18. *Spek A.L.* Structure Validation in Chemical Crystallography, *Acta Crystallographica*, 2009, D65, pp. 148-155.
19. *Zeyun Yu, Bajaj C.* Automatic Ultrastructure Segmentation of Reconstructed CryoEM Maps of Icosahedral Viruses, *IEEE Transactions on Image Processing*, 2005, 14 (9), pp. 1324-1337.
20. *Seiichi Kondo, Mark Lutwyche, Yasuo Wada.* Observation of Threefold Symmetry Images due to a Point Defect on a Graphite Surface Using Scanning Tunneling Microscope (STM), *Japanese Journal of Applied Physics*, 1994, 33 (9B), pp. 1342-1344.
21. *Markus M., Mink Kh.* Obzor po teorii matrits i matrichnykh neravenstv [Overview of the theory of matrices and matrix inequalities]. Moscow: Nauka, 1972, 232 p.
22. *Kramer G.* Matematicheskie metody statistiki [Mathematical Methods of Statistics]. Moscow: Mir, 1975, 648 p.
23. *Kibzun A.I., Goryainova E.R., Naumov A.V.* Teoriya veroyatnostey i matematicheskaya statistika. Bazovyy kurs s primerami i zadachami [Theory of Probability and Mathematical Statistics. Basic course with examples and tasks]. Moscow: Fizmatlit, 2013, 232 p.
24. *Venttsel' E.S., Ovcharov L.A.* Teoriya veroyatnostey [Theory of Probability]. Moscow: Nauka, 1969, 368 p.

Статью рекомендовал к опубликованию д.ф.-м.н. Г.В. Куповых.

Каркищенко Александр Николаевич – Научно-исследовательский институт робототехники и процессов управления ЮФУ; e-mail: karkishalex@gmail.com; 347928 г. Таганрог, Россия; тел.: +78634371694; д.ф.-м.н.; профессор; в.н.с.

Мнухин Валерий Борисович – Южный федеральный университет; e-mail: mnukhin.valeriy@mail.ru; г. Таганрог, Россия; тел.: +78634371606; к.ф.-м.н; доцент.

Karkishchenko Alexander Nikolaevich – Scientific Research Institute of Robotics and Control Processes of the Southern Federal University; e-mail: karkishalex@gmail.com; Taganrog, Russia; phone: +78634371694, dr. of math. sc.; professor; leading researcher.

Mnukhin Valeriy Borisovich – Southern Federal University; e-mail: mnukhin.valeriy@mail.ru; Taganrog, Russia; phone: +78634371606; cand. of math. sc.; associate professor.

УДК 004.89 DOI 10.18522/2311-3103-2021-2-154-167

Ю.А. Кравченко, А.М. Мансур, Ж.Х. Мохаммад

ВЕКТОРИЗАЦИЯ ТЕКСТА С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ*

В задачах интеллектуального анализа текста текстовое представление должно быть не только эффективным, но и интерпретируемым, поскольку это позволяет понять операционную логику, лежащую в основе моделей интеллектуального анализа данных. Традиционные методы векторизации текста, такие как TF-IDF и Bag-of-words, эффективны и имеют интуитивно понятную интерпретируемость, но страдают от «проклятия размерности» и не могут понимать смысл слов. С другой стороны, современные распределенные методы эффективно определяют скрытую семантику, но требуют больших вычислительных ресурсов и времени, а также им не хватает интерпретируемости. В этой статье предлагается новый метод векторизации текстов под названием Bag of weighted Concepts BoWC, который представляет документ в соответствии с содержащейся в нем информацией о концептах. Предлагаемый метод создает концепты посредством кластеризации векторов слов (т.е. встраивания слов), и использует частоты этих кластеров кон-

* Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-29-22019.

цептов для представления векторов документов. Чтобы обогатить итоговое представление документа, предлагается модифицированная весовая функция для взвешивания концептов на основе статистики, извлеченной из информации вложений слов. Векторы, сгенерированные с помощью предложенного метода, характеризуются интерпретируемостью, низкой размерностью, высокой точностью, а также низкими вычислительными затратами при использовании в задачах классификации и кластеризации. Предлагаемый метод протестирован на пяти различных наборах эталонных данных для кластеризации и классификации текстовых документов и сравнивается с несколькими базовыми методами, включая Bag-of-words, TF-IDF, Averaged GloVe, Bag-of-Concepts и VLAC. Результаты показывают, что BoWC превосходит большинство базовых методов и дает в среднем на 7 % лучшую точность.

Векторизация текста; интеллектуальный анализ данных; классификация; кластеризация; машинное обучение; концепты; семантика.

Yu.A. Kravchenko, A.M. Mansour, J.H. Mohammad

TEXT VECTORIZATION USING DATA MINING METHODS

In the text mining tasks, textual representation should be not only efficient but also interpretable, as this enables an understanding of the operational logic underlying the data mining models. Traditional text vectorization methods such as TF-IDF and bag-of-words are effective and characterized by intuitive interpretability, but suffer from the «curse of dimensionality», and they are unable to capture the meanings of words. On the other hand, modern distributed methods effectively capture the hidden semantics, but they are computationally intensive, time-consuming, and uninterpretable. This article proposes a new text vectorization method called Bag of weighted Concepts BoWC that presents a document according to the concepts' information it contains. The proposed method creates concepts by clustering word vectors (i.e. word embedding) then uses the frequencies of these concept clusters to represent document vectors. To enrich the resulted document representation, a new modified weighting function is proposed for weighting concepts based on statistics extracted from word embedding information. The generated vectors are characterized by interpretability, low dimensionality, high accuracy, and low computational costs when used in data mining tasks. The proposed method has been tested on five different benchmark datasets in two data mining tasks; document clustering and classification, and compared with several baselines, including Bag-of-words, TF-IDF, Averaged GloVe, Bag-of-Concepts, and VLAC. The results indicate that BoWC outperforms most baselines and gives 7 % better accuracy on average.

Text vectorization; data mining; classification; clustering; machine learning; concepts; semantic.

Введение. Текстовое представление данных является наиболее широко используемой формой общения и выражения среди всех различных источников данных. Кроме того, текстовые данные создаются и из других источников данных, таких как аудио и видео, где при построении автоматизированных систем респондентов голос преобразуется в текстовые данные, а также добавляются видеообъяснения для задач поиска информации, рекомендаций и оценки. Эти данные растут с каждым днем, и потребность в поиске инновационных механизмов и методов обработки, хранения, понимания и извлечения из них скрытой информации возрастает.

