Научная статья

УДК 004.021
DOI 10.25205/1818-7900-2021-19-3-61-69

# Методы Paper2vec и Cite2vec для анализа коллекций научных публикаций

**Николай Игоревич Тихонов**

Новосибирский государственный университет
Новосибирск, Россия
t.kolya54@gmail.com, https://orcid.org/0000-0001-8765-3263

*Аннотация*

Визуализации коллекций научных публикаций используются для лучшего понимания наборов данных. При построении таких визуализаций могут использоваться различные методы анализа текстовых коллекций. В данной статье речь идет о двух методах анализа – Paper2vec и Cite2vec, которые в своей работе используют информацию о цитировании и получают векторные представления документов. Чтобы продемонстрировать работу методов, были разработаны визуализации, которые описаны в данной работе.

*Ключевые слова*

визуализация коллекций документов, векторное представление документов, сети цитирования, контексты цитирования

*Благодарности*

Работа выполнена под научным руководством З. В. Апанович

*Для цитирования*

*Тихонов Н. И.* Методы Paper2vec и Cite2vec для анализа коллекций научных публикаций // Вестник НГУ. Серия: Информационные технологии. 2021. Т. 19, № 3. С. 61–69. DOI 10.25205/1818-7900-2021-19-3-61-69

# Paper2vec and Cite2vec Methods for Analyzing Collections of Scientific Publications

**Nikolay I. Tikhonov**

Novosibirsk State University
Novosibirsk, Russian Federation
t.kolya54@gmail.com, https://orcid.org/0000-0001-8765-3263

*Abstract*

Visualizations are used to better understand collections of scientific publications. Various methods of analyzing text collections can be used to build these visualizations. This article discusses two methods Paper2vec and Cite2vec that get vector representations of documents using citation information. To demonstrate a work of these techniques and an example of their application, visualizations were developed, which are described in this paper.

*Keywords*

visualization of document collections, vector representation of documents, citation networks, citation contexts

*Acknowledgements*

The work performed under the guidance of the scientific supervisor Z. V. Apanovich

*For citation*

Tikhonov N. I. Paper2vec and Cite2vec Methods for Analyzing Collections of Scientific Publications. *Vestnik NSU. Series: Information Technologies*, 2021, vol. 19, no. 3, p. 61–69. (in Russ.) DOI 10.25205/1818-7900-2021-19-3-61-69

## Введение

Коллекции научных публикаций растут быстрыми темпами. Ученым доступны порталы, содержащие большое количество документов. Например, сайт SpringerLink [1] в данный момент содержит 14,07 млн ресурсов. Исследование такого большого объема данных – трудоемкий процесс. Для сокращения трудозатрат, поиска нужных / схожих документов, оценки научного вклада определенных публикаций и обнаружения скрытых связей между документами применяются методы визуализации документов.

В основе методов визуализации документов могут лежать различные модели представления документов [1]. В последние годы чрезвычайно популярны методы векторных представлений слов для обработки естественных языков. Вслед за ними стали появляться методы анализа текстовых коллекций для получения векторных представлений документов.

Хотя существует множество систем анализа документов, новые методы могут вносить новое понимание коллекций, иметь бóльшую производительность при анализе больших коллекций документов, находить новые связи между документами.

В данной статье идет речь о двух методах анализа – Paper2vec и Cite2vec, которые в своей работе используют информацию о цитировании и получают векторные представления документов. Приведены пример использования данных представлений для построения визуализаций и описание двух рассматриваемых методов. Также описаны наши эксперименты с этими двумя методами, включающие визуализацию результатов работы методов и описание проблем при визуализации такого вида информации.

## 1. Методы анализа текстовых коллекций

Вслед за методами векторных представлений слов для обработки естественного языка [2–5] стали появляться методы для векторного представления текстовых документов.

Рассматриваемые два метода используют информацию о цитировании: первый – контексты цитирования, второй – информацию о том, кто кого цитирует в коллекции.

