УДК 004.415.2
DOI: 10.17586/0021-3454-2019-62-12-1060-1065

ВЕКТОРНОЕ ПРЕДСТАВЛЕНИЕ СЛОВ ПРИ ПОМОЩИ АППАРАТА КВАНТОВОЙ ТЕОРИИ ВЕРОЯТНОСТЕЙ

А. В. Платонов, И. А. Бессмертный, Ю. А. Королёва

Университет ИТМО, 197101, Санкт-Петербург, Россия
E-mail: bessmertny@itmo.ru

Рассмотрена проблема моделирования семантики текстовых документов на основе векторного представления слов в гильбертовом пространстве. Векторное представление отражает семантику слов, окружающих данное (контекст). Если слово встречается в исследуемом документе неоднократно, множество контекстов слова образует его обобщенный контекст. Можно рассматривать разные контексты слова как разные проекции, а обобщенный контекст — как восстановленный многомерный объект. Целью исследования является повышение качества восстановления контекста слова за счет учета дополнительных факторов, в частности, возможной неортогональности контекстов. Использована квантовая теория вероятностей. Восстановление контекста соответствует задаче квантовой томографии в квантовой физике. Задача восстановления контекста слова, или в терминах квантовой математики — матрицы плотности вероятностей, решается методом градиентного спуска с помощью машинного обучения. Набор регуляризаторов обеспечивает сходимость процесса по критерию дивергенции Кульбака—Лейблера.

Ключевые слова: *естественно-языковой текст, семантика, контекст, квантовая теория вероятностей, квантовая томография*

Получение векторного представления семантических единиц для слов естественного языка является одной из важнейших задач в автоматической обработке текстов. От свойств таких объектов и их качества, т.е. от того, насколько далеко расположены такие векторы в своем векторном пространстве для различных по смыслу слов, зависит эффективность использующих их алгоритмов машинного обучения.

Формально задача может быть поставлена в следующем виде. Для заданного набора слов (лексикона) по множеству неразмеченных текстов необходимо составить отображение из слова в некоторый вектор. При этом на целевом векторном пространстве должна быть задана метрика, отображающая близость по смыслу двух рассматриваемых слов. Алгоритмы математического представления семантики слов естественного языка подготавливают исходные данные для обработки текстов.

Состояние проблемы и текущие исследования. Существует несколько основных подходов к моделированию векторных пространств. В базовом варианте все подходы можно разделить на две большие группы — дистрибутивный [1] и структурный [2]. При структурном подходе производится синтаксический анализ текстовых данных, и полученное дерево разбора используется для выявления семантических отношений между словами. Согласно дистрибутивному подходу, применяемому в настоящей статье, смысл слова определяется через распределение его контекстов, т.е. наборов слов в некотором окне вокруг слова.

Согласно алгоритму, векторное представление слова определяется частотой появления других слов в окне сканирования фиксированного размера. Существуют различные модификации данного алгоритма, нивелирующие влияние очень редких или очень часто встречаемых слов [3].

Другой подход к моделированию семантики основан на анализе матрицы встречаемости слов в документах. В частности, при этом используются алгоритмы LSA [4] и pLSA [5].

При третьем — нейросетевом — подходе, в частности, используется алгоритм Word2Vec [6] и его производные. В этом алгоритме на вход однослойного перцептрона подается набор векторов представления слов из контекста, смоделированный, например, алгоритмом Bag-of-Words, на выходе такого перцептрона может быть получен такой же вектор для центрального слова в рассматриваемом окне. Скрытый слой такой нейросети имеет вектор небольшой размерности относительно всего лексикона — порядка нескольких сотен компонентов. Таким образом, алгоритм Word2Vec позволяет построить плотные векторы малой размерности в противовес предыдущим алгоритмам, генерирующим разреженные представления.

Подход, основанный на квантовой теории информации. Подходы, основанные на дистрибутивной гипотезе, можно отнести к классическим. В настоящей работе рассматривается альтернативный подход к моделированию семантических единиц, основанный на квантовой теории информации. Большой интерес для исследователей в области интеллектуального анализа данных представляют работы на стыке квантовой теории и информатики [7, 8]. В силу того, что квантовая математика дает удобный и легко интерпретируемый аппарат представления зависимостей между объектами, например словами, данный математический аппарат представляет большой интерес для теории информации [9], информационного поиска [10—12] и моделирования естественных языков [13, 14].

