Ranking of documents of topical corpus according to their mutual relevance in the problem of estimating of affinity of a text to the sense standard

D V Mikhaylov and G M Emelyanov
Yaрослслав-те-Wис Новгород Государственный Университет, Velikiy Novgorod, Russian Federation
E-mail: Dmitry.Mikhaylov@novsu.ru

Abstract. The offered paper is devoted to the problem of oneness and integrity of image for the semantic pattern (i.e., sense standard) revealed phrase by phrase for some text within a topical collection. One phrase corresponds here to an extended natural-language sentence. The basis of estimating affinity to the standard is the classifying of words of each phrase in a text according to the TF-IDF value relative to some text corpus. Texts to the corpus are pre-selected by an expert. The essence of the problem: for each phrase, its maximal affinity to the sense standard is achieved concerning the individual corpus document, and, consequently, it is necessary to estimate the mutual relevance of such documents concerning different phrases of the analyzed text. Based on distances between vectors of TF-IDF for words of a separate phrase obtained relative to different corpus documents, the significance estimation for each such document is entered into consideration to choose a pair of mutual relevant.

1 Introduction
The development of e-courses requires systematizing expert knowledge represented by natural-language texts in appropriate subject areas. The major requirement here is the sorting of information sources by the degree of reflection of the most significant concepts of the studied subject area at maximal compactness and non-redundancy of narration satisfies the sense standard. The affinity of a text to its standard must be estimated herewith without paraphrasing. The approach offered by us in [1] to solve the mentioned problem is based on classifying of words of each sentence by the TF-IDF values relatively to texts of corpus pre-formed by an expert. As analyzed texts, the abstracts of scientific articles together with their titles are considered. Since for each phrase, its maximal affinity to a standard is reached concerning the specific document of the topical text corpus \( D \), then the sets of text units and relationships inside of such sets from the expressing the semantic patterns of separate phrases will be revealed concerning the different documents \( d \in D \). The task is to estimate the mentioned constituents for assignment to a single image. In the current paper, the given task is solved by matching the classifications of words of each phrase of analyzed text according to TF-IDF values concerning different texts of a given topical corpus.

2. Proposed approach
Let \( \mathcal{T} \) be a group of phrases, the first of which may be interpreted as the article title, and others represent the abstract of this article. Let’s enter into consideration for each phrase \( T_s \in \mathcal{T} \) the vector of TF-IDF values of its words:
which is calculated concerning the document $d_j$ from the topical corpus $D$, where $\text{len}(T_{S_i})$ is the length of the phrase $T_{S_i}$ measured in words. The TF-IDF measure itself is the product of the ratio of the number of occurrences of a word to the total number of words in the document (term frequency, TF) and the inverse document frequency (IDF), $[2]$, \( \text{idf}(t_k, D) = \log(|D|/|D_k|) \), where $|D_k \subset D|$ is the number of those documents, in which the word $t_k$ occurred at least once. Let $X$ be a descending order sequence of components of vector (1). Let’s split $X$ into clusters $H_1, \ldots, H_r$ employing the algorithm close to algorithms of the FOREL class [3]. Further in the current paper, as applied to splitting the sequence into clusters, we will have in mind this algorithm. As the mass center of cluster $H_i$, the arithmetic mean of all elements $x_j \in H_i$ like in [1] is taken here. The rule for relating elements of $X$ to the same cluster is also identical to the one used previously. To estimate the affinity of a phrase to the sense standard the substantially interesting will be the words of clusters: $H_1$ (the terms of the source phrase which are the most unique for the analyzed document), $H_{r/2}$ (general vocabulary that ensures periphrases and synonymous terms), and also $H_r$ (terms that are prevailed in the corpus). The estimation itself was used by authors earlier in [1] and is based on the following empirical considerations. Firstly, the division of words in a phrase into general vocabulary and terms should be expressed as much as possible. Secondly, the words in clusters $H_1, \ldots, H_r$ should be distributed more or less evenly. Thirdly, the number of resulted clusters must be close to three as much as possible at a maximum of TF-IDF values for words associated with cluster $H_i$. As the estimation of affinity to the standard for a phrase here the maximum of product of estimations

\[
\text{val}_1 = -1/\log_{10}(\Sigma_{H_1}), \quad \text{val}_2 = 10^{-\sigma(|H_i:i=\{1, r/2, r\}|)}
\]

and, correspondingly,

\[
\text{val}_3 = |H_1 \setminus H_{r/2} \setminus H_r|/\text{len}(X)
\]

calculated for documents $d \in D$, is taken.

