УНИВЕРСИТЕТ ИТМО
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
март–апрель 2022 Том 22 № 2 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
March–April 2022 Vol. 22 No 2 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК УНИВЕРСИТЕТ ИТМО
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2022-22-2-287-293
УДК 004.9

Классификация коротких текстов с использованием волновой модели

Анастасия Сергеевна Груздева[1]✉, Игорь Александрович Бессмертный[2]

[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] prog.anastasia@gmail.com✉, https://orcid.org/0000-0003-4963-0823
[2] bessmertny@itmo.ru, https://orcid.org/0000-0001-6711-6399

Аннотация

Предмет исследования. Алгоритмы квантовых вычислений активно развиваются и применяются в области обработки естественного языка. В работе предложен новый квантово-подобный метод классификации коротких текстов. **Метод.** Основу метода составляет представление текста в виде ансамбля элементарных частиц. В качестве критерия классификации выбрано значение амплитуды вероятности обнаружения данного ансамбля в выбранных точках векторного пространства, описываемого при помощи дистрибутивно-семантической модели языка. Предложен один из возможных способов интерпретации параметров волновой функции описания поведения элементарной частицы, а также алгоритм расчета амплитуды вероятности с учетом этих параметров. **Основные результаты.** Выполнена экспериментальная проверка описанного метода с применением классификации интернет-сообществ по тематикам. Для расчетов использованы наименования и сведения разделов «информация» по 100 группам социальной сети «ВКонтакте» по пяти различным темам. Предложенная модель показала достаточно высокую точность классификации, которая составила 91 % в целом на наборе данных и от 75 % до 95 % в пределах отдельных классов. **Практическая значимость.** Представленная модель может быть использована для классификации отзывов пользователей о товарах, услугах и событиях, а также при определении некоторых свойств психологических портретов пользователей интернет-сообществ.

Ключевые слова

классификация, обработка естественного языка, волновая модель, интерференция, квантово-подобная модель, определение тематики текста

Благодарности

Работа выполнена в рамках магистерско-аспирантской НИР № 620164 «Методы искусственного интеллекта для киберфизических систем».

Ссылка для цитирования: Груздева А.С., Бессмертный И.А. Классификация коротких текстов с использованием волновой модели // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 2. С. 287–293. doi: 10.17586/2226-1494-2022-22-2-287-293

Classification of short texts using a wave model

Anastasia S. Gruzdeva[1]✉, Igor A. Bessmertny[2]

[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] prog.anastasia@gmail.com✉, https://orcid.org/0000-0003-4963-0823
[2] bessmertny@itmo.ru, https://orcid.org/0000-0001-6711-6399

Abstract

Quantum computing algorithms are actively developed and applied in the field of natural language processing. The authors of the paper proposed a new quantum-like method for classifying short texts. The basis of the method is the representation of the text as an ensemble of elementary particles. The value of the detection probability amplitude of a given ensemble at the selected points in space is chosen as a classification criterion. In this case, the space is understood as a vector space described using the distributive-semantic model of the language. The authors suggested one of the possible ways of interpreting the parameters of the wave function that describes the behavior of an elementary particle, as

well as an algorithm for calculating the probability amplitude taking into account these parameters. For the experimental research of the described method, authors performed the classification of Internet communities by topics. For the analysis, the names and the "information" section of communities were used. In total, 100 groups of the social network "VKontakte" belonging to five various topics were taken. The proposed model showed rather high classification accuracy (91 % in general on the data set and from 75 % to 95 % within individual classes). The proposed model is supposed to be used to classify user comments about goods, services and events, as well as to determine some properties of the psychological portraits of users of online communities.

Keywords
classification, natural language processing, wave model, interference, quantum-like model, definition of the text subject

Aknowledgements
The work was carried out within the framework of the project No. 620164 (artificial intelligence methods for cyber-physical systems).

