УДК 004.912
DOI: 10.25559/SITITO.15.201904.954-960

The Improved Algorithm for Calculation of the Contextual Words Meaning in the Text

E. A. Dorenskaya[1*], Y. A. Semenov[1,2]
[1] Alikhanov Institute for Theoretical and Experimental Physics of National Research Center "Kurchatov Institute", Moscow, Russia
25 Bolshaya Cheremushkinskaya Str., Moscow 117218, Russia
[*] dorenskaya@ itep.ru
[2] Moscow Institute of Physics and Technology, Dolgoprudny, Russia
9 Institutskij per., Dolgoprudny 141701, Moscow region, Russia

Abstract

Some modifications of the algorithm for context calculation, published in [1], are considered. A new solution for word and document context calculation is proposed. To improve a context determination it is proposed to take into consideration distances between words W1 and W2. This approach is especially important, when W2 number is >1. The results of investigations of these two formulas are presented. For efficiency comparison of these formulas calculation has been made for 100 texts. There were built distributions for C average and dispersion, which were compared with model data from [1]. The weight function $f(L_i)$ has been optimized. The versions comparison was made according to the value of σ/C_{aver}. The C dispersion was calculated for all version of the weight function. Dispersion of C appeared to be rather big because of great variation of text size, number W2 and W3, as well as wide distribution of words in the text. There is an example of L distribution for W2="компьютер".

Keywords: context recognition, weighting function, width of flat area of weight function, context value, probability density, computer analyses, artificial intellect, problem description, program error minimization.

For citation: Dorenskaya E.A., Semenov Y.A. The Improved Algorithm for Calculation of the Contextual Words Meaning in the Text. *Sovremennye informacionnye tehnologii i IT-obrazovanie* = Modern Information Technologies and IT-Education. 2019; 15(4):954-960. DOI: 10.25559/SITITO.15.201904.954-960

Улучшенный алгоритм вычисления контекстного значения слов в тексте

Е. А. Доренская[1*], Ю. А. Семенов[1,2]
[1] Институт теоретической и экспериментальной физики имени А.И. Алиханова Национального исследовательского центра «Курчатовский институт», Москва, Россия
117218, Россия, г. Москва, ул. Большая Черёмушкинская, д. 25
[*] dorenskaya@ itep.ru
[2] Московский физико-технический институт (национальный исследовательский университет), г. Долгопрудный, Россия
141701, Россия, Московская обл., г. Долгопрудный, Институтский пер., д. 9

Аннотация

Рассмотрены некоторые модификации алгоритма определения контекста, опубликованного ранее [1]. Предложена усовершенствованная формула для вычисления контекста слова и документа. Для более точного расчёта контекстного значения предложено учитывать расстояния между словами W1 и W2. Такой подход важен при наличии числа W2 >1. Приводятся результаты исследования работы этих 2-х формул. Для сравнения эффективности формул были проведены расчеты для 100 текстов. По данным расчётов построены распределения, которые были сравнены с моделью, описанной в статье [1]. Проведен оптимальный выбор весовой функции $\boldsymbol{f(L_i)}$, Сравнение вариантов проводилось по значениям отношений s/Сcред, Для всех вариантов формулы, весовой функции и ширины площадки $\boldsymbol{f(L_i)}$ была рассчитана дисперсия C. Она оказалась довольно большой за счёт разного размера текстов, количества слов W2 и W3, а также их распределения по тексту. Предоставлен пример распределения расстояний между словами W2-W3 в файле для слова W2 "компьютер".

Ключевые слова: распознавание контекста, весовая функция, ширина площадки, контекстное значение, распределение плотности вероятности, машинный анализ, искусственный интеллект, описание проблемы, минимизация программных ошибок.

Для цитирования: *Доренская Е. А., Семенов Ю. А.* Улучшенный алгоритм вычисления контекстного значения слов в тексте // Современные информационные технологии и ИТ-образование. 2019. Т. 15, № 4. С. 954-960. DOI: 10.25559/SITITO.15.201904.954-960

Introduction

Since text analyses is becoming more and more actual in particular because of computer analytics, it is so important correctly interpret context meaning of all words in the text.

There was proposed an algorithm [1,2] to calculate context of separate words and documents. This algorithm takes into account relative distance between the word and other words connected with it. It is supposed, that for any word W1, correspond some words W2, determining a context meaning of W1. To any word W2 may correspond several words W3, connected by implication with W2. For example, W2(routine(computer))=W3(algorithm,debugging,testing,iteration,CPU,recursion, variable,byte,...). The tree of connection may be spanned by adding words connected with words W3.

