The offered work is devoted to the problem of numerically estimating the mutual semantic dependence of topical texts concerning the most rational (i.e., standard) variants for describing the knowledge fragments they represent. This problem is relevant in determining the significance of information sources regarding tasks performed by the user. In this regard, the sorting of information sources by the degree of reflection of the most significant concepts of a given topical area with maximum compactness and non-redundancy of statement involves the construction of the hierarchy the upper level of which is made up by the sources from which consideration must be started. In this case, the basis for constructing the hierarchy of documents will be the interrelation of their standards in such a way that the standard of a higher-level document must redefine the standard of a directly related lower-level document. This requirement is especially relevant when forming an individual educational trajectory of a student in e-learning.

A conceptually close hierarchy of primary sources naturally arises when describing a topical area through a thesaurus or ontology, since their construction implies integration and systematization of existing sources of information on the subject under consideration. As an example here may be the “Black Square” system developed by CC RAS (see *slide 3*). The most urgent task in case of this approach was noted by *K V. Vorontsov* as building the recommended order of working with sources, including the search for an “entry point”. Here, the classification of documents can be performed using various criteria, for example, the distribution of a document by topics [Strijov V., 2014], a comparison of the frequency of occurrence of terms in the analyzed document and assigned reference corpus [Eremeev M., 2015], and many others. Sorting sources from simple to complex involves analyzing the frequency distributions of the occurrence of both individual words and their combinations. Moreover, in addition to reflecting the most significant concepts of a topical area, the minimum of terms that have an anomalously high frequency in comparison with the reference corpus, a significant role is played by linguistic expressive means, which determine the best variant among possible paraphrases of the text. From the substantial standpoint, it is required to reveal and analyze a set of text units and their relationships, which is necessary and enough to represent a unit of knowledge. It is this set that corresponds to the sense standard (i.e., semantic pattern).

In this study, the solution of this problem is based on estimations of the proximity of the topical text to the standard and the division the words of phrases of the compared texts into classes according to the value of the TF-IDF measure that underlies these estimations, which were proposed earlier by authors. The role of analyzed texts is played by annotations of scientific articles along with their titles, and the text itself is not paraphrased to search for all possible semantically equivalent linguistic forms of describing a unit of knowledge.

Basic ideas and suppositions of classification of words of initial phrase by TF-IDF as a basis of estimation of its affinity to the sense standard are represented on *slides 4–6*. To classify word combinations as key ones defining the phrase sense image, the interpretation of the TF-IDF metric presented on *slide 4*, is used in our study. This interpretation of well-known metrics estimates the number of simultaneous occurrences of all words from the analyzed combination in phrases of an individual document from the topical corpus (the value in the numerator of formula (1)). When calculating the total word number in a document (denominator of formula (1)), here we separately take into account cases of co-occurrence of words in a combination and frequency without simultaneous occurrence in a phrase. The TF-IDF value itself for the key combination (see *slide 5*) must not be lower than the minimum of the values of this measure for its words.

The implemented in current work and represented on *slide 7* the variant of search the necessary and enough constituents of the image of a phrase of subject-oriented natural language in a form of keywords and their combinations is based on the following empirical considerations. First, the division of words into general vocabulary and terms here should be expressed as greatly as possible. Another important aspect is that the words in clusters formed by the TF-IDF of words of the source phrase relative to a certain document should be distributed more or less evenly. Also, the number of resulted clusters must be close to three as much as possible at a maximum of TF-IDF values for words related to the cluster of greatest values of the mentioned measure. This requirement should be understood as the maximal relevance of term words in phrases of selected documents to the formed corpus. The corpus documents themselves are sorted descending the values of the product of estimations presented on *slide 7*. As the numerical estimation of the closeness of an individual phrase to the sense standard the greatest of the resulting values herewith is taken.

For a group of phrases, first of which is the title of scientific article and others represent its abstract, two variants for estimation of the affinity to the sense standard are used in the current work. Both variants are offered by authors earlier and equally assumed the minimum of root-mean-square deviation (RMSD) for the value of affinity to the standard for all phrases of the group.

The first variant (see *slide 8*) assumes the maximal closeness to the standard for the article title. Note, that introduced estimation does not imply sorting of phrases of the group by affinity to the sense standard and essentially corresponds to the order of selection of articles with analysis of title at first. Such a problem statement is the most adequate to the requirement which is generally accepted in scientific periodicals to reflect in the title the content of the article. Nevertheless, the apriori assumption about the maximal closeness to the standard exactly of the article title is not always performed in practice.