### 1.1. Paper2vec

Paper2vec [6] использует граф цитирования для получения векторных представлений документов, не требуя полный текст документов или контексты цитирования. Это позволяет применять его в базах данных, где полный текст или контексты цитирования не поддерживаются.

#### Алгоритм Paper2vec

**Требует**: база данных научных работ D, содержащих отношения цитирования;

**Выдает**: векторные представления W исследовательских работ, содержащихся в D;

1) построить сеть цитирования из D;

2) построить матрицу весов отношений цитирования исследовательских работ из построенной сети;

3) стохастически минимизировать функцию стоимости для получения векторов документов W;

4) вернуть W.

Набор соседних узлов в сети цитирования рассматривается авторами как «контекст цитирования», при этом считается, что контекст не ограничивается отношениями прямого цитирования, косвенные цитаты тоже несут в себе семантику рассматриваемого документа, но

[1] https://link.springer.com/

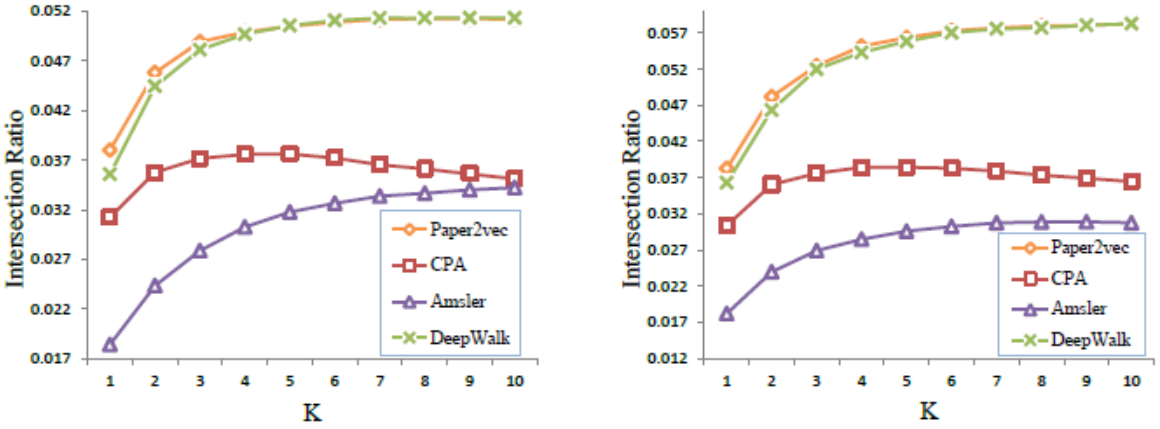
уже с меньшим весом. Таким образом, на втором шаге алгоритма определяются контексты цитирования, и задается весовая схема на сети цитирования, которая учитывает следующие характеристики:

- цитирующие и цитируемые статьи вместе помогают предсказать содержимое текущего документа;
- статьи, которые ссылаются косвенно, не так актуальны, как статьи, которые ссылаются непосредственно (т. е. имеют меньший вес);
- транзитивность. Предположим, что статья A цитирует статью B, а статья B цитирует статью C: чем меньше вес отношения цитирования между A и B или B и C, тем меньше вес отношения цитирования между A и C.

Чтобы удовлетворить вышеуказанным свойствам, весовая схема основана на вероятности случайного блуждания. Данные веса участвуют в функции стоимости, которая стохастически минимизируется.

Эксперименты по оценке работы метода проводились на основе CITREC, открытой системы оценки мер сходства на основе цитирования, которая предоставляет наборы научных данных, базовые измерения и некоторые реализации предыдущих алгоритмов, основанных на цитировании. Набор данных CITREC включает в себя коллекции PubMed Central Open Access Subset (PMCOS) и TREC Genomics. Отношения цитирования были извлечены из текста документов при помощи методов, предоставленных CITREC.

Paper2vec сравнивался с Amsler, CPA, DeepWalk (рис. 1). Каждый метод для каждого документа определял *K* похожих документов. Эти данные сравнивались с эталонными, и определялся средний коэффициент пересечения. По результатам экспериментов, Paper2vec и DeepWalk значительно превосходят модели Amsler и CPA. Paper2vec лучше, чем DeepWalk на малых значениях *K*, что означает, что он может найти лучшие результаты в первых нескольких документах.