В настоящей работе предложена идея векторного представления слов, основанная на квантовой теории вероятностей [15], которая позволяет исследовать системы как векторы в гильбертовом пространстве, а событие в этом пространстве — как некоторое измерение состояния этой системы. Состояние системы заранее неизвестно и может быть описано лишь в виде вероятностного распределения на так называемых наблюдаемых, т.е. операторов-проекторов на некоторый вектор в гильбертовом пространстве, система с некоторой вероятностью может находиться в состоянии, описываемом этой наблюдаемой. В работе проводится аналогия между состоянием квантовой системы и семантическим концептом, описывающим понятие на естественном языке.

Томография квантовых состояний. Состояние квантовой системы описывается в некотором базисе выбранного гильбертова пространства и оператора плотности (функции распределения плотности в фазовом пространстве этой системы).

В общем случае для базиса $|\Psi_1 \ldots |\Psi_n$ состояние квантовой системы может быть описано матрицей плотности по следующей формуле:

$$\rho = \sum_{i=1}^{n} p_i |\psi_i\psi_i|, \qquad (1)$$

где $\langle\psi_i|$ — вектор, сопряженный вектору $|\psi_i\rangle$, а p_i — вероятность нахождения исследуемой системы в состоянии $|\psi_i$. Если система находится в чистом состоянии, то она однозначно может быть описана лишь одним вектором в выбранном базисе и иметь матрицу плотности $\rho = |\psi\rangle\langle\psi|$. С точки зрения классической теории вероятностей чистое состояние соответствует элементарному событию в вероятностном пространстве. Если система не может быть описана одним вектором, то она описывается выражением (1), и такое состояние называется смешанным.

Наблюдаемой называется матрица-проектор на некоторое подпространство исходного гильбертова пространства, описывающая чистое состояние в этом пространстве. Зная матрицу плотности квантовой системы ρ и представление наблюдаемой A, можно записать выражение для расчета вероятности того, что система находится в состоянии, соответствующем наблюдаемой:

$$\langle A \rangle = \mathrm{tr}(A\rho), \tag{2}$$

где $\mathrm{tr}(M)$ означает след полученной матрицы $M = A\rho$.

Таким образом, можно сформулировать обратную задачу — получение матрицы плотности в результате анализа средних значений наблюдаемых. Такая процедура называется квантовой томографией. Суть ее состоит в том, что для исследуемой квантовой системы выбирается некоторый базис и для каждого базисного вектора выполняется серия измерений, в результате которой получаются средние значения наблюдаемых, по которым восстанавливается матрица плотности исходя из формулы (2) и свойств самой матрицы плотности:

$$\begin{cases} \rho_{ii} \geq 0, \\ \mathrm{tr}(\rho) = 1, \\ \mathrm{tr}(\rho^2) \leq 1, \\ \rho_{ii} = \overline{\rho}_{ji},\ i \neq j. \end{cases} \tag{3}$$

Квантовоподобное представление семантических единиц. Для этой задачи необходимо разработать алгоритм, который на вход принимает слово word и возвращает вектор $\mathrm{V}_{\mathrm{word}} = [\upsilon_1 \upsilon_2 \ldots \upsilon_L]$ размерности L в пространстве вещественных чисел. Таким образом, необходимо разработать отображение $a(w) : W \to R^L$, где W — множество всех слов, $w \in W$ — рассматриваемое слово.

Пусть A_k — матрица-проектор на подпространство, соответствующее некоторому контексту k из набора известных контекстов размерности N. Чтобы получить такой проектор, необходимо представить контекст слова как вектор по аналогии с методом, описанным выше. Так как получаемая из такого вектора матрица A_k должна обладать свойствами проектора, то должно выполняться условие нормировки такого вектора:

$$A_k = \frac{v_k \cdot v_k^T}{|v_k|^2}, \tag{4}$$

где v_k — исходный вектор контекста слова.

Вероятность P_k также известна, так как известен текстовый корпус, на котором проводится обучение. Для того чтобы получить распределение вероятностей на N контекстах для исследуемого слова, достаточно выполнить группировку векторов контекстов по их точному совпадению и выразить вероятность их появления через частоту. Если все контексты уникальны, вероятность любого такого контекста $\frac{1}{N}$.