Taking into account the mentioned above, in the *second variant* (see *slide 9*) the maximum of the found values of affinity to the standard for all phrases of analyzed text is used in the numerator of calculation formula. Herewith the maximal final rank in the collection will be designated to the article with the greatest value of the *first variant* of estimation related to the same cluster with the value of the *second variant* of estimation for the same paper. The correct application of the given statement assumes the relating to the same cluster the value of the *first* variant of estimation for an article with a maximal final rank, and maximal value of the *first variant* of estimation in the collection for paper selection. In a case of the absence of an article that meets this requirement, the maximal final rank will be designated to the article with the greatest value of the *first variant* of estimation in the analyzed collection.

Taking into account the mentioned theoretical conclusions, the ranking of articles in the collection based on the joint use of both variants of estimation of affinity to standard can be formally represented using the algorithm shown on *slide 10*. To build a hierarchy of documents of a collection at the output of this algorithm, we use an analogy with the probabilistic topic modeling problem, where the hierarchy of themes models a search strategy with a gradual focusing of user attention on subtopics. Concerning the values of the TF-IDF measure of key terms, this is expressed in our work in higher TF-IDF values of these terms in the parent document compared to the child document in the generated hierarchy.

Let $\mathbf{Ts}_i$ and $\mathbf{Ts}_j$ be the texts from those included in the sorted collection at the output of the algorithm on *slide 10*; here $i > j$, that is, the final rating of the article that

corresponds to group of phrases $\mathbf{Ts}_i$, is higher than the similar indicator for $\mathbf{Ts}_j$. Then, the measure of how the text $\mathbf{Ts}_j$ is complemented by a sense of of the text $\mathbf{Ts}_i$ concerning their sense standards has defined (see *slide 11*) by the percentage of words of clusters of greatest values of TF-IDF for phrases of text $\mathbf{Ts}_i$ not related to the clusters of greatest values of the mentioned measure for phrases of text $\mathbf{Ts}_j$, but, nevertheless, having non-zero values of TF-IDF concerning the same phrases. This assumption is naturally consistent with the distribution hypothesis known in linguistics, according to which the meaning of a word is determined by its context, that is, by which words it is surrounded by in a large corpus of texts. In our case, the role of context is played by a set of words of clusters of greatest TF-IDF values for all phrases of text $\mathbf{Ts}_i$.

The semantic images of the texts that are the closest to the standard are determined by the words with the highest TF-IDF values, which, when located next to each other in a linear row of a phrase, are most likely related by meaning and form key combinations together with the words that are close to the average value of the mentioned measure. At assessing the complementarity of text $\mathbf{Ts}_j$ by text $\mathbf{Ts}_i$ will be considered (see *slide 12*) keywords found for text $\mathbf{Ts}_i$ and contain words with non-zero TF-IDF values not appear in clusters of greatest values of the mentioned measure relatively to phrases of text $\mathbf{Ts}_j$. Herewith, for each such combination, at least one word must belong to the cluster of greatest values at least for one phrase of text $\mathbf{Ts}_i$.

To confirm the existence of a relationship between the sense standard of text $\mathbf{Ts}_j$ and the standard of text $\mathbf{Ts}_i$, in addition to estimations (9) and (10), we use the shown on *slide 13* variant of estimation of the representation of words of the analyzed phrase in the first, last and “middle” clusters of the sequence formed on the base of TF-IDF values for these words. This estimation (see formula 11) was built from geometrical considerations, namely, by means of dividing the number of words in a cluster by the phrase length the requirement of maximal concentration of words of initial phrase in three clusters most significant to the estimation of its affinity to standard was formalized. Herewith for taking into account the relationship of text $\mathbf{Ts}_j$ with the text $\mathbf{Ts}_i$ the cluster of greatest TF-IDF values for each its phrase will be extended by those words of analogous clusters for phrases of text $\mathbf{Ts}_i$ which are not related to the cluster of greatest values of mentioned measure for the analyzed phrase, but having concerning the same phrase the non-zero values of TF-IDF. The specified words are deleted from the last and “middle” clusters for the analyzed phrase (if they are presented there). The variant of estimation (11) modified in this way and represented by the formula (12) on the *slide 13*, is used further on the *slide 14* for the formalization of the criterion of choice higher-level text $\mathbf{Ts}_i$ for given text $\mathbf{Ts}_j$ in the formed hierarchy of documents. The degree of complementation of the standard of text $\mathbf{Ts}_j$ is determined directly by comparing according to *Statement 3* the values of estimations (15) and (16) represented on the *slide 14* with the estimations (13) and (14) corresponding to them. These estimations are conceptually close to estimations presented on *slides 8* and *9* for closeness to standard for a group of phrases.

The experimental material to test the proposed method is represented on *slides 15–17*. The software implementation (in Python 2.7) of the offered solutions and experimental results are presented on the website of Yaroslav-the-Wise Novgorod State University. The main criterion when choosing collections, as well as when selecting texts for

corpus, was the most complete and evident division of words into general vocabulary and terms. For the more accurate revelation of semantic context for terms the calculation of TF-IDF metrics for words of analyzed phrases was made without taking into account of prepositions and conjunctions.

