Герасименко Евгения Михайловна – Южный федеральный университет; e-mail: egerasimenko@sfedu.ru; г. Таганрог, Россия; тел.: 88634371651; кафедра систем автоматизированного проектирования; доцент.

Кравченко Даниил Юрьевич – e-mail: dkravchenko@sfedu.ru; тел.: 88634371651; кафедра систем автоматизированного проектирования; студент.

Кравченко Юрий Алексеевич – e-mail: yakravchenko@sfedu.ru; тел.: 88634371651; кафедра систем автоматизированного проектирования, д.т.н.; доцент.

Кулиев Эльмар Валерьевич – e-mail: ekuliev@sfedu.ru; тел.: 88634371651; кафедра систем автоматизированного проектирования; доцент.

Gerasimenko Evgeniya Mihailovna – Southern Federal University; e-mail: egerasimenko@sfedu.ru; Taganrog, Russia; phone: +78634371651; the department of computer aided design; associate professor.

Kravchenko Daniil Yurievich – e-mail: kravchenkodanil122@gmail.com; phone: +78634371651; the department of computer aided design; student.

Kravchenko Yury Alekseevich – e-mail: yakravchenko@sfedu.ru; phone: +78634371651; the department of computer aided design; dr. of eng. sc.; associate professor.

Kuliev Elmar Valerievich – e-mail: ekuliev@sfedu.ru; phone: +78634371651; the department of computer aided design, associate professor.

УДК 004.9 DOI 10.18522/2311-3103-2023-2-212-226

Ф.С. Булыга, В.М. Курейчик

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ВЕКТОРИЗАЦИИ ТЕКСТОВЫХ ДАННЫХ БОЛЬШОЙ РАЗМЕРНОСТИ

Представленная публикация посвящена обзору проблемы представления текстовой информации для последующего осуществления кластерного анализа в рамках обработки и управления информации большой размерности. Современные требования предъявляемые к аналитическим, поисковым и рекомендательным информационным системам демонстрируют слабую сформированность целостного решения, способного обеспечить достаточный уровень быстродействия и качества получаемых результатов в рамках функционирования текущего рынка информационных технологий. Поиск решения представленной проблемы влечет за собой необходимость в проведении объективного анализа существующих решений представления текстовой информации в векторном пространстве, с целью формирования целостного представления о достоинствах и недостатках анализируемых подходов, а также формированием критериев, позволяющих реализовать собственный подход, лишенный выявленных слабостей. Представленная работа является аналитической, и позволяет получить представление о современном состоянии и проработанности выявленной проблемы в рамках ограниченной предметной области. Кластеризация текстовых данных – автоматическое формирование подмножеств, элементами которых выступают экземпляры документов некоторой исследуемой, неструктурированной выборки фиксированной размерности. Данный процесс можно классифицировать как обучения без учителя, предполагающее, отсутствие эксперта, собственноручно присваивающего исходной выборке документов индексы классов. Однако, осуществление кластерного анализа текстовых данных без какой-либо предварительной обработки – невозможно. Для этого необходимо обеспечить стандартизацию и приведение входных данных к единому формату и виду. В рамках данного этапа осуществления кластерного анализа, в представленной публикации рассматриваются методы предварительной обработки текстовых данных. Новизна представленной публикации заключается в формировании теоретического базиса основных методов векторизации текстовых данных, путем систематизации и объективи-

зации выдвинутых предположений, путем проведения серии экспериментальных исследований. Основным отличием данной работы от уже опубликованных научных трудов является систематизации и анализ современных решений, а также выдвижение гипотезы об актуальности и эффективности собственного гибридизированного подхода предназначенного для векторизации текстовых данных.

Большие данные, кластеризация, кластерный анализ; интеллектуальный анализ данных; векторизация; кластеризация текстовых данных; k-means, Word2Vec; TF-IDF; Bag-of-Words.

F.S. Bulyga, V.M. Kureichik

COMPARATIVE ANALYSIS OF METHODS OF VECTORIZATION OF HIGH DIMENSIONAL TEXT DATA

The presented publication is devoted to an overview of the problem of presenting textual information for the subsequent implementation of cluster analysis in the framework of processing and managing high-dimensional information. Modern requirements for analytical, search and recommendation information systems demonstrate the weak formation of a holistic solution that can provide a sufficient level of speed and quality of the results obtained within the framework of the current information technology market. The search for a solution to the presented problem entails the need to conduct an objective analysis of existing solutions for representing textual information in vector space, in order to form a holistic view of the advantages and disadvantages of the analyzed approaches, as well as the formation of criteria that allow one to implement their own approach, devoid of identified weaknesses. The presented work is analytical, and allows you to get an idea of the current state and elaboration of the identified problem within a limited subject area. Clustering of text data is the automatic formation of subsets, the elements of which are instances of documents of some researched, unstructured sample of a fixed dimension. This process can be classified as unsupervised learning, which implies the absence of an expert who personally assigns class indices to the original sample of documents. However, the implementation of cluster analysis of text data without any pre-processing is impossible. To do this, it is necessary to ensure standardization and reduction of input data to a single format and form. Within the framework of this stage of the implementation of cluster analysis, the presented publication discusses methods for preprocessing text data. The novelty of the presented publication lies in the formation of the theoretical basis of the main methods of text data vectorization, by systematizing and objectifying the proposed assumptions, by conducting a series of experimental studies. The main difference of this work from the already published scientific works is the systematization and analysis of modern solutions, as well as the hypotheses about the relevance and effectiveness of our own hybridized approach designed for text data vectorization.

Big data; clustering; cluster analysis; data mining; vectorization; text data clustering; k-means; Word2Vec; TF-IDF; Bag-of-Words;

Введение. Большие данные – массивы информации большой размерности, обладающие преимущественно разветвленной и сложной структурой [1]. Подобные массивы данных встречаются в различных областях научной и коммерческой деятельности человека, связанных с необходимостью обработки и анализа входной информации. К таким областям можно отнести: анализ и обработка данных социальных сетей, экономика и менеджмент, медицина и биология, аналитика и статистика и т.п [2]. Применение вышеупомянутых массивов данных в информационных, аналитических, поисковых и т.п. систем, способствует возникновению проблемы, связанной с обработкой и анализом входящей информации, однако, за счет сложности формируемой структуры, достижение подобной цели – является нетривиальной и объемной задачей.

