TF-IDF metrics and estimation of affinity to the sense standard for titles and abstracts of scientific articles without paraphrasing

Mikhaylov D., Emelyanov G.

Yaroslav-the-Wise Novgorod State University

All-Russian Conference with International Participation
«Mathematical Methods for Pattern Recognition» (MMPR-19),

November 26–29, 2019

Moscow, Russian Federation

Optimal (i. e. standard) sense transfer

Is provided *by the set* of textual units and their links *necessary and enough* for representation of knowledge unit.

Requirements for the solution

1. 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.
2. The expert should not rephrase the text to search the semantically equivalent natural-language forms of description of knowledge unit.
3. Revelation of a set of text units and their relations necessary and enough to represent a knowledge unit and satisfies the sense standard.
4. There are no apriori restrictions on the nature of the links of text units.

Abstract and title of scientific paper

1. Reflect *the main content* and *the most important results* obtained by authors *without unnecessary methodological details*.
2. The title reflects *the name of method, model, algorithm* presented by paper, as well as the *theoretical basis* of the *proposed solutions*.

Closest researches and projects

- Hierarchical thematic modeling of major conference proceedings [Strijov V., 2014].
- Paraphrase detection using recursive neural network pre-trained on the parse trees of initial phrases [Huang E., 2011].
- Preparation of tagged text corpora for training the system of automatic paraphrasing [the ParaPhraser project].
- The ParaPlag: Russian dataset for Paraphrased Plagiarism Detection.
- Scoring the semantic equivalence of sentences by computing the edit distance between their syntactic dependency trees [HSE School of Linguistics, 2016].

Main problems

- the proper qualitative analysis of linguistic expressional means, meaningful for choosing the best variants among possible paraphrases, is not provided;
- the quality of training for paraphrase recognition system.

TF-IDF metrics and key word combinations

According to classic definition, TF-IDF is the product of two statistics:
term frequency (TF) and *inverse document frequency (IDF)*.

Term frequency estimates the significance of word t_i within the document d and can be defined as

$$\mathrm{tf}\,(t_i, d) = \frac{n_i}{\sum_k n_k}, \tag{1}$$

where n_i is the number of times that t_i occurs in document d,
and denominator contains the total number of words for d.

The value of IDF is unique for each unique word in corpus D and can be determined as follows:

$$\mathrm{idf}\,(t_i, D) = \log\left(\frac{|D|}{|D_i|}\right), \tag{2}$$

where numerator represents the total number of documents in corpus,
and $|D_i \subset D|$ is a number of documents where the word t_i appears.

Interpreting the TF-IDF for word combinations, let's identify the numerator value in (1) with the number of co-occurrences of all combination words in the phrases of given $d \in D$; when calculating the value in denominator of (1) we'll separately take into account the cases of co-occurrence of combination words and occurrence without simultaneous presence in a phrase.

Clustreing the vocabulary of initial phrase by TF-IDF: basic assumptions

1. 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.
2. The fact that the term has synonyms at the same document means the decrease of TF metrics for this word relatively to given document.
3. For words of general vocabulary and for those terms which are prevail in corpus the value of IDF tends to zero.
4. Synonyms, unique for some documents of corpus, will have a higher values of IDF.

   For example: general-vocabulary words which are define the conversive replacements, like *«приводить ⇔ являться следствием»* (in Russian).

Statement 1

The value of TF-IDF metrics for key word combination should not be less than the minimum of values of the mentioned measure for its separate words.

Let

D be an initial text set considered as a topical corpus.

X be an ordered descending sequence of $\mathrm{tf}(t_i, d) \cdot \mathrm{idf}(t_i, D)$ values for all words t_i of initial phrase relatively to document $d \in D$.

F be the sequence of clusters $H_1, \ldots H_r$ as a result of splitting the initial X by means of algorithm close to FOREL class taxonomy algorithms.

As the mass center of cluster H_i the arithmetic mean of all $x_j \in H_i$ is taken.

For estimating the phrase affinity to the sense standard
the most significant words are related to the clusters:

$H_1(X)$ — the *terms* from initial phrase which are the *most unique* for d;

$H_{r/2}(X)$ — *general vocabulary* as a basis of *synonymic paraphrases*, and those *terms* which have *synonyms*;

$H_r(X)$ — those *terms* which are *prevail* in corpus.

