УДК 004.912
DOI: 10.25559/SITITO.14.201804.896-902

МЕТОД ОПРЕДЕЛЕНИЯ КОНТЕКСТНЫХ ЗНАЧЕНИЙ СЛОВ И ДОКУМЕНТОВ

Е.А. Доренская[1], Ю.А. Семенов[1,2]
[1] Институт теоретической и экспериментальной физики имени А.И.Алиханова Национального исследовательского центра «Курчатовский институт», г. Москва, Россия
[2] Московский физико-технический институт (национальный исследовательский университет), г. Долгопрудный, Россия

THE DETERMINATION METHOD FOR CONTEXTUAL MEANINGS OF WORDS AND DOCUMENTS

Elizaveta A. Dorenskaya[1], Yuri A. Semenov[1,2]
[1] Institute for Theoretical and Experimental Physics named by A.I. Alikhanov of National Research Centre «Kurchatov Institute», Moscow, Russia
[2] Moscow Institute of Physics and Technology, Dolgoprudny, Russia

Ключевые слова

Проблема распознавания контекста; контекстное значение; машинный анализ; семантическая сеть; дерево семантических связей; искусственный интеллект; слово-характеристика; метод Монте-Карло.

Аннотация

В данной статье рассматриваются проблемы и методы программного распознавания контекста слов и документов. Даётся краткий обзор существующих методов анализа текстов, рассмотрен простой алгоритм численного определения контекста слов и документов с помощью семантической сети, которая имеет вид графа древовидной формы. Подробно описана структура семантической сети. Данная семантическая сеть необходима для того, чтобы определить контекст корневого слова W1 с помощью, связанных с ним слов-значений W2. Слова W2 представляют собой возможные значения контекста для слова W1. Словам W2 ставятся в соответствие слова-характеристики W3, которые ассоциированы с W2. При расчете контекстного значения учитываются расстояния между словами W2 и W3, измеряемые в словах, размещенных между ними. Словам W3 присваивается метрика, согласно их смысловой близости к тому или иному из слов W2. Приведена таблица слов W1, W2 и W3 и значений метрик. При контекстном анализе текста документа учитываются возможные вариации слов по числам и падежам. Представлена простая формула для расчета контекстного значения слов и документов. Описана методика проверки достоверности контекста с помощью неравенства Чебышева. Проведён анализ полученных результатов моделирования алгоритма с помощью метода Монте Карло, а также способов настройки и оптимизации параметров данного алгоритма. Приведены таблицы результатов исследования предлагаемого метода оценки контекста слов и документов. Исследования показали, что данный метод оценки контекста отдельных слов и документов применим при анализе текстов, при работе с поисковыми системами, а также для других задач, где важно распознавание контекста машинным способом.

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, кандидат физико-математических наук, 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), ORCID: http://orcid.org/0000-0002-3855-3650, semenov@itep.ru

Keywords

The problem of context recognition; contextual meaning; machine analysis; semantic network; tree of semantic links; artificial intelligence; word characteristics; Monte Carlo method.

Abstract

Problems and methods are considered for program context recognition of words and text documents. Survey of existent text processing methods is provided, simple numeric algorithm is given for determination of words and documents context with a help of semantic net, having a form of tree type graph. Semantic net structure is described in detail. Given semantic net is needed to fix basic word W1 context by means of words-meaning W2 coupled with it. Words W2 represent possible W1 context meanings. For every word W2 correspond some words-characteristics W3. At the context calculation the distances between words W2 and W3 are taken into account. The distances are measured in words between. Every word W3 has metrics, according to the concept proximity to W2. There is a table of words W1,W2 and W3 with their metrics values. At context document analyses there was taken into account case or number words variations. Simple formula for context calculation is presented. Method of results proofing with a help of Chebyshev inequality is also provided. The context analyses method was checked by Monte-Carlo simulations. Tables of investigation results are provided and some recommendation for algorithm parameters tuning and optimization are also given. The analyses showed that proposed method is quite effective for context estimation at text analyses, and for any systems, where one needs computer recognition of context.

Введение

В наше время проблема распознавания контекста слов компьютером весьма актуальна. Она важна для поисковых систем, машинного перевода, интерпретации текста при грамматическом разборе и в машинном анализе содержания документов.

Проблема определения контекста слова, на данный момент, относится к AI-полным задачам, требующим сильного искусственного интеллекта. Повышение удобства взаимодействия компьютера и человека в данной области определяет эффективность тех или иных решений.