Для решения выше сформулированной проблемы предпринимались попытки создания альтернативных методов обработки, анализа и хранения информации в противовес уже классическим методам Business Intelligence и различным СУБД. Так, к современным подходам можно отнести внедрение и применение: NoSQL

[3], библиотечные решения проекта Hadoop [4], алгоритмы семейства MapReduce [5], различные алгоритмы машинного обучения [6, 7] и т.п. Представленный список алгоритмов не является конечным и может расширяться в зависимости от условий конкретно сформулированной задачи. Однако, невзирая на столь обширный каскад методов обработки, основной проблемой данных подходов является невозможности работы с текстовыми данными в явном виде. Кластеризация текстовых данных – автоматическое формирование подмножеств, элементами которых выступают экземпляры документов некоторой исследуемой, неструктурированной выборки фиксированной размерности. Данный процесс является одним из наиболее актуальных и значимых задач в области машинного обучения и интеллектуального анализа данных в частности. Актуальность сформулированной проблемы обусловлена неконтролируемым и постоянно возрастающим объемом генерируемой информации, представляемой преимущественно в текстовом формате. Основным источником подобного рода информации, на данный момент выступает глобальная сеть Интернет, позволяющая распространять данные в любом виде, без задания какой-либо стандартизированной структуры. Следствием вышеуказанного факта является большое количество разрозненной информации, скрывающей в себе полезные данные, обладающие ценностью для того или иного пользователя.

Для современных поисковых и рекомендательных систем основной проблемой выступает сложность извлечения полезных данных из общего входного потока информации, поступающего из открытых источников. В попытке решения данной проблемы представлено большое количество методов и алгоритмов структуризации и классификации данных, вне зависимости от их содержания или формата, однако, до сих пор, какого-либо универсального и эффективного решения найдено не было. Применительно к сформулированной задаче, на сегодняшний день можно выделить два основных класса алгоритмов, позволяющих представить текстовую информацию, в виде, пригодном для осуществления процесса кластеризации:

1. Алгоритмы основанные на представлении исходной текстовой информации в векторном формате многомерного пространства признаков, применяющие при этом некоторую заранее определенную метрику вычисления коэффициентов схожести элементов исследуемого множества.

2. Алгоритмы использующие в качестве признаков – коэффициенты частоты встречаемости терминов в текстах или массивах словосочетаний, с последующим определением и формированием кластеров на основе алгоритмов частичного обучения.

Основная задача представленной публикации заключается в систематизации и обобщении информации, касающейся методов представления текстовых данных, поскольку объективизированный выбор алгоритмов, позволяет повлиять на итоговую эффективность и точность результатов проводимой кластеризации.

В рамках данной работы представлены основные теоретические сведения о наиболее популярных и часто применяемых алгоритмах представления текстовых данных, с последующим сравнением коэффициентов влияния данных алгоритмов на итоговый результат кластерного анализа заданной выборки данных. В качестве основного метода кластеризации используется алгоритм дивизимной кластеризации k-means, поскольку представленный метод обладает хорошими показателями точности и эффективности для большинства задач подобного характера. Экспериментальные исследования в данной работе проводятся на выборке текстовых документов полученных из открытой научной библиотеки Elibrary, представленной 2 тыс. экземпляров научной направленности в гуманитарной и технической области.

Формальная постановка задачи кластеризации документов, содержащих текстовую информацию. Первоначально при постановке задачи кластеризации каскада документов содержащих текстовую информацию необходимо разделить данный процесс на два характерных этапа:

- ♦ Определение и формирование первичного множества кластеров ограниченного множества экземпляров исходной выборки;
- ♦ Итоговое группирование всех подмножеств и элементов в соответствии с полученными кластерами;

Однако, стоит отметить, что, для осуществления вышеперечисленных этапов кластеризации наиболее важным параметром выступает не только размерность исходной выборки документов, но также объективизированный выбор метрики схожести элементов необходимой для выполнения кластерного анализа.

Постановка задачи определения и формирования центроидов кластеров. Пусть задано некоторое конечное множество объектов $W = \{w_1, \dots, w_m\}$, где w_m – документ, содержащий текстовые данные. При этом каждый элемент множества W формализуется n-мерным признаковым вектором вида $(t_{g1}, \dots, t_{gn})$, где $g = 1, \dots, m$, или выступает точкой, принадлежащей n-мерному пространству признаков. Следующим этапом, необходимо задать метрику расстояния удовлетворяющую условию: $dist(x^i, \omega)$, где x^i – точка признакового пространства, характеризующая центроид i-го кластера; ω – некоторая точка, принадлежащая исследуемому множеству объектов. При этом, заданному количеству кластеров $\boldsymbol{l_k}$ требуется сопоставить $\boldsymbol{l_c}$ центроидов, с соблюдением условия минимального значения метрики расстояния R (1):

$$R = \sum_{i=1}^{l} \sum_{x \in L_i} dist(x^i, \omega) \to min, \quad (1)$$

где $\boldsymbol{L_i}$ – множество экземпляров $\boldsymbol{i}$-го класса.

Стоит отметить, что для решения данной задачи, предполагается, что пользователь самостоятельно определяет итоговое количество кластеров, основываясь на характерных чертах исследуемой предметной области, или отталкиваясь от предварительного эмпирического или теоретического опыта. В случае когда количество итоговых кластеров определить заранее невозможно, определение данного значения осуществляется экспериментальным путем. В данном исследовании параметр количества кластеров определяется на основе последовательного, попарного объединения схожих экземпляров объектов с последующим усреднением характерных параметров данных объектов, применение данного подхода обосновано автоматизацией процесса подбора количества кластеров, тем самым обеспечивая одинаковые условия работы тестируемых алгоритмов.

Алгоритм группирования экземпляров исследуемого множества. В общем случае алгоритм группирования экземпляров некоторого множества можно сопоставить с задачей распределения [8], формулируемой следующим образом: пусть задано некоторое конечное множество объектов $X = \{x_i\}$, при условии, что данное множество X возможно разделить на конечное количество подмножеств $\theta_t, t = 1, \dots, r, \bigcup_{t=1}^{r} \theta_t = X$. При этом, объекты принадлежащие данному множеству определяются значениями признаков: $x_j, j = 1, \dots, m$, где общность параметров признаков x_j характеризует представление объекта $I(w) = \{x_i, \dots, x_n\}$. В то же время информация о принадлежности объекта w какому-либо классу, определяется в форме вектора $\{I_1(w), \dots, I_p(w)\}$, где $I_p(w)$ содержит в себе данные о принадлежности экземпляра w к некоторому подмножеству θ_p (2):

$$I_p(w) = \begin{cases} 1, \text{если } w \in \theta_p \\ 0, \text{если } w \notin \theta_p \\ \Delta, \text{если неопределенность} \end{cases} \quad (2)$$

Принадлежность того или иного объекта *w* к некоторому классу θ_p определяется основываясь на сравнении значений метрики расстояния между экземплярами множеств и самими множествами.