For citation: Gruzdeva A.S., Bessmertny I.A. Classification of short texts using a wave model. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 2, pp. 287–293 (in Russian). doi: 10.17586/2226-1494-2022-22-2-287-293

Введение

Квантово-подобные вычислительные методы [1] в последние десятилетия активно развиваются и достигают значительных успехов, особенно в области поиска и анализа текстовой информации [2]. При этом квантово-механический подход не только применяется самостоятельно, но и позволяет получить хорошие результаты во взаимодействии с классическими методами [3, 4]. Однако далеко не все возможности, предоставляемые математическим аппаратом квантовой механики, активно применяются в настоящее время в сфере обработки естественного языка [5], что оставляет широкие возможности для разработки новых моделей и совершенствования существующих. В данной работе предложена новая квантово-подобная модель, предназначенная для обработки и анализа текстовой информации. Модель является дополнением к классической дистрибутивно-семантической языковой модели. Предложенный подход базируется на представлении текста в виде волнового пакета. При этом на расчет важной волновой характеристики каждого элемента, а именно волнового числа, оказывают влияние все слова текста. Таким образом, волновая модель может рассматриваться как один из вариантов описания взаимосвязи слов в тексте. Кроме того, волновой подход предоставляет принципиальную возможность наблюдать такое явление как интерференция.

Разрабатываемую модель предполагается использовать в первую очередь для классификации коротких текстов. В настоящее время такие задачи являются весьма востребованными. Во-первых, данный алгоритм может применяться для структуризации и многоуровневой классификации отзывов пользователей о товарах, услугах, явлениях и событиях, что может служить расширением возможностей известных методов сентимент-анализа [6–8]. Другое потенциальное применение — определение тематик интернет-сообществ, что дает возможность заочного выявления сферы интересов пользователей и может быть интересно специалистам в сфере управления кадрами, а также сотрудникам психологических служб школ и других учебных заведений. Разрабатываемая модель станет дополнением к другим работам в области методик построения психологического портрета автора на базе анализа текста [9, 10].

Волновая модель представления текста

Для анализа текстовой информации рассмотрим волновую модель, в рамках которой текст представляется в виде ансамбля элементарных частиц. В таком ансамбле каждое слово текста, относящееся к самостоятельным частям речи, представлено как отдельная частица. В соответствии с основами квантовой механики [11], а именно с принципом корпускулярно-волнового дуализма, поведение такой частицы может быть описано при помощи волновой функции:

$$\psi_j = \frac{A_j}{r_{lj}} e^{-i(k_j v_j t - k_j r_{lj} + \varphi_{j0})}, \quad (1)$$

где A_j — амплитуда; r_{lj} — расстояние; k_j — волновое число; v_j — скорость распространения волны; φ_{j0} — начальная фаза.

Волновая функция представляет собой сферическую волну, интенсивность которой отражает вероятность обнаружения частицы в различных точках пространства в различные моменты времени.

Текст, рассматриваемый как ансамбль элементарных частиц, в то же время может быть представлен как волновой пакет, состоящий из конечного числа сферических волн, причем суммарная интенсивность такого пакета в точке пространства l отражает амплитуду вероятности обнаружения ансамбля частиц в данной точке. Для расчета интенсивности пакета используется уравнение, известное из курса волновой механики [11]

$$I_l = \sum_{j=1}^{M} \left(\frac{A_j}{r_{lj}}\right)^2 + 2 \sum_{j=1}^{M-1} \sum_{n=j+1}^{M} \frac{A_j}{r_{lj}} \frac{A_n}{r_{ln}} \cos(k_j r_{lj} - k_n r_{ln} + \varphi_{j0} - \varphi_{n0}), \quad (2)$$

где M — количество волн в пакете, что соответствует числу слов, относящихся к самостоятельным частям речи в исходном тексте.

При этом второе слагаемое уравнения, представляющее собой удвоенную сумму попарных произведений элементов, отвечает за интерференцию, которая может наблюдаться при определенных комбинациях волновых чисел, начальных фаз и расстояний от источника до точки наблюдения различных волн, представляющих отдельные слова текста. Учет интерференционного

члена дает возможность, одним словом, усиливать или ослаблять влияние других слов на расчет вероятности близости текста к выбранному классу.

