The problem considered is to find the most rational variant for transfering the meaning among experts and testees in a knowledge-control system that implements the open tests.

Development of such tests assumes formation of expert knowledge units initially represented by texts of subject-oriented natural language (NL). The most natural knowledge source here will be the scientific papers of highest-rank scholars in appropriated topical area. There we have two tasks: the thematic categorization of text documents and representation of topical areas by means of thesauruses and ontologies. The problem to seek a most rational variant to express the meaning in a knowledge unit defined by a set of NL-phrases equivalent-by-sense (i.e. semantically equivalent, SE) is the most actual here. The sense itself has to be reflected in a maximum compact volume of text data. This decision means the following expert tasks to be automated (see *slide 3*):

- to search SE-forms for description of reality fragment in the given NL;
- to compare the knowledge of given expert with the closest knowledge fragments of another experts.

It is necessary to note, that the decision of mentioned tasks is not reduced to the simple revelation of concepts and relations between them from given text with estimation of semantic affinity for pairs and groups of concepts. To search and classify the NL-forms of expression of expert knowledge it is necessary to extract the usage contexts of general vocabulary by means of which synonymic paraphrases can be formed. Problem close to given but substantially inverse is decided by paraphrase detector offered by scholars from Stanford University. For initial pair of phrases this detector judges whether these phrases are synonymic paraphrases of one another. The detection itself is carried out by neural network, the training sample for learning of which is formed from the results of syntactic analysis of phrases selected by an expert. The training sample here shall necessarily contain positive and negative examples of paraphrases, what not quite meets the requirement of possibility of comparison the different knowledge fragments, because the relations of words apart from the synonymy is not taken into account. In addition, the given approach is subjective concerning the conception of synonymy itself because the topical area of separate phrase from the pair and a degree of semantic affinity of phrases is not considered.

Taking into account the occurrence frequency of general vocabulary in texts related to various topical areas, the most naturally is to decide the tasks mentioned above by attraction the TF-IDF metrics to select both this vocabulary and terms (including multi-word terms) in initial phrase. In current paper we study the performance capabilities of vocabulary clustering by TF-IDF as a base of selecting in text corpus the closest knowledge fragments and NL-forms of their expression.

In text analysis and informational retrieval TF-IDF is a numerical statistic that is intended to reflect how important a given word is to some document being a member of some corpus.

According to classic definition mentioned on the *slide 4*, TF-IDF is the product of two statistics: term frequency (TF) and inverse document frequency (IDF). Term frequency is the quotient of number of times that the word occurs in document by total number of words in this document. The inverse document frequency is a measure of how much information the word provides, that is, whether the designated term is common or rare in corpus.

It is necessary to note that with the growth of word's occurrence frequency in corpus documents the value of IDF metrics for this word tends to zero. It is true both for general vocabulary (for example, function words) and for those terms which are prevail in corpus. At the same time, for example, the words from general vocabulary which are define the conversive replacements, like *«приводить ⇔ являться следствием»* (in Russian), will have the higher values of IDF.

It's fair to assume (see *slide 5*) that 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. The fact that the term has synonyms at the same document means the decrease of TF metrics for this word. Like the conversive replacements mentioned above each word from the synonym ring has a lower occurrence in corpus documents and a higher value of IDF in comparison with the case of synonyms absence.

Let the above arguments be the basis of required clustering of vocabulary of initial phrase by TF-IDF metrics.

The first step (see *slide 6*) is the calculation of TF-IDF for all words of initial phrase concerning each document in corpus. Each of sequences found here will be sorted descending with splitting into clusters by means of algorithm close to FOREL class taxonomy algorithms. As the mass center of cluster the arithmetic mean of all its elements is taken. The estimation of clustering quality here assumes (see *formula (3) on the slide 6*) on the one hand to maximize the number of clusters under maximization the number of words in each cluster and, on the other hand, to minimize the difference between the most and the least values of cluster width. Using this estimation we can select those corpus documents with respect to which the division of words in initial phrase between general vocabulary and terms is the most expressed.

On the next step (see *slide 7*) the documents of corpus will be sorted descending values of mentioned estimation with clustering by means of the same algorithm that was used for splitting of words of initial phrase. Here documents with the greatest values of estimation (i.e. values relating to the first cluster of the formed sequence) are selected. Let's name further these documents the best on quality. Here we have the task: from the documents which are best on quality it is necessary to select phrases with maximum number of words presented in the first, last and middle clusters of sequence formed for initial phrase according to TF-IDF of its words. The «middle» cluster here is necessary (first of all) for selection of general vocabulary and those terms which have synonyms. Indeed, the values of both TF and IDF for these two categories of vocabulary are closest to average for initial phrase. Therefore with a high probability they can be represented in the «middle» cluster.