Чтобы применить различные методы машинного обучения и интеллектуального анализа данных, необработанные документы необходимо преобразовать в формат понятный машине [1]. Первым шагом к тому, чтобы сделать текстовые документы машиночитаемыми, является векторизация, которая определяется как преобразование текстового документа в цифровой вектор, и представляет собой процесс извлечения признаков из текста для выполнения любых задач интеллектуального анализа текста и математического решения проблем [2].

Методы векторизации – это методы построения векторных представлений текстов на естественном языке [1, 3]. Они применяются во многих приложениях *Data mining*, обработки естественного языка (ОЕЯ) и поиска информации (ПИ), в таких задачах, как оценка семантического сходства и семантической близости для сопоставления текстов, классификация и кластеризация текстовых документов, обнаружение тем, генерация вопросов, ответы на вопросы, моделирование языков, машинный перевод, обобщение текста и др. Как правило, представление документа направлено на его преобразование в вектор фиксированной длины, который может описывать содержимое, чтобы уменьшить сложность документов и упростить их обработку [4].

Традиционные методы представления документов, такие как «мешок слов» (BoW, Bag-of-Words) и TF-IDF (term frequency invers document frequency), достигли многообещающих результатов во многих задачах классификации и кластеризации документов благодаря своей простоте, эффективности и точности [4–7]. Однако у этих методов есть две основные проблемы. Первая – проблема размерности, поскольку количество функций в результирующих векторах значительно увеличивается по мере увеличения размера корпуса (количества документов). Следовательно, размерность векторов документов может стать чрезвычайно большой и разреженной, и обычные метрики расстояния, такие как евклидово расстояние или косинусное расстояние, станут бессмысленными [8].

Вторая проблема заключается в том, что полученное представление не учитывает семантическое отношение между словами и игнорирует порядок слов, это означает, что разные документы будут иметь одинаковое представление, если используются одни и те же слова. Эти недостатки ограничивают способность моделей интеллектуального анализа текста фиксировать истинное сходство между документами, представленными этими методами.

Метод набора концептов BOC (Bag-of-Concepts) [8] был предложен как решение этой проблемы. Основываясь на корпусе коллекции документов, Bag-of-Concepts кодирует слова как встраиваемые векторы, а затем применяет кластеризацию для группировки похожих слов в кластеры, называемые концептами. Как и в случае метода «мешка слов», каждый вектор документа будет представлен частотами этих концептов в документе. Чтобы уменьшить влияние концептов, которые появляются в большинстве документов, применяется схема взвешивания, подобная TF-IDF, с заменой частоты термина TF на частоту концепта CF. Однако применение такой схемы взвешивания к концептам неэффективно, и снижает выразительную способность результирующих векторов. Это связано с тем, что концепт здесь представлен группой терминов, схожих по смыслу, и поэтому частота появления того или иного концепта на уровне корпуса будет намного больше, чем частота появления одного из его терминов. Математически соотношение между количеством документов, в которых присутствует концепт, к общему количеству документов будет незначительным для большинства концептов, что будет указывать на то, что все концепты являются распространенными на уровне корпуса и не содержат дискриминационной информации, а это ложный вывод.

Следуя тому же подходу, что и BOC, в работе [9] разработан метод «Векторы локально агрегированных концептов» (VLAC, Vectors of Locally Aggregated Concepts), который группирует вложения слов для генерации признаков. Однако вместо того, чтобы подсчитывать частоту кластеризованных вложения слов, VLAC берет сумму остатков каждого кластера относительно его центроида и объединяет их для создания вектора признаков. Результирующие векторы признаков содержат более ценную информацию, чем пакет концептов, благодаря дополнительному включению этих статистических данных первого порядка. Однако этот метод соз-

дает векторы относительно больших размеров с высокими вычислительными затратами. Если бы 10 концептов нужно было создать из вложений слов размером 300, то результирующий вектор документа содержал бы 10 × 300 значений.

Чтобы преодолеть вышеупомянутые недостатки, в данной работе предлагается новый метод векторизации, названный «Пакетом взвешенных концептов» (BoCW bag-of-weighted concepts), который применяет тот же подход, что и BOC, за исключением того, что он имеет несколько основных отличий:

1. Функция взвешивания, основанная на обратной частоте документа, была заменена функцией монотонного убывания, которая позволит лучше регулировать влияние частоты документа на уровне коллекции документов;

2. Представлена эвристическая функция для извлечения дополнительной дискриминирующей информации, используемая для взвешивания концептов. Эта функция вычисляет важность концепта в документе, измеряя сходство слов, образующих концепт со словами документа, которые ему принадлежат;

3. Для повышения эффективности предложенного метода при выполнении задачи классификации и кластеризации, в которых известны правильные классы документов, введен новый параметр для взвешивания вектора концептов по частоте класса документа.

Таким образом, предлагаемый метод сочетает семантику на уровне терминов с семантикой на уровне концептов. Полученные в результате векторы признаков документов содержат более ценную информацию, чем при использовании BOC и VLAC, и в то же время, эти векторы интерпретируемые и их размеры намного меньше.

1. Аналитический обзор методов векторизации текстов. *Метод набора слов «Bag-of-Words»* основан на предположении, что частота слов в документе может надлежащим образом отражать сходства и различия между документами. Следовательно, признаки векторов документов, сгенерированные методом «мешка слов», представляют вхождения каждого слова в документе. Этот метод имеет серьезный недостаток: часто встречающиеся слова могут доминировать в пространстве признаков, тогда как редкие слова могут нести больше ценной информации.

Для улучшения представления «мешка слов» предлагается метод TF-IDF, отражающий значимость слова для документа в корпусе, то есть механизм взвешивания, где *Term Frequency* (TF) – это количество раз, когда термин появляется в документе, а обратная частота документа *IDF* измеряет редкость термина во всем корпусе. Обозначая общее количество документов в коллекции $|\boldsymbol{D}|$, понятия частоты термина и обратной частоты документа объединяются, чтобы получить общий вес для каждого термина в каждом документе следующим образом:

$$\boldsymbol{tf-idf(t,d,D)} = \boldsymbol{tf(t,d)} \times \boldsymbol{idf(t,D)} = \frac{\boldsymbol{n_t}}{\sum_k \boldsymbol{n_k}} \times \boldsymbol{log}\frac{|\boldsymbol{D}|}{|\{\boldsymbol{D_i} \in \boldsymbol{D} | \boldsymbol{t} \in \boldsymbol{D_i}\}|}, \quad (1)$$

где $\boldsymbol{n_t}$ – количество вхождений слова $\boldsymbol{t}$ в документе $\boldsymbol{d}$, а $\boldsymbol{n_k}$ – общее количество слов в этом документе. $\boldsymbol{D_i}$ – количество документов из коллекции $\boldsymbol{D}$, в которых встречается $\boldsymbol{t}$.