В качестве базиса для работы волновой модели используется предобученная дистрибутивно-семантическая модель [12], в которой слова представляют собой векторы в пространстве контекстов, а расстояния между ними отражают семантическую близость между понятиями. Таким образом, высокая амплитуда вероятности обнаружения ансамбля частиц, представляющего текст, в некоторой точке такого пространства, может говорить о высокой смысловой близости данного текста к понятию, занимающему эту точку пространства. Если есть несколько точек пространства, для которых может быть вычислена соответствующая амплитуда вероятности (2), то можно предположить, что исходный текст будет ближе всего по смыслу к тому понятию, в области которого амплитуда вероятности выше. Именно такой принцип лежит в основе классификации текстов с использованием волновой модели.

Рассмотрим интерпретацию параметров волновой функции для случая представления текстовой информации. Расстояния r_{lj} и r_{ln} в уравнениях (1) и (2) отражают семантическую близость между понятиями — являются величинами, обратно пропорциональными близости. Для расчета волновых чисел k_j, k_n используются следующие соображения. У ансамбля частиц, представляющего текст, может быть найден центроид, который можно рассматривать как точку, в которой вероятность обнаружения ансамбля в целом максимальна. Следовательно, в указанной точке амплитуда вероятности, рассчитанная при помощи уравнения (2), должна быть максимальной, что может быть достигнуто, если на пути от каждого элемента ансамбля до центроида будет находиться целое количество длин волн (в простейшем случае — одна длина волны). Такие предположения дают возможность рассчитать волновое число, которое представляет собой количество длин волн в единице длины. Таким образом, для вычисления волновых чисел используется уравнение

$$k_j = \frac{1}{r_{cj}},$$

где r_{cj} — расстояние от термина j до центроида ансамбля.

Вычисление начальных фаз φ_{j0}, φ_{n0} представляет собой большую сложность. Возможны разные подходы к расчету данной величины. С одной стороны, в фазе соответствующей волны может быть учтена форма слова в тексте. При построении математической волновой модели значащие слова текста преобразуются к начальной форме, при этом теряется исходная форма слова. Возможно, данная информация могла бы найти отражение в фазе волны. С другой стороны, на начальную фазу волны может влиять ближайшее окружение слова, а именно служебные части речи и знаки препинания. Проще всего понять эту идею на примере частицы «не». Например, сравнив выражения «хлеб свежий» и «хлеб не свежий», видим, что частица «не», стоящая перед словом «свежий», меняет значение слова на противоположное, что можно рассматривать, как изменение начальной фазы на π. Если между частицей «не» и исследуемым термином присутствует другое слово, как, например «хлеб не очень свежий», то частица «не» тоже меняет значение исходного слова, но уже не на противоположное, а на некоторое среднее, что может соответствовать изменению начальной фазы, допустим на $\pi/2$. Такие рассуждения не дают возможности вывести достаточно точный алгоритм расчета начальной фазы. Если для частиц, обозначающих отрицание, возможно применение хотя бы интуитивного алгоритма, то методика учета влияния на начальную фазу прочих служебных частей речи, таких как предлоги и союзы, а также знаков препинания и форм слов в тексте, подлежит дальнейшему изучению. В данный момент расчет начальной фазы реализован только для частиц «не» и «ни» в соответствии с изложенными выше правилами. Для всех остальных слов начальная фаза считается равной нулю. Возможность использования исходной формы слова в тексте, а также других служебных слов для уточнения начальной фазы в настоящей работе рассматривается и изучается. Однако начальная фаза не определяет полностью фазу волны, а является одной из ее составляющих наряду с волновым числом и расстоянием от источника до точки наблюдения, как это видно из уравнения (1). Потому отсутствие сведений о начальной фазе оставляет, тем не менее, возможность использования волновых уравнений с учетом того, что информация о волне будет неполной.

Амплитуды A_j, A_n в данной модели не представляют трудностей при интерпретации и рассматриваются как количество вхождений данного термина в исходный текст. При этом учитывается, что одни и те же слова с разными начальными фазами являются разными волнами и их амплитуды не складываются.