Selecting the estimation for the phrase affinity to the sense standard

Basic empirical considerations

- the division into general vocabulary and terms should be expressed as much as possible;
- the words in clusters $H_1, \dots H_r$, formed by the TF-IDF of words of the source phrase relative to a certain $d \in D$, should be distributed more or less evenly;
- the number of resulted clusters on the sequence X must be close to three as much as possible at maximum of TF-IDF values for words related to the cluster H_1.

Documents of corpus D are sorted descending the product of estimations:

$$val_1 = -1/\log_{10}\left(\Sigma_{H_1}\right), \tag{3}$$

$$val_2 = 10^{-\sigma\left(\left|H_i,\, i=\left\{1,\, r/2,\, r\right\}\right|\right)}, \tag{4}$$

and, correspondingly,

$$val_3 = \left|H_1 \setminus H_{r/2} \setminus H_r\right| / \mathrm{len}\left(X\right), \tag{5}$$

where Σ_{H_1} is the sum of TF-IDF values for words related to the cluster H_1 concerning to $d \in D$;
$\sigma\left(\left|H_i,\, i = \left\{1,\, r/2,\, r\right\}\right|\right)$ is the RMSD of number of elements in $H_i \in \left\{H_1,\, H_{r/2},\, H_r\right\}$;
$\mathrm{len}\left(X\right)$ is the length of X.

Remarks

- in a case of $\Sigma_{H_1} = 0$ the value of val_1 is assumed to be zero;
- if the number of clusters TF-IDF-obtained is smaller than two, the values of $\left|H_{r/2}\right|$ and $\left|H_r\right|$ are assumed to be zero;
- in a case of only two TF-IDF-obtained clusters the value of $\left|H_r\right|$ is assumed to be zero.

Let

Ts be a *group of phrases*, first of which is *the title* of scientific article and others represent its *abstract*.

The first variant of estimation:

$$\mathrm{N}_1(\mathrm{Ts}, D) = \frac{\max\limits_{d \in D}\left(val_1(Ts_1, d) \cdot val_2(Ts_1, d) \cdot val_3(Ts_1, d)\right)}{\sigma\left(\max\limits_{d \in D}\left(val_1(Ts_i, d) \cdot val_2(Ts_i, d) \cdot val_3(Ts_i, d)\right), Ts_i \in \mathrm{Ts}\right) + 1}. \quad (6)$$

Here:

the *numerator* is the estimation of *affinity to the standard* for the *article title* (Ts_1);
the first summand in *denominator* is the RMSD for affinity to standard for all $Ts_i \in \mathrm{Ts}$.

Remarks

- the estimation (6) depends on the selection of corpus D by expert;
- the offered estimation *does not imply sorting* of phrases $Ts_i \in \mathrm{Ts}$ by *affinity to the sense standard* and corresponds essentially to the order of selection of articles with *analysis of the title at first*;
- the apriori assumption of maximal closeness to the standard exactly of the title of the article is not always performed in practice.

The second variant of estimation:

$$\mathrm{N}_2(\mathrm{Ts}, D) = \frac{\max\limits_{d \in D}\left(val_1\left(Ts_{\max}, d\right) \cdot val_2\left(Ts_{\max}, d\right) \cdot val_3\left(Ts_{\max}, d\right)\right)}{\sigma\left(\max\limits_{d \in D}\left(val_1\left(Ts_i, d\right) \cdot val_2\left(Ts_i, d\right) \cdot val_3\left(Ts_i, d\right)\right), Ts_i \in \mathrm{Ts}\right) + 1}, \quad (7)$$

where $Ts_{\max} \in \mathrm{Ts}$ is the phrase for which the affinity to the sense standard is maximal.

Statement 2

The *maximal final* rank in the collection will be designated to the article with a greatest value of estimation (6) related to the same cluster with the value of estimation (7) for the same paper.

Remarks

- the correctly application of *Statement 2* assumes the relating to the same cluster the value of estimation (6) for article with a maximal final rank, and a maximal value of estimation (6) 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 estimation (6) in analyzed collection;
- 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 estimations (6) and (7) in *Statement 2*.

