Selection of scientific articles according to the degree of proximity to the semantic pattern of the title and phrases of the abstract

D V Mikhaylov and G M Emelyanov
Yaroslav-the-Wise Novgorod State University, ul. B. St. Petersburgskaya, 41, 173003, Veliky Novgorod, Russia

E-mail: Dmitry.Mikhaylov@novsu.ru

Abstract. The article is devoted to the problem of numerical evaluation of the proximity of a thematic text to the most rational (reference) language version of the description of the piece of knowledge it represents. This problem is relevant for the implementation of targeted selection of textual information without losing the useful semantic component. Examples of practical applications here can be the selection of articles for publication in scientific journals, as well as the development of training courses and educational portals. In the proposed solution, the basis for assessing the proximity of a text to a semantic pattern (i.e. sense standard) is the division of the words of each phrase into classes according to the value of the TF-IDF measure relative to the texts of the corpus pre-formed by an expert. The analyzed texts considered in the paper are the abstracts of scientific articles along with their titles. At the same time, the semantic images of the texts closest to the standard determine the words with the highest TF-IDF values, which, being neighbors in a linear series, are most likely related by meaning and form key combinations. The proposed numerical estimate of the proximity to the standard makes it possible to rank articles according to the significance of the described fragments of knowledge with respect to a given subject area, as well as to the non-redundancy of the description itself.

1. Introduction
The development of e-learning, the introduction of a new generation of educational standards by Russian universities [1] significantly increases the requirements for the quality of educational material presented in electronic form. In the process of preparing the material, the teacher must have access to a certain section of the information space [2], the elements of which are publications or Internet pages relevant to the course. The main requirement here can be represented as sorting the sources of information according to the degree of reflection of the most significant concepts of the studied subject area with maximum compactness and non-redundancy of presentation. Ideally, information sources form a hierarchy; on the upper level of this hierarchy are those sources, from which the study process should begin. A similar opportunity 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. A substantively close task is to build and verify hierarchical thematic models of large conferences and search for the topics most relevant for a new participant [3]. At the same time, the subject of the document is determined by its terms from the terminological dictionary of the conference. The key point here is the importance of the term’s importance, which is expressed in terms of its entropy relative to expert clustering at a given level of hierarchy. Language expressive means, meaningful for
choosing the best option among possible paraphrases, with this approach remain out of consideration. Well-known solutions in the field of paraphrases detection and training in such detection [4–6] also do not provide for proper qualitative analysis of paraphrases themselves. At best, the degree of semantic proximity of sentences is calculated, for example, using the SyntaxNet parser [7] and determining the edit distance between the resulting dependency trees [8]. Interpretation of the nature of paraphrases, meaningful for choosing the most rational language variant of transferring a piece of knowledge, is not discussed here. However, even the quality of training for the paraphrase corps using a system that recognizes sentences similar in meaning, as rightly noted in [5], 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 sufficient to represent a unit of knowledge. Such a set, as shown by the authors in [9], meets the semantic standard. At the same time, a priori restrictions on the nature of the links of text units are not imposed.

In research [10], the authors proposed numerical estimates of the proximity of a phrase to a sense standard (i.e. semantic pattern) based on the division of its words into classes according to the value of the TF-IDF measure and the evaluation of the strength of the combination of words. We would like to note that Okapi BM25 as a TF-IDF-like ranking function was recommended in the project [8] for a more accurate assessment of the degree of semantic equivalence of sentences.

In this paper, we consider the possibility of using the proposed in [10] estimation of the proximity of a phrase to a semantic pattern based on the TF-IDF measure without searching for paraphrases of phrases from the analyzed text. In this case, the analyzed texts are scientific articles abstracts and their titles reflecting the main content of each work and the most significant results without unnecessary methodological details.