Таким образом, в рамках описанной модели текст представлен как волновой пакет, в котором каждое слово, относящееся к самостоятельным частям речи, формирует сферическую волну. Амплитуда волны зависит от количества вхождений данного слова в текст. Значение слова в контексте повествования отражает волновое число, а начальная фаза определяется окружением слова.

Экспериментальное исследование волновой модели

Выполним экспериментальное исследование применения волновой модели с целью классификации коротких текстов. Выберем анализ тематики открытых сообществ в социальной сети «ВКонтакте»[1]. Отберем 100 групп, принадлежащих следующим тематикам: путешествия, спорт, кулинария, гуманитарная, техническая, по которым проведена классификация текстов. При отборе отдано предпочтение «сообществам по интересам», а не группам, принадлежащим коммерческим организациям. Такое условие объясняется предположением, что информация, содержащаяся в коммерческих группах, может быть переполнена ключевыми словами,

[1] Социальная сеть «ВКонтакте» [Электронный ресурс]. URL: https://vk.com (дата обращения: 15.01.2022).

относящимися к данной тематике для улучшения условий индексации и поиска. Следовательно, классификация этих групп может дать завышенные показатели точности расчетов.

Для анализа использованы все слова, принадлежащие самостоятельным частям речи из разделов «наименование» и «информация» сообщества. Длинные тексты были обрезаны до 1500 символов, что позволило сократить время вычислений без потери точности. Использование только начальной части более длинных текстов принято допустимым, так как в описании сообществ полезная информация, относящаяся к тематике, обычно располагается в начале. Концовки текстов, как правило, посвящены регламенту поведения в группе, размещению рекламы и т. п. Преобразование исходного текста в массив слов с определением частей речи выполнено с использованием внешней библиотеки «rulemma»[1]. Расчет семантических расстояний выполнен с использованием модели «Национальный корпус русского языка (НКРЯ) и Википедия» при помощи API-методов (Application Programming Interface — программный интерфейс приложения), предоставленных интернет-ресурсом «rusvectores»[2]. Выбор данной модели [12] обусловлен главным образом тем, что она находится в открытом доступе, предоставляет словарь большого объема (249 333 слова в версии за ноябрь 2021 года), а также снабжена удобными API-методами, которые позволяют работать с моделью, не разворачивая ее полную локальную копию.

Полученные значения близости сохранены в локальной базе данных для сокращения времени расчетов. Основная часть расчетов с использованием волновой модели, а также хранение экспериментальных данных осуществлено на платформе «1С Предприятие 8». Для проведения исследований использован ноутбук Toshiba Satellite Pro 650 18-G (модель 2010 года) с процессором Pentium Dual-Core T4500 (2,3 GHz) и объемом оперативной памяти 3 ГБ. Для определения влияния предложенной методики расчета начальной фазы на точность классификации вычисления выполнены для вариантов с учетом (№ 1) и без учета (№ 2) начальной фазы. Полученные результаты приведены в табл. 1.

Распределение текстов по группам, полученное в результате классификации для вариантов начальной фазы № 1 и № 2 показано в табл. 2.

Суммарная длительность классификации, включая этапы лемматизации, расчетов волновых чисел, начальных фаз и амплитуды вероятности составила от 20 до 4422 с на один объект, в зависимости от длины классифицируемого текста (рисунок). На графике видно, что наиболее нагруженная часть алгоритма — расчет амплитуды вероятности в соответствии с уравнением (2). Основную часть нагрузки обеспечивает расчет суммы попарных произведений, организованный как вложенные циклы. При этом внешний цикл выполняется $n - 1$ раз, где n не превышает количество значащих слов в тексте, а внутренний — проходит от значения счетчика внешнего цикла, увеличенного на 1 до n. Таким образом, трудоемкость алгоритма составляет $O(n^2)$.

На основании полученных результатов сделаем следующие выводы.

