УДК 004.912

Доренская Е.А., Семенов Ю.А.

НИЦ "Курчатовский институт" – ИТЭФ, Москва, Россия

МЕТОД ОПРЕДЕЛЕНИЯ КОНТЕКСТНЫХ ЗНАЧЕНИЙ СЛОВ И ДОКУМЕНТОВ

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.

КЛЮЧЕВЫЕ СЛОВА

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

Dorenskaya E. A., Semenov Y. A.

NRC "Kurchatov Institute" – ITEP, Moscow, Russia

THE DETERMINATION METHOD FOR CONTEXTUAL MEANINGS OF WORDS AND DOCUMENTS

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 a computer recognition of context.

KEYWORDS

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

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.

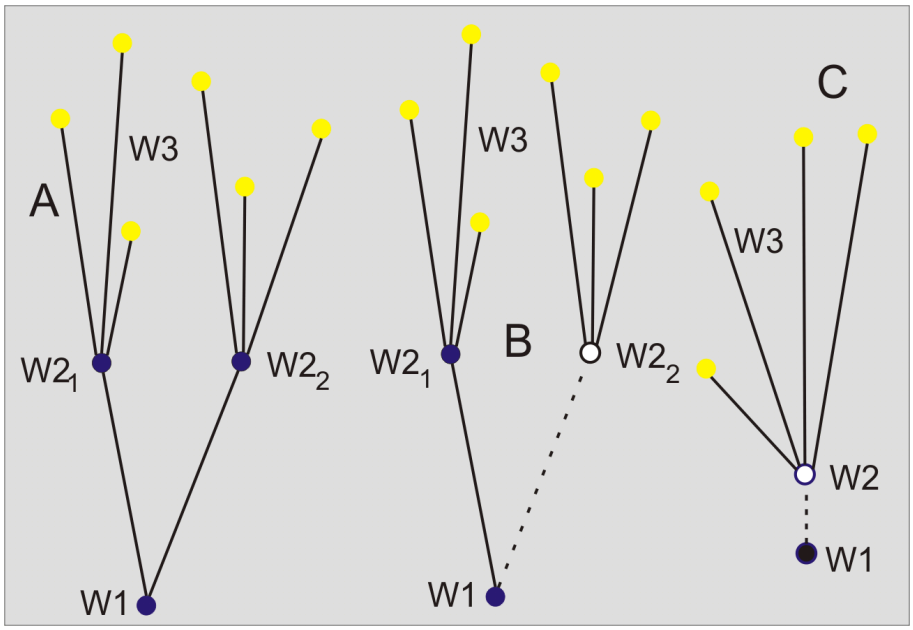

Рис.1. Варианты семантических связей в тексте

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)

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 $\mathbf{L_i}$, 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

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.

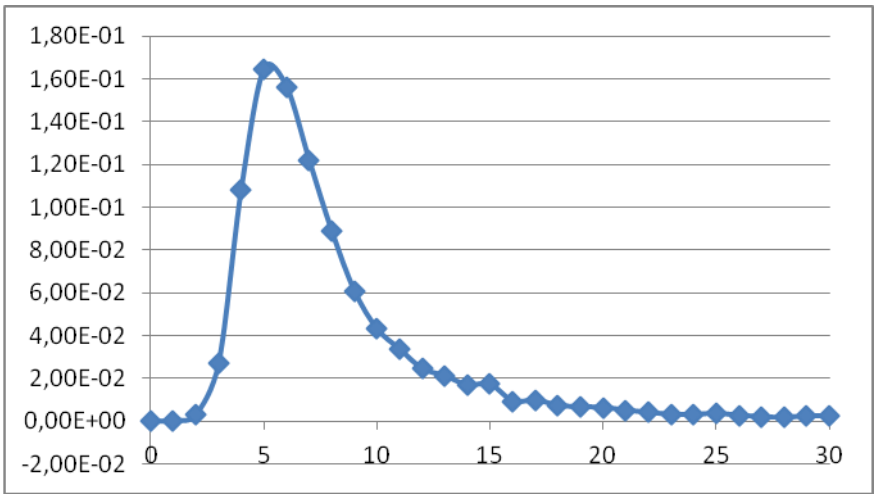

Рис. 2. Распределение плотности вероятности для значения C

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 σ/C_{avr} 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.