Метод набор концептов (BOC, Bag-of-concepts), описанный в работе [8], генерирует векторы слов документа с помощью совокупности моделей *word2vec*, которые встраивают семантически похожие слова в соседнюю область. Это позволяет сгруппировать соседние слова в один общий кластер концептов. *Bag-of-Concepts* генерирует кластеры слов, применяя сферические *k*-средства к вложениям слов. Полученные кластеры содержат слова с похожим значением и поэтому называются концептами. Подобно принципу метода «мешок слов», каждый вектор документа представляется частотой каждого кластера концептов в документе. Чтобы смягчить влияние концептов, которые появляются в большинстве докумен-

тов, используется схема взвешивания, аналогичная *TF_IDF*, с заменой термина частота *TF* на частоту концепта *CF*. Следовательно, он называется *CF-IDF* (частота концепта с обратной частотой документа) и рассчитывается на основе следующего уравнения.

$$\boldsymbol{CF-IDF}\left(\boldsymbol{c_i}, \boldsymbol{d_j}, \boldsymbol{D}\right) = \frac{\boldsymbol{n_c}}{\sum_k \boldsymbol{n_k}} \times \boldsymbol{log}\frac{|\boldsymbol{D}|}{|\{\boldsymbol{d}\in\boldsymbol{D}|\boldsymbol{c_i}\in\boldsymbol{d}\}|}, \quad (2)$$

где $|\boldsymbol{D}|$ – число документов в коллекции, а в знаменателе это число документов из коллекции $\boldsymbol{D}$, в которых встречается концепт $\boldsymbol{c}$; $\boldsymbol{n}_{\text{c}}$ – число вхождений концепта $\boldsymbol{c}$ в документ $\boldsymbol{d}$, а $\boldsymbol{n_k}$ – общее число концептов в данном документе.

Этот метод решает проблему больших размеров в методе набора слов, поскольку он нелинейно уменьшает размеры при преобразовании пространства слов в пространство концептов на основе семантического сходства. Кроме того, в [8] показано, что этот метод обеспечивает лучшее представления документа, чем *Bag-of-Words* и *TF-IDF*, в задаче классификации для поиска двух наиболее похожих документов среди троек документов. Однако в задаче прогнозирования правильной метки для каждого документа *BOC* не удалось превзойти *TF-IDF* на двух из трех наборов данных.

2. Постановка задачи. Пусть дан $\boldsymbol{\mathcal{D}}$ текстовый словарь, т.е. список уникальных слов, которые появляются в коллекции текстовых документов. Пусть $\boldsymbol{x_i \epsilon R^D}$ – вектор вложения $\boldsymbol{i}$-го слова словаря $\boldsymbol{\mathcal{D}}$, где $\boldsymbol{D}$ - размерность вложения слова. Множество всех векторов вложения слов обозначается $\boldsymbol{\varepsilon} = \{\boldsymbol{x_i}, \boldsymbol{i} = \boldsymbol{1}, \dots, |\boldsymbol{D}|\}$.

Кроме того, пусть $\boldsymbol{N}$ – количество текстовых документов, которые должны быть закодированы с использованием предложенного метода. Каждый документ описывается векторами вложения своих слов $\boldsymbol{N_i}: \boldsymbol{x_{ij}} \in \boldsymbol{\varepsilon}, (\boldsymbol{i} = \boldsymbol{1}, \dots, \boldsymbol{N}, \boldsymbol{j} = \boldsymbol{1}, \dots, \boldsymbol{N_i})$, где $\boldsymbol{N_i}$ – количество слов $\boldsymbol{i}$-го документа. Каждое слово в документе имеет вектор вложения, то есть $\boldsymbol{x_{ij}}$ – это вектор вложения $\boldsymbol{j}$-го слова из $\boldsymbol{i}$-го документа. Количество слов варьируется от одного документа к другому.

Таким образом, ставится задача найти преобразование $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}): \boldsymbol{R^D} \rightarrow \boldsymbol{R^C}$, такое, что преобразованный вектор признаков $\boldsymbol{y_i \epsilon R^C}$ сохраняет (большую часть) преобразование или структуру в $\boldsymbol{R^D}$**.** Оптимальное преобразование $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ будет таким, которое не приводит к увеличению минимальных ошибок вероятности.

Соответственно, требуется найти преобразование из пространства слов в пространство концептов, которое позволяет каждому документу быть представленным вектором фиксированной длины. $\boldsymbol{N_i}: \boldsymbol{c_{ij}} = (\boldsymbol{i} = \boldsymbol{1}, \dots, \boldsymbol{K}, \boldsymbol{j} = \boldsymbol{1}, \dots, \boldsymbol{N_i})$, где $\boldsymbol{k}$ – количество извлеченных концептов, а $\boldsymbol{c_{ij}}$ – это признак $\boldsymbol{j}$-го концепта $\boldsymbol{i}$-го документа.

3. Разработка метода «Bag of weighted concepts». Для реализации поставленной задачи разрабатывается принципиально новый метод для представления текстовых документов в виде числовых векторов фиксированной длины. Предлагаемый метод преобразует документ в вектор в соответствии с содержащейся в нем информацией о концептах. Для этого, аналогично методу «пакета концептов», создаётся словарь концептов $\boldsymbol{T}$, а затем векторы документа создаются на основе статистики частотности концептов в документе.

Все векторы вложения кластеризуются в $\boldsymbol{N_k}$ кластеров. Используется сферический алгоритм $\boldsymbol{k}$-средних, в котором косинусное подобие применяется в качестве метрики расстояния. Для заранее определенного значения $\boldsymbol{K}$ алгоритм итеративно назначает каждую точку данных одному из $\boldsymbol{k}$ центроидов и обновляет каждый центроид с учетом принадлежности точек данных. Слово может быть перегруппировано более чем в один кластер, это фактически означает, что различные значения слова могут быть объединены с их синонимами.