а
PMCOS

б
TREC Genomics

*Рис. 1.* Сравнение метода Paper2vec на двух наборах данных (PMCOS и TREC Genomics) с Amsler, CPA и DeepWalk
*Fig. 1.* Comparison of the Paper2vec method on two datasets (PMCOS and TREC Genomics) with Amsler, CPA and DeepWalk

### 1.2. Cite2vec

Система Cite2vec [7] предназначена для векторного представления документов и их динамического исследования. Для обучения векторных моделей используются контексты цитирования научных публикаций.

Для представления документов в векторном пространстве Cite2vec использует дополненную модель Skip-gram из Word2vec. Модель Skip-gram принимает последовательность слов и назначает точку в многомерном пространстве каждому слову, так чтобы слова, которые встречаются вместе в предложениях, были близки в пространстве. Разработчики Cite2vec дополняют эту модель словарем документов, которые цитируются в рассматриваемой коллекции, чтобы исследовать слова и документы в одном и том же пространстве.

Каждый документ, упоминаемый в списках цитируемой литературы, получает в Cite2vec уникальный идентификатор. Таким образом, с каждым документом сопоставляется уникальное слово, что позволяет представлять и слова, и документы в едином векторном пространстве. При этом векторы, соответствующие словам-идентификаторам статей, оказываются в окружении векторов-слов, используемых авторами цитирующих публикаций для описания этих документов.

Полученные таким образом векторы слов и документов обладают рядом полезных характеристик. Во-первых, слова несут семантику тем документов. Кроме того, единое пространство наследует линейную структуру, это позволяет связывать слова и документы при помощи простой арифметики.

Система включает в себя визуализацию и позволяет пользователю динамически просматривать документы на основе информации о том, как их используют остальные документы. Пользователь имеет возможность сформулировать любое понятие, состоящее из нескольких слов. При выборе такого понятия все слова, входящие в понятие, суммируются со словами, присутствующими в визуализации, и исследуемый документ перемещается ближе к тому слову, чья сумма со словами выбранного понятия наиболее точно отражает смысл рассматриваемого документа. В качестве понятий можно указывать сами документы, они несут в себе богатый набор значений.

## 2. Эксперименты

Рассмотрим пример визуализации полученных результатов анализа текстов и визуально оценим работу двух указанных методов на подготовленном наборе данных. Опишем проблемы, возникающие при визуализации векторов большой размерности.

### 2.1. Набор данных

Данные для экспериментов были взяты с сайта CiteSeerX [2] и содержат 3 811 контекстов цитирования из 363 статей. Стати были отобраны поиском по месту публикации, содержащим слово *visualization*. Информация о каждой статье содержит: уникальный идентификатор статьи DOI, уникальные идентификаторы цитируемых статей и контексты цитирования, название, автора, ключевые слова и место публикации.