Like the quality of clustering, the numerical estimation of representation of words of phrase related to the best-on-quality documents in three above-mentioned clusters (see *formula (4) on the slide 7*) can be defined from geometrical considerations, but instead of the difference between the most and least values of cluster width the root-mean square deviation of number of phrase's words presented in cluster is taken. The phrases selected are clustering according to the value of this estimation. As a result a set of phrases corresponding to the cluster of greatest values is returned.

Note that the offered method does not take into account the word's syntactic context. Indeed, such context orientation would make impossible the search of phrases synonymic to initial from the point of view of both syntax and vocabulary (like the *conversive replacements* mentioned before).

The test corpus for proposed method includes papers in Russian and has the structure presented on the *slide 8*. The scope of selected papers is shown on the *slide 9*. Russian phrases which describe the facts of topical area «Mathematical methods for learning by precedents» and were used as a basis for knowledge units formation, are presented on the *slide 10*.

The software implementation (in Java) of the offered method and experimental results are presented on the website of Yaroslav-the-Wise Novgorod State University.

As an example on the *slide 11* the search of phrases closest to the phrase №9 concerning the described knowledge fragment and its Russian expressional means can be given. From corpus documents according to the above-mentioned criteria of quality of clustering of words of initial phrase as the best two articles of Konstantin Vorontsov were selected. The first was published in «Taurida journal of computer science theory and mathematics» (№1, 2004 г.) and the second – in the proceedings of All-Russian Conference with International Participation on Mathematical Methods for Pattern Recognition (MMPR-15). These two documents became the source for selection of phrases according to the requirement of maximization the value of representation of words of phrase in three most significant clusters which were formed by TF-IDF metrics.

The first from resulted phrases presented on the *slide 11* can be an example of relation between the knowledge fragment of given expert with summary knowledge of another experts, namely, the definition of generalizing capability of algorithm mentioned in initial phrase relates here with the concepts of error probability and rate (frequency) of errors in control sample. It is of importance for selection of paraphrases for initial phrases №6 and 7 from presented on the *slide 10*. Another resulted phrase on the *slide 11* is the example of Russian paraphrasing tool for expert, compare: *«ведёт к ⇔ является результатом»*.

The next example presented on the *slides 12* and *13* illustrates the search of synonyms for term in initial phrase №4. Note, that the word *«переподгонка» (overfitting)* has the synonym *«переобучение» (overtraining)* in corpus texts, and concerning the first document from the table shown on the *slide 12*, TF-IDF value for this word denotes to the «middle» cluster, see tables on the *slide 13* for compare. The resulted phrase on the *slide 12* also contains the variant of conversive replacement for initial phrase: *«причина ⇔ результат» (reason ⇔ result)*.

It is necessary to note that if the occurrence of term is the least in corpus then for the most best-on-quality documents this word will be related to the last cluster from sequence of formed by TF-IDF. As a result we have a quite low co-occurrence of this word with other terms of interest and general vocabulary. That will mean an absence of phrases close to the initial (from the point of view of the expert) concerning the described reality fragment or forms of its NL-expression.

As an example the word *«заниженность»* can be given concerning the search of phrases closest to the initial phrase №8. As evident from experiment results presented on the *slide 14*, there were no phrases where along with maximization of criteria of words representation in three most significant TF-IDF-clusters the requirement of presence of the mentioned word would be fulfilled.

One of the possible ways to improve the search for the considered case can be the usage of summary TF-IDF value for initial phrase's words occurring at the phrase from the document of bests on quality, as an alternative of the most significant TF-IDF-clusters. But as shown in the experiment with the same initial phrase №8, this way is only lead to growth the number of selected phrases without seeking of neither synonymic paraphrases, nor phrases which would relate the concepts mentioned in initial phrase with other concepts of given topical area.

One more way to improve the search here is to use the additional knowledge about semantic relations and their textual expressional forms. But as shown in the experiments with well-known system «Serelex» for initial phrases №8 and №9 (see *slide 15*), a fixed set of such knowledge is usually not enough. Actual problem of such decisions lies in the dependence of quality of work of lexico-syntactic patterns implemented by system from subject orientation of vocabulary of analyzed text. *The method offered in current paper* allows to reveal concepts and relations between them irrespective of topical area and types of links of words in initial phrases. It is necessary to note that the result given by offered method is significantly depends on selection of initial corpus by expert. One of the open problems here is (see *slide 17*) the elaboration of numerical estimation which respects simultaneously the most significant criteria to select documents for adding to corpus, namely: the quality of extraction of themes (topics), the singularity of term distribution for a topic and the singularity of topic distribution for a document.