Матрица плотности ρ представляет собой описание „системы“, соответствующей исследуемому слову. Именно эту матрицу необходимо восстановить из выражения (2), получив, таким образом, матричное представление обобщенного контекста слова. В отличие от квантовой теории далее все матрицы рассматриваются как объекты над полем вещественных чисел.

Таким образом, задача получения представления слова сводится к отысканию матрицы ρ, удовлетворяющей уравнению (2). Запишем более подробно матрицы плотности и проектора:

$$\rho = \begin{bmatrix} \rho_{11} & \cdots & \rho_{1L} \\ \vdots & \ddots & \vdots \\ \rho_{L1} & \cdots & \rho_{LL} \end{bmatrix},\ A_k = \begin{bmatrix} a_{11} & \cdots & a_{1L} \\ \vdots & \ddots & \vdots \\ a_{L1} & \cdots & a_{LL} \end{bmatrix}. \tag{5}$$

Получив выражение для градиента целевой функции, можно применить методы градиентного спуска с целью оптимизации и получения матрицы плотности слова:

$$\mathrm{tr}\left(A_k\rho\right)=\sum_{i=1}^{L}\sum_{j=1}^{L}a_{ij}\rho_{ji}=P_k\,. \tag{6}$$

При этом на матрицу плотности, исходя из квантовой теории вероятностей, налагается ряд ограничений (3).

Восстановление матриц плотности. В работе Пивоварского [16] подход к получению матриц плотности основан на взвешенной сумме матриц-проекторов на контексты слова:

$$\rho=\sum_{k=1}^{L}v_k A_k\ \ ,\ \sum_{k=1}^{L}v=1\ ,v_k\geq 0\,. \tag{7}$$

В такой сумме вес соответствует частоте появления контекстов слова. Данный подход вычислительно прост, он позволяет сохранить все свойства матрицы плотности из формулы (3). Однако он непригоден для неортогональных контекстов слов, а также для систем в смешанном состоянии. Такой алгоритм не восстанавливает целевое распределение для контекстов. В настоящей работе предложено применять подход, основанный на решении оптимизационной задачи методом градиентного спуска. При восстановлении параметров дискретных распределений вероятностей в машинном обучении часто используется дивергенция Кульбака—Лейблера:

$$\mathrm{D_{KL}}\left(\rho\,\|\,P\right)=-\sum_{k=1}^{L}\mathrm{tr}\left(\rho A_k\right)\log\frac{v_k}{\mathrm{tr}\left(\rho A_k\right)}\,. \tag{8}$$

Если метрику (8) использовать в качестве основной целевой функции, а $\lambda_i\left(\rho\right)$ рассматривать как набор регуляризаторов для сохранения свойств матрицы плотности (3), то получится следующая оптимизационная задача:

$$\min_{\rho}Q\left(\rho,P\right)=\min_{\rho}D_{KL}\left(\rho\,||P\right)+\sum_{i=1}^{R}\lambda_i\left(\rho\right)\to\min\,, \tag{9}$$

где R — число регуляризаторов. Градиент по параметрам матрицы плотности для формулы (9) получен из следующего выражения:

$$\nabla Q\left(\rho\right)=-\sum_{k=1}^{L}\left(\log\left(\frac{\mathrm{tr}\left(\rho A_k\right)}{v_k}\right)+1\right)A_k^T+\sum_{i=1}^{R}\nabla\lambda_i\left(\rho\right)\,. \tag{10}$$

Получив выражение градиента целевой функции, можно применять методы градиентного спуска для оптимизации и построения матрицы плотности слова.

Заключение. В настоящей работе векторные представления семантики слова формируются из контекстов, полученных из множества окружений данного слова. Основная задача данного исследования — восстановление обобщенного контекста слова на основе всех частных контекстов, решаемая методом квантовой томографии. Математический аппарат квантовой теории вероятностей позволяет работать с неортогональными контекстами слов. Дальнейшие исследования будут направлены на программную реализацию предложенного подхода.

СПИСОК ЛИТЕРАТУРЫ

1. *Harris Z.* Distributional structure // Word. 1954. P. 146—162.

2. *Chomsky N.* Three models for the description of language // IRE Transactions on Information Theory. 1956. P. 113—124.