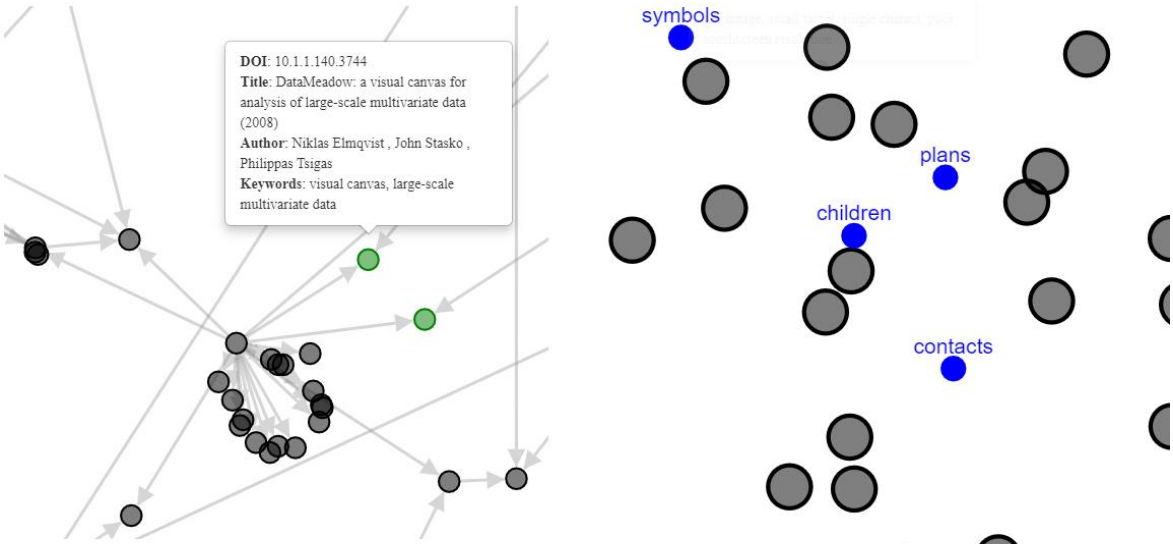
Далее были подготовлены входные данные для каждого из методов (Paper2vec, Cite2vec) и получены векторные представления документов, а также, в случае метода Cite2vec, векторные представления слов из контекстов цитирования.

Далее подготовка данных визуализации для каждого метода имеет различия. На полученных векторах документов метода Paper2vec применялись сначала иерархическая кластериза-

[2] https://citeseerx.ist.psu.edu/

ция (чтобы сохранить часть семантики полученных векторных представлений), а затем T-SNE [8] для уменьшения размерности до двух.

В случае метода Cite2vec сначала находились репрезентативные слова документов, которые отражают семантику тем, аналогично подходу оригинальной статьи [7]. А именно в цикле на каждой итерации определялся самый дальний документ от уже обработанных и для него находилось ближайшее слово с некоторыми ограничениями. Если ближайшее уже выбрано или расстояние до уже выбранных меньше некоторого порогового значения, то отбрасываем это слово и рассматриваем следующее ближайшее. При этом число слов ограничено сверху, чтобы выбранные слова были значимы для пользователя. Затем для выбранных слов применялся метод T-SNE, и все документы рассеивались в двумерном пространстве около самого близкого репрезентативного слова. Все расстояния считались по косинусной мере.




а б

*Рис. 2.* Визуализация документов:
*а* – на примере Paper2vec (в виде прозрачных дисков, описание документа во всплывающем окне, ребрами представлены отношения цитирования); *б* – на примере Cite2vec (ребра скрыты)
*Fig. 2.* Visualization of documents:
*a* – using the visualization Paper2vec as an example (in the form of transparent disks, document description in a pop-up window, the edges represent citation relationships);
*b* – using the visualization Cite2vec as an example (the edges are hidden)

## 2.2. Визуализация

Результаты обоих методов (Paper2vec и Cite2vec) визуализировались при помощи JavaScript библиотеки Cytoscape [3]. Визуализации обоих методов имеют как сходства, так и различия в функционале.

На обоих изображениях каждый документ представляется в виде прозрачного диска (рис. 2, *а*), чтобы хорошо были видны наложения одного документа на другой. В визуализации Cite2vec слова показываются в виде непрозрачной точки синего цвета меньшего размера, возле точки отображается само слово (рис. 2, *б*)

