The offered work is devoted to the problem of numerical estimation for the closeness of a topical text to the most rational linguistic variant of description of knowledge fragment represented by it in a given natural language. This variant satisfies the sense standard (i.e. semantic pattern). The problem is of importance when implementing targeted lossless-in-sense selection of text information at maximization of useful output concerning the tasks solved by user. As examples of practical applications here may be selection of papers for publishing, designing of training courses and educational programs. So, when preparing the material, the teacher must have access to a certain section of the information space, the elements of which are publications or Internet pages relevant to the course. The major requirement here may be formulated as the sorting 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 (see *slide 2*). Ideally, the information sources form a hierarchy at a top level of which will be placed the start points for study. A similar possibility should be given to the student in the process of independent work, it is especially important as part of teaching students to prepare and implement projects in the professional sphere.

In current paper by splitting of words of each phrase of analyzed text into classes according to the value of the TF-IDF measure the problem of numerical estimation for the text closeness to the most rational linguistic variant of sense transfer without paraphrasing is solved. Herewith as the analyzed texts the abstracts of scientific articles together with their titles are considered. These parts of articles reflect the main content of each paper and the most important results without unnecessary methodological details.

Essentially close problem (see *slide 3*) is the construction and verification of thematic models of major conferences with the finding of most relevant themes for a new participant. Here the theme of a document is defined by its terms from the terminological dictionary of the conference. The key point here is the value of the term's importance, which is expressed in terms of its entropy relative to expert clustering at a given level of hierarchy. Linguistic expressional means, meaningful for choosing the best variant among possible paraphrases, with this approach will be remain out of consideration. Well-known solutions in the field of paraphrases detection and training in such detection also do not provide for proper qualitative analysis of paraphrases themselves. In a best case, the degree of semantic proximity of sentences is calculated, for example, using the SyntaxNet parser and measuring the edit distance between the resulting dependency trees. Interpretation of the nature of paraphrases, meaningful for choosing the most rational linguistic variant of transferring a piece of knowledge, is not discussed here. However, even the quality of preparation for the paraphrase corpus using a system that recognizes sentences similar in meaning depends on the accuracy of its training. The primary role here is played by the selection of a set of text units and their relations, necessary and enough to represent a unit of knowledge. Just such set meets the sense standard.

In offered solution the base for estimation of the closeness of a text to the semantic pattern is the splitting of words of each its phrase into classes by the value of the TF-IDF metric relative to the corpus pre-formed by expert (see *slides 4–6*). To select keyword combinations from defining the semantic image of phrase the interpretation of TF-IDF metrics which is presented on the *slide 4* and respects a number of simultaneous occurrences of all words of analyzed combination in the phrases of separate document of corpus (the value in numerator of formula (1)) is entered into consideration in current paper. When calculating the total number of document words (the value in denominator of formula (1)) we'll separately take into account the cases of co-occurrence of combination words and occurrence without simultaneous presence in a phrase. Herewith (see slide 5) the value of TF-IDF metrics for key word combination should not be less than the minimum of values of mentioned measure for its separate words.

The implemented in current work and represented on the slide 7 the variant of search the necessary and enough constituents of image of a phrase of subject-oriented natural language in a form of key words 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. In addition, the number of resulted clusters must be close to three as much as possible at maximum of TF-IDF values for words related to the cluster of greatest values of 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 product of estimations presented on the 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 introduced in current paper. Both variants are equally assumed the minimum of root-mean-square deviation (RMSD) for value of affinity to the standard for all phrases of 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 problem statement is the most adequate to the requirement general accepted in scientific periodicals to reflect in the title the content of the article. Nevertheless, the a priori assumption about the maximal closeness to the standard exactly of the title of article 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 a 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 correctly application of given statement assumes the relating to the same cluster the value of the *first* variant of estimation for article with a maximal final rank, and a maximal value of the *first variant* of estimation 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 the *first variant* of estimation in analyzed collection.

As can be seen from definition, the both variants of estimation are depend on the selection of corpus by expert. In addition, 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 considered variants of estimation in *Statement 2* on the slide 9.

The experimental material to test the proposed method is represented on the slides 10–12. 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 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.