3. *Levy O., Goldberg Y. and Dagan I.* Improving distributional similarity with lessons learned from word embeddings // Transactions of the Association for Computational Linguistics. 2015. Vol. 3. P. 211—225.

4. *Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K. and Harshman R.* Indexing by latent semantic analysis // J. of the American Society for Information Science. 1990. Vol. 41, N 6. P. 391—407.

5. *Hofmann T.* Probabilistic latent semantic indexing // Proc. of the 22nd Annual Intern. ACM SIGIR Conf. on Research and Development in Information Retrieval. NY, USA, 1999. P. 50—57.

6. *Mikolov T., Chen K., Corrado G. and Dean J.* Efficient estimation of word representations in vector space // CoRR. arXiv:1301.3781v3. 2013.

7. *Aerts D., Czachor M. and Sozzo S.* Quantum Interaction Approach in Cognition, Artificial Intelligence and Robotics // CoRR. arXiv:1104.3345v1. 2011.

8. *Barros J., Toffano Z., Meguebli Y., Doan B.-L.* Contextual Query Using Bell Tests // Quantum Interaction. Berlin—Heidelberg: Springer, 2014. P. 110—121.

9. *Хренников А. Ю.* Введение в квантовую теорию информации. М.: Физматлит, 2008. 284 с.

10. *Frommholz I., Larsen B., Piwowarski B., Lalmas M., Ingwersen P. and van Rijsbergen K.* Supporting polyrepresentation in a quantum-inspired geometrical retrieval framework // Proc. of the 3rd Symp. on Information interaction in context. ACM, New Brunswick, USA, 2010. P. 115—124. ISBN: 978-1-4503-0247-0.

11. *Piwowarski B., Frommholz I., Lalmas M., and van Rijsbergen K.* What can quantum theory bring to information retrieval // Proc. of the 19th ACM Intern. Conf. on information and knowledge management. CIKM'10. 2010. NY, USA, 2010. P. 59—68.

12. *Melucci M. and Piwowarski B.* Quantum mechanics and information retrieval: from theory to application // Proc. of the 2013 Conf. on the Theory of Information Retrieval. NY, USA, 2013.

13. *Sadrzadeh M. and Grefenstette E.* A compositional distributional semantics, two concrete constructions, and some experimental evaluations // Quantum Interaction. Berlin—Heidelberg: Springer, 2011. P. 35—47. Electronic ISBN: 978-3-642-24971-6

14. *Sordoni A., Nie J. and Bengio Y.* Modeling Term Dependencies with Quantum Language Models for IR // Proc. of the 36th Intern. ACM SIGIR Conf. on Research and Development in Information Retrieval. SIGIR '13. NY, USA, 2013. P. 653—662.

15. *Khrennikov A.* Classical and quantum probability for biologists-introduction // Quantum Probability and White Noise Analysis. 2010. P. 179—192.

16. *Piwowarski B. and Lalmas M.* A quantum-based model for interactive information retrieval // Advances in information retrieval theory. Berlin—Heidelberg: Springer, 2009. P. 224—231. ISBN: 9781450300995

Сведения об авторах

Алексей Владимирович Платонов — аспирант; Университет ИТМО, факультет программной инженерии и компьютерной техники; E-mail: avplatonov@ itmo.ru

Игорь Александрович Бессмертный — д-р техн. наук, доцент; Университет ИТМО, факультет программной инженерии и компьютерной техники; E-mail: bessmertny@itmo.ru

Юлия Александровна Королёва — канд. техн. наук; Университет ИТМО, факультет программной инженерии и компьютерной техники; E-mail: jakoroleva@itmo.ru

Поступила в редакцию
03.10.19 г.

Ссылка для цитирования: *Платонов А. В., Бессмертный И. А., Королёва Ю. А.* Векторное представление слов при помощи аппарата квантовой теории вероятностей // Изв. вузов. Приборостроение. 2019. Т. 62, № 12. С. 1060—1065.