Research aim and used methods

The disadvantage of the method, described in article [1], is that, it does not take into account distances between words W1 and W2. The aim of our investigation is an improvement of the formula, described in [1] and selection of weight function, permitting to analyze context most precisely.

If there is one word W1 and several W2, defining different context meaning of W1, than to calculate real context meaning of W1 one may use the following formula:

$C_{k,n} = \sum_{i=1}^{m} (M_i \times f(L_i))$; **[1]**,

where C – number, defining context of W1, L - distance between word, e.g.,, "computer" (W2) and "debugging" (W3), M_i - metrics for word-characteristic W3 (M=1÷100), m - number of words W3, semantically corresponding to definite word W2, $f(L_i)$ - weight function of L_i, i - number of particular word W3 [1]. The meaning of M is taken from the table, see [1], their values were optimized for test texts.

At simplest case $f(L_i) = 1/L_i$, and for small documents $f(L_i) = 1$. **L** is defined by word number **N** situated between word **W2** and one of the words **W3 (L=N+1).** weight function $f(L_i)$ is used to decrease influence of distant words to the context estimate of W1.

If there are several W1, and text volume is big enough, we may estimate the context meaning of every word W1 separately, fixing definitional domain size.

It is supposed, that W2 follows after W1, and W3 – after W2 and only contributions from these words influence on the C value, i.e. a contribution all previous word is ignored.

Let there are **n** identical words W1 ($W1_1$,$W1_2$,...$W1_i$,...,$W1_n$). Let also there are **k** identical words W2 (see figure 1), determining context of W1 ($W2_1$,$W2_2$,...,$W2_j$,...,$W2_k$). Generally speaking, every of words W2 (and also W3) may met in the text more than once. Let there are **m** words W3, corresponding these W2. L[$W1_i$,$W2_j$] – distance between words $W1_i$ and $W2_j$. In W2[1,1] the first index corresponds to the number of word W1, and the second – to the number of W2. For W3[1,1,3] first index corresponds to the word number of word W1, the second – to the number of word W2, and the third– to the number of W3. $\mathbf{P}_{L[1,1]}$ – probability, that distance between W1 and W2[1,1] is **L[1,1].**

On figure 1 N corresponds to a number of words W1 in the text.

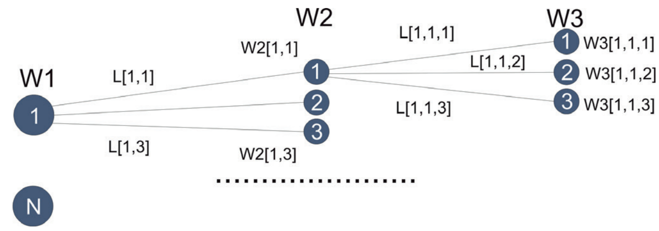


Fig. 1. Parameter determination scheme

To any word W1 must correspond two or more words W2, presented in the text. If there is only one word W1 and one word W2, we need to calculate nothing as in the case context of W1 is defined by W2.

If there is more than one sample of word-meaning W2, defining context of W1, then the formula for W1 context calculation should be changed[1].

$$\mathbf{C}_{k,n} = \sum_{j=1}^{K} \frac{\sum_{i=1}^{m} (M_i \times f(L_i))}{\mathbf{D}_j} \quad \mathbf{[2]},$$

where D_j defines a distance from given W1 to one of the words W2 with number j; K – z number of specific word-meaning W2 in the document. If there are 4 words W2="code", then K=4. $D_j = |L_j - L_0| + 1$. L_0 – minimal distance from W1 to the particular word W2.

Let word W1 may have **m** context meanings, determined by W2. Cardinality of set W2 **ω** is equal **m** (it contains **m** elements). Actually there are **k** words W2 (k ≤ m). 0 ≤ k ≤ m, i.e. W2 are present not to all possible contexts of W1.

If there is only one word W2 and number of W1>1, then at big text volume one may attempt to determine the best regional context meaning for any W1.

For every W2 from the list (W2 number > 1) $\mathbf{C}_{k,n}$ is calculated. Evidently, that these values will be different, as they correspond to different values of distances W1-W2.

