$\mathrm{N}'\Big(Ts_{jk}, Cl\left(Ts_{jk}\right), \mathbf{Ts}_i\Big) = \frac{\sqrt{\sum_{\mathrm{m}\in\{1,\mathrm{r}/2,\mathrm{r}\}}\Big(\left|\mathbf{H}'_{\mathrm{m}}\left(Ts_{jk}, \mathbf{Ts}_i\right)\right|/\mathrm{len}\left(Ts_{jk}\right)\Big)^2}}{\sigma\Big(\left|\mathbf{H}'_{\mathrm{m}}\left(Ts_{jk}, \mathbf{Ts}_i\right)\right|/\mathrm{len}\left(Ts_{jk}\right)\Big)+1}, \qquad (12)$$

Choosing higher-level document in the formed hierarchy

$$\mathrm{N}_3\left(\mathbf{Ts}_j\right) = \frac{\mathrm{N}\left(Ts_{j1}, Cl\left(Ts_{j1}\right)\right)}{\sigma\left(\left\{\mathrm{N}\left(Ts_{jk}, Cl\left(Ts_{jk}\right)\right) : Ts_{jk} \in \mathbf{Ts}_j\right\}\right) + 1}, \tag{13}$$

$$\mathrm{N}_4\left(\mathbf{Ts}_j\right) = \frac{\max\left[\left\{\mathrm{N}\left(Ts_{jk}, Cl\left(Ts_{jk}\right)\right) : Ts_{jk} \in \mathbf{Ts}_j\right\}\right]}{\sigma\left(\left\{\mathrm{N}\left(Ts_{jk}, Cl\left(Ts_{jk}\right)\right) : Ts_{jk} \in \mathbf{Ts}_j\right\}\right) + 1}, \tag{14}$$

$$\mathrm{N}'_3\left(\mathbf{Ts}_j, \mathbf{Ts}_i\right) = \frac{\mathrm{N}'\left(Ts_{j1}, Cl\left(Ts_{j1}\right), \mathbf{Ts}_i\right)}{\sigma\left(\left\{\mathrm{N}'\left(Ts_{jk}, Cl\left(Ts_{jk}\right), \mathbf{Ts}_i\right) : Ts_{jk} \in \mathbf{Ts}_j\right\}\right) + 1}, \tag{15}$$

$$\mathrm{N}'_4\left(\mathbf{Ts}_j, \mathbf{Ts}_i\right) = \frac{\max\left[\left\{\mathrm{N}'\left(Ts_{jk}, Cl\left(Ts_{jk}\right), \mathbf{Ts}_i\right) : Ts_{jk} \in \mathbf{Ts}_j\right\}\right]}{\sigma\left(\left\{\mathrm{N}'\left(Ts_{jk}, Cl\left(Ts_{jk}\right), \mathbf{Ts}_i\right) : Ts_{jk} \in \mathbf{Ts}_j\right\}\right) + 1}, \tag{16}$$

Statement 3

The criterion for choosing higher-level text $\mathbf{Ts}_i$ for given text $\mathbf{Ts}_j$ in the formed hierarchy is the non-decrease of values of estimations (15) and (16) with respect to their corresponding estimations (13) and (14) at maximizing of estimations (9) and (10).

Remark

Let's name further the first terms in the denominators of formulas (13) and (14) as the RMSD of estimation (11), the same for (15) and (16) — as the RMSD of estimation (12), respectively.

Initial data for experiment: the structure of text corpus

- 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 documents included in 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 experiment: collections for selecting the articles

- 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

- Estimations (4)–(8) 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

Table 1. Ranking of articles according to algorithm on *Slide 10* concerning estimation (7).

№	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

Table 2. Ranking of articles according to algorithm on *Slide 10* concerning estimation (8).

№	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

Table 3. Complementarity of texts in sense without taking keyword combinations into account[1].