Initial data for experiment: the structure of text corpus

- Taurida journal of computer science theory and mathematics (TJCSTM, 3 papers);
- Proceedings of International conferences «Intelligent Information Processing» IIP-8 and IIP-9 of the years 2010 and 2012 (2 papers);
- Proceedings of All-Russian Conference with International Participation on Mathematical Methods for Pattern Recognition (MMPR-15, 1 paper);
- Proceedings of the Conference MMPR-13 (2007, 2 papers);
- Proceedings of the Conference MMPR-16 (2013, 14 papers);
- Proceedings of the Conference IIP-10 (2014, 2 papers);
- the text of a scientific report prepared in 2003 by Dmitry Mikhaylov.

Remark

The number of words in documents of corpus varied here from 218 to 6298, and the number of phrases per document varied between 9 and 587.

Initial data for experiment: the scope of documents included in the corpus

- mathematical methods for learning by precedents (K. Vorontsov, M. Khachay, E. Djukova, N. Zagoruiko, Yu. Dyulicheva, I. Genrikhov, A. Ivakhnenko);
- methods and models of pattern recognition and forecasting (V. Mottl, O. Seredin, A. Tatarchuk, P. Turkov, M. Suvorov, A. Maysuradze);
- intelligent processing of experimental information (S. Dvoenko, N. Borovykh);
- image processing, analysis, classification and recognition (A. Zhiznyakov, K. Zhukova, I. Reyer, D. Murashov, N. Fedotov, V. Martyanov, M. Kharinov).

Initial data for experiment: collections for selecting the articles

- proceedings of «Intelligent Information Processing» conference of the year 2012, section «Theory and Methods of Pattern Recognition and Classification» (14 articles);
- proceedings of the 14^{th} All-Russian conference «Mathematical Methods for Pattern Recognition», section «Methods and Models of Pattern Recognition and Forecasting» (2009, 35 articles);
- proceedings of the 15^{th} All-Russian conference «Mathematical Methods for Pattern Recognition» (2011), section «Theory and Methods of Pattern Recognition and Classification» (18 articles) and «Statistical Learning Theory» (10 articles).

Some technical details

- Estimations (3)–(7) are calculated disregard of prepositions and conjunctions.
- Text extraction from a PDF file was implemented using the functions of the *pdfinterp*, *converter*, *layout* and *pdfpage* classes as part of the *PDFMiner* package.
- In order to be correctly recognized, all formulas from the analyzed documents here were translated by an expert manually into a format close to 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.
- Lemmatization of words was performed using the morphological analyzer *PyMorphy2*.
- If a word has more than one parsing variant when determining its initial form (lemma), the closest one issued by the n-gram tagger from the *nltk4russian* library is taken.

software implementation (in Python 2.7) and experimental results

The result: articles with a maximum value of estimation (6) 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* [In Russian]
The maximum affinity to the sense standard for the title is achieved relative to the document	*Vorontsov K. V. 2011. Combinatorial theory of overfitting: results, applications and open problems* [In Russian] // *MMPR-15*
Estimation of affinity to the sense standard for the article title:	0,0729
RMSD of affinity to the standard for all phrases from abstract and title:	0,0252
Value of estimation (6):	0,0711
Value of estimation (7):	0,0711
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 classification tasks by precedents* [In Russian]
The maximum affinity to the sense standard for the title is achieved relative to the document	*Vorontsov K. V. 2011. Combinatorial theory of overfitting: results, applications and open problems* [In Russian] // *MMPR-15*
Estimation of affinity to the sense standard for the article title:	0,1253
RMSD of affinity to the standard for all phrases from abstract and title:	0,0489
Value of estimation (6):	0,1194
Value of estimation (7):	0,1194

The result: articles with a maximum value of estimation (6) in collections

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* [In Russian]
The maximum affinity to the sense standard for the title is achieved relative to the document	*Vorontsov K. V. 2011. Combinatorial theory of overfitting: results, applications and open prob lems* [In Russian] // *MMPR-15*
Estimation of affinity to the sense standard for the article title:	0,1359
RMSD of affinity to the standard for all phrases from abstract and title:	0,0498
Value of estimation (6):	0,1295
Value of estimation (7):	0,1295
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 matrices of pairwise comparisons* [In Russian]
The maximum affinity to the sense standard for the title is achieved relative to the document	*Dvoenko S. D., Pshenichny D. O. 2013. Metrical correction of matrices of pairwise comparisons* [In Russian] // *MMPR-16*
Estimation of affinity to the sense standard for the article title:	0,0952
RMSD of affinity to the standard for all phrases from abstract and title:	0,0353
Value of estimation (6):	0,0920
Value of estimation (7):	0,0920