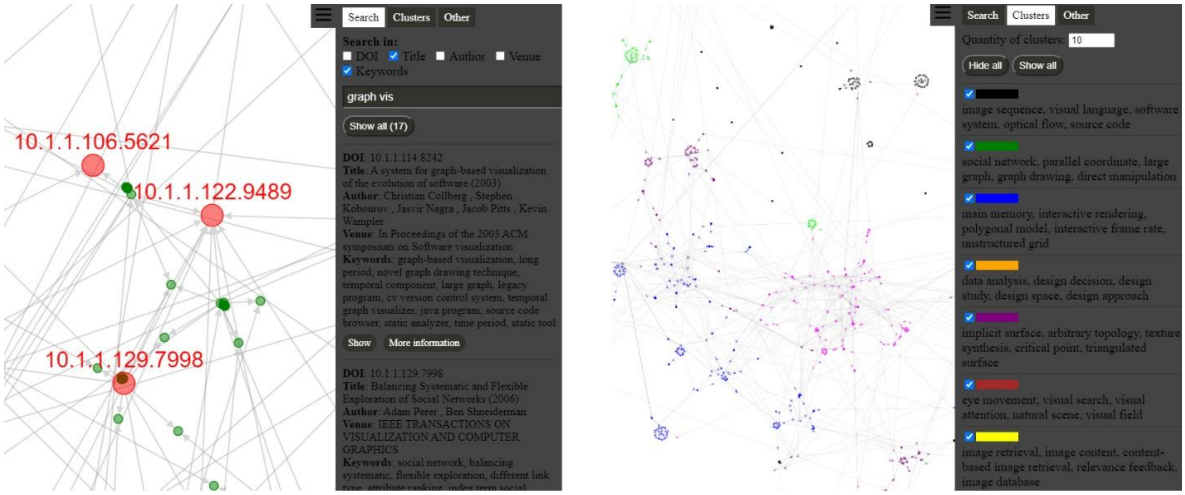
Диски документов Paper2vec из одного и того же кластера окрашиваются в одинаковый цвет. Следует отметить, что из-за аппроксимационной природы t-SNE документы из одного кластера не обязательно размещаются близко друг к другу на 2D-карте.

[3] https://js.cytoscape.org/

Пользователю доступен базовый набор методов взаимодействия с визуализациями, который позволяет перемещать точки, масштабировать изображение, смотреть описания документов во всплывающих окнах, выполнять поиск, и управлять выводом дополнительных элементов визуализации. Кроме этого в Paper2vec доступна настройка кластеризации.

### *2.2.1 Поиск*

Чтобы быстрее находить интересующие документы, был добавлен поиск статей по DOI, автору, названию, ключевым словам, месту публикации (рис. 3, *а*) и, в случае Cite2vec, по репрезентативным словам документов. Можно настраивать, где именно производится поиск. В результатах поиска показывается информация (DOI, название, автор, место публикации, ключевые слова, ссылка на статью в CiteSeerX) из подходящих документов с возможностью выделить все найденные документы (документы и слова в визуализации Cite2vec) на карте либо один, выбранный из списка результатов.




а б

*Рис. 3.* (*а*) Поиск статей в визуализации. Выделены статьи из списка результата поиска «graph vis» в названиях документов и ключевых словах. На примере визуализации Paper2vec.
(*б*) Иерархическая кластеризация векторов Paper2vec. Количество кластеров выбрано равное десяти
*Fig. 3.* (*a*) Search articles in visualization. Selected articles from the list of “graph vis” search results in document titles and keywords. Using the visualization Paper2vec as an example.
(*b*) Hierarchical clustering of Paper2vec vectors. The number of clusters was chosen equal to ten.

### *2.2.2 Кластеры*

Для сохранения исходной структуры векторов Paper2vec применяется иерархическая кластеризация до применения метода снижения размерности T-SNE. На изображении каждый кластер окрашен в уникальный цвет, можно скрывать / показывать каждый из кластеров, управлять их количеством (рис. 3, *б*). В настройках кластеризации возле каждого кластера отображаются пять самых часто встречаемых ключевых слов среди его документов, исключая те, которые содержатся более чем в $^1/_3$ кластеров.

### *2.2.3 Дополнительные элементы визуализации*

Для явного отражения семантики данных используются дополнительные элементы визуализации.

Отношения цитирования в обоих визуализациях показаны ориентированными ребрами от цитирующей статьи к цитируемой (см. рис. 2, *а*). Вы можете в любой момент отключить показ ребер, если они загромождают изображение. Кроме этого в визуализации Cite2vec дополнительными элементами являются слова, которые также можно скрыть.