To compare results for formulas [1] and [2] 100 files were taken with more than 500 words each. In every of these files there was one word "programm" (W_1), one or more word "computer" (W_2) and arbitrary number of words W_3, coresponding it. For formula 2 also were taken 100 files with more than 500 words each. In every of these files there was one word "programm" (W_1), one or more word "computer" (W_2) and arbitrary naumber of words W_3.

Dispersion, root mean-square deviation σ, average value of C and ratio σ/C_{avr} were calculated for weight functions 1/L, 1, 1/(1+Log(L)) and for formulas [1] and [2] at a=100 and 300. Figure 1 explanes form weight function and its parameters. f(L=1) for all variants =1.

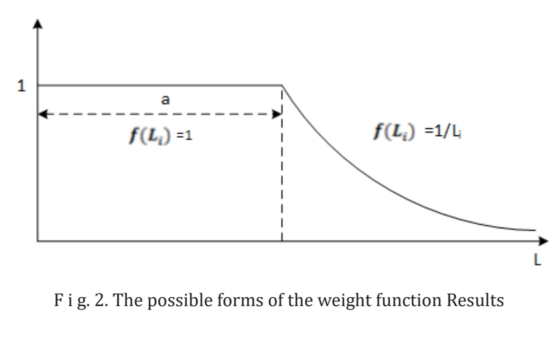


Fig. 2. The possible forms of the weight function Results

[1] *Dorenskaya E.A.* Raschyot kontekstnogo znacheniya zadannogo slova v tekstovom fajle [Calculation of the Context Value of a Given Word in a Text File]. Patent RF, no. 2018615758, 2018. (In Russ.); Dorenskaya E.A. "Wordcontext" v1.0 dlya raschyota kontekstnogo znacheniya slova s uchyotom rasstoyanij ["Wordcontext" v1.0 for Calculating the Contextual Meaning of a Word with Distance]. Patent RF, no. 2019660440, 2019. (In Russ.); Dorenskaya E.A., Semenov Yu.A. Sposob opredeleniya konteksta slova i tekstovogo fajla [The Way to Determine the Context of a Word and a Text File]. Patent RF, no. 2685044. 2018. (in Russ.)

The calculation results are presented in tables 1 and 2 below. In the table 1 results for formula [1] are presented. In the table 2 results for formula [2] are presented. In every table different variants of weight function $f(L_i)$ are considered.

In the first function variant $1/L_i$ was used. In the second also used function $1/L_i$, but at low values of L there is flat area of width **a,** where $f(L_i)$ =1 (see. fig. 2). The calculations were made for a=100 and a=300 (see tables 1 and 2).

The same calculations were made for $f(L_i)$ = $1/(1+\log(L_i))$. This variant of weight function was used to mitigate influence of big L_i on C value.

The best result was achieved for $f(L_i)$ = $1/(1+\log(L_i))$ with a=300 for formula 2 and 1 both. This is related to increasing influence of far words close to this context. This gives preference in comparison with variants, where $f(L_i)$ = 1/L.

However a distance between words is taken into account and if word is far away its influence will be negligible, but it happens for big files only.

The least ratio σ/C_{avr} was resulted for formula 2 (see tables 1 and 2). The less this ratio, the more probability of correct word context identification. That is why we may say, that formula 2 provides better result, than 1. An advantage of formula [2] is connected with that it takes into consideration distances between W1 and W2.

Table 1 for the formula 1

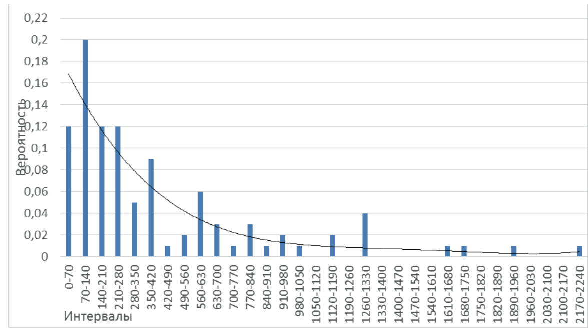
Weight function version	σ	C_{avr}	σ/C_{avr}
$1/L_i$	13,34	7,22	1,85
$1/L_i$ a=100	89,11	58,17	1,53
$1/L_i$ a=300	198,80	149,87	1,33
$1/(1+\log(L_i))$	278,02	221,41	1,26
$1/(1+\log(L_i))$ a=100	300,93	253,45	1,19
$1/(1+\log(L_i))$ a=300	369,72	317,29	**1,17**
1	1061,53	774,58	1,37