Выбор метрики расстояния. Первоначально, необходимо сформулировать полное и всеобъемлющее определение термина «метрики расстояния». Метрика расстояния – величина $R(X_i, X_j)$, удовлетворяющая следующим положениям:

- $R(X_i, X_j) \geq 0$ для всякого X_i и X_j из E_z (расстояние не должно быть отрицательным);
- $R(X_i, X_j) = R(X_j, X_i)$ (симметричность расстояния, вне зависимости от порядка точек измерения);
- $R(X_i, X_j) \leq R(X_j, X_i) + R(X_k, X_j)$, где X_k, X_i, X_j – векторы принадлежащие E_z (удовлетворение правилу неравенства треугольника);
- $R(X_i, X_j) = 0$ тогда и только тогда, когда $X_i = X_j$ (удовлетворение правилу неразличимости тождественных объектов);

Параметр $R(X_i, X_j)$ для исследуемых X_i и X_j именуется как величина расстояния между X_i и X_j и эквивалентна расстоянию между U_i и U_j соответственно выбранным атрибутам $G=(G_1, \ldots, G_z)^T$ [9].

В классической и современной литературе посвященной исследованиям в области интеллектуального анализа данных, в частности алгоритмам и методам кластерного анализа, определено более пятидесяти различных метрик, предназначенных для вычисления параметра метрики расстояния. Наиболее классическим и популярным методом вычисления расстояния является «евклидово расстояние» или «евклидова метрика», вычисляемая в соответствии с формулой [10] (3):

$$d(X_i, X_j) = \left(\sum_{k=1}^{z} (x_{ki} - x_{kj})^2\right)^{1/2}. \quad (3)$$

При этом метрика расстояния «Евклида» достаточно схожа с метрикой расстояния «Минковского» [11], с отличием в степени. В общем случае в метрике «Минковского» степень уравнения определяется при помощи параметра ***p.*** Обобщенно, данную метрику можно представить в виде (4):

$$d(X_i, X_j) = \left(\sum_{k=1}^{z} (x_{ki} - x_{kj})^2\right)^{1/p}. \quad (4)$$

Однако, несмотря на вышеописанные подходы вычисления расстояния, также существуют и другие, часто встречаемые и применяемые в исследованиях метрики: «Расстояние городских кварталов» [12], «Махаланобиса» [13] и т.п.

Также, следует отметить, важность объективного выбора метрики расчета расстояния между элементами исследуемого множества, поскольку корректность выбранного подхода влияет на конечные результаты скорости работы предлагаемого решения, а также на точность получаемых значений. Исходя из представленного краткого обзора метрик вычисления расстояния между объекта, в качестве основного метода расчета в проводимых исследованиях выбрано «Евклидово расстояния».

Векторизация. Для осуществления кластерного анализа применительно к текстовой информации, существующие алгоритмы машинного обучения оперируют в пространстве числовых атрибутов, т.е. для их функционирования, необходимо на вход предоставить двумерный массив данных, представляющий из себя матрицу, столбцами которой являются признаки данных, при этом строки матрицы, содержат в себе конкретные экземпляры данных. Векторизация – процесс представления исходной текстовой информации в векторном пространстве признаков для осуществления дальнейшей кластеризации [14]. Выполнение данного процесса – первый и наиболее важный этап анализа информации представленной в текстовом формате. Подобное преобразование документов, позволяет анализировать, обрабатывать и создавать экземпляры данных, с которыми впоследствии смогут

работать существующие подходы кластерного анализа. К преимуществам подхода векторизации, также можно отнести то, что изначально на вход поступают документы различной размерности, однако, формируемые в ходе преобразования векторы, всегда будут одинаковой длины, что позволяет стандартизировать входные данные и оптимизировать дальнейший процесс кластеризации.

В векторном пространстве каждое слово принадлежащие обрабатываемому тексту представляется в виде признаков, в последствии формирующих признаковое пространство, поступающее на вход алгоритму кластеризации. Таким образом, осуществляется переход от отдельных предложений и словосочетаний к точкам в многомерном семантическом пространстве, расположенным на расстоянии друг от друга, в соответствии со смысловой схожестью анализируемых предложений. Следовательно, при векторизации текстовых документов, с последующим вычислением величины меры схожести множества точек, возможно удостовериться или опровергнуть принадлежность того или иного текста к имеющимся кластерам.

Частота векторизации (Bag-of-Words). Одним из способов векторизации входного текстового массива данных является вычисление частоты появления термина в каждом предложении с последующим определением зависимости данного значения к набору терминов исходного текста [15]. Первоначально, необходимо сформировать словарь терминов, принадлежащих исследуемому массиву данных, с присвоением каждому слову собственного индекса. Данное действие необходимо выполнить для последующей векторизации предложений, при этом итоговый вектор каждого предложения впоследствии равен длине сформированного словаря, включающего в себя количество повторений того или иного термина в конкретном предложении. Массив тестовых предложений для выполнения частотной векторизации представлен на рис. 1.

```
source_text_array = [
'Кластеризация - задача группировки множества объектов на подмножества (кластеры) таким
образом, чтобы объекты из одного кластера были более похожи друг на друга, чем на объекты
из других кластеров по какому-либо критерию',

'Кластерный анализ - многомерная статистическая процедура, выполняющая сбор данных,
содержащих информацию о выборке объектов, и затем упорядочивания объекты в сравнительно
однородные группы',

'Кластер - объединение нескольких однородных элементов, которое может рассматриваться
как самостоятельная единица, обладающая определенными свойствами',

'Кластер - это графическая форма организации информации, когда выделяются основные
смысловые единицы, которые фиксируются в виде схемы обозначением всех связей между ними']
```

Рис. 1. Массив тестовых предложений

После формирования массива исследуемых предложений, для осуществления частотной векторизации, возможно использование готовой библиотеки scikit-learn, обладающей соответствующими методами, такими как «countvectorizer».

Результатом выполнения частотной векторизации выступает словарь уникальных терминов, принадлежащих исходному множеству предложений. В дальнейшим возможно осуществление преобразования полученного множества терминов в векторы исходных предложений, отображающие частоту вхождений того или иного элемента сформированного словаря в каждом конкретном предложении исходного массива.

Подобный подход является достаточно распространенным в области предварительной обработки текстовых массивов данных для осуществления последующей кластеризации. В технической литературе подобные методы называются «Bag-of-Words» (т.е. каждый текстовый документ исходного массива текстовых данных представляется в виде вектора).