The value under logarithm in the denominator of formula (2) is the sum of TF-IDF values for words associated with the cluster $H_1$ by the value of this measure relative to considered document $d \in D; \sigma(|H_i, i = \{1, r/2, r\}|)$ in formula (3) is the root-mean-square deviation of the number of elements in the cluster from list $\{H_1, H_{r/2}, H_r\}$; $\text{len}(X)$ in the denominator of expression (4) is the length of sequence $X$. In the case when $\Sigma_{H_1} = 0$, the value of $\text{val}_1$ is assumed to be equal to zero. If the number of TF-IDF clusters is less than two, the values of $|H_{r/2}|$ and $|H_r|$ are taken equal to zero. In the case of exactly two clusters, the value of $|H_1|$ is considered to be zero.

Let $T_{S_{\text{mx}(d,i)}}$ be the vector (1) for the document $d_{\text{mx}(i)} \in D$ concerning which the maximum of product of estimations (2), (3), and (4) is reached for $T_{S_i} \in T_i$. Let’s designate as $T_i$ the sequence of vectors (1) obtained for $T_{S_i}$ concerning documents $d_j \in D: d_j \neq d_{\text{mx}(i)}$. The sequence is sorted by descending the Euclidean distance to $T_{S_{\text{mx}(d,i)}}$. Let’s split $T_i$ into clusters $H_1^{E}, \ldots, H_r^{E}$, where $H_r^{E}$ according to the definition will correspond to documents with the least distance to $d_{\text{mx}(i)}$.

\textbf{Definition 1.} We’ll assume, that the classification of words of a phrase $T_{S_i} \in T_i$ according to the value of TF-IDF measure concerning some document $d_j \in D$ can be considered as comparable with the analogous classification concerning $d_{\text{mx}(i)}$ at the fulfillment of one of two conditions:

- $d_j \in H_r^{E}$
- $\exists T_{S_j} \in T_j: T_{S_j} \neq T_{S_i}, d_j = d_{\text{mx}(j)}$, herewith $\exists d_k \in D: d_k \neq d_j, d_k \neq d_{\text{mx}(i)}$, and $d_k$ itself simultaneously relates to both $H_r^{E}$ and $H_r^{E}$. 

\[
\text{val}_4 = \frac{\text{val}_1 \cdot \text{val}_2 \cdot \text{val}_3}{\text{val}_4}
\]

\[
\text{val}_5 = \text{val}_1 \cdot \text{val}_2 \cdot \text{val}_3
\]
Let’s name further such documents $d_j$ and $d_{\text{min}(j)}$ as mutually relevant by TF-IDF.

Let’s enter into consideration the graph, which vertexes are corresponded to those documents concerning which the maximum of affinity to a sense standard is reached at least for one phrase $Ts_i \in Ts$, and each edge connects vertexes for a pair of documents mutually relevant by TF-IDF. We’ll name further such a graph as the relevancy graph. Entering the mentioned graph into consideration allows us to make an important practical conclusion concerning the building of a hierarchy of texts with the application of techniques offered by authors in the paper [4].

**Statement 1.** If relevancy graphs of all texts of the analyzed collection will be united, then for higher-level text $Ts_j$ and lower-level text $Ts_j$ directly related with it in the formed hierarchy their relevancy graphs will be subgraphs of some connectivity component of the united relevancy graph for collection.