Table 2 for the formula 2

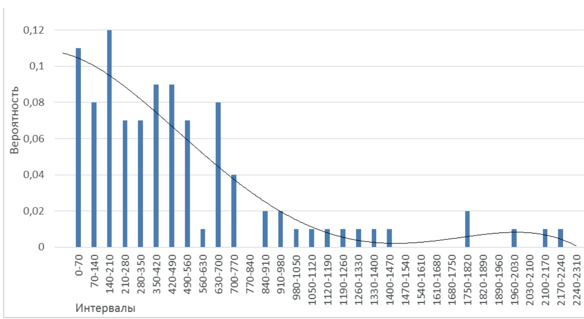
Weight function version	σ	C_{avr}	σ/C_{avr}
$1/L_i$	9,66	7,64	1,26
$1/L_i$ a=100	108,35	83,19	1,30
$1/L_i$ a=300	245,36	183,66	1,34
$1/(1+\log(L_i))$	328,40	276,49	1,19
$1/(1+\log(L_i))$ a=100	351,13	324,26	1,08
$1/(1+\log(L_i))$ a=300	411,93	394,16	**1,05**
1	1434,12	1000,43	1,43

For different types of formula the distributions of probability density distributions were built. That was done for the version with minimal ratio σ/Cavr .

On figures 3 and 4 the abscissa scale shows values of C, and scale of ordinate indicates a probability C to have value in corresponding interval. It is seen that probability is growing at the beginning and then is falling down. Therefore one may say, that it has Gaussian like form. Similar distribution we got at simulation for C in [1]. That means, that the results of calculations of C for real texts confirm the data, got from simulations.



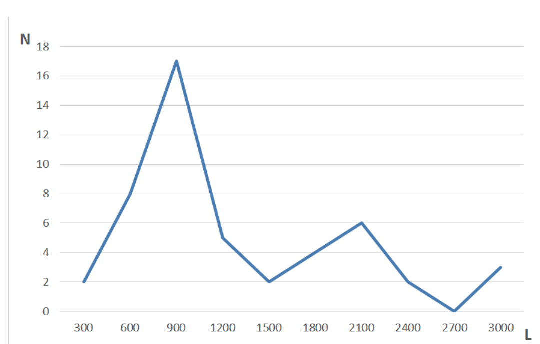
F i g. 3. Probability density distribution for C (formula 2) at weight function 1/(1+log(L)) and flat area width a=300



F i g. 4. . Probability density distribution for C (formula 1) at weight function 1/(1+log(L)) and flat area width a=300

From figures 3 and 4 it is seen, that C dispersion is big enough. It is contributed from variations of file sizes, word numbers W2 and W3, and also their allocation in the text. In [1] it was shown, that even at such dispersion a probability of correct context determination is higher than 90%. Nevertheless, any methods of a dispersion lowering should be welcome.

In [1] at simulation it was supposed a homogeneous word distribution in file and even under this assumption we got rather high dispersion. In reality L variations might be very big, on fig. 5 L-distribution is shown for words W2-W3 for one of the files. The ordinate scale shows event number with given L for ΔL=300. The file contained 75041 bytes.



F i g. 5. An example of distance distribution between W2 and W3

The L scatter lead to an additional dispersion rise, that is partially compensated by averaging.
The dependence of σ/C_{avr} from text size was investigated. The dependence appeared to be week.
Conclusion
As investigations show a proposed formula improve results, gotten in [1,30]. An optimizing weight function was found. This technics may be used at text analyses, and to minimize software errors, by using at software development a problem but not an algorithm description [2].

References

[1] Dorenskaya E.A., Semenov Y.A. The Determination Method for Contextual Meanings of Words and Documents. *Sovremennye informacionnye tehnologii i IT-obrazovanie* = Modern Information Technologies and IT-Education. 2018; 14(4):896-902. (In Russ., abstract in Eng.) DOI: 10.25559/SITITO.14.201804.896-902

[2] Dorenskaya E.A., Semenov Y.A. About the Programming Techniques, Oriented to Minimize Errors. *Sovremennye informacionnye tehnologii i IT-obrazovanie* = Modern Information Technologies and IT-Education. 2017; 13(2):50-56. (In Russ., abstract in Eng.) DOI: 10.25559/SITITO.2017.2.226

[3] Dorenskaya E.A., Semenov Y.A. New Methods of Minimizing the Errors in the Software. In: *CEUR Workshop Proceedings: Proceedings of the VIII International Conference "Distributed Computing and Grid-technologies in Science and Education" (GRID 2018), Dubna, Moscow region, Russia, September 10 -14, 2018*, vol. 2267. 2018, pp. 150-154. Available at: http://ceur-ws.org/Vol-2267/150-154-paper-27.pdf (accessed 15.08.2019). (In Eng.)