From collections for paper selection shown on *slide 15*, *slides 18–26* further show the results of collection experiments for the "Statistical Learning Theory" section of the proceedings of the 15th All-Russian Conference "Mathematical Methods of Pattern Recognition" MMPR, 2011). *Slide 18* shows the result of ranking the collection by the algorithm presented on *slide 10* concerning *the first* variant of estimation of affinity to standard (see *slide 8*). To give a more demonstrable illustration of the working of the mentioned algorithm, *slide 19* shows the result of ranking the same collection by it, but concerning *the second* variant of estimation the affinity to standard (see *slide 9*). In tables on the *slide 18* and *19* cells for documents with mismatched values of the first and second variants of estimation of affinity to standard, are highlighted by yellow. In subsequent tables and on figures relationships of documents within the hierarchy in the direction from lower to higher are interpreted as $j \to i$, where i and j are the serial numbers of the documents in the ranked list from the *slide 18*. In a case of an empty set of keyword combinations for the higher-level document, the relationship is excluded from consideration (see *slide 22*) in a case of zero value of estimation for the complementarity of text $\mathbf{Ts}_j$ by text $\mathbf{Ts}_i$, see *slide 11*. A relationship is out of consideration also in a case of simultaneous zero values of complementarity estimations both with and without respect of keyword combinations. In *tables 3*, *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. The obtained results generally correspond to our hypothesis about the nature of the redistribution of words in the clusters, which underlies estimation (12) shown on *slide 13*. Here, one of the variants for a more accurate analysis can be based on the use of quantiles of empirical distributions of the frequencies of occurrence of words in the first, last and "middle" clusters of the sequence formed on the base of TF-IDF values for different texts of the analyzed collection. The above is claimed in the analysis in addition to the title and annotation, for example, the analysis of the review part of a scientific article.

An example of using *Statement 3* to select a higher-level text from several variants for the given in the formed hierarchy are relationships $8 \to 4$ and $8 \to 7$. In the collection under consideration it is the choice of a higher-level document for the article by *Botov P.V. (8)* from two alternatives: the article by *Frei A.I. (4)* and the article by *Nedelko V.M. (7)*. For relationship $8 \to 4$, the value of estimation (9) on *slide 11* is 0.4, and estimation (10) on *slide 12* takes the value 0.(3). For relationship $8 \to 7$, we have the value of estimation (9) that is 0.(3), and estimation (10) for this relationship is not calculated because for the higher-level document keyword combinations were not found; in fact, estimation (10) would coincide by definition with estimation (9) in this case. Thus, under otherwise equal conditions of nondecreasing estimations (15) and (16) concerning their corresponding estimations (13) and (14) on *slide 14*, preference is given to the relationship $8 \to 4$. We also note that for the mentioned relationship words of clusters of greatest TF-IDF values for phrases of higher-level document not related to the same clusters for lower-level but having non-zero values of the mentioned measure concerning its phrases, will include the Russian words *"минимизация"* and *"риск"* and for relationship $8 \to 7$ it will include only *"риск"*. This fact serves as additional confirmation of the above hypothesis about the nature of the redistribution of words in clusters.

The main result of this study may be denoted as *a method for the hierarchization of texts of a subject-oriented natural language based on estimations of the proximity of a topical text to the sense standard*.

The effectiveness of the proposed solution can be estimated by the number and type of connectivity components of the graph obtained from the graph of found relationships between the documents of the analyzed collection by replacing the oriented edges with non-oriented ones. Ideally, each connectivity component in case of restoring the orientation of the edges will be a directed tree, where the maximum and minimum heights of a child subtree for any vertex differ by no more than 1 (by analogy with the height-balanced binary search tree), and the number of connectivity components themselves must be as close as possible to 1. After removing from the original graph those relationships that either do not meet the condition of *Statement 3,* or for which the keyword combinations for higher-level document were not found (see *slides 23* and *25*), the subgraph that corresponds to the maximum connectivity component always contains the vertex for the article with the maximum total rating for the collection among the vertices with the maximum degree. For comparison: the variant without taking into account keyword combinations for the considered collection of ten articles gives the maximum connectivity component of eight vertices, and in the case of taking into account key combinations it gives the maximum connectivity component of six vertices. Meanwhile, at the expense of articles that are not reflected in the maximum connectivity component, we have an additional (by at least 20%) reduction in the number of documents that should be primarily familiarized with when studying a given topical area, for example, by students.

To increase the accuracy of dividing words into general vocabulary and terms, it is of interest *to study the relationship* between the frequency distributions of the occurrence of words in the clusters of the highest values of the TF-IDF measure for phrases of different texts of the analyzed collection and the cases of achieving the maximal closeness of phrases to the standard relative to specific documents of the given text corpus by the value of the product of estimations presented on *slide 7*.