All other things stipulated in [4] being equal, when choosing a higher-level text for the given text $Ts_j$ in the formed hierarchy the preference will be given to those text $Ts_i$, which meets the condition of Statement 1. Let $S$ be the sequence of texts of the initial collection. $U_S$ be the united set of phrases for all texts $Ts$ within $S$. Then the estimation of the significance of document $d \in D$ for the formation of the relevancy graph of text $Ts$ can be defined from geometrical considerations (by analogy with estimations of affinity to the standard for $Ts$, which were developed earlier by us and applied in paper [4]) as

$$N(d) = \frac{|D| - \min_{d'} \left\{ \sum_{Ts \in U_S} \text{dist}(d, Ts) \right\}}{\sigma_d \left( \sum_{Ts \in U_S} \text{dist}(d, Ts) \right) + 1}.$$  \hspace{1cm} (5)

The first summand in the denominator of formula (5) is the root-mean-square deviation of the number of elements for clusters that are obtained according to Euclidean distance relatively to document $d \in D$ for different texts of the given collection $S$. The subtrahend in the numerator here is the minimum of the number of elements of a cluster of least Euclidean distances relative to the abovementioned $d$, also for different texts of the collection $S$.

As yet mentioned by us in [5], in practical applications of FOREL-family algorithms an a priori knowledge about the width (i.e. diameter) of a cluster is required to minimize computational costs of recalculation of values of the quality measure. Regarding the considered clustering of documents $d \in D$ the quality of splitting of documents into classes assumes, from one side, as much number of clusters as possible at minimizing the diameter of a cluster of least distances for separate phrases, and from another side – a minimum of variation of the number of cluster elements in general. Essentially, the estimation (5) allows revealing those documents of text corpus, concerning which the considered classification according to Euclidean distance is the most expressed.

### 3. Experimental research

To test the proposed approach, the material from paper [4] was involved as the corpus $D$ formed by an expert. The collections from which papers are selected also coincide with those used in [4] and include:

- Proceedings of the Conference “Intelligent Information Processing” (IIP-9), section “Theory and Methods of Pattern Recognition and Classification” (14 papers);
- Proceedings of the All-Russian Conference with International Participation “Mathematical Methods for Pattern Recognition” (MMPR-14, 2009), section “Methods and Models of Pattern Recognition and Forecasting” (35 papers);
- Proceedings of the MMPR-15 conference (2011), sections “Theory and Methods of Pattern Recognition and Classification” (18 papers) and “Statistical Learning Theory” (10 papers).

The software implementation (in Python 2.7) of the offered solutions and experimental results are presented on the NovSU website at http://www.novsu.ru/file/1752845 in a ZIP archive (the version dated April 9, 2021). The archive includes a directory with the results of experiments and a directory
of the Test_Python_PyDev workspace of the Eclipse environment [6], where the TestMorpho/src subdirectory contains the TestMorpho.py file of the project’s main module, text corpora directory (my_corpora), the example of original phrases (text.txt) and the result of their processing (issue.txt). Also, the TestMorpho/src subdirectory contains the max_affinity_to_standard_achieved.txt file with the results of calculation of estimation (5) for documents of corpus which was chosen by a user. The directory with the results of the experiments includes (inside subdirectories for variants of TF-IDF calculation) directories with the results of processing of individual topical collections. The structure of these directories is similar to the one described in the paper [4]. Besides the described in [4], for each collection, such catalog contains the connected_total.txt file with the results of connectivity estimating for the united relevancy graph for collection.

Further in the tables, the experimental results are represented for the “Statistical Learning Theory” section of the MMPR-15 proceedings. The variant of the calculation formula for TF-IDF is “classic” and was used before in papers [1, 4, 5]. In addition to the publications in proceedings of the MMPR conferences, in Table 1 three papers from “Taurida journal of computer science theory and mathematics” (TJCSTM) were published in a period from 2002nd to 2004th years, are represented.