2. The choice of assessing the proximity of the semantic pattern for a phrase and a group of phrases

In the problems of text analysis and information retrieval TF-IDF there is a statistical measure used to assess the importance of a word in the context of a document included in a text corpus. According to the definition, this measure is the product of the ratio of the number of occurrences of a word to the total number of words in the document and the inverse of the frequency of the document

\[
\text{idf}(t_i, D) = \log\left(\frac{|D|}{|D_i|}\right),
\]

where \(D\) is the corpus, \(D_i \subseteq D\) – number of documents of corpus \(D\), in which a word \(t_i\) was noticed at least once.

Let \(X\) be an ordered sequence of TF-IDF measure values in descending order for the words of the original phrase relative to a given text as part of a certain set (corpus) previously formed by the expert.

We divide \(X\) into clusters \(H_1, \ldots, H_r\) using an algorithm that is meaningfully close to the class algorithms FOREL [11] and used previously by us [9, 10]. Further, as applied to the splitting of a sequence into clusters, we will mean just this algorithm. As the center of mass of the cluster \(H_t\) here, as in [9, 10], the arithmetic average of all its elements is taken. From the result of splitting the sequence \(X\) to assess the proximity of the phrase to the semantic pattern, the substantial interest present:

- cluster \(H_1\), to which the terms from the source phrase correspond, the most unique for the analyzed text document;
- “middle” cluster \(H_{r/2}\), formed with the general vocabulary, providing paraphrases, and synonymous terms;
- cluster \(H_r\), to which the words-terms prevailing in the corpus correspond.

The assessment of the proximity of a single phrase to a standard, which does not require the search for equivalent forms of the linguistic description of the corresponding unit of knowledge, is based on the following empirical considerations. First, as follows from the reasoning in [10], the division into general vocabulary and terms here should be expressed to the greatest extent possible, whereas the words in clusters \(H_1, \ldots, H_r\), formed by the TF-IDF of the original phrase with respect to some \(d \in D\) should be distributed more or less evenly. But unlike the estimation from [10], the number of the resulting clusters should be as close to three as possible with the maximum TF-IDF values for words from cluster \(H_1\). The latter requirement should be understood as the maximum relevance of terms within the
phrases of the selected document to the formed corpus. Substantially the above can be thought of as maximizing the values

\[
val_1 = -\frac{1}{\log_{10}(\Sigma_{H_1})},
\]

\[
val_2 = 10^{-\sigma(|H_1,i=\{1/r/2,r\}|)}
\]  

(1)  
and, correspondingly,

\[
val_3 = |H_1 \setminus H_{r/2} \setminus H_r|/\text{len}(X).
\]  

(2)  

(3)

The estimates (1) and (2) are modifications of the corresponding estimates from [10]. The logarithmic value in the denominator of the formula (1) is the sum of the TF-IDF values of words assigned to the cluster \(H_1\) by the value of the specified measure relative to the document \(d \in D\) under consideration; \(\sigma(|H_i, i = \{1/r/2,r\}|)\) in the formula (2) there is a standard deviation (RMS) of the number of elements in the cluster from the list \(\{H_1, H_{r/2}, H_r\}\); \(\text{len}(X)\) in the denominator of the formula (3) is the length of the sequence \(X\). In case \(\Sigma_{H_1} = 0\) value \(val_1\) is assumed to be zero. If the number of TF-IDF clusters obtained is less than two, then the values \(\left[H_{r/2}\right] \cup \left[H_r\right]\) are assumed to be zero. With exactly two clusters of TF-IDF, the value \(\left[H_r\right]\) is considered to be zero. The estimation of the proximity of the phrase to the standard with respect to \(d \in D\) on the basis of the quantities (1), (2) and (3) from geometric considerations can be represented as the volume of a rectangular parallelepiped with measurements equal to \(val_1, val_2 \text{ and } val_3\).

Documents \(d \in D\) are sorted by descending product of ratings (1), (2) and (3). In this case, the largest of the resulting values is taken as a numerical estimate of the proximity of the phrase to the standard.