$j = 2, i = 1$	Estimation (9) 0,42857143
$\left(\mathbf{H}_{\overline{\mathbf{Z}}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\right) \cap \mathbf{H_1}(\mathbf{Ts}_i)$	*обобщать, монотонный, способность*
$\mathbf{H_1}(\mathbf{Ts}_j)$	*контроль, скользящий, выборка*
$\mathbf{H_1}(\mathbf{Ts}_i)$	*контроль, монотонный, скользящий, обобщать, способность, близкий, сосед*
$j = 6, i = 1$	Estimation (9) 0,28571429
$\left(\mathbf{H}_{\overline{\mathbf{Z}}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\right) \cap \mathbf{H_1}(\mathbf{Ts}_i)$	*обобщать, способность*
$\mathbf{H_1}(\mathbf{Ts}_j)$	*комбинаторный, семейство, обучать, завышенность*
$j = 9, i = 1$	Estimation (9) 0,14285714
$\left(\mathbf{H}_{\overline{\mathbf{Z}}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\right) \cap \mathbf{H_1}(\mathbf{Ts}_i)$	*монотонный*
$\mathbf{H_1}(\mathbf{Ts}_j)$	*связность, переобучение*
$j = 4, i = 3$	Estimation (9) 1,00000000
$\left(\mathbf{H}_{\overline{\mathbf{Z}}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\right) \cap \mathbf{H_1}(\mathbf{Ts}_i)$	*эмпирический*
$\mathbf{H_1}(\mathbf{Ts}_j)$	*минимизация, обобщать, способность, риск, комбинаторный*
$\mathbf{H_1}(\mathbf{Ts}_i)$	*эмпирический*

[1] Here and further i and j are the serial numbers of the documents by *Table 1*.

Continuation of table 3.

$j = 6, i = 4$	Estimation (9) 0,40000000
$\left(\mathbf{H}_{\overline{\mathbf{Z}}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\right) \cap \mathbf{H_1}(\mathbf{Ts}_i)$	*обобщать, способность*
$\mathbf{H_1}(\mathbf{Ts}_j)$	*комбинаторный, семейство, обучать, завышенность*
$\mathbf{H_1}(\mathbf{Ts}_i)$	*минимизация, обобщать, способность, риск, комбинаторный*
$j = 8, i = 4$	Estimation (9) 0,40000000
$\left(\mathbf{H}_{\overline{\mathbf{Z}}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\right) \cap \mathbf{H_1}(\mathbf{Ts}_i)$	*минимизация, риск*
$\mathbf{H_1}(\mathbf{Ts}_j)$	*семейство, статистический, способность, комбинаторный, обобщать, вероятность*
$j = 9, i = 4$	Estimation (9) 0,20000000
$\left(\mathbf{H}_{\overline{\mathbf{Z}}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\right) \cap \mathbf{H_1}(\mathbf{Ts}_i)$	*комбинаторный*
$\mathbf{H_1}(\mathbf{Ts}_j)$	*связность, переобучение*
$j = 9, i = 5$	Estimation (9) 0,20000000
$\left(\mathbf{H}_{\overline{\mathbf{Z}}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\right) \cap \mathbf{H_1}(\mathbf{Ts}_i)$	*комбинаторный*
$\mathbf{H_1}(\mathbf{Ts}_j)$	*связность, переобучение*
$\mathbf{H_1}(\mathbf{Ts}_i)$	*комбинаторный, связность, контроль, скользящий, выборка*

Here and further in *Tables 5* and *6* cells for relationships that meet the condition of *Statement 3* are highlighted *by green*; in a case of partially meeting this condition, the highlighting color is *yellow*.

End of table 3.

$j = 9, i = 6$	Estimation (9) 0,25000000
$\left(\mathbf{H}_{\overline{\mathbf{Z}}}\left(\mathbf{Ts}_j\right) \setminus \mathbf{H_1}\left(\mathbf{Ts}_j\right)\right) \cap \mathbf{H_1}\left(\mathbf{Ts}_i\right)$	*комбинаторный*
$\mathbf{H_1}\left(\mathbf{Ts}_j\right)$	*связность, переобучение*
$\mathbf{H_1}\left(\mathbf{Ts}_i\right)$	*комбинаторный, семейство, обучать, завышенность*
$j = 8, i = 7$	Estimation (9) 0,33333333
$\left(\mathbf{H}_{\overline{\mathbf{Z}}}\left(\mathbf{Ts}_j\right) \setminus \mathbf{H_1}\left(\mathbf{Ts}_j\right)\right) \cap \mathbf{H_1}\left(\mathbf{Ts}_i\right)$	*риск*
$\mathbf{H_1}\left(\mathbf{Ts}_j\right)$	*семейство, статистический, способность, комбинаторный, обобщать, вероятность*
$\mathbf{H_1}\left(\mathbf{Ts}_i\right)$	*достоинство, риск, эмпирический*
$j = 9, i = 8$	Estimation (9) 0,33333333
$\left(\mathbf{H}_{\overline{\mathbf{Z}}}\left(\mathbf{Ts}_j\right) \setminus \mathbf{H_1}\left(\mathbf{Ts}_j\right)\right) \cap \mathbf{H_1}\left(\mathbf{Ts}_i\right)$	*комбинаторный, вероятность*
$\mathbf{H_1}\left(\mathbf{Ts}_j\right)$	*связность, переобучение*
$\mathbf{H_1}\left(\mathbf{Ts}_i\right)$	*семейство, статистический, способность, комбинаторный, обобщать, вероятность*