В целом волновая модель обладает высокой точностью классификации коротких текстов по тематикам. Отметим класс «гуманитарный», точность классификации для которого оказалась значительно ниже, чем для остальных классов. Этот факт говорит не столько о погрешности самой волновой модели, сколько о возможной недостаточной точности определения позиции вектора, соответствующего понятию «гуманитарный» в базисной дистрибутивно-семантической модели. Данный результат возможен, так как существуют примеры пар слов, противоположных по значению с точки зрения человеческой логики, которые определяются моделью «НКРЯ и Википедия» почти как синонимы. Это, например, пара «черный–белый», для которой семантическая близость равна 0,732 или «много–мало» с семантической близостью 0,808. Оба значения близки к единице, что соответствует словам, близким по значению. Следовательно, вычисление расстояний в исходной дистрибутивно-семантической модели может быть недостаточно точным, что оказывает негативное влияние на точность работы волновой модели. Для повышения точности расчетов целесообразно проводить предварительную оценку качества представления терминов, обозначающих классы в дистрибутивно-семантической модели и, возможно, замену исходных терминов синонимами с более точным представлением. Также необходимо провести аналогичные расчеты с использованием других языковых корпусов для выбора оптимальной базисной модели, таких как, например, RDT [13] и DISCO [14, 15].

Полученные результаты не дают возможность сделать вывод о влиянии предложенного алгоритма определения начальной фазы. При данном подходе учет начальной фазы не вызвал значительного изменения точности классификации в целом, немного ухудшив результат для класса «гуманитарный» и улучшив — для класса «технический». Очевидно, этот вопрос требует дальнейшей разработки и исследования.

Таблица 1. Точность классификации с использованием волновой модели, %

Table 1. Accuracy of classification using the wave model, %

Класс	Точность классификации для вариантов начальной фазы	
	№ 1	№ 2
Путешествия	95	95
Кулинария	95	95
Спорт	95	95
Гуманитарный	75	80
Технический	95	85
Все классы	91	90

[1] Лемматизатор для русскоязычных текстов [Электронный ресурс]. URL: https://github.com/Koziev/rulemma (дата обращения: 05.01.2022).

[2] RusVectōrēs: семантические модели для русского языка [Электронный ресурс]. URL: https://rusvectores.org/ (дата обращения: 05.01.2022).

Таблица 2. Распределение текстов по реальным классам, полученные для вариантов с учетом и без учета начальной фазы, %
Table 2. Distribution of texts by real classes obtained for variants with and without taking into account the initial phase, %

Класс	Путешествия		Кулинария		Спорт		Гуманитарный		Технический	
	№ 1	№ 2	№ 1	№ 2	№ 1	№ 2	№ 1	№ 2	№ 1	№ 2
Путешествия	95	95	—	—	—	—	—	—	5	5
Кулинария	5	5	95	95	—	—	—	—	—	—
Спорт	5	5	—	—	95	95	—	—	—	—
Гуманитарный	5	5	15	15	—	—	75	80	5	—
Технический	—	—	5	10	—	5	—	—	95	85