[4] Semenov Y.A., Ovsyannikov A.P., Ovsyannikova T.V. Development of the algorithm bank and basics of the language for problem description to minimize a number of program errors. *Proceedings of NIISI RAS*. 2016; 6(2):96-100. Available at: https://elibrary.ru/item.asp?id=29798446 (accessed 15.08.2019). (In Russ., abstract in Eng.)

[5] Semenov Y.A. IT-Economy in 2016 and in 10 Years. Economic Strategies. 2017; 19(1):126-135. Available at: https://elibrary.ru/item.asp?id=28409521 (accessed 15.08.2019). (In Russ., abstract in Eng.)

[6] Rishel T., Perkins L.A., Yenduri S., Zand F. Determining the context of text using augmented latent semantic indexing. *Journal of the American Society for Information Science and Technology*. 2007; 58(14):2197-2204. (In Eng.) DOI: 10.1002/asi.20687

[7] Chen J., Scholz U., Zhou R., Lange M. LAILAPS-QSM: A RESTful API and JAVA library for semantic query suggestions. *PLoS Computational Biology.* 2018; 14(3):e1006058. (In Eng.) DOI: 10.1371/journal.pcbi.1006058

[8] Yang L., Zhang J. Automatic transfer learning for short text mining. *Eurasip Journal on Wireless Communications and Networking.* 2017;

[9] Yan E., Williams J., Chen Z. Understanding disciplinary vocabularies using a full-text enabled domain-independent term extraction approach. *PLoS ONE.* 2017; 12(11):e0187762. (In Eng.) DOI: 10.1371/journal.pone.0187762

[10] Arras L., Horn F., Montavon G., Müller K.-R., Samek W. What is relevant in a text document?": An interpretable machine learning approach. *PLoS ONE.* 2017; 12(8):e0181142. (In Eng.) DOI: 10.1371/journal.pone.0181142

[11] Eidlin A.A., Eidlina M.A., Samsonovich A.V. Analyzing weak semantic map of word senses. *Procedia Computer Science.* 2018; 123:140-148. (In Eng.) DOI: 10.1016/j.procs.2018.01.023

[12] Samsonovich A.V. Weak Semantic Map of the Russian Language: Preliminary Results. *Procedia Computer Science*. 2016; 88:538-543. (In Eng.) DOI: 10.1016/j.procs.2016.08.001

[13] Wei T., Lu Y., Chang H., Zhou Q., Bao X. A semantic approach for text clustering using WordNet and lexical chains. *Expert Systems with Applications.* 2015; 42(4):2264-2275. (In Eng.) DOI: 10.1016/j.eswa.2014.10.023

[14] Zhan J., Dahal B. Using deep learning for short text understanding. *Journal of Big Data.* 2017; 4(1):34. (In Eng.) DOI: 10.1186/s40537-017-0095-2

[15] 15.Khenner E., Nasraoui O. A bilingual semantic network of computing concepts. *Procedia Computer Science.* 2016; 80:2392-2396. (In Eng.) DOI: 10.1016/j.procs.2016.05.460

[16] Yu B. Research on information retrieval model based on ontology. *EURASIP Journal on Wireless Communications and Networking.* 2019; 2019(1):30. (In Eng.) DOI: 10.1186/s13638-019-1354-z

[17] Yelkina E.E., Kononova O.V., Prokudin D.E. Typology of Contexts and Contextual Approach Principles in Multidisciplinary Scientific Research. *Sovremennye informacionnye tehnologii i IT-obrazovanie* = Modern Information Technologies and IT-Education. 2019; 15(1):141-153. (In Russ., abstract in Eng.) DOI: 10.25559/SITITO.15.201901.141-153

[18] Komrakov A.A. Using Ontologies to Describe the Structure of Arrays of Information Exchange. *Sovremennye informacionnye tehnologii i IT-obrazovanie* = Modern Information Technologies and IT-Education. 2019; 15(1):182-189. (In Russ., abstract in Eng.) DOI: 10.25559/SITITO.15.201901.182-189