Remarks

- the relationship of the documents is excluded from consideration, if the values of estimations (9) and (10) are simultaneously zero;
- estimation (10) is calculated only when $|\mathbf{Kw}\left(\mathbf{Ts}_i\right)| > 0$;
- when $|\mathbf{Kw}\left(\mathbf{Ts}_i\right)| = 0$, the relationship is not considered in a case of zero value of (9).

Table 4. Complementarity of texts in sense taking keyword combinations into account.

$j = 2, i = 1$	Estimation (10)	0,40000000
$\mathbf{Kw}\,(\mathbf{Ts}_i)$	*ближайший сосед, скользящий контроль, обобщающая способность, разделяющая поверхность*	
$\mathbf{Kw}'\,(\mathbf{Ts}_j, \mathbf{Ts}_i)$	*обобщающая способность*	
$j = 6, i = 1$	Estimation (10)	0,20000000
$\mathbf{Kw}'\,(\mathbf{Ts}_j, \mathbf{Ts}_i)$	*обобщающая способность*	
$j = 9, i = 1$	Estimation (10)	0,20000000
$\mathbf{Kw}'\,(\mathbf{Ts}_j, \mathbf{Ts}_i)$	—	
$j = 6, i = 4$	Estimation (10)	0,16666667
$\mathbf{Kw}\,(\mathbf{Ts}_i)$	*обобщающая способность, комбинаторная теория, минимизация эмпирического риска*	
$\mathbf{Kw}'\,(\mathbf{Ts}_j, \mathbf{Ts}_i)$	*обобщающая способность*	
$j = 8, i = 4$	Estimation (10)	0,33333333
$\mathbf{Kw}'\,(\mathbf{Ts}_j, \mathbf{Ts}_i)$	—	
$j = 9, i = 4$	Estimation (10)	0,16666667
$\mathbf{Kw}'\,(\mathbf{Ts}_j, \mathbf{Ts}_i)$	—	
$j = 9, i = 5$	Estimation (10)	0,16666667
$\mathbf{Kw}\,(\mathbf{Ts}_i)$	*скользящий контроль*	
$\mathbf{Kw}'\,(\mathbf{Ts}_j, \mathbf{Ts}_i)$	—	
$j = 9, i = 8$	Estimation (10)	0,40000000
$\mathbf{Kw}\,(\mathbf{Ts}_i)$	*обобщающая способность*	
$\mathbf{Kw}'\,(\mathbf{Ts}_j, \mathbf{Ts}_i)$	—	

Table 5. Assessment of the representation of words in the clusters that are the most significant for the standard.

j	$\mathrm{N}_3\,(\mathbf{Ts}_j)$	$\mathrm{N}_4\,(\mathbf{Ts}_j)$	RMSD of estimation (11)
2	0,442059165587	0,502119662613	0,048678362277
4	0,376446598212	0,529404830202	0,066634500491
6	0,362818302898	0,504283491203	0,106699761390
8	0,452293583860	0,452293583860	0,058355201581
9	0,346816072806	0,346816072806	0,021096101739

Table 6. Assessment of the representation of words in the clusters that are the most significant for the standard taking relationships of documents into account.

$j \to i$	$\mathrm{N}'_3\,(\mathbf{Ts}_j, \mathbf{Ts}_i)$	$\mathrm{N}'_4\,(\mathbf{Ts}_j, \mathbf{Ts}_i)$	RMSD of estimation (12)
$2 \to 1$	0,528411029776	0,537246561748	0,0387975635841
$6 \to 1$	0,365859769250	0,508510844834	0,0974995421573
$9 \to 1$	0,346022664877	0,346022664877	0,0234374100578
$4 \to 3$	0,457707643226	0,554500713867	0,0458362979168
$6 \to 4$	0,365859769250	0,508510844834	0,0974995421573
$8 \to 4$	0,457175036510	0,457175036510	0,0470546921663
$9 \to 4$	0,346022664877	0,346022664877	0,0234374100578
$9 \to 5$	0,346022664877	0,346022664877	0,0234374100578
$9 \to 6$	0,346022664877	0,346022664877	0,0234374100578
$8 \to 7$	0,454613088142	0,454613088142	0,0529553143232
$9 \to 8$	0,341968227624	0,376375190389	0,0355714693817