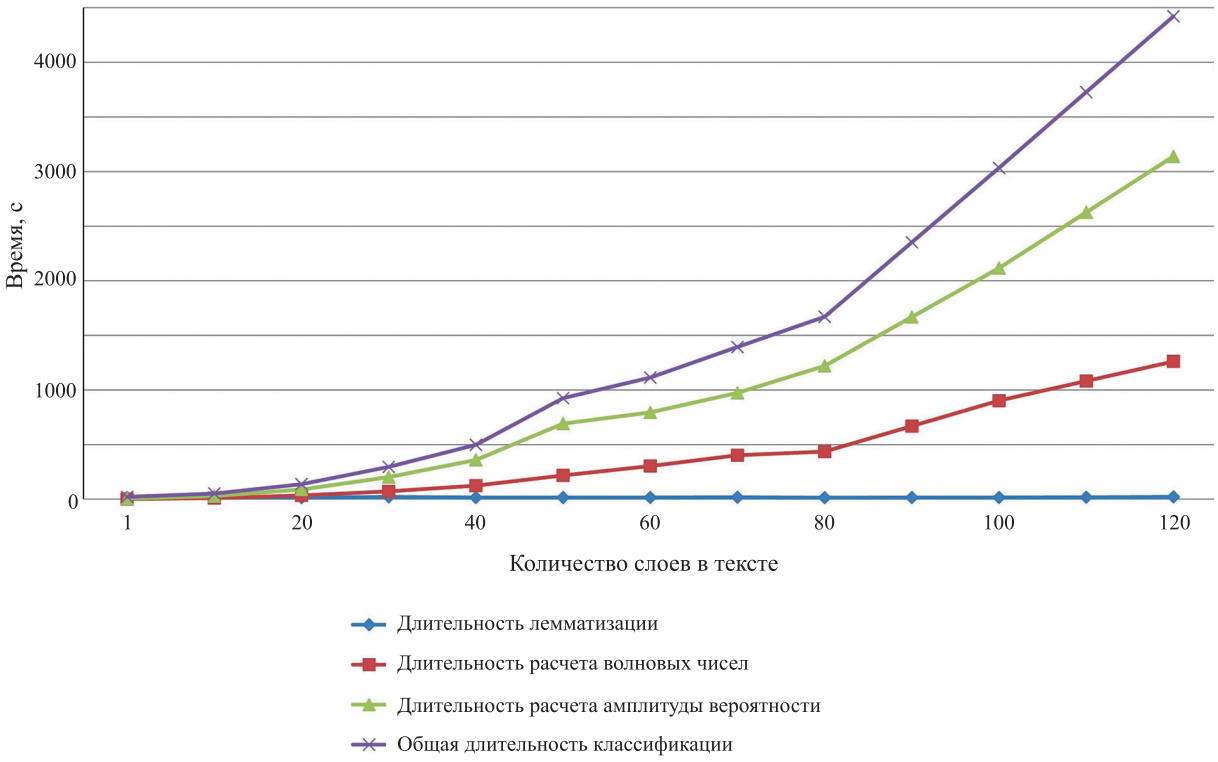

Рисунок. Длительность классификации в зависимости от длины текста
Figure. Duration of classification depending on the text length

Время, затрачиваемое на выполнение вычислений при использовании волновой модели, оказалось достаточно большим. При этом, если длительность этапа лемматизации практически не зависит от длины классифицируемого текста, то для этапов расчета волновых чисел и амплитуд вероятностей наблюдается резкий рост временных затрат при росте количества слов в тексте. Отметим, что при данном варианте реализации вычислительный алгоритм располагает значительными ресурсами для оптимизации. В настоящее время для расчетов используется интеграция трех систем: основная расчетная часть в среде 1С Предприятие 8, внешний модуль лемматизации, обращение к языковой модели при помощи методов API. Реализация полного цикла расчетов в рамках одной системы, хранение полностью локальной версии дистрибутивно-семантической модели с использованием для этих целей оптимальной системы управления базами данных, а также применение более высокопроизводительной вычислительной техники могут значительно снизить временные затраты. Тем не менее, принципиальная возможность применения волновой модели для классификации длинных текстов вызывает большие сомнения.

Заключение

Результаты экспериментального исследования волновой модели позволяют рассматривать ее как перспективный инструмент для классификации коротких текстов, заслуживающий дальнейшего изучения и развития. Одно из важных достоинств предложенной модели — универсальность. Она не требует специального обучения для работы с определенными классами, ее работа основана на использовании универсальной дистрибутивно-семантической языковой модели. Второе достоинство модели — достаточно высокая точность. При этом существуют возможности для дальнейшей работы над повышением точности классификации.

Возможно использование не единичных понятий, представляющих классы, а группы синонимов, а также уточнение алгоритмов определения волновых чисел и начальных фаз волнового пакета. Еще одно важное направление работы — оптимизация алгоритмов и сокращение временных затрат при работе описанной модели.