Estimation of proximity to the standard for a group of phrases \(T_s\), the first of which is the title of a scientific article, and the rest represent an abstract, implies maximum proximity to the standard for the title with the minimum RMS of the proximity to the standard for all \(T_{s_1} \in T_s\):

\[
N(T_s, D) = \frac{\max_{d \in D} (val_1(T_{s_1},d)\cdot val_2(T_{s_1},d)\cdot val_3(T_{s_1},d))}{\sigma\left(\max_{d \in D} (val_1(T_{s_1},d)\cdot val_2(T_{s_1},d)\cdot val_3(T_{s_1},d))\right)}+1.
\]  

(4)

The numerator is an estimate of the proximity to the standard of the article title \((T_{s_1})\). In order to prevent possible division by zero, a one is added to the RMS value in the denominator of formula (4).

As can be seen from the definition, the estimate (4), as proposed earlier in [10], depends on the selection of the corpus \(D\) by the expert. One should note, that the entered estimate does not imply sorting of phrases \(T_{s_1} \in T_s\) by proximity to the standard and meaningfully corresponds to the order of selection of articles, starting with the analysis of the title. Such problem statement is the most adequate to the requirement generally accepted in scientific periodicals to reflect the content of the article in the title. The rule of “good manners” of some publications on computer science and computer technology is, for example, to display in the title the name of the method, model, algorithm presented in the work, as well as the theoretical basis of the proposed solutions [12]. However, the a priori assumption of maximum proximity to the standard of the title of the article is not always implemented in practice.

3. The Experiment

For approbation of the assessment (4), as a corpus \(D\); formed by an expert, a variant from experiments in [10] was involved, including:

- 3 articles from the journal *Taurida journal of computer science theory and mathematics*;
- 2 articles in the proceedings of the international conferences “Intelligent Information Processing” of the years 2010 and 2012;
- 1 article in the collection of works of the 15th All-Russian Conference “Mathematical Methods for Pattern Recognition” (2011);
- 2 reports theses at the 13th All-Russian Conference “Mathematical Methods for Pattern Recognition” (2007);
- theses of 14 reports at the 16th All-Russian Conference “Mathematical Methods for Pattern Recognition” (2013);
- theses of 2 reports at the conference “Intelligent Information Processing” (2014);
- a scientific report written by D.V. Mikhailov in 2003.
Subjects included:
- mathematical methods for learning by precedents (K.V. Vorontsov, M.Yu. Khachay, E.V. Djukova, N.G. Zagoruiko, Yu.Yu. Dyulicheva, I.E. Genrikhov, A.A. Ivakhnenko);
- methods and models of pattern recognition and forecasting (V.V. Mottl, O.S. Seredin, A.I. Tatarchuk, P.A. Turkov, M.A. Suvorov, A.I. Maysuradze);
- intelligent processing of experimental information (S.D. Dvonenko, N.I. Borovykh);
- image processing, analysis, classification and recognition (A.L. Zhiznyakov, K.V. Zhukova, I.A. Reyer, D.M. Murashov, N.G. Fedotov, V.Yu. Martyanov, M.V. Kharinov).

The number of words in the documents ranged from 218 to 6298, the number of phrases ranged from 9 to 587. Selection of articles was made from:
- proceedings of the international conference “Intelligent Information Processing” IIP -9 (2012), section “Theory and Methods of Pattern Recognition and Classification” (14 articles);
- proceedings of the international conference “Mathematical Methods for Pattern Recognition” MMPR-14 (2009), section “Methods and Models of Pattern Recognition and Forecasting” (35 articles);
- proceedings of the international conference “Mathematical Methods for Pattern Recognition” MMPR-15, section “Theory and Methods of Pattern Recognition and Classification” (18 articles) and “Statistical Learning Theory” (10 articles).

The main criterion when choosing collections, as well as when selecting texts for corpus $D$, was the most complete and clear division of words of the analyzed texts into general vocabulary and terms.

Software implementation in Python 2.7 and the results of experiments are presented on the NovSU web-site at [http://www.novsu.ru/file/1493710](http://www.novsu.ru/file/1493710) in a ZIP archive (version of 19.04.2019). The archive includes a directory with the experiment results and a directory of the Test_Python_PyDev workspace of the Eclipse environment [13], where the subdirectory TestMorpho/src contains the TestMorpho.py file of the project head module, the text package catalog (my_corpora), an example of source phrases (text.txt) and the processing results (issue.txt).