[19] Barakhnin V.B., Kozhemyakina O.Yu., Rychkova E.V., Pastushkov I.S., Borzilova Y.S. The extraction of lexical and metrorhythmic features which are characteristic for the genre and the style and for their combinations within the process of automated processing of texts in Russian. *Sovremennye informacionnye tehnologii i IT-obrazovanie* = Modern Information Technologies and IT-Education. 2018; 14(4):888-895. (In Russ., abstract in Eng.) DOI: 10.25559/SITITO.14.201804.888-895

[20] Krassovitsky A.M., Ualiyeva I.M., Meirambekkyzy Z., Mussabayev R.R. Lexicon-based approach in generalization evaluation in Russian language media. *Sovremennye informacionnye tehnologii i IT-obrazovanie* = Modern Information Technologies and IT-Education. 2018; 14(3):567-572. (In Russ., abstract in Eng.) DOI: 10.25559/SITITO.14.201803.567-572

[21] Kogalovsky M.R., Parinov S.I. Semantic Annotation of Information Resources by Taxonomies in Scientific Digital Library. In: CEUR Workshop Proceedings: Selected Papers of the XIX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2017). Moscow, Russia, October 9-13, 2017, vol. 2022. 2017, pp. 301-310. Available at: http://ceur-ws.org/Vol-2022/paper47.pdf (accessed 15.08.2019). (In Russ., abstract in Eng.)

[22] Tsukanova Z.V. *Strukturnye i semanticheskie osobennosti zagolovkov sovremennyh nauchnyh statej (na materiale russkogo i anglijskogo jazykov)* [Structural and semantic features of the headings of modern scientific articles (by the material of Russian and English languages)]. *Modern scientific researches and innovations.* 2018; (5):33. Available at: https://elibrary.ru/item.asp?id=35050179 (accessed 15.08.2019). (In Russ.)

[23] Chapaykina N.E. *Semanticheskij analiz tekstov. Osnovnye polozhenija* [Semantic analysis of texts. Fundamentals]. *Young Scientist.* 2012; (5):112-115. Available at: https://elibrary.ru/item.asp?id=20470090 (accessed 15.08.2019). (In Russ.)

[24] Batura T.V. *Metody i sistemy semanticheskogo analiza tekstov* [Methods and systems of semantic text analysis]. Software Journal: Theory and Applications. 2016; (4). (In Russ.) DOI: 10.15827/2311-6749.21.220

[25] Bessmertny I.A. Knowledge visualization based on semantic networks. *Programming and Computer Software.* 2010; 36(4):197-204. (In Eng.) DOI: 10.1134/S036176881004002X

[26] Ayusheeva N.N., Dikikh A.Yu. Model of constructing a semantic network of scientific text. *Modern High Technologies.* 2018; (6):9-13. Available at: https://www.elibrary.ru/item.asp?id=35197327 (accessed 15.08.2019). (In Russ., abstract in Eng.)

[27] Ustalov D.A., Sozykin A.V. A Software System for Automatic Construction of a Semantic Word Network. *Bulletin of the South Ural State University. Series: Computational Mathematics and Software Engineering.* 2017; 6(2):69-83. (In Russ., abstract in Eng.) DOI: 10.14529/cmse170205

Submitted 15.08.2019; revised 30.10.2019; published online 23.12.2019

About the authors:

Elizaveta A. Dorenskaya, Software Engineer, Alikhanov Institute for Theoretical and Experimental Physics of National Research Center "Kurchatov Institute" (25 Bolshaya Cheremushkinskaya Str., Moscow 117218, Russia), ORCID: http://orcid.org/0000-0002-4249-5131, dorenskaya@ itep.ru

Yuri A. Semenov, Lead Researcher, Alikhanov Institute for Theoretical and Experimental Physics of National Research Center "Kurchatov Institute" (25 Bolshaya Cheremushkinskaya Str., Moscow 117218, Russia); Deputy Head of the Chair for Computer Science, Institute of Nano-, Bio-, Information, Cognitive and Socio-humanistic Sciences and Technologies, Moscow Institute of Physics and Technology (9 Institutskij per., Dolgoprudny 141701, Moscow region, Russia), Ph.D. (Phys.-Math.), ORCID: http://orcid.org/0000-0002- 3855-3650, semenov@itep.ru

All authors have read and approved the final manuscript.