Table 1. Documents $d \in D$ and maximum of affinity to the standard for separate phrases.

| No. | Author(s), paper title, and output data (in Russian and English) |
|-----|------------------------------------------------------------------|
| 1   | Воронцов К.В. Обзор современных исследований по проблеме качества обучения алгоритмов // ТВИМ. 2004. №1 / Vorontsov K.V. The review of contemporary investigations at the problem of quality of learning of algorithms // TJCSTM. 2004. No.1 |
| 2   | Воронцов К.В. Комбинаторная теория переобучения: результаты, приложения и открытые проблемы // ММРО-15 / Vorontsov K.V. The combinatorial theory of overfitting: results, applications, and open problems // MMPR-15 |
| 3   | Дюличева Ю.Ю. Стратегии редукции решающих деревьев (обзор) // ТВИМ. 2002. №1 / Dyulicheva Yu.Yu. Pruning strategies of decision trees (review) // TJCSTM. 2002. No.1 |
| 4   | Дюкова Е.В., Песков Н.В. Об алгоритме классификации на основе полного решающего дерева // ММРО-13 / Dyukova E.V., Peskov N.V. About classification algorithm based on complete decision tree // MMPR-13 |
| 5   | Дюличева Ю.Ю. О программной реализации и апробации алгоритма DFBSA синтеза эмпирического решающего леса // ТВИМ. 2003. №2 / Dyulicheva Yu.Yu. On software implementation and approbation of the DEBSA algorithm of synthesis of empirical decision forest // TJCSTM. 2003. No.2 |
| 6   | Ишкина Ш.Х., Ивакненко А.А. Комбинаторные оценки переобучения пороговых решающих правил // ММРО-2013 / Ishkina Sh.Kh., Ivakhnenko A.A. Combinatorial estimations for the overfitting of threshold decision rules // MMPR-2013 |

Table 1 represents those documents $d \in D$, concerning which the maximum of affinity to a standard was reached at least for one phrase. Relationships of documents within the hierarchy in the direction from lower to higher are interpreted as $j \rightarrow i$, where $i$ and $j$ are the ordinal numbers of the documents in the ranked list from Table 2. Herewith in Table 2 for each phrase $T S_i$ of a separate text $T S$ its length measured in words is given, and further in brackets – the document number according to Table 1 for the document $d_{mx(i)} \in D$, concerning which the maximum affinity to standard was reached for a given phrase. The hierarchy of texts itself is built using the techniques proposed by authors in the paper [4]. Table 3 represents those relationships $j \rightarrow i$, for which values of complementarity of texts that correspond to them, are distinct from zero. The complementarity itself of text $T S_j$ by text $T S_i$ relatively of their sense standards has defined, as shown in [4], by the percentage of words of clusters of greatest values of TF-IDF for phrases of text $T S_i$, not related to the clusters of greatest values of the mentioned measure for phrases of text $T S_j$, but, nevertheless, having nonzero values of TF-IDF concerning the same phrases. Connectivity components of the united relevancy graph for texts $T S_j$ and $T S_i$ are represented in Table 3 by lists of vertexes, and each vertex – by the corresponding number from Table 1. For those relationships, where the united relevancy graph for texts $T S_j$ and $T S_i$ consists of some connectivity components, the corresponding rows of Table 3 have a darker background.
Nevertheless, note, that the resulted components are subgraphs of the same connectivity component of the united relevancy graph for collection.

Table 2. Documents of analyzed collection.