One should note that the phrase with the maximum proximity to the sense standard for the analyzed group in all four experiments in table 1 was the title of the article. The result obtained from the collection “MMPR-15, Theory and Methods of Pattern Recognition and Classification” fully agrees with the theoretical conclusion made in [14] about the relativity of the notion “general vocabulary”. In fact, the method proposed in [14] attributes each of the collections presented in Table 1 to a specific topic described by a discrete distribution on the set of terms. The classification by TF-IDF considers the words “review” and “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.

| Table 1. Articles with a maximum rating value (4) in collections. |
|---------------------------------------------------------------|
| **MMPR -15, Statistical Learning Theory**                     |
| Author (s)                                                    |
| K.V. Vorontsov, G.A. Makhina                                  |
| Title of the article                                          |
| The principle of gap maximization for nearest neighbor monotonic classifier |
| The maximum rating of proximity to the standard for the title is achieved relative to the document | K.V. Vorontsov Combinatorial theory of overfitting: results, applications and open problems // MMPR-15 |
| Words from clusters of highest TF-IDF values for individual phrases | monotonic, neighbor, close, cross, validation, generalize, ability |
| Combinations formed from them: | the nearest neighbor, cross-validation, generalization ability |
| Rating of proximity to the standard of the article title | 0.0729 |
| RMS of proximity to the standard for all phrases from abstracts and title | 0.0252 |

**MMPR-15, Theory and Methods of Pattern Recognition and Classification**

| Author (s) | I.E. Genrikhov, E.V. Djukova |
| Title of the article | Complete decision trees in a tasks of classification by precedents |
| The maximum rating of proximity to the standard for the title is achieved relative to the document | K.V. Vorontsov Combinatorial theory of overfitting: results, applications and open problems // MMPR-15 |
| Words from clusters of highest TF-IDF values for individual phrases | classification, complete, decision, tree, precedent, procedure, description, review, give |
| Combinations formed from them: | (Complete) decision tree, give a review |
| Rating of proximity to the standard of the article title | 0.1253 |
| RMS of proximity to the standard for all phrases from abstracts and title | 0.0489 |

**MMPR-14, Methods and Models of Pattern Recognition and Forecasting**

| Author (s) | O.V. Barinova, D.P. Vetrov |
| Title of the article | Estimates of the generalization ability for boosting with a probabilistic entries |
| The maximum rating of proximity to the standard for the title is achieved relative to the document | K.V. Vorontsov Combinatorial theory of overfitting: results, applications and open problems // MMPR-15 |
| Words from clusters of highest TF-IDF values for individual phrases | generalize, ability, composition, error, probability, classification, sample, top, boosting |
| Combinations formed from them: | generalization ability, classification error |
| Rating of proximity to the standard of the article title | 0.1359 |
| RMS of proximity to the standard for all phrases from abstracts and title | 0.0498 |

**IIP-9, Theory and Methods of Pattern Recognition and Classification**

| Author (s) | S.D. Dvoenko, D.O. Pshenichny |
| Title of the article | On negative eigenvalues removing from matrixes of pairwise comparisons |
The maximum rating of proximity to the standard for the title is achieved relative to the document Dvoenko S.D., Pshenichny D.O. Metrical correction of matrices of pairwise comparisons // MPR-16

Words from clusters of highest TF-IDF values for individual phrases:
- pairwise
- matrix
- numerical
- measurement
- correct
- processing
- follow

Combinations formed from them:
- matrix of pairwise

Rating of proximity to the standard of the article title: 0.0952
RMS of proximity to the standard for all phrases from abstracts and title: 0.0353

Table 2. Subordinate words from clusters with highest TF-IDF values.