Необходимо обеспечить минимальное количество слов в каждом кластере, представляющем концепт, для этого проверяется размер результирующего кластера и расширяется его. После этого слова располагаются в каждом кластере в соответствии с близостью к центру кластера, а затем выбираются $\boldsymbol{M}$ слова, наиболее близкие к центроиду кластера. В итоге получается группа концептов (понятий), каждое из которых представлено $\boldsymbol{M}$ словами принадлежащие к одному общему понятию или имеющие общий гипероним. Словарь концептов представляется следующим образом $\boldsymbol{T} = (\boldsymbol{w_1^1}, \boldsymbol{w_2^1}, \dots, \boldsymbol{w_M^2}, \boldsymbol{w_1^2}, \boldsymbol{w_2^2}, \dots, \boldsymbol{w_M^2}, \dots, \boldsymbol{w_1^K}, \boldsymbol{w_2^K}, \dots, \boldsymbol{w_M^K})$, где $\boldsymbol{w_i^j}$ – $\boldsymbol{i}$-е слово $\boldsymbol{j}$-го кластера.

Алгоритм построения словаря концептов

Ввод Набор $\boldsymbol{D} = \{\boldsymbol{d_1}, \boldsymbol{d_2}, \dots, \boldsymbol{d_N}\}$ из $\boldsymbol{N}$ документов
$\boldsymbol{M_{desired}}$ // минимальное количество слов кластера
Вывод Словарь концептов $\boldsymbol{T} = [\,]$

1: Отсканировать документы и создать словарь $\boldsymbol{\mathcal{D}}$
2: Инициализировать вложение слова $\boldsymbol{\mathcal{E}}$ с помощью $\boldsymbol{\mathcal{D}}$
3: Инициализировать словарь $\boldsymbol{T}$, запустив $\boldsymbol{k}$-means на $\boldsymbol{\mathcal{E}}$
4: $\boldsymbol{C} = \boldsymbol{count}(\boldsymbol{T})$ //количество концептов
5: $\boldsymbol{While}\ i < C\ \boldsymbol{do}$
6: $\quad \boldsymbol{M_{real}} = \boldsymbol{count}(\boldsymbol{clusters}[\boldsymbol{i}])$
7: $\quad \boldsymbol{L} = \boldsymbol{M_{desired}} - \boldsymbol{M_{real}}$ //Количество слов, которыми следует расширить кластер
8: $\quad \boldsymbol{if}(L > 0)\ \boldsymbol{then}$
9: $\quad\quad \boldsymbol{for}\ j = 0\ to\ L\ \boldsymbol{do}$
10: $\quad\quad\quad w \coloneqq similarTo(T[i, j])$ //получить синоним слово слова j
11: $\quad\quad\quad w^* = embedding(w)$ //встраивание слова
12: $\quad\quad\quad T[i] \leftarrow w^*$ // Добавить этот слово в текущий кластер
$\quad\quad \boldsymbol{end}$
$\quad \boldsymbol{else}$
13: $\quad\quad sort(clusters[i])$
14: $\quad\quad T[\,] \leftarrow clusters[M_{desired}]$
$\quad \boldsymbol{end}$
$\boldsymbol{end}$
15: Сохранить кластеры в виде словаря $\boldsymbol{T}$.

Для рассматриваемого документа $\boldsymbol{d}$ создается вектор признаков размера $\boldsymbol{k}$, равного количеству концептов $\boldsymbol{V^d} = (c_1^d, c_2^d, \dots, c_K^d)$, где признак c_i^d выражает степень значимости $\boldsymbol{i}$-го концепта (его вес) в документе. Значимость концепта в векторе документа рассчитывается аналогично методу *BOC*, где слова документа сравниваются с концептами, т.е. измеряется степень косинусного сходства между вектором слова документа и вектором центроида кластера и записывается появление концептов, превышающих определенный порог $\boldsymbol{\Theta}$, определяемый экспериментально. Частота концепта задается следующей формулой:

$$\boldsymbol{CF}(\boldsymbol{c_i}, \boldsymbol{d_j}, \boldsymbol{D}) = \frac{\boldsymbol{n_c}}{\sum_k \boldsymbol{n_k}}, \tag{3}$$

где $\boldsymbol{n_k}$ – общее количество концептов в документе, а $\boldsymbol{n}_c$ – количество вхождений концепта $\boldsymbol{c}$ в документ, которое вычисляется следующей двоичной функцией $\boldsymbol{g}(\boldsymbol{s})$:

$$\boldsymbol{g}(\boldsymbol{s}) = \begin{cases} \boldsymbol{1}, & \boldsymbol{s} > \theta \\ \boldsymbol{0}, & \boldsymbol{otherwize} \end{cases}, \tag{4}$$

где сходство s между двумя векторами слов X и Y вычисляется следующим образом:

$$s = similarity(X,Y) = cos(\theta) = \frac{X.Y}{\|X\|.\|Y\|} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}}, \quad (5)$$

где X_i и Y_i – компоненты вектора слов X и Y, соответственно.

Чтобы уменьшить влияние концептов, которые появляются в большинстве документов, в методе *BOC* применяется схема взвешивания, подобная *TF-IDF*, в которой частота концепта *CF* взвешивается по обратной частоте документа *IDF* (2). Однако использование этой формулы на уровне концепта неэффективно, потому что частота появления концепта на уровне коллекции документов намного выше, чем частота одного из его терминов. Математически это приводит к очень малым значениям логарифма – близкими к нулю, что указывает ложный вывод.

Для преодоления этой проблемы авторы предлагают новую функцию взвешивания, основанную на обратной частоте документа ранее установленной функцией монотонного убывания [10], которая позволяет повысить качество регулировки влияния частоты документа на уровне коллекции документов. Функция задается следующей формулой:

$$f(F) = e^{-\alpha \times F}, \quad (6)$$

где α – константа, а F – частота документа, вычисляемая по формуле:

$$F = \left(\frac{|\{d \in D | c_i \in d\}|}{|D|}\right). \quad (7)$$

Экспоненциальная функция была выбрана для гарантии, что значение f находится в диапазоне [0,1]. Тогда весовая функция, которая получила название *CF-EDF* (concept frequency – exponential document frequency), принимает следующий вид:

$$CF - EDF(c_i, d_j, D) = \frac{n_c}{\sum_k n_k} \times exp\left(-\frac{|\{d \in D | c_i \in d\}|}{|D|}\right). \quad (8)$$

Чтобы полученный вектор документа содержал более значимую (ценную) информацию, авторы предлагают эвристическую функцию, которая извлекает статистику, характеризующую отношение каждого концепта c документом, путем вычисления сходства слов, образующих концепт, со словами документа, которые ему принадлежат. Для документа d с N словами, принадлежащими c-му концепту, сходство вычисляется следующим образом:

$$S_c = \frac{1}{N}\sum_{i=1}^{N} \underset{1 \le j \le M}{maxsim}(u_i, v_j), \quad (9)$$

где u_i – i-е слово документа, принадлежащее концепту; v_j – j-е слово концепта. Функция *Max* возвращает наивысшую оценку сходства, записанную для каждого слова документа, принадлежащего концепту c.