Благодаря существованию полисемии, одно и то же слово может употребляться в разных значениях. Например, слово «ключ» может иметь значения ключ от замка или ключ родник или криптоключ. Человек может определить контекстное значение слова, анализируя соседние слова в предложении и сам текст в целом.

Одной из причин, почему для описания алгоритма не используется естественный язык, является контекстная многозначность многих слов.

Человек относительно легко определяет контекстные значения слов в тексте. Для решения задачи он использует много критериев, иногда даже достаточно интуитивно

Для распознавания контекста слов с помощью компьютера часто используют семантические сети, онтологии и тезаурусы.

Мы предлагаем упрощённый легко реализуемый метод анализа контекста.

Цель исследования

Главными недостатками существующих методов является сложность их применения, а также то, что они требуют часто больших вычислительных ресурсов [1-8]. Поэтому целью нашего исследования является создание упрощённого метода машинного определения контекстного значения отдельных слов, частей текста и текстовых файлов.

Основная часть

В данном исследовании считалось, что контекстное значение слова зависит от расстояния L между этим словом и другими словами, задающими контекст. Расстояние между словами определяется числом слов N, размещенных между ними (L=N+1). Предполагалось, что контекст конкретного слова можно определить по положению некоторых семантически связанных с ним слов, содержащихся в тексте.

Корневое слово W1 может иметь два или более значений, зависящих от контекста и определяемых словами W2. Слова W2 могут и отсутствовать в тексте документа. Контекстное значение слова W1 в этом случае может определяться семантически связанными с ним словами W3. Варианты семантических сетей показаны на рис.1. Вариант **A** предполагает наличие в тексте документа корневого слова W1, которое может иметь разные контекстные значения, определяемые словами W2. Некоторые слова-значения W2 (например, $W2_2$) могут в документе отсутствовать (рис. 1B). Предполагается, что каждому из слов W2 соответствует некоторое число слов W3 (слова-характеристики), именно они и определяют выбор контекстного значения слова W1. Секция рис. 1C иллюстрирует вариант оценки контекста документа в отсутствии слова W2.

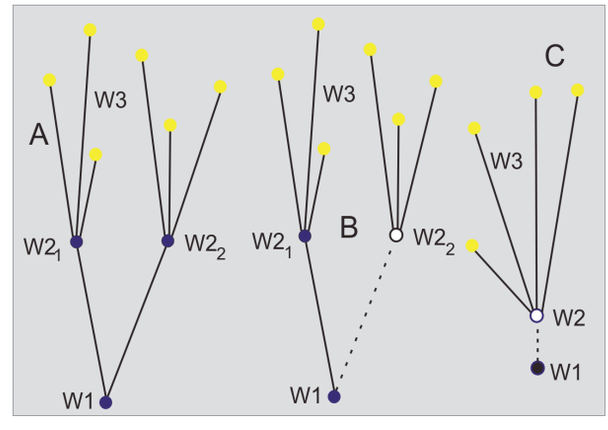

Рис. 1. Варианты семантических связей в тексте

Fig. 1. Variants of semantic links in the text

Рассмотрим это на примере разделения контекстных значений слова «программа»: **компьютер** и **обучение**. W1= программа; $W2_1$ = компьютер; $W2_2$= обучение. Если имеется в виду компьютерная программа, то в тексте могут встретиться слова: *подпрограмма, цикл, файл, библиотека, прерывание, память,*

код, трансляция; цикл; метка; исполнение; исключение; наследование; скрипт; накопитель; синтаксис; присвоение; комментарий; итерация и т.д. Эти слова в таблицу не были включены из-за экономии места. Если имеется в виду программа обучения, в тексте могут встретиться слова: *учитель, лектор, студент, тестирование, ЕГЭ, зачет, экзамен и т.д.* Эти слова также не были включены в таблицу из-за экономии места. (см. таблицу 1). Таблица должна быть создана заранее и никак не зависит от исследуемого текста.

Следует иметь в виду, что слова могут встретиться в разных падежах, числах и пр.