| Subordinate words | Combination |
|-------------------|-------------|
| nearest neighbor  | The principle of gap maximization for nearest neighbor monotonic classifier |
| cross-validation  | Accurate estimates of complete cross-validation for monotonic classifiers based on the nearest neighbor principle are obtained. |
| generalization ability | It is shown that a monotonic classifier has the best generalization ability, in which the separating surface passes in the middle of the gap between the classes. |
| complete          | Complete decision trees in a tasks of classification by precedents |
| (complete) decision trees | The report presents the results obtained by the authors, the development of classification algorithms based on complete decision trees. |
| give an overview  | An overview of the main results obtained by the authors earlier in this field is given. |
| generalization ability | Estimates of the generalization ability for boosting with a probabilistic entries |
| classification error | In this paper, we propose a new upper bound for the classification error for compositions of simple classifiers, based on the reduction of the binary classification problem with overlapping class distributions to the classification problem with non-overlapping classes. |
| matrices of paired | In data mining, the results of experiments are often presented in the form of matrices of paired comparisons of the elements of the set being analyzed among themselves. |

MMPR-15, Statistical Learning Theory, K.V. Vorontsov, G.A.Makhina.

MMPR-15, Theory and Methods of Pattern Recognition and Classification, I.E. Genrikhov, E.V. Djukova

MMPR-14, O.V.Barinova,D.P.Vetrov
To check the presence in the analyzed phrases of links between the words related to the clusters of the highest TF-IDF values, MaltParser [15] was used, i.e., a tool for parsing the phrases of natural languages and working with dependency trees. As can be seen from Table 1 and Table 2, the syntactic relation corresponds to the location of the words of the specified clusters in the neighborhood in the linear phrase series, which also indicates the unity of the component of the semantic image of the text.

Since the title and phrases of the article abstract (by definition) represent a certain single semantic image, it is entirely acceptable to analyze the occurrence of words assigned to a cluster of the highest TF-IDF values for one phrase in relation to words relative to other phrases. In the example from Table 2 in the above-mentioned article by I.E. Genrikh and E.V. Djukova the word “complete” was not included in the cluster of the highest TF-IDF values, but it is syntactically submitted to the word “tree” from the specified cluster and, therefore, forms the desired key combination.

4. Some technical details
Taking into account the conclusions of [10] regarding the semantic context of term words, the evaluation of estimates (1) - (4) was made without taking into account prepositions and conjunctions. Text extraction from a PDF file was performed using the functions of the pdfinterp, converter, layout, and pdfpage classes as part of the PDFMiner package [16]. In order to correctly recognize formulas, as in [10], all formulas from the analyzed documents here were translated by an expert manually into a format that is close to that 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 [17]. Lemmatization of words was performed using the morphological analyzer pymorphy2 [18]. If a word has more than one parsing variant when determining its initial form (lemma), in order to calculate the TF-IDF measure, the closest one issued by the n-gram tagger from the nltk4russian library is taken [19].

5. Conclusions
The main result of this paper is a proposed method for assessing the proximity of a text to a semantic pattern relative to a thematic text corpus.

The effectiveness of the proposed method can be estimated by splitting the collection’s texts into clusters by the evaluation value (4) and the ratio of the number of texts assigned to the cluster of the highest evaluation values to the total number of texts in the collection. So, on the basis of the collections mentioned in Table 1 and Table 2, we have at least a threefold reduction in the number of documents (scientific articles) that should be read first when studying a given subject area, for example, by students.

It should be noted that the last example in Table 2 takes into account the transitivity of the syntactic relation within the sequence of coordinated words, cf. “Matrices - comparisons – paired”. A topic for separate study here is the dynamics of the change in the TF-IDF measure when we extend our consideration from discrete words to L-grams (according to C. Shannon, [20]). The arrangement of the cluster words of the highest TF-IDF values in the neighborhood in the linear phrase structure should be considered as a necessary but not sufficient condition for assignment to the key combinations that determine the semantic image of the text.

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 pattern”, it is also of interest to use formula (4) in the numerator of the maximum of the obtained values of the proximity to the standard for all the phrases of the text under consideration. To assess the accuracy of identifying key combinations of words, hereafter, it may be necessary to integrate estimates of the form (4) in the entire collection under analysis.