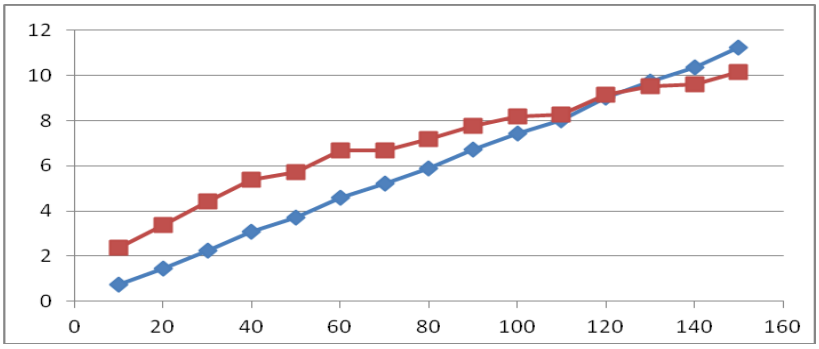

Рис. 3. Зависимость C (ромбики) и σ_C (квадратики) от числа слов-характеристик в документе (10-150)

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:

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

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.

$\Delta C = 9{,}12\text{-}2{,}38 = 6{,}74$ (разница между математическими ожиданиями взятых нами распределений.

Неравенство Чебышева для этого случая имеет вид:

$$P(|X\text{-}2{,}38| \geq (9{,}12\text{-}2{,}38)) \leq 4{,}73^2/(9{,}12\text{-}2{,}38)^2$$
$$P(|X\text{-}2{,}38| \geq 6{,}74) \leq 4{,}73^2/6{,}74^2$$

Исходя из этого получается что:

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

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.