The experimental results represented further in the tables on the slides 13–17 are confirm the rule of «good manners» of some periodicals on information science and computer engineering to display in the title the name of method, model, algorithm presented by paper, as well as the theoretical basis of the proposed solutions. For the collection *«MMPR-15, Statistical Learning Theory»* the maximums of both variants of estimation of affinity to the standard took place relatively to the same article, the similar result was reached for the collection *«MMPR-15, Theory and Methods of Pattern Recognition and Classification»*. As can be seen from the tables on the slides 13–16, the values of estimation variants for affinity to standard for mentioned articles are coincided. So, according to the condition of *Statement 2*, the papers *«Принцип максимизации зазора для монотонного классификатора ближайшего соседа» (The principle of gap maximization for nearest neighbor monotonic classifier)* by *K.V. Vorontsov* and *G.A. Makhina*, and *«Полные решающие деревья в задачах классификации по прецедентам» (Complete decision trees in classification tasks by precedents)* by *I.E. Genrikhov* and *E.V. Djukova* will have a maximal final rank each in its collection.

The result obtained for the collection *«MMPR-14, Methods and Models of Pattern Recognition and Forecasting»* and represented on the slide 16 illustrates the case when an article with a greatest-in-collection value of *second estimation variant* for affinity to standard has the value of *first variant* of this *estimation* not relates to the same cluster with it. Indeed, for the article *«Выбор опорного множества при построении устойчивых интегральных индикаторов» (Support set selection when constructing of stable integral indicators)* by *D.I. Mel'nikov, V.V. Strijov, E.Yu. Andreeva and G. Edenharter* the values of the *first* and *second variants* of estimation equal, correspondingly, to 0.0129 and 0.1426, are form two independent clusters. By this virtue the maximal final rank in the collection will be designated to the article by *O.V. Barinova and D.P. Vetrov* having the maximal value of the *first variant* of estimation for affinity to standard relatively to the considering collection (see slide 17).

The similar situation is also take place for the collection *«IIP-9, Theory and Methods of Pattern Recognition and Classification»*. Here the maximal value of the *second variant* of estimation equal to 0.1336 will be belonged to the article *«Критерии точности комбинаторных оценок обобщающей способности» (The exactness criteria of combinatorial generalization bounds)* by *N.K. Zhivotovskiy and K.V. Vorontsov*. The value of the *first variant* of estimation here is equal to 0.0600 and related to the same cluster with the maximal value equal to 0.0920 for this estimation variant in collection, but, nevertheless, not lies in the same cluster with the value of the *second variant* of estimation of affinity to standard for this article. Therefore the maximal final rank obtains the article by *S.D. Dvoenko and D.O. Pshenichny* having the greatest value of the *first variant* of estimation in the considered collection.

In a case of change between each other of variants of affinity-to-standard estimation in *Statement 2* on the slide 9, in considered examples for the both collections by *MMPR-15* conference the maximal final ranks herewith will be designated to the same articles (see slide 17). For the collection *«MMPR-14, Methods and Models of Pattern Recognition and Forecasting»* the maximal rank here the article by *O.V. Barinova and D.P. Vetrov* obtains again. Indeed, the maximal value of the *second variant* of estimation of affinity to standard in this collection will be for the article by *D.I. Mel'nikov, V.V.*

Strijov, E.Yu. Andreeva and G. Edenharter. But as we showed earlier, the values of the *first* and *second variants* of estimation of affinity to standard for this article in the considered collection are related to different clusters. Therefore, according to the condition of *Statement 2* the maximal final rank obtains the article having among the remaining articles (except the article mentioned above) the maximal value of the *second variant* of estimation relating to the same cluster with the value of the *first variant* of affinity-to-standard estimation for itself, i.e. the article by *O.V. Barinova and D.P. Vetrov*.

The single exclusion in considered series of experiments will be the result for collection *«IIP-9, Theory and Methods of Pattern Recognition and Classification»*. As in the previous example, the maximal final rank in the collection may be designated to the article by *S.D. Dvoenko and D.O. Pshenichny* as having the maximal value of the *second variant* of estimation which relates to the same cluster with the value of the *first variant* of given estimation for this paper. But the value of the *second variant* of estimation for it not relates to the same cluster with the maximal value of this estimation variant in the collection. So, the maximal final rank together with the mentioned value in this collection here obtains the article by *N.K. Zhivotovskiy and K.V. Vorontsov*.