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

[1] Website of Federal State Educational Standards of Higher Education Available at: http://fgosvo.ru (accessed 12.03.2019)
[2] Emelyanov G M and Smirnova E I 1999 Logical Model Of Hypertext Image Database Pattern Recognition and Image Analysis 9 (3) 458–491
[3] Kuzmin A, Aduenko A and Strijov V 2014 Thematic Classification Using Expert Model for Major Conference Abstracts Informational Technologies 6 22–26
[4] Huang E 2011 Paraphrase Detection Using Recursive Autoencoder (accessed 14.03.2019) Available at: http://nlp.stanford.edu/courses/cs224n/2011/reports/ehhuang.pdf
[5] ParaPhraser: Paraphrasing and Text Synonymization (accessed 14.03.2019) Available at: http://paraphraser.ru/
[6] Sochenkov I V, Zubarev D V and Smirnov I V 2017 The ParaPlag: Russian Dataset for Paraphrased Plagiarism Detection Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference «Dialogue» 16 (23) 1 284–297 (accessed 01.05.2019) Available at: http://www.dialog-21.ru/media/3982/dialogue2017_v1.pdf
[7] SyntaxNet: Neural Models of Syntax (accessed 01.05.2019) Available at: https://github.com/tensorflow/models/tree//m
[8] Russian Paraphrase Detection Task (accessed 01.05.2019) Available at: https://ainconf.ru/2016/paraphraser
[9] Emelyanov G M, Mikhailov 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) 1089–1106
[10] Emelyanov G M, Mikhailov D V and Kozlov A P 2018 Relevance of a Set of Topical Texts to a Knowledge Unit and the Estimation of the Closeness of Linguistic Forms of Its Expression to a Semantic Pattern Pattern Recognition and Image Analysis 28 (4) 771–782. DOI: 10.1134/S1054661818040090
[11] Zagoruiko N G 1999 Prikladnye Metody Analiza Dannyy I Znanii [Applied Methods of Data and Knowledge Analysis] (Novosibirsk: Sobolev Institute of Mathematics of the Siberian Branch of the Russian Academy of Sciences)
[12] Proceedings of the South-West State University. Series: Control, Computer engineering, Information science. Medical instruments engineering (accessed 18.03.2019) Available at: https://swwsu.ru/izvestiya/seriesivt/eng/
[13] The Eclipse Foundation (accessed 21.03.2019) Available at: https://www.eclipse.org
[14] Chirkova N, Aysina R and Vorontsov K 2015 Hierarchical Additively Regularized Topic Model of a Scientific Conference Russian National Conference MMPR-17 (Svetlogorsk, Kaliningrad region, Russia) (Moscow: TORUS PRESS) 231
[15] MaltParser – a data-driven dependency parser (accessed 23.03.2019) Available at: http://www.maltparser.org/
[16] PDFMiner – Python PDF parser and analyzer (accessed 25.03.2019) Available at: https://euske.github.io/pdfminer/
[17] Natural Language Toolkit (accessed 25.03.2019) Available at: http://www.nltk.org/
[18] Korobov M 2015 Morphological Analyzer and Generator for Russian and Ukrainian Languages Analysis of Images, Social Networks and Texts: 4th Int. Conf. (Yekaterinburg, Russia) (Springer) 320-332 DOI: 10.1007/978-3-319-26123-2_31
[19] Moskvina A, Orlova D, Panicheva P and Mitrofanova O 2016 Development of the Core for Syntactic Parser for the Russian Language On the Basis of NLTK Libraries Computational Linguistics and Ontology, Proceedings of XIX International Conference «Internet and Modern Society» (St. Petersburg, Russia) 44–54 (accessed 07.04.2019) Available at: https://openbooks.ifmo.ru/ru/file/4103/4103.pdf
[20] Shannon C E 1951 Prediction and Entropy of Printed English BSTJ 30 (1) 50–64