Example: MMPR-15, Statistical Learning Theory

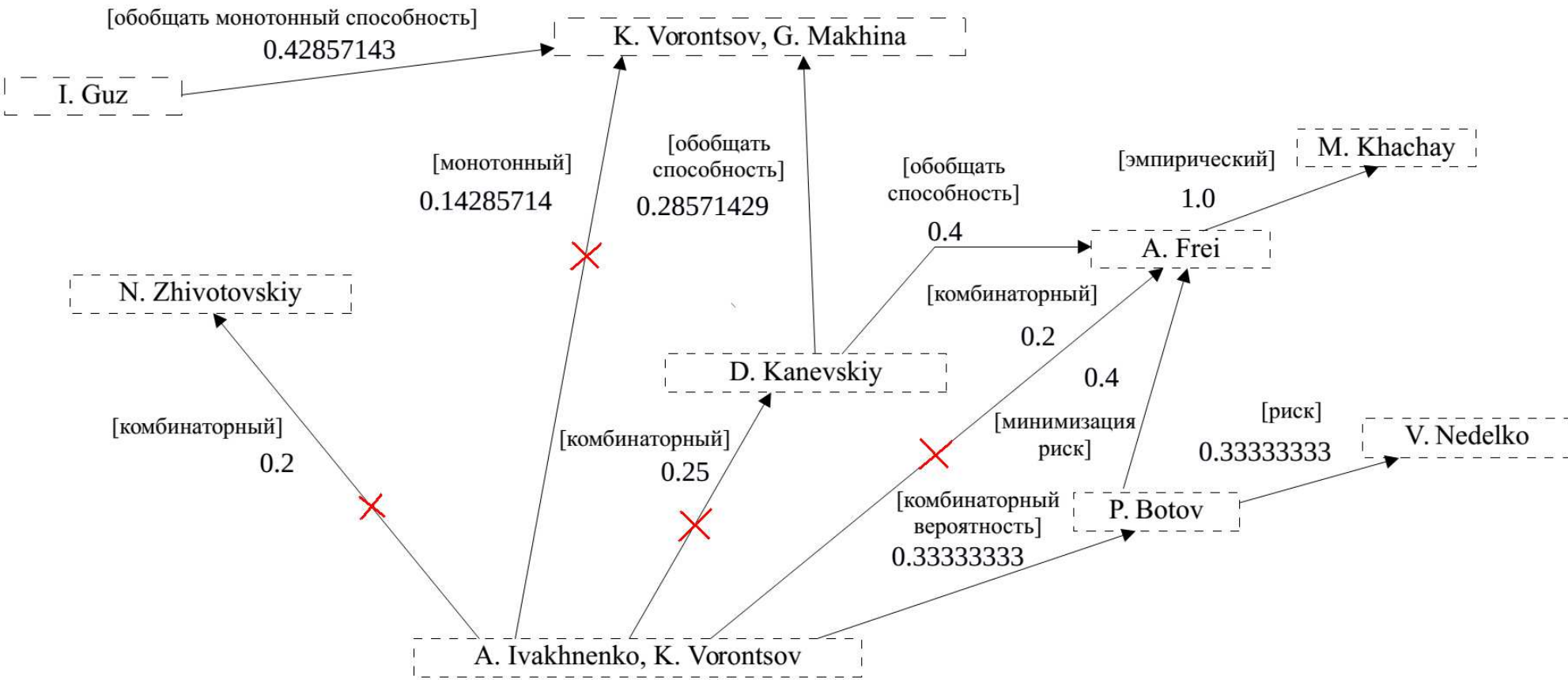

Fig. 1. Hierarchization of documents without taking keyword combinations into account [2, 3].

[2] For each relationship in "[]" the words from $\left(\mathbf{H}_{\overline{\mathbf{Z}}}\left(\mathbf{Ts}_j\right) \setminus \mathbf{H_1}\left(\mathbf{Ts}_j\right)\right) \cap \mathbf{H_1}\left(\mathbf{Ts}_i\right)$ are shown.

[3] Edges for relationships that do not meet the condition of *Statement 3* are marked by "×".