Из недостатков данного подхода стоит отметить зависимость размерности формируемых векторов исходного множества от размерности текстов принадлежащих входному массиву данных. Таким образом, формируемые векторы будут обладать высокой размерностью и сильной разрежённостью, как следствие увеличивая объемы памяти для их хранения, а также повышая вычислительные затраты. Для исправления подобного недостатка необходимо осуществить предварительную обработку текстовых данных, устранив «стоп-слова», а также осуществив процесс лемматизации и т.п. Выполнение подобных действий позволит сократить размерность исследуемого множества, как следствие сократив размерность итоговых векторов.

Метрика TF-IDF. Метрика TF-IDF – применяется для вычисления коэффициента весомости некоторого термина, принадлежащего исходному множеству терминов $\boldsymbol{X}$ [16]. Данная метрика определяется для каждого термина принадлежащего исследуемой выборке, при этом, чем выше показатель данной метрики для некоторого термина, тем весомей данное слово является в контексте исследуемого массива или множества.

Основная идея данного подхода заключается в нормализации и структуризации частоты слов некоторого текстового документа, с учетом множества терминов всего исследуемого множества текстовых файлов. Таким образом, если некоторый термин j часто встречается в некотором документа X_m, однако редко встречается в оставшихся текстовых документах, тогда, термин j обладает высоким коэффициентом значимости для текста X_m и является более приоритетным в сравнении с остальными терминами принадлежащими общему корпусу документов.

Метрика TF-IDF рассчитывается в соответствии со следующими положениями: первоначально осуществляется вычисление параметра TF – величина отношения количества вхождений исследуемого термина в рамках одного конкретного множества (документа). Данный параметр рассчитывается в соответствии с формулой (5):

$$TF = \frac{m_j}{\sum_n m_n}, \tag{5}$$

где n_t – количество вхождений исследуемого термина j; $\sum_n m_n$ – общее количество терминов в исследуемом документе.

Параметр IDF – величина отношения количества вхождений исследуемого термина в рамках всего корпуса терминов исходного множества документов. Данный параметр при этом рассчитывается по формуле (6):

$$IDF = log\frac{I}{DF_j}, \tag{6}$$

где I – количество исследуемых документов в исходном множестве; DF – количество документов принадлежащих исходному множеству где встречается исследуемый термин j.

После вычисления двух вышеописанных параметров, осуществляется вычисление метрики TF-IDF в соответствии с формулой (7):

$$TF - IDF = TF * IDF. \tag{7}$$

В рамках демонстрации работы данного подхода на исходном множестве предложений представленных в предыдущем методы осуществим вычисление метрики TF-IDF. Результаты полученные метрикой TF-IDF приведены на рис. 2.

	упорядочивающая	задача	и	-	кластерный	кластеризация	объектов	чем
0	0.000000	0.044719	0.000000	0.0	0.000000	0.044719	0.044719	0.044719
1	0.013076	0.000000	0.013076	0.0	0.013076	0.000000	0.000000	0.000000
2	0.013076	0.000000	0.013076	0.0	0.013076	0.000000	0.000000	0.000000
3	0.013076	0.000000	0.013076	0.0	0.013076	0.000000	0.000000	0.000000

	содержащих	группировки	образом	по	данных	множества	друга	из
0	0.000000	0.044719	0.044719	0.044719	0.000000	0.044719	0.044719	0.089438
1	0.013076	0.000000	0.000000	0.000000	0.013076	0.000000	0.000000	0.000000
2	0.013076	0.000000	0.000000	0.000000	0.013076	0.000000	0.000000	0.000000
3	0.013076	0.000000	0.000000	0.000000	0.013076	0.000000	0.000000	0.000000

	Объектов,	одного	(кластеры)	друг
0	0.000000	0.044719	0.044719	0.044719
1	0.013076	0.000000	0.000000	0.000000
2	0.013076	0.000000	0.000000	0.000000
3	0.013076	0.000000	0.000000	0.000000

Рис. 2. Результаты полученные метрикой TF-IDF

К преимуществам данного подхода можно отнести игнорирование проблемы удаления стоп-слов, поскольку данные термины зачастую в большом количестве присутствуют во всех исследуемых массивах текстовых документах, тем самым присваиваемые значения метрики TF-IDF данным терминам будут достаточно низкими.

К недостаткам данного подхода можно отнести:

1. Отсутствие возможности отслеживания уровней вложенности документов.
2. Ненадежность первоначальной идеи данного подхода. Значение коэффициента встречаемости, особенно в русском языке достаточно низкое, за счет большого количества синонимов.
3. Получаемая оценка является статичной.

Word2Vec. Основной подход вышеизложенных методов заключается в представлении экземпляра исходного множества в векторной форме, тем самым формируя при этом единый вектор объектов, однако, зачастую оправданно осуществить векторизацию основанную не только на внутреннем сходстве элементов некоторого документа, принадлежащего исходному множеству файлов, но также учесть показатели внешнего сходства между экземплярами, в контексте самого векторного пространства.

Представленные ранее методы формируют векторы исключительно с положительными элементами, тем самым не позволяя осуществлять сравнение документов, не обладающих общими терминами [17]. Данное ограничение обосновано тем, что в случае если значение косинуса угла между парой векторов равно единице, это не позволит однозначно констатировать схожесть исследуемых документов.

В ситуации, когда схожесть документов обладает высоким приоритетом для проводимого исследования, возможно провести векторизацию текстовых данных при помощи подхода распределенного представления. В данном подходе вектор способен не только отобразить позиции терминов, но также сформировать каскад признаков, способных задать сходство тех или иных слов, при этом, размерность пространства признаков и как следствие длина вектора определяется качеством проводимого обучения алгоритма, и не зависима от размерности исследуемого документа.

В данном разделе публикации рассматривается алгоритм, за основу которого взят подход распределенного представления – Word2Vec. Word2Vec – нейронная сеть, основное назначение которой заключается в осуществлении обработки и анализа текстовой информации. Данная нейронная сеть разработана группой исследователей компании Google в 2013 году [18]. Невзирая на то, что ранее уже существовали схожие векторно-семантические методы, Word2Vec является первым наиболее популярным и часто применяемым алгоритмом представления текстовой информации в векторном пространстве. Данная нейронная сеть представляет из себя гибрид двух моделей обучения: K-Skip-n-Gram и Continuous Bag of Words.