Итоговая формула взвешивания концепта выглядит следующим образом:

$$CF - EDF(c_i, d_j, D) = \frac{n_c}{\sum_k n_k} \times exp\left(-\frac{|\{d \in D | c_i \in d\}|}{|D|}\right) \times exp(S_{c_i}). \quad (10)$$

Для повышения эффективности предложенного метода при выполнении задачи классификации и кластеризации, в которых известны правильные классы документов, введен новый параметр для представления внутриклассовых характеристик, который называется частотой класса. Аналогично принципу весовой функции *TF-IDF-CF*, предложенному в [11] в данной работе вычисляется частота концептов в документах в пределах одного класса. Итоговая формула взвешивания концепта выглядит следующим образом:

$$\boldsymbol{f(c_i, d_j, D) = CF - EDF(c_i, d_j, D) \times \frac{\mu_{k_{ij}}}{N_{k_j}}}, \quad (11)$$

где $\boldsymbol{\mu_{k_{ij}}}$ – количество документов, в которых концепт $\boldsymbol{c_i}$ появляется в том же классе $\boldsymbol{k}$, которому принадлежит $\boldsymbol{j}$-ый документ; $\boldsymbol{N_{k_j}}$ – количество документов в том же классе $\boldsymbol{k}$, к которому принадлежит $\boldsymbol{j}$-ый документ.

4. Вычислительный эксперимент и анализ полученных результатов. Эффективность метода *BoWC* оценивалась по нескольким базовым критериям посредством их оценки при выполнении задачи кластеризации. Приведем описание использованных наборов данных и метрик оценки.

Для оценки предложенного метода использовались пять наборов данных представленных в табл. 1.

Таблица 1

Наборы использованных данных

Набор данных	***BBC***	***R8***	***OH***	***20Newsgroup***	***WebKB***
Количество документов	2225	8491	5380	18821	4199
Количество категорий	5	8	7	20	4
Среднее количество слов в документе	2262	742	1008	1902	909
Среднее количество словарных слов в одном документе	207	66	79	135	98

BBC содержит 2225 документов, принадлежащих к 5 различным классам. *Reuters (R8)* содержит статьи из новостной ленты *Reuters*. В этой работе используется разделение *R8* набора данных *Reuters*, которое содержит 8491 документ, где документы принадлежат 8 различным классам. *20Newsgroups*, содержит 18821 документ, где документы принадлежат 20 различным категориям групп новостей. *OHSUMED (OH)* содержит 5380 документов, принадлежащих 7 различным классам. *WebKB* содержит веб-страницы из различных разделов информатики, которые были разделены на 7 разных классов: студенты, преподаватели, сотрудники и т.д. В этой работе использовался предварительно обработанный набор данных *WebKB*, который содержит 4 различных класса и всего 4199 документов [12, 13].

Чем больше объём корпуса, на котором обучается модель вложений, тем выше репрезентативная способность полученных векторов слов. Следовательно, обучение модели вложений на данных разного размера было бы неправильным вариантом, поэтому используется предварительно обученная модель, а именно *GloVe* из-за ее преимуществ перед другими [14]. Модель *GloVe* была обучена на наборе данных *Common Crawl* и содержит векторы для 1,9 миллиона английских слов.

Такая же предварительная обработка была применена ко всем наборам данных путем перевода текста в нижний регистр и разбивки на самые длинные непустые последовательности буквенно-цифровых символов, которые содержат не менее трехбуквенных символов. Стоп-слова, редкие слова и слова без вложений были удалены. Поскольку используется предварительно обученное встраивание (*embeddings*), выделение корней слов (*stemming*) не выполнялось, чтобы уменьшить количество слов вне словарного запаса, за исключением набора данных *WebKB*, где исходные данные получить не удалось, поэтому использовалась предварительно обработанная версия.

В задаче кластеризации, зная присвоения эталонных данных классам образцов, определена интуитивно понятная метрика с помощью анализа условной энтропии, одной из которых является *V-мера*.

V-мера – мера, основанная на энтропии [15], которая явно измеряет, насколько успешно были удовлетворены критерии однородности (*homogeneity*), когда каждый кластер содержит только членов одного класса, и полноты (*completeness*), когда все члены данного класса относятся к одному кластеру. *V-мера* вычисляется как среднее гармоническое для различных оценок однородности и полноты, точно так же, как точность и полнота обычно объединяются в *F-меру* [16].

$$\boldsymbol{V}-\text{мера}\ =\frac{(\mathbf{1}+\boldsymbol{\beta})\times\text{Однородность}\times\text{Полнота}}{(\boldsymbol{\beta}\times\text{Однородность}+\text{Полнота})}. \quad (12)$$

Фактор $\boldsymbol{\beta}$ может быть откорректирован, чтобы обеспечить либо однородность, либо полноту алгоритма кластеризации.

В проведенном вычислительном эксперименте предложенный метод BoWC оценивался по результатам решения задачи кластеризации путем сравнения с пятью проанализированными ранее методами, а именно, Bag-of-Words, TF-IDF, GloVe, Bag-of-Concepts и VLAC. Результаты эксперимента показаны в табл. 2.

Таблица 2

Результаты кластеризации по *V-мере*

	BBC	***R8***	***OHSUMED***	***20NG***	***WebKB***
TF-IDF	0,663	0,513	0,122	0,362	**0,313**
Количество признаков	29821	29290	40179	173446	7632
Bag of Words	0,209	0,248	0,027	0,021	0,021
Количество признаков	17350	14446	17481	92718	7637
Averaged GloVe	0,774	0,481	0,109	0,381	0,219
Количество признаков	300	300	300	300	300
Bag-of-Concepts	0,638	0,131	0,1	0,394	0.084
Количество признаков	200	200	200	200	200
VLAC	0,808	0,456	0.118	–	0.286
Количество признаков	9000	9000	9000	–	9000
BoWC	**0,889**	**0,534**	**0,151**	**0.511**	0.158
Количество признаков	100	55	75	65	70

Эффективность метода *BoWC* была проанализирована с постоянно увеличивающимся количеством концептов от 5 до 100 концептов, чтобы обеспечить справедливое сравнение со всеми методами. Признаки усредненных вложений слов *GloVe*, *Bag-of-Words* и *TF-IDF* максимизированы по умолчанию. Результаты сравнения предложенного метода векторизации текста *BoWC* с рассмотренными каноническими при решении задачи кластеризации документов на наборах данных (*BBC, Reuters, 20Newsgroups, OHSUMED, WebKB*) с использованием *V-меры* представлены на рис. 1.