Таблица 1. Фрагмент таблицы корневых слов (W1), слов-значений (W2) и слов-характеристик (W3)
Table 1. Fragment of the table of root words (W1), word-values (W2) and word-characteristics (W3)

Корневое слово W1	Слова-значения W2	Слова-характеристики (W3)	Метрика [M]
Программа	компьютер	программирование	**70**
		отладка	**60**
		тестирование	**40**
		подпрограмма/subroutine	**30**
		объект	**15**
		файл	**26**
		прерывание	**40**
		Оперативная память	**70**
		переменная	**30**
		константа	**20**
		SSD	**30**
		массив/array	**50**
		библиотека (программ)	**15**
		язык (программирования - название)	**60**
	обучение	пособие	**45**
		преподаватель	**50**
		учащийся	**95**
		учебник	**90**
		дистанционное	**70**

В таблицу заносятся только слова, имеющие два или более контекстных значений (W2). Полная таблица даже для отдельной области знаний может быть в сотни раз больше. Содержимое таблицы должно храниться в банке данных, что облегчит доступ к хранящимся в ней словам.

В первой колонке таблицы размешаются слова, которые могут иметь несколько контекстных значений (корневые слова - W1) и могут также определять контекст документа в целом. Во второй колонке (W2) помещаются слова, которые обозначают возможные контекстные значения слов из первой колонки. В третьей колонке (W3) записаны слова, конкретизирующие значения слов из второй и первой колонки. Слова из этих трех колонок образуют древовидный граф. Значения метрики M относятся к словам из третьей колонки таблицы.

Значения метрик может настраиваться с помощью контрольных текстов на стадии отладки системы. Слово в первой колонке является корнем дерева семантических связей. Любое из слов первой колонки (W1), второй -W2 и третьей -(W3) может встретиться в документе больше одного раза. Слово из колонки W1 должно присутствовать в документе обязательно, в противном случае не возникает задачи определения его контекстного значения. Слову из второй колонки, если оно встретилось в документе, присваивается метрика M=100. Но это должно учитываться лишь при определении контекстного значения всего документа. Слово из второй колонки, определяющее контекстное значение слова из первой колонки, может и не встречаться в документе вовсе.

При отсутствии в тексте слова из второй колонки, но при наличии слов из третьей колонки, сопряженных с ним семантически, можно однозначно определить контекстное значение слова из первой колонки (W1).

Можно предположить, что чем ближе слово-характеристика к слову из вышестоящей вершины графа, тем с большей вероятностью оно определяет контекст этого слова. Наличие слова из третьей колонки, размещенного в тексте ближе к слову из второй колонки, должно влиять на выбор контекстного значения слова сильнее, чем в случае слов, размещенных дальше. Одним из возможных методов оценки контекстного значения слова может быть формула [1].

После того как положение слов W1, W2 и W3 определено, производится вычисление суммы **C.**

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

где C – мера, определяющая контекстное значение слова W1, L - расстояние между словом, например, «компьютер» и «отладка» (см. табл. 1), M_i - метрика слова-характеристики W3 (M=1÷100), m - число семантически связанных слов W3 (см. таблицу 1), $f(L_i)$ - весовая функция от L_i, i - номер встретившегося слова из колонки 3. В простейшем случае $f(L_i)=1/L_i$, а для небольших документов $f(L_i)=1$. L определяется числом слов N размещенных между словом W2 и одним из слов W3 (L=N+1). Весовая функция $f(L_i)$ нужна для ослабления влияния удаленных слов на оценку контекстного значения слова W1**.** Если в тексте присутствует две или более копий слова W2, формула [1] может быть модифицирована.

Для больших документов контекст каждого конкретного слова **W1** может оказаться разным для разных областей документа. Размер области может быть настраиваемым, с дискретом в одну страницу (~400 слов). При этом можно варьировать начало и размер области и отслеживать вариации значений C и контекстного значения конкретного слова W1.

Индекс k для **C** определяет, к какому из возможных значений W2 относится данная мера (k=1,..n). смотри вторую колонку таблицы 1. **n** – число возможных значений слова W1 (чаще всего n=2÷3). Значение слова W2 с большим значением **C** в контекстном смысле считается предпочтительным.

Значения M_i выбираются при настройке с использованием тестовых документов.

В таблице 2 представлены данные анализа контекста в конкретных файлах. Расчеты контекста были проведены для более чем 10 файлов. Значения C вычислены по формуле [1]. В скобках приведено число слов W1, W2 и W3, обнаруженных в конкретном документе.