Continuous Bag of Words (CBOW) – нейросетевая архитектура, предназначенная для определения основных принципов обучения и хранения представления терминов исследуемого множества. Данная модель выступает подходом Bag-of-Words, способного учитывать две пары ближайших соседей исследуемого термина (пара предшествующих и пара последующих терминов), пренебрегая при этом порядком следования.

K-Skip-n-Gram – нейросетевая архитектура, позволяющая формировать предсказания контекста, основываясь на исследуемом термине. В общем виде, данный подход можно описать следующим образом: пусть рассматриваемая модель представляет из себя последовательность размерности ***n,*** в которой всякий принадлежащий ей элемент располагается на расстоянии ***k*** от соседнего элемента.

В общем виде метод Word2Vec можно представить в виде гибридизированной нейронной сети, основанной на двух основных нейронных архитектурах. Визуализация данного подхода представлена на рис. 3.

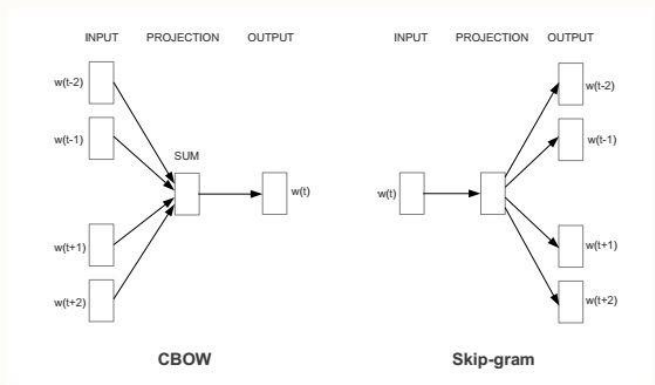

Рис. 3. Общая архитектура метода Word2Vec

Данная нейронная сеть в качестве входных данных принимает корпус текстовых документов размерности X в котором каждому термину t, $t \in X$ сопоставляется вектор v_t, при этом, алгоритм работы данного подхода, выглядит следующим образом: первоначально выполняется формирование словаря терминов, включающего в себя термины некоторого исследуемого документа, далее формируется векторное пространство слов, основанное на контекстной близости терминов, вычисляемых при помощи расчета частоты встречаемости терминов в тексте. Близость терминов в данном алгоритме определяется за счет косинусного сходства, и рассчитывается в соответствии с формулой (8):

$$similarity(A,B) = \cos(\vartheta) = \frac{A*B}{||A||*||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2}\sqrt{\sum_{i=1}^{n} {B_i}^2}}. \quad (8)$$

С целью ускорения обучения представленных моделей, в данной работе применяется модификация **softmax** – **hierarchical softmax** и **negative sampling.** Данные модификации позволяют обеспечить прирост вычисления значения распределения вероятностей.

Однако, для полноценного функционирования представленного метода, необходимо осуществить обучение нейронной сети. Процесс обучения алгоритма Word2Vec осуществляется следующим образом:

1. Выполняем предварительную обработку документа ***T*** , исключая из данных «шумовые» символы (знаки пунктуации и т.п.).

2. Формируем словарь терминов W основываясь на полученном тексте $\tilde{T}$;

3. Для каждого слова $w_i \in T$ формируем контекст, т.е. набор слов $C_i \subset T$, удалённых от w_i не более чем на s позиций в последовательности слов T. $C_i = \{w_j \in T: (i-s) \leq j \leq (i+s),\ j \neq i\}$.

4. Осуществляем унитарное кодирование (one-hot encoding) словаря W, т.е. каждому слову $w_i \in W$ ставится в соответствие вектор $u_i \in U$, длина вектора u_i равна размеру словаря W, позиция единицы в векторе u_i соответствует номеру слова в словаре W.

5. Заменяем слова в тексте T и контекстах C соответствующими кодами P и Q принадлежащими U.

Таким образом, получаем два множества – кодированный текст P и наборы кодированных слов контекста Q.

Для эффективной генерации векторов слов необходимо применить модель Skip-gram, поскольку данная модель обладает улучшенными показателями к обучению в сравнении с моделью CBOW (без учета скорости вычисления). На вход модели skip-грамм подается одно слово w_i , а на выходе мы получаем слова w_i в контексте $\{w_{O,1}, \dots, w_{O,c}\}$, определяемом размером окна слов. После обучения каждому слову сопоставляется вектор, с последующим построением матрицы большой размерности, где каждая строка представляет каждый пример обучения, а столбцы выступают сгенерированными векторами слов. Как следствие, термин обладает несколькими степенями подобия.

Алгоритм кластеризации. В качестве основного алгоритма кластерного анализа применяемого в рамках проводимого исследования выбран алгоритма – K-Means. Данный алгоритм является одним из наиболее популярных и модифицируемых подходов для осуществления кластерного анализа. Основная идея данного подхода заключается в произвольном разбиении исходного множества элементов на группы (кластеры), с последующим итеративным перерасчетом центроидов для каждого кластера, сформированного на предыдущей итерации алгоритма.

Математическое описание алгоритма формулируется так: пусть дано некоторое множество элементов X разделяемое на k подмножеств $X = \{G_1, \dots, G_k\}$, где G_k – подмножество исходного множества X (кластер), таким образом, чтобы суммарное значение квадратов расстояний от экземпляра кластера к его центроиду являлось минимальным [19] (9):

$$\arg min_G \sum_{i=1}^{k} \sum_{x \in G_i} \rho(x, \mu_i)^2, \tag{9}$$

где, μ_i – центроиды кластеров, $i \in \{1, \dots, k\}$; $\rho(x, \mu_i)$ – метрика вычисления расстояния между x и μ_i;

Этапы выполнения данного алгоритма можно представить в виде следующей последовательности:

1. *Формирование кластеров*
   Случайным образом выбирается некоторое количество точек $\mu_i, i \in \{1, \dots, k\}$, равное итоговому количеству кластеров. Данные точки рассматриваются в качестве начальных центроидов будущих кластеров: $\mu_i^{(0)} = \mu_i, i \in \{1, \dots, k\}$;
2. *Выполнение группирование векторов в соответствующие кластеры*
   $\forall x_i \in X, i = 1, \dots, n: x_i \in G_j \Leftrightarrow j = \arg min_k \rho(x_i, {\mu_k}^{(t-1)})^2$.
3. *Пересчет центроидов кластеров*
   На данном этапе пересчет центроидов кластеров осуществляется в соответствии с формулой вычисления центров масс (10):

$$\mu_i = \frac{1}{|G|} \sum_{x \in G} x. \tag{10}$$

4. *Проверка условия остановки алгоритма*

   Если $\exists i \in \overline{1,k}: \mu_i^{(t)} \neq \mu_i^{(t-1)}$ тогда $t = t + 1$, необходимо вернуться к шагу 2, в противном случае, остановить выполнение алгоритма;