The result: articles with a maximum value of estimation (7) 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* [In Russian]		
Phrase closest to the standard [In Russian]	*Принцип максимизации зазора для монотонного классификатора ближайшего соседа*		
The maximum affinity to the sense standard for the phrase is achieved relative to the document	*Vorontsov K. V. 2011. Combinatorial theory of overfitting: results, applications and open prob lems* [In Russian] // *MMPR-15*		
Estimation of affinity to the sense standard for the article title:		0,0729	
RMSD of affinity to the standard for all phrases from abstract and title:		0,0252	
Value of estimation (7):	0,0711	Value of estimation (6):	0,0711

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 classification tasks by precedents* [In Russian]		
Phrase closest to the standard [In Russian]	*Полные решающие деревья в задачах классификации по прецедентам*		
The maximum affinity to the sense standard for the phrase is achieved relative to the document	*Vorontsov K. V. 2011. Combinatorial theory of overfitting: results, applications and open prob lems* [In Russian] // *MMPR-15*		
Estimation of affinity to the sense standard for the article title:		0,1253	
RMSD of affinity to the standard for all phrases from abstract and title:		0,0489	
Value of estimation (7):	0,1194	Value of estimation (6):	0,1194

The result: articles with a maximum value of estimation (7) in collections

MMPR-14, Methods and Models of Pattern Recognition and Forecasting			
Author(s)	*D. I. Mel'nikov, V. V. Strijov, E. Yu. Andreeva and G. Edenharter*		
Title of the article	*Support set selection when constructing of stable integral indicators* [In Russian]		
Phrase closest to the standard [In Russian]	*Объекты описаны в линейных шкалах*		
The maximum affinity to the sense standard for the phrase is achieved relative to the document	*Abramov V. I., Seredin O. S., Sulimova V. V., Mottl V. V. 2010. Equivalence of kernel functions and linear-space representations of arbitrary real-world objects // IIP-8*		
Estimation of affinity to the sense standard for the article title:	0,0137		
RMSD of affinity to the standard for all phrases from abstract and title:	0,0639		
Value of estimation (7):	0,1426	Value of estimation (6):	0,0129
IIP-9, Theory and Methods of Pattern Recognition and Classification			
Author(s)	*N. K. Zhivotovskiy, K. V. Vorontsov*		
Title of the article	*The exactness criteria of combinatorial generalization bounds* [In Russian]		
Phrase closest to the standard [In Russian]	*Комбинаторная теория переобучения даёт точные оценки вероятности переобучения для некоторых нетривиальных семейств алгоритмов классификации*		
The maximum affinity to the sense standard for the phrase is achieved relative to the document	*Vorontsov K. V. 2011. Combinatorial theory of overfitting: results, applications and open prob lems* [In Russian] *// MMPR-15*		
Estimation of affinity to the sense standard for the article title:	0,0634		
RMSD of affinity to the standard for all phrases from abstract and title:	0,0578		
Value of estimation (7):	0,1336	Value of estimation (6):	0,0600

Articles with a maximal final rank in collections

Concerning the estimation (6)

MMPR-15, Statistical Learning Theory			
Author(s)	*K. V. Vorontsov, G. A. Makhina*		
Value of estimation (6):	0,0711/0,0711	Value of estimation (7):	0,0711/0,0711[1]
MMPR-15, Theory and Methods of Pattern Recognition and Classification			
Author(s)	*I. E. Genrikhov, E. V. Djukova*		
Value of estimation (6):	0,1194/0,1194	Value of estimation (7):	0,1194/0,1194
MMPR-14, Methods and Models of Pattern Recognition and Forecasting			
Author(s)	*O. V. Barinova, D. P. Vetrov*		
Value of estimation (6):	0,1295/0,1295	Value of estimation (7):	0,1295/0,1426
IIP-9, Theory and Methods of Pattern Recognition and Classification			
Author(s)	*S. D. Dvoenko, D. O. Pshenichny*		
Value of estimation (6):	0,0920/0,0920	Value of estimation (7):	0,0920/0,1336