Таблица 2. Примеры результатов контекстного анализа

Table 2. Examples of context analysis results

URL файла	Число слов	Корневые слова (W1)	Слова- значения (W2)	Слова-характе-ристики (W3)	Значения **C**
http://book.itep.ru/4/6/blockchain.htm «Технология blockchain»	5180	Программа (7)	Компьютер (3)	Объект (5) файл (24) код (6)	8,69
			Реализация проекта (9)	Этап (1) Инновация (2)	4,04
			План (0)	Годовой (1)	0,045
http://book.itep.ru/6/i2p.htm **«Стек протоколов I2P и немного о TOR»**	10812	Программа (5)	Компьютер (2)	Метка (30) Объект (7) Файл (5) тестирование (9) код (19) html (13) сайт (6) бит (6)	9,51
			Реализация проекта (16)	Этап (9)	1,58
			План (0)	Обслуживание (1)	0,022
http://book.itep.ru/4/6/set_66.htm **«SET и другие системы осуществления платежей»**	40631	Программа (62)	Компьютер (0)	Объект (33) код (146) бит (14) массив (5) метка (4) переменная (5) исключение (6)	9,12
			Реализация проекта (18)	Этап (14) Стоимость (12)	2,38
			План (0)	Обслуживание (9)	0,059

Если бы для таблицы 1 в семантической цепочке слова “программа” среди слов-характеристик присутствовало слово blockchain (статья «Технология blockchain»), то значение C для слова-значения «компьютер» было бы равно 32,54, а не 8,69. Из этого следует, что полнота семантической сети (таблицы 1) существенно влияет на результаты оценки контекстного значения слова или документа.

Механизм распознавания контекста моделировался по методу Монте-Карло. Предполагалось, что в документе имеется N слов. При моделировании считалось, что положение слов в документе имеет постоянную плотность вероятности (слова размещены в документе статистически равномерно, что не всегда справедливо).

Для анализа в документ засевались случайным образом слова «программа» и слова-характеристики.

На рис. 2 представлено распределение вероятности значений C при фиксированном положении слова «программа» и случайном распределении положений слов-характеристик (n=213) в документе, содержащем 40000 слов.

По вертикальной оси отложено значение вероятности, а по горизонтальной - значение суммы C. Для выявления статистического распределения **C** расчет повторяется 10000 раз. Распределение C имеет гауссоподобную форму, но имеет относительно длинный «хвост» в сторону больших значений C.

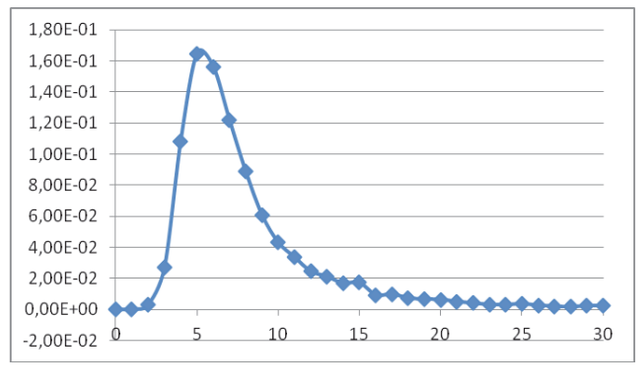

Рис. 2. Распределение плотности вероятности для значения C

Fig. 2. Distribution in frequency for the value of C

Распределение плотности вероятности позволяет оценить эффективность идентификации контекстных значений слов и документов.

Опробовались варианты, где вместо весовой функции $1/L_i$ используется $1/L^2$ или $\exp(-\alpha L)$, где α – постоянный коэффициент <1. Варианты сравнивались по отношению σ/C_{avr}, где C_{avr} - среднее значение C, вычисленное по формуле [1], а σ - средкеквадратичная

ошибка определения C. Зависимость отношения $\sigma/C_{авг}$ от формы весовой функции оказалась слабой. Для определенных классов документов могут использоваться специальные весовые функции, где при малых значениях L весовая функция характеризуется константой, а в области больших L быстро стремится к нулю.

Полученные результаты

На рис. 3 показана зависимость значения C (ромбики) и его среднеквадратичного отклонения (квадратики - σ) от числа слов-характеристик в документе (10÷150). Документ содержал 40000 слов.