Представленная выше информация касающаяся основных аспектов работы алгоритма k-means позволяет получить полное представление о базовых принципах и методах работы данного алгоритма, однако стоит также упомянуть некоторые свойства данного алгоритма. Рассматриваемый алгоритм является итерационным, т.е. численность итераций алгоритма не фиксирована и зависит исключительно от изначальной локализации экземпляров исследуемого множества в пространстве, величины $\boldsymbol{k}$ (заданного количества кластеров), а также от изначального значения метрики расстояния центроидов кластера $\mu_1, \ldots, \mu_k$. Исходя из вышеперечисленного каскада параметров, можно констатировать высокую вариативность качества получаемых результатов по окончанию работы данного алгоритма. В случае неудачного подбора стартовых параметров итерационный процесс способен сойтись к локальному оптимуму, в связи с этим данный алгоритм является слабо детерминированным.

Однако, несмотря на вышесказанное, к преимуществам данного алгоритма можно отнести: хорошее соотношение показателей эффективности, относительно к скорости выполнения; высокий уровень качества получаемых результатов кластеризации; способность работы алгоритмы в параллельном режиме; перспективы и широкие возможности модификации и модернизации;

Экспериментальные исследования. Основная задача данного исследования заключается в определении наиболее оптимального и эффективного подхода для векторизации текстовых данных с последующим осуществлением кластерного анализа. В данном блоке представленной публикации рассматриваются результаты экспериментальных исследований, позволяющие оценить итоговые показатели скорости работы реализованных алгоритмов, а также конечные результаты кластеризации.

В качестве тестового набора данных из открытой научной библиотеки Elibrary получены публикации технической и гуманитарной направленности в количестве 2 тыс. экземпляров, в качестве основного метода кластеризации выступает алгоритм K-means. Для обучения Word2Vec применялся набор данных, полученный из открытых источников [20], данный каскад документов включает в себя новостные публикации различной тематики, в том числе публикации научной направленности.

Основными методами векторизации текстовых данных сравниваемыми в данном исследовании выступают методы, рассмотренные ранее в публикации: TF-IDF, Bag-of-Words и Word2Vec. Основными критериями оценки эффективного того или иного подхода выступают показатели скорости формирования векторного представления исследуемого множества документов, а также показатели эффективности кластеризации, выраженные в процентном соотношении.

Экспериментальные исследования проводимые в данной работе осуществлялись следующим образом: на вход методам векторизации текстовых данных передавались сформированные множества текстовых документов, предварительно проанализированные с помощью модуля предварительной обработки [19]. Далее, преобразованные множества передаются алгоритму k-means, с помощью которого формируются кластеры документов, основываясь на информации, содержащейся внутри документов.

Экспериментальные исследования проводились на наборах данных различной размерности, что позволяет обеспечить анализ эффективности работы того или иного подхода на массивах данных большой и малой размерности. Результаты скорости выполнения векторизации текстовых документов приведены в табл. 1.

Таблица 1

Скорость векторизации входного множества документов

Размер данных	Скорость векторизации (мс.)		
	TF-IDF	Word2Vec	Bag-of-Words
2000	$4 * 10^3$	$9 * 10^3$	$7 * 10^3$
1500	$12 * 10^2$	$8 * 10^2$	$10 * 10^2$
1000	$5 * 10^2$	$6 * 10^2$	$4 * 10^2$
500	$6 * 10^2$	$4 * 10^2$	$8 * 10^2$
250	$5 * 10^2$	$3 * 10^2$	$6 * 10^2$

Исходя из данных представленных в табл. 1 можно констатировать эффективность предложенного алгоритма Word2Vec при векторизации множества документов большой размерности, однако при этом, во множествах малой размерность, показатели скорости работы выше у подхода TF-IDF. Как говорилось ранее, еще одним, немаловажным показателем при проведении данной серии экспериментов является оценка точности проводимой кластеризации. Результаты оценки точности кластеризации представлены в табл. 2.

Таблица 2

Показатели эффективности кластеризации

Размер данных	Эффективность кластеризации (%)		
	k-means (TF-IDF)	k-means (Word2Vec)	k-means (Bag-of-Words)
2000	60.127	64.452	64.125
1500	68.462	69.784	68.584
1000	71.985	75.987	76.123
500	72,654	74.254	71.965
250	75,771	77.320	75.786

Основываясь на данных представленных в табл. 2, можно констатировать влияние векторизации текстовых данных на конечный результат кластеризации. В среднем метод Word2Vec позволил повысить процент точности предсказания от 2 до 4 %.

Заключение. Проведенные исследования посвящены рассмотрению и анализу проблемы векторизации текстовых данных для последующего осуществления кластерного анализа. Итогом написания данной работы является определение основных положения и задач существующих методов векторизации текстовых данных. Серия экспериментальных исследований демонстрирует эффективность применения подхода Word2Vec для осуществления векторизации данных, поскольку показатели скорости и качество итоговой кластеризации, в сравнении с иными методами оказались выше.

Резюмируя все вышесказанное, можно выделить следующие основные положения: подходы и алгоритмы предварительной обработки текстовой информации оказывают непосредственное влияние на конечный результат кластеризации; подбор методов векторизации текстовых данных, необходимо осуществлять ответственно, в силу оказания влияния на конечный результат работы; гибридизация подхода Word2Vec демонстрирует прирост показателей по всем выделенным пунктам оценки качества, что подтверждает предположения об увеличении показателей эффективности.

Новизна представленной работы заключается в систематизации и обработки информации касающейся современных и популярных методов кластерного анализа, а также в представлении авторского подхода к модернизации, уже существующих решений.