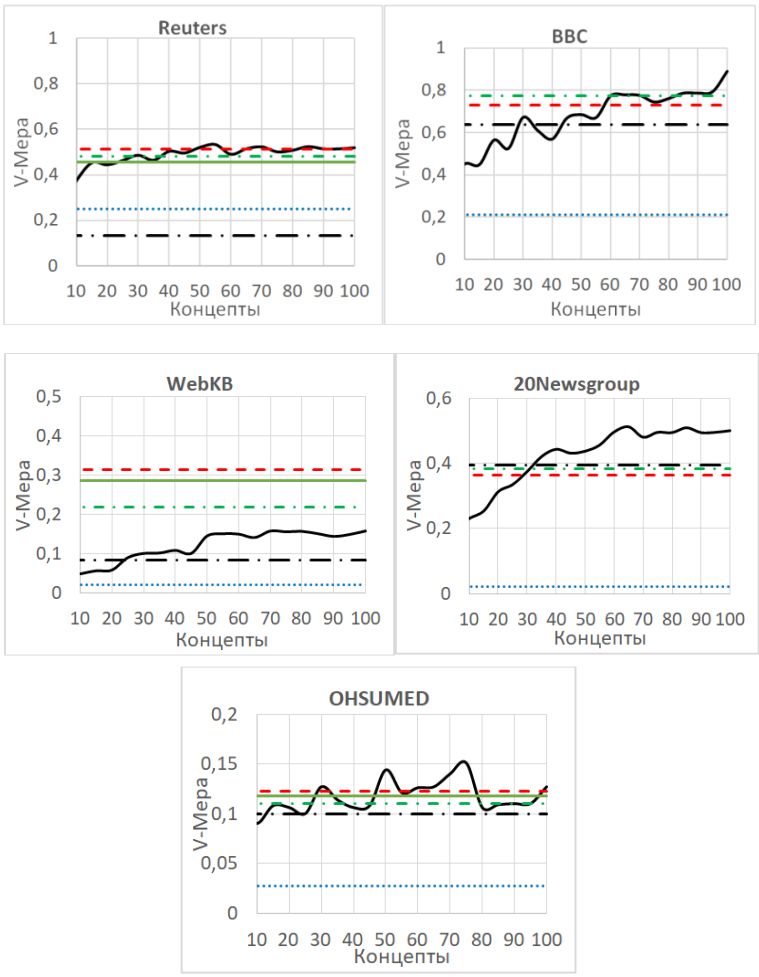

Рис. 1. Результаты сравнения предложенного метода векторизации текста с другими методами при выполнении задачи кластеризации документов на наборах данных (BBC, Reuters, 20Newsgroups, OHSUMED, WebKB) с применением V-меры

Для метода *VLAC*, прилагаемый к работе API был использован при повторной реализации экспериментов на наборах данных из-за разницы в источнике данных и критериях оценки. Было принято 30 концептов (9000 признаков), учитывая, что большее количество концептов требует больших вычислительных затрат [9]. Существует значительная разница в количестве признаков между *BoWC* и методом *VLAC*, однако они будут сравниваться с точки зрения информативности первых тридцати концептов, а не только с точки зрения количества признаков.

Учитывая, что *BOC* является частным случаем предложенного метода *BoWC*, он был реализован через код метода *BoWC*. В [8] показано, что *Bag-of-Concepts* по сравнению с *TF-IDF* требуется не менее 100 концептов для достижения конкурентоспособных результатов. Итак, чтобы максимизировать точность *Bag-of-Concepts*, количество концептов было установлено равным 200.

Чтобы сделать сравнение между реализациями *BoWC* возможным, все вложения слов имели размер 300. Метод *K*-средних использовался в этом эксперименте в качестве классификатора методов генерации признаков.

Табл. 1 показывает, что предложенный метод, превзошел большинство других методов, с уменьшением количества признаков в среднем на 50%, сохранив высокую точность классификации. Значения *V-меры* на наборе данных *OHSUMED* низкие для всех методов. Вероятно, это связано с природой этих данных, поскольку они содержат большое число медицинских терминов, связанных с точки зрения темы, и, следовательно, на уровне термина требуется значительное количество признаков для различения каждого документа, а на уровне концептов требуются дополнительные ресурсы для построения соответствующего концептуального словаря. Тем не менее, предложенный метод дал наилучшее значение для *V-меры* с 75 концептами.

Для набора данных *WebKB* методы *TF-IDF* и *VLAC* превзошли *BoWC*. Этот результат логичен, учитывая, что взята предварительно обработанная версия этого набора данных, что увеличивает количество слов вне словарного запаса и влияет на качество и количество результирующих концептов [17–21].

Эксперимент *VLAC* с набором данных *20Newsgroup* был исключен из-за его высоких вычислительных затрат, так как алгоритм кластеризации завершился неудачно из-за полной загрузки 12,72 ГБ ОЗУ. Причина этого в том, что этот набор данных огромен, а векторы документов, сгенерированные *VLAC*, длинные и, следовательно, требуют большого объема памяти для хранения и обработки.

Колебания кривых на рис. 1 обусловлены разным качеством концептов, извлекаемых при каждом запуске метода, и это весьма логично, поскольку алгоритм построения словаря концептов при каждом запуске дает разные концепты, это связано с типом используемой функции расстояния и порога схожести, по которому определяется появление того или иного концепта в документе. Соответственно, сложно определить конкретное количество признаков, обеспечивающих стабильную производительность метода, что является одним из недостатков предложенного метода, требующим дополнительных исследований.

Заключение. В данной работе представлена разработка метода векторизации текстов *BoWC*, который представляет документ в соответствии с содержащейся в нем информацией о концептах. Предлагаемый метод создает концепты посредством кластеризации векторов слов (т.е. встраивания слов), и использует частоты этих кластеров концептов для представления векторов документов. Чтобы обогатить итоговое представление документа, предлагается новая модифицированная весовая функция для взвешивания концептов на основе статистики, извлеченной из информации вложений слов.

Векторы, сгенерированные с помощью предложенного метода, характеризуются интерпретируемостью, низкой размерностью, высокой точностью, а также низкими вычислительными затратами при использовании в задачах кластеризации.