| No. | Author(s) and paper title (in Russian and English) | $\ln(T_{S_i})$ and $d_{\text{max}}(i)$ for separate $T_{S_i} \in T_S$ |
|-----|---------------------------------------------------|-------------------------------------------------|
| 1   | Воронцов К.В., Махина Г.А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа / Vorontsov K.V., Makhina G.A. The principle of gap maximization for nearest neighbor monotonic classifier | 7(2), 12(1), 13(1), 9(1) |
| 2   | Гуз И.С. Гибридные оценки полного скользящего контроля для монотонных классификаторов / Guz I.S. Hybrid estimations of complete cross-validation for monotonic classifiers | 6(1), 10(1), 18(1) |
| 3   | Хачай М.Ю. Сходимость эмпирических случайных процессов, порождаемых процедурами обучения / Khachay M.Yu. The convergence of empirical random processes generated by procedures of learning | 7(1), 14(1) |
| 4   | Фрей А.И. Метод порождающих и запрещающих множеств для рандомизированной минимизации эмпирического риска / Frei A.I. The method of generating and destroying sets for randomized minimization of empirical risk | 7(2), 14(6), 8(2) |
| 5   | Животовский Н.К. Комбинаторные оценки вероятности отклонения тестовой ошибки от ошибки скользящего контроля / Zhivotovskiy N.K. Combinatorial estimations for the probability of test error deviation from the cross-validation error | 8(1), 15(1), 11(2) |
| 6   | Каневский Д.Ю. Переобучение и комбинаторная радемахеровская сложность в задачах восстановления регрессии / Kanevskiy D.Yu. Overfitting and combinatorial Rademacher complexity in regression recovery tasks | 7(1), 20(1), 10(3), 18(1), 9(1) |
| 7   | Неделько В.М. Эмпирические доверительные интервалы для условного риска в задаче классификации / Nedelko V.M. Empirical confidence intervals for conditional risk in the classification problem | 7(1), 16(1), 9(1), 5(4) |
| 8   | Ботов П.В. Уменьшение вероятности переобучения итерационных методов статистического обучения / Botov P.V. Reducing the probability of overfitting for iterative methods of statistical learning | 7(1), 25(2), 14(1) |
| 9   | Ивахненко А.А., Воронцов К.В. Критерии информативности пороговых логических правил с поправкой на переобучение порогов / Ivakhnenko A.A., Vorontsov K.V. Informativity criteria for thresholded logical rules with the correction for overfitting of thresholds | 8(2), 13(2), 10(2) |
| 10  | Сенко О.В., Кузнецова А.В. Системы достоверных эмпирических закономерностей в моделях оптимальных разбиений и методы их анализа / Senko O.V., Kuznetsova A.V. Systems of reliable empirical regularities in models of optimal partitionings and methods to analyze them | 10(5), 18(1), 15(2), 11(1), 18(1), 10(2) |

Table 3. Connectivity components of the united relevancy graph for $T_S_j$ and $T_S_i$.

| Connectivity components | Relationships, $j \to i$ |
|-------------------------|--------------------------|
| {1,2}                  | 2 $\to$ 1, 9 $\to$ 1, 9 $\to$ 5, 9 $\to$ 8 |
| {1,2,3}                | 6 $\to$ 1 |
| {1}, {2}, {6}          | 4 $\to$ 3 |
| {2}, {6}, {1,3}        | 6 $\to$ 4 |
| {1,2,6}                | 8 $\to$ 4 |
| {2,6}                  | 9 $\to$ 4 |
| {1,3}, {2}             | 9 $\to$ 6 |
| {1,2,4}                | 8 $\to$ 7 |
Table 4. Significance of documents $d \in D$ at the selection of a pair of mutual relevant.

| No. according to Table 1 | Value of estimation (5) | No. of cluster |
|--------------------------|-------------------------|----------------|
| 1                        | 3.22591590939           | 1              |
| 2                        | 2.89575917000           | 2              |
| 3                        | 2.30983870471           |                |
| 5                        | 2.09090909091           |                |
| 6                        | 2.09090909091           |                |
| 4                        | 1.00000000000           | 3              |

Estimation (5) can be a basis of selection of a variant of higher-level text for the given $T \mathcal{S}_i$ in the hierarchy under formation according to the degree of complementarity of the sense standard, and, so, be the alternative of complementarity estimation by analysis of the occurrence of words with greatest TF-IDF values in different texts of collection, [4]. Indeed, if documents represented by Table 1 be split into clusters according to the value of estimation (5), then when all other things being equal the least priority will have a relationship with those text $T \mathcal{S}_i$ for which a maximum of affinity to the standard at least for one phrase is achieved concerning some $d \in D$ related to the cluster of least values of estimation (5). An example in the experiments under consideration can be relationships and . In the document with the serial number 7 (according to Table 2) relative to one of the phrases the greatest affinity to standard was achieve concerning document No. 4 (according to Table 1) which relating (see Table 4) to the cluster of least values of estimation (5). Hence, when all other things being equal preference is given to the relationship $8 \rightarrow 4$. The given result fully agrees with obtained one in paper [4] for the mentioned pair of relationships. Indeed, here we have $|H^E_{r(\mathcal{W}_i)}| = 24$ at $|D| = 25$, that is not guarantees higher TF-IDF values in the parent document in comparison with the child document in the hierarchy under formation, and, hence, the navigation through the collection with the gradual focusing of user attention on subtopics [7].