Список использованных источников

[1] *Доренская Е. А., Семенов Ю. А.* Метод определения контекстных значений слов и документов // Современные информационные технологии и ИТ-образование. 2018. Т. 14, № 4. С. 896-902. DOI: 10.25559/SITITO.14.201804.896-902

[2] *Доренская Е. А., Семенов Ю. А.* О технологии программирования, ориентированной на минимизацию программных ошибок // Современные информационные технологии и ИТ-образование. 2017. Т. 13, № 2. С. 50-56. DOI: 10.25559/SITITO.2017.2.226

[3] *Доренская Е. А., Семенов Ю. А.* Новые методы минимизации ошибок в программном обеспечении // CEUR Workshop Proceedings. Материалы 8-й Международной конференции «Распределенные вычисления и GRID-технологии в науке и образовании». 10-14 Сентября 2018 г. Дубна. Т. 2267. 2018. С. 150-154. URL: http://ceur-ws.org/Vol-2267/150-154-paper-27.pdf (дата обращения: 15.08.2019).

[4] *Семенов Ю. А., Овсянников А. П., Овсянникова Т. В.* Разработка банка алгоритмов и основ языка описания проблем с целью минимизации числа программных ошибок // Труды научно-исследовательского института системных исследований Российской академии наук. 2016. Т. 6, № 2. С. 96-100. URL: https://elibrary.ru/item.asp?id=29798446 (дата обращения: 15.08.2019).

[5] *Семенов Ю. А.* IT-экономика в 2016 году и через 10 лет // Экономические стратегии. 2017. Т. 19, № 1. С. 126-135. URL: https://elibrary.ru/item.asp?id=28409521 (дата обращения: 15.08.2019).

[6] *Rishel T., Perkins L. A., Yenduri S., Zand F.* Determining the context of text using augmented latent semantic indexing // Journal of the American Society for Information Science and Technology. 2007. Vol. 58, № 14. Pp. 2197-2204. DOI: 10.1002/asi.20687

[7] *Chen J., Scholz U., Zhou R., Lange M.* LAILAPS-QSM: A RESTful API and JAVA library for semantic query suggestions // PLoS Computational Biology. 2018. Vol. 14, № 3. Article number: e1006058. DOI: 10.1371/journal.pcbi.1006058

[8] *Yang L., Zhang J.* Automatic transfer learning for short text mining // Eurasip Journal on Wireless Communications and Networking. 2017. Vol. 2017, № 1. Article number: 42. DOI: 10.1186/s13638-017-0815-5

[9] *Yan E., Williams J., Chen Z.* Understanding disciplinary vocabularies using a full-text enabled domain-independent term extraction approach // PLoS ONE. 2017. Vol. 12, № 11. Article number: e0187762. DOI: 10.1371/journal.pone.0187762

[10] *Arras L., Horn F., Montavon G., Müller K.-R., Samek W.* What is relevant in a text document?": An interpretable machine learning approach // PLoS ONE. 2017. Vol. 12, № 8. Article number: e0181142. DOI: 10.1371/journal.pone.0181142

[11] *Eidlin A. A., Eidlina M. A., Samsonovich A. V.* Analyzing weak semantic map of word senses // Procedia Computer Science. 2018. Vol. 123. Pp. 140-148. DOI: 10.1016/j.procs.2018.01.023

[12] *Samsonovich A. V.* Weak Semantic Map of the Russian Language: Preliminary Results // Procedia Computer Science. 2016. Vol. 88. Pp. 538-543. DOI: 10.1016/j.procs.2016.08.001

[13] *Wei T., Lu Y., Chang H., Zhou Q., Bao X.* A semantic approach for text clustering using WordNet and lexical chains // Expert Systems with Applications. 2015. Vol. 42, №4. Pp. 2264-2275. DOI: 10.1016/j.eswa.2014.10.023

[14] *Zhan J., Dahal B.* Using deep learning for short text

understanding // Journal of Big Data. 2017. Vol. 4, № 1. Article number: 34. DOI: 10.1186/s40537-017-0095-2

[15] *Khenner E., Nasraoui O.* A bilingual semantic network of computing concepts // Procedia Computer Science. 2016. Vol. 80. Pp. 2392-2396. DOI: 10.1016/j.procs.2016.05.460

[16] *Yu B.* Research on information retrieval model based on ontology // EURASIP Journal on Wireless Communications and Networking. 2019. Vol. 2019, № 1. Article number: 30. DOI: 10.1186/s13638-019-1354-z