It's necessary to note, that the obtained results confirm the hypothesis relatively to the semantic load of title for scientific paper on information science and computer engineering. For disputable cases similar to the aforementioned for the collection *«IIP-9, Theory and Methods of Pattern Recognition and Classification»*, depending on the subject area of analyzed texts it's possible to give preference to the requirement of relating to the cluster of maximal value of either *first* or *second variant* of estimation of affinity to standard.

The semantic image of article having the maximal value of used estimation of affinity to standard in the collection for paper selection essentially will be defined by words which being a neighbors in a corresponding phrase of analyzed group and related to the cluster of the greatest values of TF-IDF concerning the document with the greatest value of product of estimations presented on the *slide 7*. For more exact identification of multi-word terms among the words of general vocabulary let's extend the key combinations revealed in a phrase, by words of «median» cluster of sequence formed on the basis of TF-IDF of words of analyzed phrase relatively to the given document (see *slides 18–19*).

To check the presence in analyzed phrases the links between the words related to the clusters of greatest TF-IDF values, MaltParser was used, i.e. a tool for parsing the phrases of natural languages and working with dependency trees. As can be seen from the tables on the *slides 18–22,* the syntactic relation corresponds to the location of words of the specified clusters in the neighborhood in the linear phrase structure, what indicates the unity of the component of text's semantic image.

The result obtained from the collection *«MMPR-15, Theory and Methods of Pattern Recognition and Classification»* fully agrees with the theoretical conclusion about the relativity of the notion «general vocabulary». In fact, each of the collections presented in the tables is related to a specific topic described by a discrete distribution on the set of terms. The classification by TF-IDF considers the words *«обзор» (review)* and *«дать» (to give)* as terms, these words are treated as general vocabulary for the language as a whole, but as unique ones to the article by *I.E. Genrikhov* and *E.V. Djukova*. Note, that the combination of mentioned words does not meet the condition of *Statement 1* on the *slide 5* and, therefore, does not be related to the key combinations.

Since the title and phrases of the article abstract satisfy to a single semantic image, it is quite correctly to analyze the occurrence of words related to the cluster of greatest

TF-IDF values for one phrase, in the links of words relatively to another phrases. In current work the following condition is used for associating a set of such links with the key word combination: a connected sub-graph of the parse syntactic tree (undirected) is corresponded to the set of considered word combinations and at least one word combination must satisfy the condition of *Statement 1* on the *slide 5*. In the example for abovementioned article by *I.E. Genrikhov* and *E.V. Djukova* the word *«полный» (complete)* was not included to the cluster of greatest TF-IDF values, but it is syntactically submitted to the word *«дерево» (tree)* from the specified cluster and, therefore, forms the desired key combination. It should be noted that the last example on the *slide 21* takes into account the transitivity of the syntactic relation within the sequence of coordinated words, cf. *«матриц – сравнений – парных» (matrices – comparisons – paired)*. Relating it to the key word combinations by the proposed interpretation of TF-IDF metrics may be considered as the additional proof of link presence. A topic for separate study here is the dynamics of the change in the TF-IDF measure when we extent our consideration from discrete words to *L*-grams (according to C. Shannon). The second example on the top of the same slide illustrates that the placement of words of cluster of greatest TF-IDF values in the neighborhood in the linear phrase structure should be considered as a necessary but not enough condition for relating to the key combinations that determine the semantic image of the text. Enough conditions are defined by the technique proposed in paper to implement the offered-by-authors interpretation of TF-IDF metrics.

The main *result* of current work is the *proposed method for estimating the affinity of a text to the sense standard relative to a topical text corpus*.

The effectiveness of the proposed method can be estimated by splitting the collection's texts into clusters by the value of used estimation for the closeness to a standard and the ratio of the number of texts assigned to the cluster of the greatest evaluation values to the total number of texts in the collection. So, on the material of collections mentioned on the *slides 13–22*, we have at least a threefold reduction in the number of documents (i.e. scientific articles) that should be read first when studying a given subject area, for example, by students.

Taking into account the evaluated degree of division of its words into general vocabulary and terms, when a phrase is referred to as a «representative of the standard», it is also of interest to integrate estimations of affinity to standard for phrase groups in the entire article collections under analysis. Herewith the presence of key word combinations in abstracts and titles can be a basis for designating in disputable cases the final rank and hierarchization of articles according to significance when studying a given subject area.