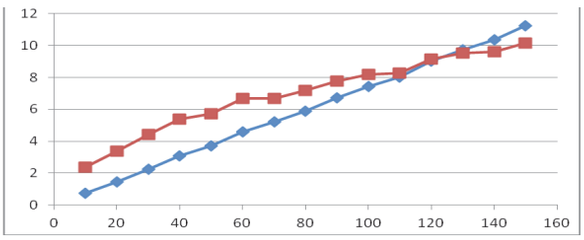

Рис. 3. Зависимость C (ромбики) и σ_C (квадратики) от числа слов-характеристик в документе (10-150)

Fig. 3. Dependence of C (diamonds) and σ_C (squares) on the number of word-characteristics in the document (10-150)

Из рисунка видно, что значение среднеквадратичного отклонения C (σ_C) практически всегда больше C. Для нас важно уметь определить, какова вероятность того, что полученное значение C задает корректно то или иное контекстное значение слова из первой колонки (W1).

Вероятность **p**, например, получения определенного значения C может быть оценена на основе распределения плотности вероятности. Вероятность P получения C=9,12 (см. рис. 2) равна 0,06, при этом вероятность C=2,38 <0,001.

В случае использования неравенства Чебышева [9] имеем:

$p(|x-\overline{C}| \geq \Delta C) \leq (\sigma^2/(\Delta C)^2)$ [2]

Это неравенство определяет верхнюю границу вероятности того, что разность случайной величины x и $\overline{C}$ превышает определенный порог ΔC для произвольного распределения с дисперсией σ^2 и средним значением $\overline{C}$.

Рассмотрим третий пример из таблицы 2. При 62 словах «программа» в документе «SET и другие системы осуществления платежей» можно вычислить значение для слова «компьютер» $\overline{C}$ = 9,12 и σ =14,0. Для слова «реализация» (программы) $\overline{C}$ = 2,38, а σ=4,73.

ΔC = 9,12-2,38 = 6,74 (разница между математическими ожиданиями взятых нами распределений.

Неравенство Чебышева для этого случая имеет вид:

$P(|X-2{,}38| \geq (9{,}12-2{,}38)) \leq 4{,}73^2/(9{,}12-2{,}38)^2$

$P(|X-2{,}38| \geq 6{,}74) \leq 4{,}73^2/6{,}74^2$

Исходя из этого получается что:

$P(|X-2{,}38| \geq 6{,}74) \leq 0{,}49$

Это вполне согласуется с оценкой по плотности вероятностей при моделировании (рис. 2) и подтверждает корректность распознавания контекста. Во всех полутора десятках документов, подвергнутых программному анализу, контекст был определен корректно.

Неравенство Чебышева удобно использовать, когда число слов W1 в документе достаточно велико.

Заключение

Предложенный метод оценки контекстных значений слов и документов нельзя считать универсальным. В нем, в частности, не учитываются смысловые связи. Но предложенный алгоритм легко реализовать, он не требует сложной программной реализации, серьезных вычислительных ресурсов и в большинстве случаев дает правильную оценку значения контекста.

Список использованных источников

[1] *Усталов Д.А.* Модели, методы и алгоритмы построения семантической сети слов для задач обработки естественного языка: дис.....канд. физ.-мат.н. Екатеринбург, 2017, 129 с.

[2] *Бондарчук Д.В.* Определение семантической близости термов с помощью контекстного множества // Компьютерный анализ изображений: Интеллектуальные решения в промышленных сетях (CAI-2016): сборник научных трудов по материалам I Международной конференции 5-6 мая 2016 г./ Под общ. ред. А.Г. Тягунова. Екатеринбург: Изд-во УМЦ УПИ, 2016. С. 175-179. URL: https://elibrary.ru/item.asp?id=28549507 (дата обращения: 12.06.2018).

[3] *Добрынин В.Ю., Клюев В.В., Некрестьянов И.С.* Оценка тематического подобия текстовых документов // Тр. второй всероссийской научной конф. «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции». Протвино, 2000. С. 204-210. URL: http://rcdl.ru/doc/2000/069.pdf (дата обращения: 12.06.2018).

[4] *Ильвовский Д.А.* Модели, алгоритмы и программные комплексы обработки текстовых данных на основе решеток замкнутых описаний: дис. канд. тех.н. Москва, 2014. 158 с.

[5] *Малахов Д.А., Серебряков В.А.* Модель семантического поиска на базе тезауруса // CEUR Workshop Proceedings. 2017. Vol. 2022. С. 191-196. URL: http://ceur-ws.org/Vol-2022/paper32.pdf (дата обращения: 12.06.2018).