В дальнейшем планируется уделить внимание исследованию возможностей повышения точности и производительности предложенной модели. Предполагается изучить перспективы ее применения для решения других задач, связанных с классификацией коротких текстов, например, определение эмоциональной окраски сообщений, а также структуризация и обобщение отзывов и комментариев пользователей о товарах, услугах и событиях.

Литература

1. Nielsen M.A., Chuang I.L. Quantum Computation and Quantum Information. Cambridge University Press, 2010. 704 p. https://doi.org/10.1017/CBO9780511976667
2. Melucci M. Introduction to Information Retrieval and Quantum Mechanics. Berlin, Heidelberg: Springer-Verlag, 2015. 247 p. https://doi.org/10.1007/978-3-662-48313-8
3. Blacoe W., Kashefi E., Lapata M. A Quantum-theoretic approach to distributional semantics // Proc. of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT). 2013. P. 847–857.
4. Jaiswal A.K., Holdack G., Frommholz I., Liu H. Quantum-like Generalization of complex word embedding: a lightweight approach for textual classification // CEUR Workshop Proceedings. 2018. V. 2191. P. 159–168.
5. Surov I.A., Semenenko E., Platonov A.V., Bessmertny I.A., Galofaro F., Toffano Z., Khrennikov A.Y., Alodjants A.P. Quantum semantics of text perception // Scientific Reports. 2021. V. 11. N 1. P. 4193. https://doi.org/10.1038/s41598-021-83490-9
6. Pang B., Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts // Proc. of the 42nd Annual Meeting Association for Computational Linguistics (ACL). 2004. P. 271–278. https://doi.org/10.3115/1218955.1218990
7. Клековкина М.В., Котельников Е.В. Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Материалы XIV Всероссийской научной конференции (RCDL-2012). 2012. С. 118–123.
8. Меньшиков И.Л. Анализ тональности текста на русском языке при помощи графовых моделей // Доклады всероссийской научной конференции АИСТ'2013 «Анализ Изображений, Сетей и Текстов». Екатеринбург, 2013. С. 151–155.
9. Татарникова Т.М., Богданов П.Ю. Построение психологического портрета человека с применением технологий обработки естественного языка // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21. № 1. С. 85–91. https://doi.org/10.17586/2226-1494-2021-21-1-85-91
10. Литвинова Т.А., Загоровская О.В., Середин П.В., Лантюхова Н.Н., Шевченко И.С. Профилирование автора письменного текста: подходы, методы и их оптимизация // Филология, искусствоведение и культурология: актуальные вопросы и тенденции развития: материалы международной. научно-практической конференции (13 мая 2013 г.). Новосибирск: СибАК, 2013. С. 69–79.
11. Френкель Я.И. Волновая механика. Ч. 1. Элементарная теория. Квантовая физика. М.: URSS, 2019. 392 с.
12. Kutuzov A., Kuzmenko E. WebVectors: A toolkit for building web interfaces for vector semantic models // Communications in Computer and Information Science. 2017. V. 661. P. 155–161. https://doi.org/10.1007/978-3-319-52920-2_15
13. Panchenko A., Ustalov D., Arefyev N., Paperno D., Konstantinova N., Loukachevitch N., Biemann C. Human and machine judgements about russian semantic relatedness // Communications in Computer and Information Science. 2017. V. 661. P. 221–235. https://doi.org/10.1007/978-3-319-52920-2_21
14. Kolb P. Experiments on the difference between semantic similarity and relatedness // Proc. of the 17th Nordic Conference of Computational Linguistics (NODALIDA '09). 2009. P. 81–88.
15. Kolb P. DISCO: A multilingual database of distributionally similar words // Proc. of the KONVENS-2008. Berlin. 2008. P. 6–12.