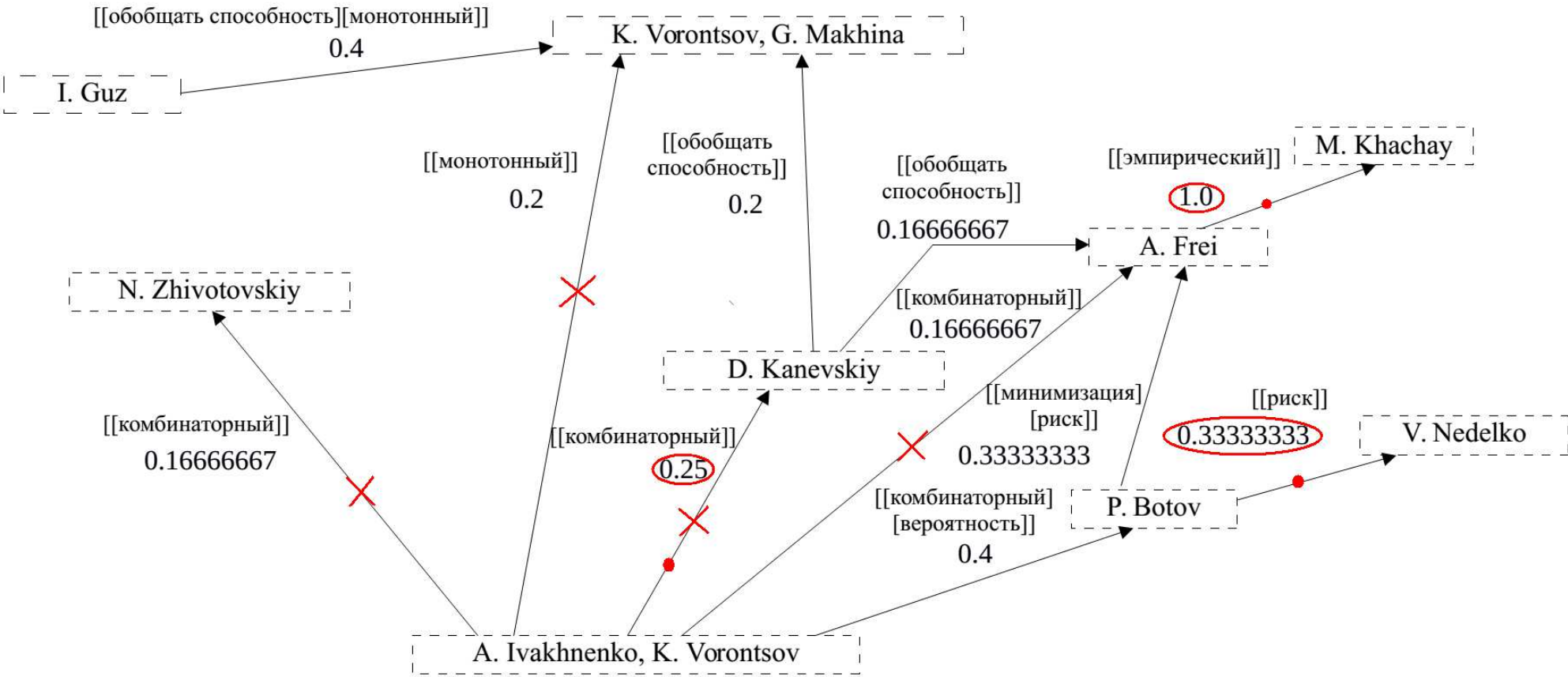

Fig. 2. Hierarchization of documents taking keyword combinations into account [4, 5].

[4] In a case of $|\mathbf{Kw}(\mathbf{Ts}_i)| = 0$, edges show the values of estimation **(9)** and marked by "●".

[5] Edges for relationships that do not meet the condition of *Statement 3* are marked by "×".

Conclusions

1. The main *result* of current work is the proposed *method* for hierarchization of the texts of a subject-oriented natural language based on estimations of the closeness of a topical text to the sense standard.
2. Its *effectiveness* can be *estimated* by the number and type of connectivity components of the graph obtained from the graph of found relationships for the collection by replacing the oriented edges with non-oriented ones.
3. After removing those links that do not meet the condition of *Statement 3* from the original graph, the *subgraph* that corresponds to the maximum connectivity component *contains* the vertex for the article with the maximum total rating for the collection *among the vertices* with the maximum degree.
4. At the expense of articles that are not reflected in the maximum connectivity component, we have *at least a* 20% reduction *in the number of documents* that should be primarily considered when studying a given topical area.
5. It is of interest *to study the relationship* between
   - *the distributions* of frequencies of occurrence of words in the clusters of the highest TF-IDF for phrases of different texts of the analyzed collection;
   - *the cases* of achieving the maximal product of estimations (4), (5) and (6) relative to specific documents of the given text corpus.