Литература

1. Усталов Д.А. Модели, методы и алгоритмы построения семантической сети слов для задач обработки естественного языка // диссертация на соискание уч. степени кфмн, Институт математики и механики им. Н.Н.Красовского, УО РАН. 2017, 129 стр. http://www.susu.ru/sites/default/files/dissertation/dissertation_0.pdf
2. Бондарчук Д.В. Определение семантической близости термов с помощью контекстного множества // Уральский государственный университет путей сообщения, Екатеринбург. 2016, стр. 175-179 http://elar.urfu.ru/bitstream/10995/43751/1/cai-2016-41.pdf
3. Добрынин В.Ю., Клюев В.В., Некрестьянов И.С. Оценка тематического подобия текстовых документов // СПбГУ. 2000, стр. 204-210 http://web.ihep.su/library/pubs/aconf00/dconf00/ps/069.pdf
4. Ильвовский Д.А. Модели, алгоритмы и программные комплексы обработки текстовых данных на основе решеток замкнутых описаний // диссертация на соискание уч. степени ктн, НИУ ВШЭ, Москва. 2014, 158 стр. https://www.hse.ru/data/2014/09/24/1315819304/dis.pdf
5. Малахов Д.А., Серебряков В.А. Модель семантического поиска на базе тезауруса // МГУ. 2017 , стр. 191-196 http://ceur-ws.org/Vol-2022/paper32.pdf
6. Воронина И.Е., Кретов А.А., Попова И.В. Алгоритмы определения семантической близости ключевых слов по их окружению в тексте // Воронежский Государственный Университет. 2010, стр. 148-153 http://www.vestnik.vsu.ru/pdf/analiz/2010/01/2010-01-25.pdf
7. Крейнес М.Г., Модели текстов и текстовых коллекций для поиска и анализа информации // труды МФТИ, 2017, том 9, №3. стр. 132-142 https://mipt.ru/upload/medialibrary/067/16_kreines_132_142.pdf
8. Турдаков Д.Ю. Методы и программные средства разрешения лексической многозначности терминов на основе сетей документов // диссертация на соискание уч. степени кфмн, МГУ. 2010, 138 стр. http://www.ispras.ru/upload/iblock/3ea/3ea4b70757395519f5799222c1189fe9.pdf
9. Прохоров Ю.В., Розанов Ю.А., Теория вероятностей. Основные понятия, предельные теоремы, случайные процессы", изд. "Наука", 1967, 495 стр.
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, №. 14. Pp. 2197-2204. DOI: https://doi.org/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, №. 3 : e1006058. DOI: https://doi.org/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, №1:42 DOI: https://doi.org/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, №11: e0187762 DOI: https://doi.org/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, №8: e0181142 DOI: https://doi.org/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: https://doi.org/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: https://doi.org/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, №4. Pp 2264-2275 DOI: https://doi.org/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, №1:34 DOI: https://doi.org/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: https://doi.org/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. стр. 45-57 DOI: https://doi.org/10.15827/0236-235X.116.045-057
21. Мозговой М.В. Машинный семантический анализ русского языка и его применения // диссертация на соискание уч. степени кфмн, Санкт-Петербургский Государственный Университет. 2006 , 116 стр. http://web-ext.u-aizu.ac.jp/~mozgovoy/homepage/papers/amcp_dissertation.pdf
22.Надеждин Е.Н. Прикладные задачи семантического анализа текстовых документов // Фундаментальные исследования. 2017, № 1. стр 94-100 https://fundamental-research.ru/ru/article/view?id=41321
23. Боярский К.К. Введение в компьютерную лингвистику // учебное пособие, Университет ИТМО Санкт-Петербург. 2013, 73 стр. http://books.ifmo.ru/file/pdf/1470.pdf
24. Шелманов А.О. Исследование методов автоматического анализа текстов и разработка интегрированной системы семантико-синтаксического анализа // диссертация на соискание уч. степени ктн, Институт системного анализа ФИЦ ИУ РАН, Москва. 2015,182 стр. http://www.ipiran.ru/announce/dissertation_Shelmanov.pdf
25. Батура Т.В. Математическая лингвистика и автоматическая обработка текстов на естественном языке // учебное пособие, НГУ Новосибирск. 2016, 166 стр. https://www.iis.nsk.su/files/book/file/Batura_Matlingvistika_i_avtomat._obrabotka_tekstov.pdf
26. Марченко А.А., Никоненко А.А., Контекстный семантический анализ текста. Система текстового мониторинга и качественного оценивания фокусного объекта // Искусственный интеллект, Киевский национальный университет имени Тараса Шевченко, г. Киев, Украина. 2008, №3 стр. 808-813 http://dspace.nbuv.gov.ua/bitstream/handle/123456789/7155/02-Marchenko.pdf?sequence=1
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 // учебное пособие, НИУ ВШЭ, Москва. 2017, 269 стр. https://www.hse.ru/data/2017/08/12/1174382135/NLP_and_DA.pdf
28. Орлова Ю.А. Автоматизация семантического анализа текста технического задания // диссертация на соискание уч. степени ктн, ВГТУ, Волгоград. 2008, 190 стр.
29. Святогор Л., Гладун В. Семантический анализ текстов естественного языка: цели и средства // International Book Series "Information Science and Computing". 2009, стр. 9-18 http://www.foibg.com/ibs_isc/ibs-15/ibs-15-p01.pdf

References

1. Ustalov D. A. Models, methods and algorithms for constructing a semantic network of words for natural language processing problems // dissertation for the degree of Ph.D, Institute of mathematics and mechanics. N. N. Krasovsky, UO Russian Academy of Sciences. 2017, 129 p. http://www.susu.ru/sites/default/files/dissertation/dissertation_0.pdf
2. Bondarchuk D. V. determination of semantic proximity of terms by means of context set //Ural state University of railway engineering, Ekaterinburg. 2016, Pp. 175-179. http://elar.urfu.ru/bitstream/10995/43751/1/cai-2016-41.pdf
3. Dobrynin V.Yu., Klyuev B.B., Nekrestyanov I.S. Evaluation of the thematic similarity of text documents // SPbSU. 2000, Pp. 204-210 http://web.ihep.su/library/pubs/aconf00/dconf00/ps/069.pdf
4. Ilovski D. A. Models, algorithms and software systems for processing text data based on lattices of closed descriptions // dissertation for the degree of Ph.D, HSE, Moscow. 2014, 158 p. https://www.hse.ru/data/2014/09/24/1315819304/dis.pdf
5. Malakhov D. A., Serebryakov V. A. model of semantic search based on the thesaurus // Moscow state University. 2017, Pp. 191-196 http://ceur-ws.org/Vol-2022/paper32.pdf
6. Voronina E. I., Kretov A. A., Popova I. V. Algorithms for determining the semantic proximity of key words in their setting in the text // of Voronezh State University. 2010, Pp. 148-153