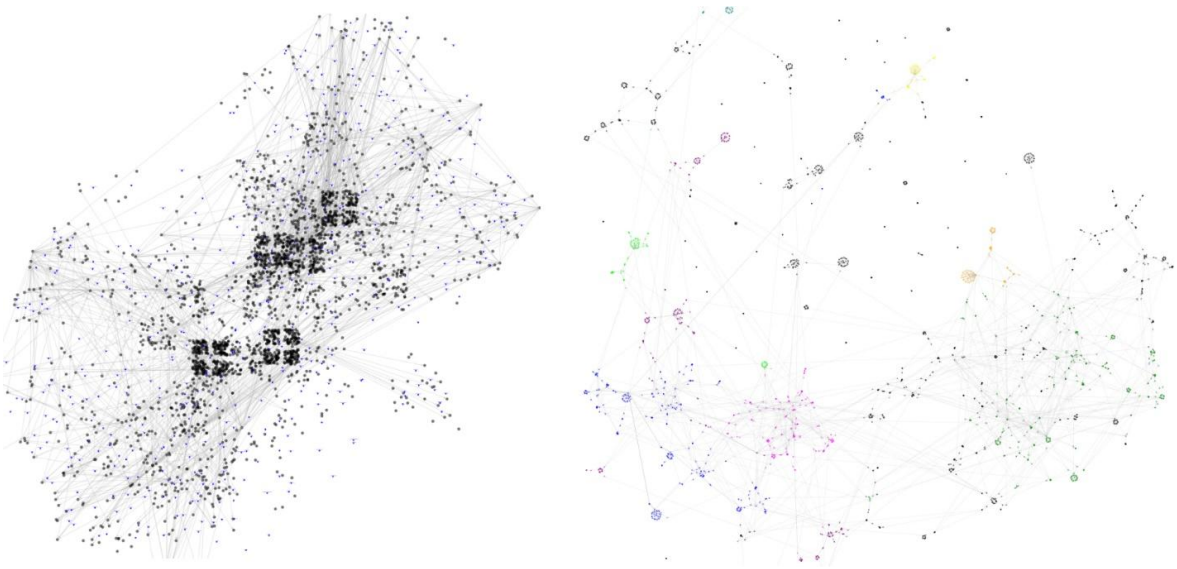
### 2.3 Проблемы при визуализации векторов большой размерности

Векторные представления коллекций научных публикаций могут быть очень эффективными, но их трудно интерпретировать. Методы уменьшения размерности теряют множественные признаки схожести / разности векторов, поэтому нужны дополнительные средства визуализации, которые позволяли бы лучше ухватывать семантику этих представлений, и, значит, улучшать эти представления.

В визуализации Cite2vec используются репрезентативные слова документов, которые несут семантику тем документов. В визуализации Paper2vec применяется иерархическая кластеризация перед уменьшением размерности, чтобы сохранить часть семантики векторных представлений.

### 2.4 Вывод

На рис. 4 представлены визуализации результатов методов Paper2vec и Cite2vec. В Cite2vec слова и документы моделируются в едином векторном пространстве, что позволяет добавить дополнительные элементы визуализации (документы в окружении слов, которые несут семантику тем документов; возможность связывать слова и документы при помощи простой арифметики). Для лучших результатов нужны большие наборы данных. В статье [7] авторами был показан функционал для ввода ключевых фраз. После их ввода документы перемещаются к их наиболее подходящей композиции репрезентативных слов и введенных пользователем фраз.



а б

*Рис. 4.* Визуализация результатов: *а* – Cite2vec; *б* – Paper2vec
*Fig. 4.* Visualization of results: *a* – Cite2vec; *b* –Paper2vec

Paper2vec получает представления для цитирующих и цитированных документов, в отличие от Cite2vec, где только документы из контекстов цитирования (только цитированные). В Paper2vec цитирующие и цитируемые статьи, как правило, расположены близко. На изображении отчетливо видно множество небольших скоплений документов.