Not obtains the maximal final rank

MMPR-14, Methods and Models of Pattern Recognition and Forecasting			
Author(s)	*D. I. Mel'nikov, V. V. Strijov, E. Yu. Andreeva and G. Edenharter*		
Value of estimation (6):	0,0129/0,1295	Value of estimation (7):	0,1426/0,1426

Concerning the estimation (7)

MMPR-14, Methods and Models of Pattern Recognition and Forecasting			
Author(s)	*O. V. Barinova, D. P. Vetrov*		
IIP-9, Theory and Methods of Pattern Recognition and Classification			
Author(s)	*N. K. Zhivotovskiy, K. V. Vorontsov*		
Value of estimation (6):	0,0600/0,0920	Value of estimation (7):	0,1336/0,1336

[1] After fraction bar is the maximum of estimation for collection

Key word combinations for articles with a maximum value of estimation (6)

MMPR-15, Statistical Learning Theory	
Author(s)	*K. V. Vorontsov, G. A. Makhina*
Words from clusters of greatest TF-IDF values for individual phrases [In Russian]	*монотонный, сосед, близкий, скользящий, контроль, обобщать, способность*
Combinations formed from them satisfying the condition of *Statement 1*	*ближайший сосед, скользящий контроль, обобщающая способность*
including the words of «median» clusters	*разделяющая поверхность*

MMPR-15, Theory and Methods of Pattern Recognition and Classification	
Author(s)	*I. E. Genrikhov, E. V. Djukova*
Words from clusters of greatest TF-IDF values for individual phrases [In Russian]	*классификация, полный, решающий, дерево, прецедент, процедура, описание, обзор, дать*
Combinations formed from them satisfying the condition of *Statement 1*	*решающее дерево*
including the words of «median» clusters	*распознающая процедура*

MMPR-14, Methods and Models of Pattern Recognition and Forecasting	
Author(s)	*O. V. Barinova, D. P. Vetrov*
Words from clusters of greatest TF-IDF values for individual phrases [In Russian]	*обобщать, способность, композиция, ошибка, вероятность, классификация, выборка, верхний, бустинг*
Combinations formed from them satisfying the condition of *Statement 1*	*обобщающая способность*

IIP-9, Theory and Methods of Pattern Recognition and Classification	
Author(s)	*S. D. Dvoenko, D. O. Pshenichny*
Words from clusters of greatest TF-IDF values for individual phrases [In Russian]	*парный, матрица, численный, измерение, корректный, обработка, следовать*
Combinations formed from them satisfying the condition of *Statement 1*	*матрица парных*
including the words of «median» clusters	*матрица парных сравнений*

Key word combinations for articles with a maximum value of estimation (7)

MMPR-14, Methods and Models of Pattern Recognition and Forecasting	
Author(s)	*D. I. Mel'nikov, V. V. Strijov, E. Yu. Andreeva and G. Edenharter*
Words from clusters of greatest TF-IDF values for individual phrases [In Russian]	*опорный, описание, линейный, помощь, учитель, основный, предложить, получение*
Combinations formed from them satisfying the condition of *Statement 1* including the words of «median» clusters	*выбор опорного, описаны (в) линейных*
Estimation of affinity to the sense standard for the article title:	0,0137
Maximum of affinity to the sense standard for phrase:	0,1517
RMSD of affinity to the standard for all phrases from abstract and title:	0,0639

IIP-9, Theory and Methods of Pattern Recognition and Classification	
Author(s)	*N. K. Zhivotovskiy, K. V. Vorontsov*
Words from clusters of greatest TF-IDF values for individual phrases [In Russian]	*обобщать, комбинаторный, вероятность, семейство*
Combinations formed from them satisfying the condition of *Statement 1* including the words of «median» clusters	
Estimation of affinity to the sense standard for the article title:	0,0634
Maximum of affinity to the sense standard for phrase:	0,1413
RMSD of affinity to the standard for all phrases from abstract and title:	0,0578