В целях контроля качества концептов, генерируемых разработанным методом, планируется провести дополнительные исследования для внесения улучшений в алгоритм формирования словаря концептов с целью создания более качественных концептов для преодоления неопределенности в результатах, которая проявлялась в колебаниях значений *V-меры*, отражающей точность процесса кластеризации. Кроме того, необходимо протестировать возможность применения процессов обрезки или слияния концептов, чтобы контролировать качество полученных концептов. С другой стороны, эксперименты будут расширены, чтобы включить анализ влияния других типов вложений на эффективность предложенного метода, включая создание вложений на основе самих данных.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Бенгфорт Б., Билбро Р., Охеда Т.* Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка. – СПб.: Питер, 2019. – 368 с.
2. *Лапшин С.В., Лебедев И.С., Спивак А.И.* Классификация коротких сообщений с использованием векторизации на основе elmo // Известия ТулГУ. Технические науки. – 2019. – № 10. – С. 410-418.
3. *Киреев В.С., Федоренко В.И.* Использование методов векторизации текстов на естественном языке для повышения качества контентных рекомендаций фильмов // Современные наукоемкие технологии. – 2018. – № 3. – С. 102-106.
4. *Lin Y., Liu Z., Sun M.* Representation Learning for Natural Language Processing. – Singapore: Springer Nature, 2020. – 334 p.
5. *Baeza-Yates R., Ribeiro-Neto B.* Modern Information Retrieval. – New York: ACM Press, 1999. – 501 p.
6. *Jones K.S.* A Statistical Interpretation of Term Specificity and its Application in Retrieval // Journal of Documentation. – 1972. – Vol. 28, No. 1. – P. 11-21.
7. *Hoi S., Wu L., Yu N.* Semantics-Preserving Bag-of-Words Models and Applications // IEEE Transactions on Image Processing. – 2010. – Vol. 19, No. 7. – P. 1908-1920.
8. *Kim, H.K., Kim H.-j.* Bag-of-Concepts: Comprehending Document Representation through Clustering Words in Distributed Representation // Neurocomputing. – 2017. – Vol. 266. – P. 336-352.
9. *Grootendorst M., Vanschoren J.* Beyond Bag-of-Concepts: Vectors of Locally Aggregated Concepts // Joint European Conference on Machine Learning and Knowledge Discovery in Databases. – 2019. – P. 681-696.
10. *Bandar Z., Crockett K., Li Y. et al.* Sentence Similarity Based on Semantic Nets and Corpus Statistics // IEEE Transactions on Knowledge. – 2006. – Vol. 18. – P. 1138-1150.
11. *Liu M., Yang J.* An Improvement of TFIDF Weighting in Text Categorization // International Proceedings of Computer Science Information Technology. – 2012. – Vol. 47. – P. 44-47.
12. *Cardoso-Cachopo, A.L., Oliveira A.* Semi-Supervised Single-Label Text Categorization Using Centroid-Based Classifiers // Proceedings of the 2007 ACM Symposium on Applied Computing. – 2007. – P. 844-851.
13. *Lang, K., Rennie J.* The 20 Newsgroups Data Set. – 2008.
14. *Manning C.D., Pennington J., Socher R.* Glove: Global Vectors for Word Representation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). – 2014. – P. 1532-1543.
15. *Hirschberg J., Rosenberg A.* V-measure: A Conditional Entropy-Based External Cluster Evaluation Measure // Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). – 2007. – P. 410-420.
16. *Van Rijsbergen C.J.* Information Retrieval. – Butterworth-Heinemann, 1979. – 224 p.
17. *Bova V., Zaporozhets D., Kureichik V.* Integration and Processing of Problem-Oriented Knowledge Based on Evolutionary Procedures // Advances in Intelligent Systems and Computing. – 2016. – Vol. 450. – P. 239-249.
18. *Kureichik V.M., Semenova A.V.* Ensemble of Classifiers for Ontology Enrichment // Journal of Physics: Conference Series. – 2018. – Vol. 1015. – Issue 3. – Article id. 032123.
19. *Bova V.V., Nuzhnov E.V., Kureichik V.V.* The Combined Method of Semantic Similarity Estimation of Problem Oriented Knowledge on the Basis of Evolutionary Procedures // Advances in Intelligent Systems and Computing. – 2017. – Vol. 573. – P. 74-83.
20. *Pulyavina N., Taratukhin V.* The Future of Project-Based Learning for Engineering and Management Students: Towards an Advanced Design Thinking Approach // ASEE Annual Conference and Exposition, Conference Proceedings. – 2018. – No. 125.
21. *Becker J., Pulyavina N., Taratukhin V.* Next-Gen Design Thinking. Using Project-Based and Game-Oriented Approaches to Support Creativity and Innovation // Proceedings of the 1st International Conference of Information Systems and Design. – 2020.