VECTOR REPRESENTATION OF WORDS USING THE APPARATUS OF QUANTUM PROBABILITY THEORY

A. V. Platonov, I. A. Bessmertny, Yu. A. Koroleva

ITMO University, 197101, St. Petersburg, Russia
E-mail: bessmertny@itmo.ru

The problem of modeling the semantics of text documents based on vector representation of words in a Hilbert space is considered. The vector representation of a word reflects the words surrounding the given word (the context of the word). If a word is found in a document more than once, the set of contexts of a word forms its generalized context, i.e. meaning of the word. Different contexts of a word may be considered as different projections, and the generalized context — as a reconstructed multidimensional object. The purpose of the presented study is to improve the quality of the restoration of a word the context by considering additional factors, in particular, the possible non-orthogonality of contexts. To achieve the goal, quantum probability theory is used here, and the context recovery procedure corresponds to the problem of quantum tomography in quantum physics. The task of restoring a word context or, in terms of quantum mathematics, the probability density matrix, is solved by the method of gradient descent using machine learning. Restrictions on the learning process are implemented by a set of regularizers that ensure the convergence of the process according to the Kullback—Leibler divergence criterion.

Keywords: natural language texts, document semantics, context, quantum theory of probabilities, quantum tomography

REFERENCES

1. Harris Z. *Word*, 1954, pp. 146–162.
2. Chomsky N. *IRE Transactions on Information Theory*, 1956, pp. 113–124.
3. Levy O., Goldberg Y. and Dagan I*. Transactions of the Association for Computational Linguistics*, 2015, vol. 3, pp. 211–225.
4. Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K. and Harshman R. *Journal of the American Society for Information Science*, 1990, no. 6(41), pp. 391–407.
5. Hofmann T. *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, NY, 1999, pp. 50–57.
6. Mikolov T., Chen K., Corrado G. and Dean J. *CoRR*, arXiv:1301.3781v3, 2013.
7. Aerts D., Czachor M. and Sozzo S. *CoRR*, arXiv:1104.3345v1, 2011.
8. Barros J., Toffano Z., Meguebli Y., Doan B.-L. *Quantum Interaction*, Berlin, Heidelberg, Springer, 2014, pp. 110–121.
9. Hrennikov A.J. *Vvedenie v kvantovuju teoriju informacii* (Introduction to Quantum Information Theory), Moscow. 2008. (in Russ.)
10. Frommholz I., Larsen B., Piwowarski B., Lalmas M., Ingwersen P. and van Rijsbergen K. *Proc. of the 3rd symp. on Information interaction in context*, ACM, 2010, pp. 115–124.
11. Piwowarski B., Frommholz I., Lalmas M., Mounia and van Rijsbergen K. *Proceedings of the 19th ACM international conference on information and knowledge management, CIKM '10*, NY, USA, 2010, pp. 59–68. ISBN: 9781450300995
12. Melucci M. and Piwowarski B. *Proceedings of the 2013 Conference on the Theory of Information Retrieval,* 2013, NY, USA, 2013. ISBN: 978-1-4503-2107-5.
13. Sadrzadeh M. and Grefenstette E. *Quantum Interaction*, Berlin, Heidelberg, Springer, 2011, pp. 35–47. Electronic ISBN: 978-3-642-24971-6
14. Sordoni A., Nie J. and Bengio Y. *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13*, NY, USA, 2013, pp. 653–662.
15. Khrennikov A*. Quantum Probability and White Noise Analysis*, 2010, pp. 179–192. ISBN-10: 9814295426.
16. Piwowarski B. and Lalmas M. *Advances in information retrieval theory*, Berlin, Heidelberg, Springer, 2009, pp. 224–231. Electronic ISBN: 978-3-642-04417-5.

Data on authors

Alexey V. Platonov — Post-Graduate Student; ITMO University, Faculty of Software Engineering and Computer Technique; E-mail: avplatonov@ itmo.ru

Igor A. Bessmertny — Dr. Sci., Associate Professor; ITMO University, Faculty of Software Engineering and Computer Technique; E-mail: bessmertny@itmo.ru

Yulia A. Koroleva — PhD; ITMO University, Faculty of Software Engineering and Computer Technique; E-mail: jakoroleva@itmo.ru

For citation: Platonov A. V., Bessmertny I. A., Koroleva Yu. A. Vector representation of words using the apparatus of quantum probability theory. *Journal of Instrument Engineering*. 2019. Vol. 62, N 12. P. 1060—1065 (in Russian).

DOI: 10.17586/0021-3454-2019-62-12-1060-1065