Co-ordinated words from clusters of greatest TF-IDF values

MMPR-15, Statistical Learning Theory, K. V. Vorontsov, G. A. Makhina	
ближайший сосед	*«Принцип максимизации зазора для монотонного классификатора ближайшего соседа»*
скользящий контроль	*«Получены точные оценки полного скользящего контроля для монотонных классификаторов, основанных на принципе ближайшего соседа»*
обобщающая способность	*«Показано, что наилучшей обобщающей способностью обладает монотонный классификатор, в котором разделяющая поверхность проходит посередине зазора между классами»*
including the words of «median» clusters	
монотонный классификатор	*«Принцип максимизации зазора для монотонного классификатора ближайшего соседа»*
MMPR-15, Theory and Methods of Pattern Recognition and Classification, I. E. Genrikhov, E. V. Djukova	
полный	*«Полные решающие деревья в задачах классификации по прецедентам»*
(полное) решающее дерево	*«В докладе представлены результаты, полученные авторами, разработки алгоритмов классификации на основе полных решающих деревьев»*
дан обзор	*«Дан обзор основных результатов, полученных авторами ранее в данной области»*
including the words of «median» clusters	
распознающая процедура	*«Построены модели распознающих процедур, нацеленные на решение задач с неполными данными (с пропусками в признаковых описаниях объектов) и с неравномерным распределением обучающих объектов по классам»*

Co-ordinated words from clusters of greatest TF-IDF values

MMPR-14, O. V. Barinova, D. P. Vetrov	
обобщающая способность	*«Оценки обобщающей способности бустинга с вероятностными входами»*
including the words of «median» clusters	
ошибка классификации	*«В данной работе предлагается новая верхняя оценка ошибки классификации для композиций простых классификаторов, основанная на сведении бинарной задачи классификации с перекрывающимися распределениями классов к задаче классификации с неперекрывающимися классами»*
MMPR-14, D. I. Mel'nikov, V. V. Strijov, E. Yu. Andreeva and G. Edenharter	
including the words of «median» clusters	
выбор опорного	*«Выбор опорного множества при построении устойчивых интегральных индикаторов»*
описание объекта	*«Исследуется задача построения интегрального индикатора множества объектов, устойчивого к выбросам в описаниях объектов»*
описаны (в) линейных	*«Объекты описаны в линейных шкалах»*
IIP-9, S. D. Dvoenko, D. O. Pshenichny	
матрица парных	*«В интеллектуальном анализе данных результаты экспериментов часто представлены в виде матриц парных сравнений элементов анализируемого множества между собой»*
including the words of «median» clusters	
матрица парных сравнений	*«Об устранении отрицательных собственных значений матриц парных сравнений»*

Co-ordinated words from clusters of greatest TF-IDF values

IIP-9, N. K. Zhivotovskiy, K. V. Vorontsov	
including the words of «median» clusters	
точность комбинаторных	*«Критерии точности комбинаторных оценок обобщающей способности»*
комбинаторная теория	*«Комбинаторная теория переобучения даёт точные оценки вероятности переобучения для некоторых нетривиальных семейств алгоритмов классификации»*

Remarks

- since the title and phrases of the article abstract represent a uniform semantic image, it is entirely acceptable to analyze the occurrence of words related to the cluster of greatest TF-IDF values concerning the given phrase, in a word relationships of other phrases;
- in current work a set of aforementioned links will be associated with the key word combination, if a connected subgraph of the phrase syntactic tree (undirected) is corresponded to it, and at least one word combination satisfies the condition of *Statement 1*.

Some technical details

To reveal the links of words in analyzed phrases, *MaltParser* was used, i. e. a tool for parsing the phrases of natural languages and working with dependency trees.

Conclusions

1. The main *result* of current work is the proposed *method* for estimating the closeness of a text to the sense standard relatively to a topical text corpus.
2. Its *effectiveness* can be *estimated* by splitting the collection into clusters by the closeness to a standard and the ratio of number of texts assigned to the cluster of highest values to the total number of texts in the collection.
3. The offered method gives *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.
4. The transitivity of syntactic relation within the sequence of co-ordinated words *requires to study* the dynamics of TF-IDF change when we extent our consideration from discrete words to L-grams (according to C. Shannon).
5. The neighborhood of the words of greatest TF-IDF values in the phrase should be considered as *necessary but not enough* condition for assignment to the key combinations that determine the semantic image of the text.
6. It is of interest the *search* of key word combinations in abstracts and titles *as a basis* for designating in disputable cases the final rank and hierarchization of articles according to significance when studying a given subject area.