Благодарность:

Исследование выполнено за счет гранта Российского научного фонда № 22-21-00316, https://rscf.ru/project/22-21-00316/ в Южном федеральном университете.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Parkhomenko D.A.* Data vizualization makes sense of Big data // Big Data and Advanced Analytics. – 2021. – No. 7-1. – P. 416-417.
2. *Есауленко А.С., Никоненко Н.Д.* Большие данные. Реальность и перспективы // Управление инновациями: теория, методология, практика. – 2016. – № 17. – С. 74-79.
3. *Гродель Ю.В., Лагун Д.А.* Проблема Big Data и NoSQL подход к её решению // Наука, образование, общество: тенденции и перспективы: Сб. научных трудов по материалам Международной научно-практической конференции: в 5 ч. – М., 2014. – С. 31-32.
4. *Абашин В.Г., Жолобова Г.Н., Горохова Р.И., Никитин П.В., Семенов А.М., Зараев Р.Э.* Подготовка студентов к работе с большими данными с применением кластера Hadoop // Современные наукоемкие технологии. – 2022. – № 6. – С. 78-82.
5. *Денисенко В.В., Евтеева К.С., Савченко И.И., Скрыпников А.В., Берестовой А.* Распределенные вычислительные модели Mapreduce и Mapreduce-алгоритма // Системный анализ и моделирование процессов управления качеством в инновационном развитии агропромышленного комплекса: Матер. V Международной научно-практической конференции,в рамках реализации Ассоциации «Технологическая платформа «Технологии пищевой». – 2021. – С. 319-326.
6. *Мамедова Г.А., Зейналова Л.А., Меликова Р.Т.* Технологии больших данных в электронном образовании // Открытое образование. – 2017. – Т. 21, № 6. – С. 41-48.
7. *Трофимов И.Е.* Распределенные вычислительные системы для машинного обучения // Информационные технологии и вычислительные системы. – 2017. – № 3. – С. 56-69.
8. *Журавлёв Ю.И.* Об алгебраическом подходе к решению задачи распознавания или классификации // Проблемы кибернетики. – 1978. – Т. 33. – С. 5-68.
9. *Рабинович Ю.И.* Кластерный анализ детализации телефонных переговоров // Системы и средства информатики. – 2007. – Т. 17, № 1. – С. 52-78.
10. *Лушников Н.Д., Исмагилова А.С.* Евклидово расстояние как основа программного комплекса по многофакторной биометрической аутентификации // Математическое моделирование процессов и систем: Матер. XI Международной молодежной научно-практической конференции. – Стерлитамак, 2021. – С. 53-55.
11. *Рузибаев О.Б., Эшметов С.Д.* Исследование и анализ алгоритмов на основе нечеткого метода k ближайших соседей с применением различных метрик при диагностике рака молочной железы // Наука и мир. – 2016. – № 5-1 (33). – С. 102-107.
12. *Ле Минь Таун, Шукуров И.С., Нгуен Тхи Май.* Исследование интенсивности городского острова тепла на основе городской планировки // Строительство: наука и образование. – 2019. – Т. 9, № 3. – С. 54-65.
13. *Шумская А.О.* Оценка эффективности метрик расстояния Евклида и расстояния Махаланобиса в задачах идентификации происхождения текста // Доклады Томского государственного университета систем управления и радиоэлектроники. – 2013. – № 3 (29). – С. 141-145.
14. *Шерстнев П.А.* Исследование методов векторизации документов на основе векторизации слов // Актуальные проблемы авиации и космонавтики: Сб. материалов VII Международной научно-практической конференции, посвященной Дню космонавтики: в 3 т. – Красноярск, 2021. – С. 216-218.
15. *Tian L., Huang R., Wang Y.* Metric learning in codebook generation of bag-of-words for person re-identification // ICPRAM 2019 - Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods. – Prague. 2019. – P. 298-306.

16. *Булыга Ф.С., Курейчик В.М.* Алгоритмы агломеративной кластеризации применительно к задачам анализа лингвистической экспертной информации // Известия ЮФУ. Технические науки. – 2021. – № 6 (223). – С. 73-88.
17. *Nartsev A.D.* Text classification by means of word2vec model and a convolutional neural network // Presenting Academic Achievements to the World. Natural Science: Матер. X научной конференции молодых ученых. Саратов, 16 апреля 2019 года. – Saratov, 2020. – Vol. 9. – P. 71-77.
18. *Левченко С.В.* Разработка метода кластеризации слов по смысловым характеристикам с использованием алгоритмов Word2Vec // Новые информационные технологии в автоматизированных системах. – 2017. – № 20. – С. 44-46.
19. *Булыга Ф.С., Курейчик В.М.* Кластеризация корпуса текстовых документов при помощи алгоритма k-means // Известия высших учебных заведений. Северо-Кавказский регион. Технические науки. – 2022. – № 3 (215). – С. 33-40.
20. Newsgroups // Qwone: [сайт]. – 2019. – URL: http://qwone.com/~jason/20Newsgroups/ (дата обращения: 15.08.2022).

REFERENCES

1. *Parkhomenko D.A.* Data vizualization makes sense of Big data, *Big Data and Advanced Analytics,* 2021, No. 7-1, pp. 416-417.
2. *Esaulenko A.S., Nikonenko N.D.* Bol'shie dannye. Real'nost' i perspektivy [Big data. Reality and prospects], *Upravlenie innovatsiyami: teoriya, metodologiya, praktika* [Management of innovations: theory, methodology, practice], 2016, No. 17, pp. 74-79.
3. *Grodel' Yu.V., Lagun D.A.* Problema Big Data i NoSQL podkhod k ee resheniyu [The problem of Big Data and NoSQL approach to its solution], *Nauka, obrazovanie, obshchestvo: tendentsii i perspektivy: Sb. nauchnykh trudov po materialam Mezhdunarodnoy nauchno-prakticheskoy konferentsii* [Collection of scientific papers based on the materials of the International Scientific and Practical Conference]: in 5 part. Moscow, 2014, pp. 31-32.
4. *Abashin V.G., Zholobova G.N., Gorokhova R.I., Nikitin P.V., Semenov A.M., Zaraev R.E.* Podgotovka studentov k rabote s bol'shimi dannymi s primeneniem klastera Hadoop [Preparing students to work with big data using the Hadoop cluster], *Sovremennye naukoemkie tekhnologii* [Modern high technologies], 2022, No. 6, pp. 78-82.
5. *Denisenko V.V., Evteeva K.S., Savchenko I.I., Skrypnikov A.V., Berestovoy A.* Raspredelennye vychislitel'nye modeli Mapreduce i Mapreduce-algoritma [Distributed computational models of Mapreduce and Mapreduce-algorithm], *Sistemnyy analiz i modelirovanie protsessov upravleniya kachestvom v innovatsionnom razvitii agropromyshlennogo kompleksa: Mater. V Mezhdunarodnoy nauchno-prakticheskoy konferentsii,v ramkakh realizatsii Assotsiatsii «Tekhnologicheskaya platforma «Tekhnologii pishchevoy»* [System analysis and modeling of quality management processes in the innovative development of agro-industrial complex: Materials of the V International Scientific and Practical Conference, within the framework of the Association "Technological Platform" Food Technologies"], 2021, pp. 319-326.
6. *Mamedova G.A., Zeynalova L.A., Melikova R.T.* Tekhnologii bol'shikh dannykh v elektronnom obrazovanii [Big data technologies in e-education], *Otkrytoe obrazovanie* [Open education], 2017, Vol. 21, No. 6, pp. 41-48.
7. *Trofimov I.E.* Raspredelennye vychislitel'nye sistemy dlya mashinnogo obucheniya [Distributed Computing Systems for Machine Learning], *Informatsionnye tekhnologii i vychislitel'nye sistemy* [Information Technologies and Computing Systems], 2017, No. 3, pp. 56-69.
8. *Zhuravlev Yu.I.* Ob algebraicheskom podkhode k resheniyu zadachi raspoznavaniya ili klassifikatsii [On the algebraic approach to solving the problem of recognition or classification], *Problemy kibernetiki* [Problems of Cybernetics], 1978, Vol. 33, pp. 5-68.
9. *Rabinovich Yu.I.* Klasternyy analiz detalizatsii telefonnykh peregovorov [Cluster analysis of the details of telephone conversations], *Sistemy i sredstva informatiki* [Systems and means of informatics], 2007, Vol. 17, No. 1, pp. 52-78.
10. *Lushnikov N.D., Ismagilova A.S.* Evklidovo rasstoyanie kak osnova programmnogo kompleksa po mnogofaktornoy biometricheskoy autentifikatsii [Euclidean distance as the basis of a software package for multi-factor biometric authentication], *Matematicheskoe modelirovanie protsessov i sistem: Mater. XI Mezhdunarodnoy molodezhnoy nauchno-prakticheskoy konferentsii* [Mathematical modeling of processes and systems: Proceedings of the XI International Youth Scientific and Practical Conference]. Sterlitamak, 2021, pp. 53-55.