References

1. Nielsen M.A., Chuang I.L. *Quantum Computation and Quantum Information*. Cambridge University Press, 2010, 704 p. https://doi.org/10.1017/CBO9780511976667
2. Melucci M. *Introduction to Information Retrieval and Quantum Mechanics*. Berlin, Heidelberg, Springer-Verlag, 2015, 247 p. https://doi.org/10.1007/978-3-662-48313-8
3. Blacoe W., Kashefi E., Lapata M. A Quantum-theoretic approach to distributional semantics. *Proc. of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, 2013, pp. 847–857.
4. Jaiswal A.K., Holdack G., Frommholz I., Liu H. Quantum-like Generalization of complex word embedding: a lightweight approach for textual classification. *CEUR Workshop Proceedings*, 2018, vol. 2191, pp. 159–168.
5. Surov I.A., Semenenko E., Platonov A.V., Bessmertny I.A., Galofaro F., Toffano Z., Khrennikov A.Y., Alodjants A.P. Quantum semantics of text perception. *Scientific Reports*, 2021, vol. 11, no. 1, pp. 4193. https://doi.org/10.1038/s41598-021-83490-9
6. Pang B., Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Proc. of the 42nd Annual Meeting Association for Computational Linguistics (ACL)*, 2004, pp. 271–278. https://doi.org/10.3115/1218955.1218990
7. Kotelnikov E., Klekovkina M. The automatic sentiment text classification method based on emotional vocabulary. *Digital Libraries: Advanced Methods and Technologies. Proc. of the RCDL-2012*, 2012, pp. 118–123. (in Russian)
8. Menshikov I.L. Sentiment analysis of a text in russian using graph models. *Proc. of the Conference AIST'2013 "Analysis of Images, Social Networks, and Texts"*. Ekaterinburg, 2013, pp. 151–155. (in Russian)
9. Tatarnikova T.M., Bogdanov P.Yu. Human psyche creation by application of natural language processing technologies. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2021, vol. 21, no. 1, pp. 85–91. (in Russian). https://doi.org/10.17586/2226-1494-2021-21-1-85-91
10. Litvinova T.A., Zagorovskaia O.V., Seredin P.V., Lantiukhova N.N., Shevchenko I.C. Author profiling of a written text: approaches, methods, and their optimization. *Philology, Art Criticism, and Cultural Studies: Topical Issues and Development Trends. Proceedings of the International Research-to-Practice Conference. May, 13, 2013*. Novosibirsk, SibAK, 2013, pp. 69–79. (in Russian)
11. Frenkel J. *Wave Mechanics: Elementary Theory. The Quantum Physics*. Moscow, URSS, 2019, 392 p. (in Russian)
12. Kutuzov A., Kuzmenko E. WebVectors: A toolkit for building web interfaces for vector semantic models. *Communications in Computer and Information Science*, 2017, vol. 661, pp. 155–161. https://doi.org/10.1007/978-3-319-52920-2_15
13. Panchenko A., Ustalov D., Arefyev N., Paperno D., Konstantinova N., Loukachevitch N., Biemann C. Human and machine judgements about russian semantic relatedness. *Communications in Computer and Information Science*, 2017, vol. 661, pp. 221–235. https://doi.org/10.1007/978-3-319-52920-2_21
14. Kolb P. Experiments on the difference between semantic similarity and relatedness. *Proc. of the 17th Nordic Conference of Computational Linguistics (NODALIDA '09)*, 2009, pp. 81–88.
15. Kolb P. DISCO: A multilingual database of distributionally similar words. *Proc. of the KONVENS-2008*, Berlin, 2008, pp. 6–12.

Авторы

Груздева Анастасия Сергеевна — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0000-0003-4963-0823, prog.anastasia@gmail.com
Бессмертный Игорь Александрович — доктор технических наук, профессор, профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 36661767800, https://orcid.org/0000-0001-6711-6399, bessmertny@itmo.ru

Authors

Anastasia S. Gruzdeva — PhD student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0000-0003-4963-0823, prog.anastasia@gmail.com
Igor A. Bessmertny — D.Sc., Full Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 36661767800, https://orcid.org/0000-0001-6711-6399, bessmertny@itmo.ru

Статья поступила в редакцию 25.01.2022
Одобрена после рецензирования 11.02.2022
Принята к печати 17.03.2022

Received 25.01.2022
Approved after reviewing 11.02.2022
Accepted 17.03.2022