## Заключение

В данной работе представлены два современных метода анализа текстовых коллекций – Paper2vec и Cite2vec, которые получают представления документов (и слов из контекстов цитирования в случае Cite2vec) в векторном пространстве. Для Paper2vec не важен язык, на котором написаны статьи, и он применим к базам данных, которые не поддерживают полные тексты документов, метод получает векторные представления для цитируемых и цитирующих документов. Для лучшей работы Cite2vec нужны большие наборы данных, данный метод получает векторные представления слов и документов в едином векторном пространстве, при этом документы оказываются в окружении слов, которые авторы используют в контекстах цитирования для описания данных документов.

Чтобы продемонстрировать работу методов были созданы визуализации результатов работы каждого метода. При визуализации такого вида информации (многомерных векторов) возникают проблемы с потерей семантики данных представлений в связи с уменьшением размерности. В статье были описаны примеры того, как можно сохранить часть семантики.

## Список литературы

1. **Апанович З. В.** Эволюция методов визуализации коллекций научных публикаций // Электронные библиотеки. 2018. Т. 21, № 1. С. 2–42.
2. **Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J.** Distributed Representations of Words and Phrases and Their Compositionality. *Advances in Neural Information Processing Systems*, 2013, vol. 26, pp. 3111–3119.
3. **Pennington J., Socher R. D., Manning C.** Glove: Global vectors for word representation. In: Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014), 2014, pp. 1532–1543. DOI 10.3115/v1/D14-1162
4. **Bojanowski P., Grave E., Joulin A., Mikolov T.** Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 2017, vol. 5, pp. 135–146. DOI 10.1162/tacl_a_00051
5. **Peters M., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L.** Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018, vol. 1, pp. 2227–2237. DOI 10.18653/v1/N18-1202
6. **Tian H., Zhuo H. H.** Paper2vec: Citation-Context Based Document Distributed Representation for Scholar Recommendation. ArXiv. abs/1703.06587, 2017.
7. **Berger M., McDonough K., Seversky Lee M.** Cite2vec: Citation-Driven Document Exploration via Word Embeddings. *IEEE Transactions on Visualization and Computer Graphics*, 2017, vol. 23, no. 1, pp. 691–700. DOI 10.1109/TVCG.2016.2598667
8. **Maaten L. van der, Hinton G.** Viualizing data using t-SNE. *Journal of Machine Learning Research*, 2008, vol. 9, pp. 2579–2605.

## References

1. **Apanovich Z. V.** Evolution of Visualization Methods for Research Publication Collections. *Elektronnye biblioteki*, 2018, vol. 21, no. 1, pp. 2–42. (in Russ.)

2. **Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J.** Distributed Representations of Words and Phrases and Their Compositionality. *Advances in Neural Information Processing Systems*, 2013, vol. 26, pp. 3111–3119.
3. **Pennington J., Socher R. D., Manning C.** Glove: Global vectors for word representation. In: Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014), 2014, pp. 1532–1543. DOI 10.3115/v1/D14-1162
4. **Bojanowski P., Grave E., Joulin A., Mikolov T.** Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 2017, vol. 5, pp. 135–146. DOI 10.1162/tacl_a_00051
5. **Peters M., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L.** Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018, vol. 1, pp. 2227–2237. DOI 10.18653/v1/N18-1202
6. **Tian H., Zhuo H. H.** Paper2vec: Citation-Context Based Document Distributed Representation for Scholar Recommendation. ArXiv. abs/1703.06587, 2017.
7. **Berger M., McDonough K., Seversky Lee M.** Cite2vec: Citation-Driven Document Exploration via Word Embeddings. *IEEE Transactions on Visualization and Computer Graphics*, 2017, vol. 23, no. 1, pp. 691–700. DOI 10.1109/TVCG.2016.2598667
8. **Maaten L. van der, Hinton G.** Viualizing data using t-SNE. *Journal of Machine Learning Research*, 2008, vol. 9, pp. 2579–2605.

## Информация об авторе

**Николай Игоревич Тихонов,** аспирант

## Information about the Author

**Nikolay I. Tikhonov**, Graduate Student

*Статья поступила в редакцию 20.07.2021*;
*одобрена после рецензирования 21.08.2021*; *принята к публикации 21.08.2021*
*The article was submitted 20.07.2021*;
*approved after reviewing 21.08.2021*; *accepted for publication 21.08.2021*