4. Conclusions
The main result of this paper is the improvement of the technique proposed by authors in [4] for the hierarchization of texts of subject-oriented natural language by entering into consideration estimations of mutual relevance between documents concerning which the maximum affinity of separate phrases to standard is achieved.

It’s necessary to note, that when reasoning about mutual relevance of documents compared by Definition 1, in the current paper we have not considered their possible swapping by places with maintaining the feasibility of the first of the conditions of given definition. In a real text, the probability of an existence of a pair of mutual different phrases satisfying this situation essentially depends on the length of the text itself. For the given reason, the possibility of such interchange is not put forward here as the mandatory requirement concerning the mutual relevancy of documents.

The topic for separate research is the quality of clustering of documents of corpus $D$ according to the value of estimation (5). In the future, of interest here may be to study the distributions of frequencies of occurrence of documents in a cluster of least values of estimation (5) for different collections related to the same subject, particularly, using quantiles of the empirical distributions of mentioned frequencies, [8]. The above is in demand, for example, to reveal “noise” documents in the corpus $D$ at using it as a reference corpus in the problem of estimating the cognitive complexity of a text [8].

The estimation (5) itself should be considered as a basis of revelation in a corpus the necessary and enough set of documents for estimation of affinity to the sense standard both of separate phrases as whole texts within an analyzed collection. The revelation of the “standard” range of clusters for values of estimation (5) here is also required separate research. The most promising here represents the search
of documents $d \in D$, the least frequently changing their location in clusters obtained according to estimation (5) when we change one collection to another on a given subject.

Acknowledgments
This work was supported by the Russian Foundation for Basic Research, project no. 19-01-00006.

References
[1] Mikhaylov D V and Emelyanov G M 2019 Estimation of the closeness to a semantic pattern of a topical text without construction of periphrases Pattern Recognition and Image Analysis 29 (4) 647–53. DOI: 10.1134/S1054661819040114
[2] Jones K S 2004 A statistical interpretation of term specificity and its application in retrieval Journal of Documentation 60 (5) 493–502
[3] Zagoruiko N G 1999 Applied Methods of Data and Knowledge Analysis (Novosibirsk: Sobolev Institute of Mathematics of the Siberian Branch of the Russian Academy of Sciences)
[4] Mikhaylov D V and Emelyanov G M 2020 Hierarchization of topical texts based on the estimate of proximity to the semantic pattern without paraphrasing Pattern Recognition and Image Analysis 30 (3) 440–49 DOI: 10.1134/S1054661820030207
[5] Emelyanov G M, Mikhaylov D V and Kozlov A P 2014 Formation of the representation of topical knowledge units in the problem of their estimation on the basis of open tests Machine learning and data analysis 1 (8) 1089–106
[6] The Eclipse Foundation Available at: https://www.eclipse.org (accessed 10.07.2021)
[7] Ianina A and Vorontsov K 2019 Regularized multimodal hierarchical topic model for document-by-document exploratory search Mathematical Methods for Pattern Recognition: Book of Abstracts of the 19th All-Russian Conf with Int Participation MMPR-2019 (Moscow) (Moscow: Russian Academy of Sciences) 256–58 Available at: http://www.machinelearning.ru/wiki/images/b/bf/MMPR2019.pdf (accessed 12.07.2021)
[8] Eremeev M and Vorontsov K 2019 Lexical quantile-based text complexity measure Proc of Recent Advances in Natural Language Processing (September, 2–4, 2019, Varna, Bulgaria) 270–75 Available at: https://doi.org/10.26615/978-954-452-056-4_031 (accessed 12.07.2021)