[6] *Воронина И.Е., Кретов А.А., Попова И.В.* Алгоритмы определения семантической близости ключевых слов по их окружению в тексте // Вестник ВГУ. Серия: Системный анализ и информационные технологии. 2010. № 1. С. 148-153. URL: https://elibrary.ru/item.asp?id=15199663 (дата обращения: 12.06.2018).

[7] *Крейнес М.Г.* Модели текстов и текстовых коллекций для поиска и анализа информации // Труды Московского физико-технического института. 2017. Том 9, № 3. С. 132-142. URL: https://elibrary.ru/item.asp?id=32736043 (дата обращения: 12.06.2018).

[8] *Турдаков Д.Ю.* Методы и программные средства разрешения лексической многозначности терминов на основе сетей документов: дис.....канд. физ.-мат.н. Москва, 2010. 138 с.

[9] *Прохоров Ю.В., Розанов Ю.А.* Теория вероятностей. Основные понятия, предельные теоремы, случайные процессы. М.: Наука, 1967. 495 с.

[10] *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, issue 14. Pp. 2197-2204. DOI: 10.1002/asi.20687

[11] *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, issue 3. Art. e1006058. DOI: 10.1371/journal.pcbi.1006058

[12] *Yang L., Zhang J.* Automatic transfer learning for short text mining // EURASIP Journal on Wireless Communications and Networking. 2017. Vol. 2017, issue 1:42. 8 p. DOI: 10.1186/s13638-017-0815-5

[13] *Yan E., Williams J., Chen Z.* Understanding disciplinary vocabularies using a full-text enabled domain-independent term extraction approach // PLoS ONE. 2017. Vol. 12, issue 11. Art. e0187762. DOI: 10.1371/journal.pone.0187762

[14] *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, issue 8. Art. e0181142. DOI: 10.1371/journal.pone.0181142

[15] *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

[16] *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

[17] *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, issue 4. Pp. 2264-2275. DOI: 10.1016/j.eswa.2014.10.023

[18] *Zhan J., Dahal B.* Using deep learning for short text understanding // Journal of Big Data. 2017. Vol. 4, issue 34. 15 p. DOI: 10.1186/s40537-017-0095-2

[19] *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

[20] *Батура Т.В.* Семантический анализ и способы представления смысла текста в компьютерной лингвистике // Программные продукты и системы. 2016. № 4. С. 45-57. DOI: 10.15827/0236-235X.116.045-057

[21] *Мозговой М.В.* Машинный семантический анализ русского языка и его применения: дис.....канд. физ.-мат.н. СПб, 2006. 116 с.

[22] *Надеждин Е.Н.* Прикладные задачи семантического анализа текстовых документов // Фундаментальные исследования. 2017. № 1. С. 94-100. URL: https://elibrary.ru/item.asp?id=28307282 (дата обращения: 12.06.2018).

[23] *Боярский К.К.* Введение в компьютерную лингвистику. СПб: НИУ ИТМО, 2013. 72 с. URL: http://books.ifmo.ru/file/pdf/1470.pdf (дата обращения: 12.06.2018).

[24] *Шелманов А.О.* Исследование методов автоматического анализа текстов и разработка интегрированной системы семантико-синтаксического анализа: дис.....канд. тех.н. Москва, 2015. 182 с.

[25] *Батура Т.В.* Математическая лингвистика и автоматическая обработка текстов на естественном языке. Новосибирск: РИЦ НГУ, 2016. 166 с. URL: https://www.iis.nsk.su/files/book/file/Batura_Matlingvistika_i_avtomat._obrabotka_tekstov.pdf (дата обращения: 12.06.2018).

[26] *Марченко А.А., Никоненко А.А.* Контекстный семантический анализ текста. Система текстового мониторинга и качественного оценивания фокусного объекта // Искусственный интеллект. 2008. № 3. С. 808-813. URL: http://dspace.nbuv.gov.ua/bitstream/handle/123456789/7155/02-Marchenko.pdf?sequence=1 (дата обращения: 12.06.2018).

[27] *Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С.* Автоматическая обработка текстов на естественном языке и анализ данных. Москва: Изд-во НИУ ВШЭ, 2017. 269 с. URL: https://www.hse.ru/data/2017/08/12/1174382135/NLP_and_DA.pdf (дата обращения: 12.06.2018).