[17] *Елькина Е. Е., Кононова О. В., Прокудин Д. Е.* Типология контекстов и принципы контекстного подхода в междисциплинарных научных исследованиях // Современные информационные технологии в ИТ-образовании. 2019. Т. 15, № 1. С. 141-153. DOI: 10.25559/SITITO.15.201901.141-153

[18] *Комраков А. А.* Использование онтологий для описания структуры массивов информационного обмена // Современные информационные технологии в ИТ-образовании. 2019. Т. 15, № 1. С. 182-189. DOI: 10.25559/SITITO.15.201901.182-189

[19] *Барахнин В. Б., Кожемякина О. Ю., Рычкова Е. В., Пастушков И. С., Борзилова Ю. С.* Извлечение лексических и метроритмических признаков, характерных для жанра и стиля и их комбинаций в процессе автоматизированной обработки текстов на русском языке // Современные информационные технологии в ИТ-образовании. 2019. Т. 14, № 4. С. 888-895. DOI: 10.25559/SITITO.14.201804.888-895

[20] *Красовицкий А. М., Уалиева И. М., Мейрамбеккизы Ж., Мусабаев Р. Р.* Основанный на лексиконе подход в оценке обобщений в русскоязычных СМИ // Современные информационные технологии в ИТ-образовании. 2019. Т. 14, № 3. С. 567-572. DOI: 10.25559/SITITO.14.201803.567-572

[21] *Когаловский М. Р., Паринов С. И.* Семантическое аннотирование информационных ресурсов в научной электронной библиотеке средствами таксономий // CEUR Workshop Proceedings. Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10-13 октября 2017 г. 2017. Т. 2022. С. 301-310. URL: http://ceur-ws.org/Vol-2022/paper47.pdf (дата обращения: 15.08.2019).

[22] *Цуканова Ж. В.* Структурные и семантические особенности заголовков современных научных статей (на материале русского и английского языков) // Современные научные исследования и инновации. 2018. № 5. С. 33. URL: https://elibrary.ru/item.asp?id=35050179 (дата обращения: 15.08.2019).

[23] *Чапайкина Н. Е.* Семантический анализ текстов. Основные положения // Молодой ученый. 2012. № 5. С. 112-115. URL: https://elibrary.ru/item.asp?id=20470090 (дата обращения: 15.08.2019).

[24] *Батура Т. В.* Методы и системы семантического анализа текстов // Программные продукты, системы и алгоритмы. 2016, № 4. DOI: 10.15827/2311-6749.21.220

[25] *Бессмертный И. А.* Визуализация знаний на основе семантической сети // Программирование. 2010. Т. 36, № 4. С. 16-24. URL: https://www.elibrary.ru/item.asp?id=15136916 (дата обращения: 15.08.2019).

[26] *Аюшеева Н. Н., Диких А. Ю.* Модель построения семантической сети научного текста // Современные наукоемкие технологии. 2018. № 6. С. 9-13. URL: https://www.elibrary.ru/item.asp?id=35197327 (дата обращения: 15.08.2019).

[27] *Усталов Д. А., Созыкин А. В.* Комплекс программ автоматического построения семантической сети слов // Вестник Южно-Уральского государственного университета. Серия: Вычислительная математика и информатика. 2017. Т. 6, № 2. С. 69-83. DOI: 10.14529/cmse170205

Поступила 15.08.2019; принята к публикации 30.10.2019; опубликована онлайн 23.12.2019.

Об авторах:

Доренская Елизавета Александровна, инженер-программист, Институт теоретической и экспериментальной физики имени А.И. Алиханова Национального исследовательского центра «Курчатовский институт» (117218, Россия, г. Москва, ул. Большая Черемушкинская, д. 25), ORCID: http://orcid.org/0000-0002-4249-5131, dorenskaya@itep.ru

Семёнов Юрий Алексеевич, ведущий научный сотрудник, Институт теоретической и экспериментальной физики имени А.И. Алиханова Национального исследовательского центра «Курчатовский институт» (117218, Россия, г. Москва, ул. Большая Черемушкинская, д. 25); заместитель заведующего кафедрой информатики и вычислительных сетей, Институт нано-, био-, информационных, когнитивных и социогуманитарных наук и технологий, Московский физико-технический институт (национальный исследовательский университет) (141701, Россия, Московская обл., г. Долгопрудный, Институтский пер., д. 9), кандидат физико-математических наук, ORCID: http://orcid.org/0000-0002-3855-3650, semenov@itep.ru

Все авторы прочитали и одобрили окончательный вариант рукописи.