http://www.vestnik.vsu.ru/pdf/analiz/2010/01/2010-01-25.pdf
7. Kreines M. G., models of texts and text collections for information search and analysis // proceedings of MIPT, 2017, volume 9, №3. Pp. 132-142 https://mipt.ru/upload/medialibrary/067/16_kreines_132_142.pdf
8. Turdakov D. Y. Methods and software tools for the resolution of lexical ambiguity of terms based on networks of documents // dissertation for the degree of Ph.D, Moscow state University. 2010, 138 p. http://www.ispras.ru/upload/iblock/3ea/3ea4b70757395519f5799222c1189fe9.pdf
9. Prohorov U. V., Rozanov U. A. Theory of probability. Basic concepts, limit theorems, random processes", ed. "Science", 1967, , 495 p.
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, №. 14. Pp. 2197-2204. DOI: https://doi.org/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, №. 3 : e1006058. DOI: https://doi.org/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, №1:42 DOI: https://doi.org/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, №11: e0187762 DOI: https://doi.org/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, №8: e0181142 DOI: https://doi.org/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: https://doi.org/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: https://doi.org/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, №4. Pp 2264-2275 DOI: https://doi.org/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, №1:34 DOI: https://doi.org/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: https://doi.org/10.1016/j.procs.2016.05.460
20. Batura T. V., Semantic analysis and ways of representing the meaning of the text in computer linguistics // Software products and systems, 2016 №4. Pp.45-57 DOI: https://doi.org/10.15827/0236-235X.116.045-057
21. Mozgovoy M.V. Machine semantic analysis and its applications of the Russian language // dissertation for the degree of Ph.D, St. Petersburg state University. 2006, 116 p .http://web-ext.u-aizu.ac.jp/~mozgovoy/homepage/papers/amcp_dissertation.pdf
22. Nadezhdin E. N. Applied problems of semantic analysis of text documents // Fundamental research. 2017, № 1. Pp 94-100 https://fundamental-research.ru/ru/article/view?id=41321
23. Boyarsky K. K. Introduction to computer linguistics // textbook, ITMO University St. Petersburg. 2013, 73 p. http://books.ifmo.ru/file/pdf/1470.pdf
24. Shelmanov A. O. Research of methods of automatic text analysis and development of an integrated system of semantic and syntactic analysis // dissertation for the degree of Ph.D, Institute of system analysis, Russian Academy of Sciences, Moscow. 2015,182 p. http://www.ipiran.ru/announce/dissertation_Shelmanov.pdf
25. Batura T. V., Mathematical linguistics and automatic processing of natural language texts // teaching aid, NSU, Novosibirsk. 2016, 166 p. https://www.iis.nsk.su/files/book/file/Batura_Matlingvistika_i_avtomat._obrabotka_tekstov.pdf
26. Marchenko A. A., Nikonenko A. A., Contextual semantic analysis of the text. System of text monitoring and qualitative assessment of the focal object // Artificial intelligence, Taras Shevchenko national University of Kiev, Ukraine. 2008, №3 Pp. 808-813http://dspace.nbuv.gov.ua/bitstream/handle/123456789/7155/02-Marchenko.pdf?sequence=1
27. Bolshakova E. I., Vorontsov K. V., Efremova N. E., Klysinski E. S., Lukashevich N. V. Sayapin, A. S. Automatic text processing in natural language and data analysis // teaching aid, HSE, Moscow. 2017, 269 p. https://www.hse.ru/data/2017/08/12/1174382135/NLP_and_DA.pdf
28. Orlova Yu.A. automation of semantic analysis of the text of the technical task // dissertation for the degree of Ph.D, VSTU, Volgograd. 2008, 190 p.

29. Svyatogor L., Gladun V. Semantic analysis of natural language texts: goals and instruments // International Book Series "Information Science and Computing". 2009, Pp. 9-18 http://www.foibg.com/ibs_isc/ibs-15/ibs-15-p01.pdf

Об авторах:

Доренская Елизавета Александровна, инженер-программист ИТЭФ, dorenskaya@itep.ru; тел. 8-916-239-97-91
Семёнов Юрий Алексеевич, ведущий научный сотрудник ИТЭФ, заместитель заведующего кафедрой «Информатики» ФНБИК, МФТИ, кандидат физико-математических наук, semenov@itep.ru; 8-916-962-67-71