[28] *Орлова Ю.А.* Автоматизация семантического анализа текста технического задания: дис.....канд. тех.н. Волгоград, 2008. 228 с. URL: https://elibrary.ru/item.asp?id=16191917 (дата обращения: 12.06.2018).

[29] *Святогор Л., Гладун В.* Семантический анализ текстов естественного языка: цели и средства // International Book Series «Information Science and Computing». Knowledge – Dialogue – Solution. Supplement to International Journal «Information Technologies and Knowledge». 2009. Vol. 3. С. 9-18. URL: http://www.foibg.com/ibs_isc/ibs-15/ibs-15-p01.pdf (дата обращения: 12.06.2018).

Поступила 12.06.2018; принята в печать 10.09.2018;
опубликована онлайн 10.12.2018.

References

[1] Ustalov D.A. Modeli, metody i algoritmy postroeniya semanticheskoj seti slov dlya zadach obrabotki estestvennogo yazyka. Diss. kand. fiz.-mat. nauk [Models, methods and algorithms for constructing a semantic network of words for natural language processing problems]. Ekaterinburg, 2017. 129 p. (In Russian)

[2] Bondarchuk D.V. Calculating the semantic relatedness of terms with the context set. *Proceedings of the First Computer Image Analysis: Intelligent Solutions in Industrial Networks.* Ekaterinburg. 2016, pp. 175-179. Available at: https://elibrary.ru/item.asp?id=28549507 (accessed 12.06.2018). (In Russian)

[3] Dobrynin V.Yu., Klyuev B.B., Nekrestyanov I.S. Evaluation of the thematic similarity of text documents. *Digital Libraries: Advanced Methods and Technologies*. Protvino, 2000, pp. 204-210. Available at: http://rcdl.ru/doc/2000/069.pdf (accessed 12.06.2018). (In Russian)

[4] Ilvovski D.A. Modeli, algoritmy i programmnye kompleksy obrabotki tekstovyh dannyh na osnove reshetok zamknutyh opisanij. Diss. kand. tekh. nauk [Models, algorithms and software systems for processing text data based on lattices of closed descriptions]. Moscow, 2014. 158 p. (In Russian)

[5] Malakhov D.A., Serebryakov V.A. The Semantic Search Model Based on the Thesaurus. *CEUR Workshop Proceedings*. 2017; 2022:191-196. Available at: http://ceur-ws.org/Vol-2022/paper32.pdf (accessed 12.06.2018). (In Russian)

[6] Voronina E.I., Kretov A.A., Popova I.V. Algorithms of semantic proximity assessment based on the lexical environment of the keywords in a text. *Proceedings of Voronezh State University. Series: Systems analysis and information technologies*. 2010; 1:148-153. Available at: https://elibrary.ru/item.asp?id=15199663 (accessed 12.06.2018). (In Russian)

[7] Kreines M.G. Text and text corpora models for information retrieval and analysis. *Proceedings of MIPT.* 2017; 9(3):132-142. Available at: https://elibrary.ru/item.asp?id=32736043 (accessed 12.06.2018). (In Russian)

[8] Turdakov D.Y. Metody i programmnye sredstva razresheniya leksicheskoj mnogoznachnosti terminov na osnove setej dokumentov. Diss. kand. fiz.-mat. nauk [Methods and software tools for the resolution of lexical ambiguity of terms based on networks of documents]. Moscow, 2010. 138 p. (In Russian)

[9] Prokhorov U.V., Rozanov U.A. Teoriya veroyatnostey. Osnovnye ponyatiya, predel'nye teoremy, sluchajnye process [Theory of probabilities. Basic concepts, limit theorems, random processes]. 2nd ed. Moscow: Nauka, 1973. 494 p. (In Russian)

[10] 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. DOI: 10.1002/asi.20687

[11] 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. DOI: 10.1371/journal.pcbi.1006058

[12] Yang L., Zhang J. Automatic transfer learning for short text mining. *EURASIP Journal on Wireless Communications and Networking*. 2017; 2017(1):42. 8 p. DOI: 10.1186/s13638-017-0815-5

[13] 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. DOI: 10.1371/journal.pone.0187762

[14] 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. DOI: 10.1371/journal.pone.0181142

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

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

[17] 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. DOI: 10.1016/j.eswa.2014.10.023