11. *Ruzibaev O.B., Eshmetov S.D.* Issledovanie i analiz algoritmov na osnove nechetkogo metoda k blizhayshikh sosedey s primeneniem razlichnykh metrik pri diagnostike raka molochnoy zhelezy [Research and analysis of algorithms based on the fuzzy k nearest neighbors method using various metrics in the diagnosis of breast cancer], *Nauka i mir* [Nauka i mir], 2016, No. 5-1 (33), pp. 102-107.
12. *Le Min' Taun, Shukurov I.S., Nguen Tkhi May.* Issledovanie intensivnosti gorodskogo ostrova tepla na osnove gorodskoy planirovki [Study of the intensity of the urban heat island based on urban planning], *Stroitel'stvo: nauka i obrazovanie* [Construction: science and education], 2019, Vol. 9, No. 3, pp. 54-65.
13. Shumskaya A.O. Otsenka effektivnosti metrik rasstoyaniya Evklida i rasstoyaniya Makhalanobisa v zadachakh identifikatsii proiskhozhdeniya teksta [Estimation of Efficiency Metrics of Euclid Distance and Mahalanobis Distance in Problems of Identification of Text Origin], *Doklady Tomskogo gosudarstvennogo universiteta sistem upravleniya i radioelektroniki* [Reports of Tomsk State University of Control Systems and Radioelectronics], 2013, No. 3 (29), pp. 141-145.
14. Sherstnev P.A. Issledovanie metodov vektorizatsii dokumentov na osnove vektorizatsii slov [Investigation of document vectorization methods based on word vectorization], *Aktual'nye problemy aviatsii i kosmonavtiki: Sb. materialov VII Mezhdunarodnoy nauchno-prakticheskoy konferentsii, posvyashchennoy Dnyu kosmonavtiki* [Actual problems of aviation and astronautics: Collection of materials of the VII International scientific and practical conference dedicated to Cosmonautics Day]: in 3 vol. Krasnoyarsk, 2021, pp. 216-218.
15. *Tian L., Huang R., Wang Y.* Metric learning in codebook generation of bag-of-words for person re-identification, *ICPRAM 2019 - Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*. Prague. 2019, pp. 298-306.
16. *Bulyga F.S., Kureychik V.M.* Algoritmy aglomerativnoy klasterizatsii primenitel'no k zadacham analiza lingvisticheskoy ekspertnoy informatsii [Algorithms of agglomerative clustering in relation to the problems of analysis of linguistic expert information], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Technical science], 2021, No. 6 (223), pp. 73-88.
17. *Nartsev A.D.* Text classification by means of word2vec model and a convolutional neural network, *Presenting Academic Achievements to the World. Natural Science*: *Mater. X nauchnoy konferentsii molodykh uchenykh. Saratov, 16 aprelya 2019 goda* [Presenting Academic Achievements to the World. Natural Science: Proceedings of the X scientific conference of young scientists, Saratov, April 16, 2019]. Saratov, 2020, Vol. 9, pp. 71-77.
18. *Levchenko S.V.* Razrabotka metoda klasterizatsii slov po smyslovym kharakteristikam s ispol'zovaniem algoritmov Word2Vec [Development of a method for clustering words by semantic characteristics using Word2Vec algorithms], *Novye informatsionnye tekhnologii v avtomatizirovannykh sistemakh* [New information technologies in automated systems], 2017, No. 20, pp. 44-46.
19. *Bulyga F.S., Kureychik V.M.* Klasterizatsiya korpusa tekstovykh dokumentov pri pomoshchi algoritma k-means [Clusterization of text document corpus using the k-means algorithm], *Izvestiya vysshikh uchebnykh zavedeniy. Severo-Kavkazskiy region. Tekhnicheskie nauki* [Izvestia of higher educational institutions. North Caucasian region. Technical science], 2022, No. 3 (215), pp. 33-40.
20. Newsgroups, *Qwone*. 2019. Available at: http://qwone.com/~jason/20Newsgroups/ (accessed 15 August 2022).

Статью рекомендовал к опубликованию д.т.н., профессор Н.Н. Прокопенко.

Булыга Филипп Сергеевич – Южный федеральный университет; e-mail: bulyga@sfedu.ru; г. Таганрог, Россия; тел.: +79001330866; кафедра САПР; аспирант.

Курейчик Виктор Михайлович – e-mail: vmkureychik@sfedu.ru; тел.: +79282132730; кафедра САПР; д.т.н.; профессор.

Bulyga Philip Sergeevich – Southern Federal University; e-mail: bulyga@sfedu.ru; Taganrog, Russia; phone: +79001330866; the department of computer-aided design; graduate student.

Kureichik Viktor Mikhailovich – e-mail: vmkureychik@sfedu.ru; phone: +79282132730; the department of computer-aided design; dr. of eng. sc; professor.