REFERENCES

1. *Bengfort B., Bilbro R., Okheda T.* Prikladnoy analiz tekstovykh dannykh na Python. Mashinnoe obuchenie i sozdanie prilozheniy obrabotki estestvennogo yazyka [Applied analysis of text data in Python. Machine learning and building natural language processing applications]. Saint Petersburg: Piter, 2019, 368 p.
2. *Lapshin S.V., Lebedev I.S., Spivak A.I.* Klassifikatsiya korotkikh soobshcheniy s ispol'zovaniem vektorizatsii na osnove elmo [Classification of short messages using elmo-based vectorization], *Izvestiya TulGU. Tekhnicheskie nauki* [News of TulSU. Technical sciences], 2019, No. 10, pp. 410-418.
3. *Kireev V.S., Fedorenko V.I.* Ispol'zovanie metodov vektorizatsii tekstov na estestvennom yazyke dlya povysheniya kachestva kontentnykh rekomendatsiy fil'mov [Using methods of vectorization of texts in natural language to improve the quality of content recommendations of films], *Sovremennye naukoemkie tekhnologii* [Modern science-intensive technologies], 2018, No. 3, pp. 102-106.
4. *Lin Y., Liu Z., Sun M.* Representation Learning for Natural Language Processing. Singapore: Springer Nature, 2020, 334 p.
5. *Baeza-Yates R., Ribeiro-Neto B.* Modern Information Retrieval. New York: ACM Press, 1999, 501 p.
6. *Jones K.S.* A Statistical Interpretation of Term Specificity and its Application in Retrieval, *Journal of Documentation*, 1972, Vol. 28, No. 1, pp. 11-21.
7. *Hoi S., Wu L., Yu N.* Semantics-Preserving Bag-of-Words Models and Applications, *IEEE Transactions on Image Processing*, 2010, Vol. 19, No. 7, pp. 1908-1920.
8. *Kim H.K., Kim H.-j.* Bag-of-Concepts: Comprehending Document Representation through Clustering Words in Distributed Representation, *Neurocomputing*, 2017, Vol. 266, pp. 336-352.
9. *Grootendorst M., Vanschoren J.* Beyond Bag-of-Concepts: Vectors of Locally Aggregated Concepts, *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2019, pp. 681-696.
10. *Bandar Z., Crockett K., Li Y. et al.* Sentence Similarity Based on Semantic Nets and Corpus Statistics, *IEEE Transactions on Knowledge*, 2006, Vol. 18, pp. 1138-1150.
11. *Liu M., Yang J.* An Improvement of TFIDF Weighting in Text Categorization, *International Proceedings of Computer Science Information Technology*, 2012, Vol. 47, pp. 44-47.
12. *Cardoso-Cachopo, A.L., Oliveira A.* Semi-Supervised Single-Label Text Categorization Using Centroid-Based Classifiers, *Proceedings of the 2007 ACM Symposium on Applied Computing*, 2007, pp. 844-851.
13. *Lang, K., Rennie J.* The 20 Newsgroups Data Set., 2008.
14. *Manning C.D., Pennington J., Socher R.* Glove: Global Vectors for Word Representation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532-1543.
15. *Hirschberg J., Rosenberg A.* V-measure: A Conditional Entropy-Based External Cluster Evaluation Measure, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007, pp. 410-420.
16. *Van Rijsbergen C.J.* Information Retrieval. Butterworth-Heinemann, 1979, 224 p.
17. *Bova V., Zaporozhets D., Kureichik V.* Integration and Processing of Problem-Oriented Knowledge Based on Evolutionary Procedures, *Advances in Intelligent Systems and Computing*, 2016, Vol. 450, pp. 239-249.
18. *Kureichik V.M., Semenova A.V.* Ensemble of Classifiers for Ontology Enrichment, *Journal of Physics: Conference Series*, 2018, Vol. 1015, Issue 3, Article id. 032123.
19. *Bova V.V., Nuzhnov E.V., Kureichik V.V.* The Combined Method of Semantic Similarity Estimation of Problem Oriented Knowledge on the Basis of Evolutionary Procedures, *Advances in Intelligent Systems and Computing*, 2017, Vol. 573, pp. 74-83.
20. *Pulyavina N., Taratukhin V.* The Future of Project-Based Learning for Engineering and Management Students: Towards an Advanced Design Thinking Approach, *ASEE Annual Conference and Exposition, Conference Proceedings*, 2018, No. 125.
21. *Becker J., Pulyavina N., Taratukhin V.* Next-Gen Design Thinking. Using Project-Based and Game-Oriented Approaches to Support Creativity and Innovation, *Proceedings of the 1st International Conference of Information Systems and Design*, 2020.

Статью рекомендовал к опубликованию к.т.н., доцент С.Г. Буланов.

Мансур Али Махмуд – Южный федеральный университет; e-mail: mansur@sfedu.com; г. Таганрог, Россия; тел.: 88634371651; кафедра систем автоматизированного проектирования; аспирант.

Мохаммад Жуман Хуссайн – e-mail: zmohammad@sfedu.ru; кафедра систем автоматизированного проектирования; аспирант.

Кравченко Юрий Алексеевич – e-mail: yakravchenko@sfedu.ru; тел.: +79289080151; кафедра систем автоматизированного проектирования, доцент.

Mansour Ali Mahmoud – Southern Federal University; e-mail: mansur@sfedu.com; Taganrog, Russia; phone: +78634371651; the department of computer aided design; graduate student.

Mohammad Juman Hussain – e-mail: zmohammad@sfedu.ru; the department of computer aided design, graduate student.

Kravchenko Yury Alekseevich – e-mail: yakravchenko@sfedu.ru; phone: +79289080151; the department of computer aided design, associate professor.

УДК 519.224.22 DOI 10.18522/2311-3103-2021-2-167-181

А.К. Мельников

ОГРАНИЧЕНИЕ КОЛИЧЕСТВА РАЗЛИЧНЫХ ОПРОБУЕМЫХ ВЕКТОРОВ ДЛЯ ПОЛУЧЕНИЯ ВСЕХ РЕШЕНИЙ СИСТЕМЫ ЛИНЕЙНЫХ УРАВНЕНИЙ ВТОРОЙ КРАТНОСТИ НА МНОГОПРОЦЕССОРНОЙ ВЫЧИСЛИТЕЛЬНОЙ СИСТЕМЕ

Статья посвящена нахождению всех целочисленных неотрицательных решений системы линейных уравнений второй кратности типов, далее с.л.у., методом последовательного опробования векторов на принадлежность к решениям системы. Рассматривается количество различных векторов, опробование которых на принадлежности к решениям с.л.у. приведет к получению всех решений с.л.у. Вектор опробований с.л.у. состоит из элементов определяющих число знаков алфавита, имеющих одинаковое число вхождений в выборку. С.л.у. связывает между собой число вхождений элементов всех типов в рассматриваемую выборку, мощность алфавита, объём выборки и ограничение на максимальное число вхождений знаков алфавита в выборку. Решение с.л.у. является основой расчета точных распределений вероятностей значений статистик и их точных приближений методом второй кратности, где в качестве точных приближений выступают Δ–точные распределения, отличающиеся от точных распределений не более чем на заранее заданную, сколь угодно малую величину Δ. Величина, выражающая количество опробуемых векторов, является одной из величин определяющих алгоритмическую сложность метода второй кратности, без знания значения которой нельзя определить параметры выборок, для которых при ограничениях на вычислительный ресурс могут быть рассчитаны точные распределения и их точные приближения. Количество различных опробуемых векторов рассматривается в условиях ограничения на максимальное значение числа вхождений элементов алфавита в выборку, так и без ограничений. Найдены аналитические выражения, позволяющие для любых значений мощности алфавита, объёма выборки и ограничения на значение максимального числа вхождений знаков алфавита в выборку вычислять количество опробований различных векторов для получения всех целочисленных неотрицательных решений системы линейных уравнений второй кратности типов. Вид полученного аналитического выражения для количества опробований векторов позволяет использовать его при изучении алгоритмической сложности расчетов точных распределений и их точных приближений с заранее указанной точностью Δ.

Вероятность; точное распределение; точное приближение; система линейных уравнений; алгоритмическая сложность; многопроцессорная вычислительная система.