[18] Zhan J., Dahal B. Using deep learning for short text understanding. *Journal of Big Data*. 2017; 4(34). 15 p. DOI: 10.1186/s40537-017-0095-2

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

[20] Batura T.V. Semantic analysis and methods of text meaning representation in computer linguistics. *Programmnye produkty i sistemy = Software & Systems*. 2016; 4:45-57. (In Russian) DOI: 10.15827/0236-235X.116.045-057

[21] Mozgovoy M.V. Mashinnyj semanticheskij analiz russkogo yazyka i ego primeneniya. Diss. kand. fiz.-mat. nauk [Machine semantic analysis and its applications of the Russian language]. St. Petersburg, 2006. 116 p. (In Russian)

[22] Nadezhdin E.N. Applied problems of semantic analysis of text documents. *Fundamental research*. 2017; 1:94-100. Available at: https://elibrary.ru/item.asp?id=28307282 (accessed 12.06.2018). (In Russian)

[23] Boyarsky K.K. Introduction to Computer Linguistics. St. Petersburg, ITMO University. 2013, 73 p. Available at: http://books.ifmo.ru/file/pdf/1470.pdf (accessed 12.06.2018). (In Russian)

[24] Shelmanov A.O. Issledovanie metodov avtomaticheskogo analiza tekstov i razrabotka integrirovannoj sistemy semantiko-sintaksicheskogo analiza. Diss. kand. tekh. nauk [Research of methods of automatic text analysis and development of an integrated system of semantic and syntactic analysis]. Moscow, 2015. 182 p. (In Russian)

[25] Batura T.V. Mathematical linguistics and automatic processing of natural language texts. Novosibirsk, NSU, 2016. 166 p. Available at: https://www.iis.nsk.su/files/book/file/Batura_Matlingvistika_i_avtomat._obrabotka_tekstov.pdf (accessed 12.06.2018). (In Russian)

[26] Marchenko O.O., Nikonenko A.A. The Contextual Semantic Analysis of Natural Language Text. System of Text Monitoring and Qualitative Estimation of the Focus Object. *Artificial intelligence*. 2008; 3:808-813. Available at: http://dspace.nbuv.gov.ua/bitstream/handle/123456789/7155/02-Marchenko.pdf?sequence=1 (accessed 12.06.2018). (In Russian)

[27] Bolshakova E.I., Vorontsov K.V., Efremova N.E., Klyshinsky E.S., Lukashevich N.V. Sayapin A.S. Automatic text processing in natural language and data analysis. Moscow, HSE, 2017. 269 p. Available at: https://www.hse.ru/data/2017/08/12 /1174382135/NLP_and_DA.pdf (accessed 12.06.2018). (In Russian)

[28] Orlova Yu.A. Avtomatizaciya semanticheskogo analiza teksta tekhnicheskogo zadaniya. Diss. kand. tekh. nauk [Automation of semantic analysis of the text of the technical task]. Volgograd, 2008. 228 p. Available at: https://elibrary.ru/item.asp?id=16191917 (accessed 12.06.2018). (In Russian)

[29] Svyatogor L., Gladun V. Semantic analysis of natural language texts: goals and instruments International. *Book Series «Information Science and Computing». Knowledge – Dialogue – Solution. Supplement to International Journal «Information Technologies and Knowledge».* 2009; 3:9-18. Available at: http://www.foibg.com/ibs_isc/ibs-15/ibs-15-p01.pdf (accessed 12.06.2018). (In Russian)

Submitted 12.06.2018; revised 10.09.2018;
published online 10.12.2018.

About the authors:

Elizaveta A. Dorenskaya, software engineer, Institute for Theoretical and Experimental Physics named by A.I. Alikhanov of National Research Centre «Kurchatov Institute» (25 Bolshaya Cheremushkinskaya Str., Moscow 117218, Russia), ORCID: http://orcid.org/0000-0002-4249-5131, dorenskaya@itep.ru

Yuri A. Semenov, Candidate of Physical and Mathematical Sciences, Lead Researcher, Institute for Theoretical and Experimental Physics named by A.I. Alikhanov of National Research Centre «Kurchatov Institute» (25 Bolshaya Cheremushkinskaya Str., Moscow 117218, Russia); Deputy Head of the Department of Informatics and Computer Networks, 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), ORCID: http://orcid.org/0000-0002-3855-3650, semenov@itep.ru