DOI: 10.12731/2227-930X-2022-12-4-81-95
УДК 681.3

АНАЛИЗ ВОЗМОЖНОСТЕЙ ИСПОЛЬЗОВАНИЯ АЛГОРИТМОВ ФИЛЬТРАЦИИ ИНФОРМАЦИИ И ЗАПОЛНЕНИЯ ПРОБЕЛОВ

Я.Е. Львович, А.П. Преображенский, Т.В. Аветисян

Анализ показывает, что реальные информационные потоки, характеризуются: произвольной функцией распределения; наличием как количественных, так и качественных показателей; отсутствием части содержательных фрагментов данных; недостоверностью данных, связанной с ошибками при регистрации и их вводе; параметрической избыточностью.

Не все существующие средства, включенные в состав большинства инструментальных систем, дают возможность решать ряд перечисленных задач.

В статье рассматривается задача, связанная с фильтрацией информации. Представлена иллюстрация схемы алгоритма, позволяющего вести информационную фильтрацию. Указаны подробно этапы, используемые в ходе построения алгоритма. Применяется геометрический подход. Он дает возможности для того, чтобы проводить рассмотрение информационных сообщений в виде «созвездий», которые связаны с в і-мерным гиперпространством признаков. Предложен алгоритм, который позволяет реализовать заполнение пробелов.

Представленные в работе результаты могут быть полезными при рассмотрении различных потоков информации в современных информационных системах.

Ключевые слова: *алгоритм; фильтрация информации; заполнение пробелов; показатель*

 Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International

ANALYSIS OF THE USE OF ALGORITHMS TO FILTER INFORMATION AND FILL GAPS

Ya.E. Lvovich, A.P. Preobrazhensky, T.V. Avetisyan

The analysis shows that the real information flows are characterized by: arbitrary distribution function; the presence of both quantitative and qualitative indicators; the absence of part of the meaningful data fragments; unreliability of data associated with errors in registration and their entry; parametric redundancy.

Not all of the existing tools included in most of the instrumental systems, make it possible to solve a number of these problems. The article deals with the problem related to information filtering. The illustration of an algorithm scheme, which allows to conduct information filtering, is presented. The stages used in the construction of the algorithm are specified in detail. The geometric approach is applied. It gives an opportunity to lead consideration of information messages in the form of "constellations" which are connected with the i-dimensional feature hyperspace. An algorithm that allows to implement gap filling is proposed. The results presented in the paper can be useful when considering different information flows in modern information systems.

Keywords: *algorithm; information filtering; gap filling; indicator*

Введение

Сейчас можно наблюдать активное использование различных систем, позволяющих осуществлять статистическую обработку информации в различных сферах. Для тех данных, которые применяются в практических приложениях, можно отметить произвольные функции распределения, наличие пропусков в данных, влияние различных помех, данные могут быть недостоверными, в них могут встречаться ошибки.

Не во всех случаях существуют возможности для того, чтобы по произвольным наборам данных осуществлять процессы каче-

ственной обработки. Это ведет к необходимости разработки соответствующих алгоритмов.

Цель работы связана с разработкой подхода, направленного на фильтрацию информации.

Предложения по процедурам фильтрации информации

При рассмотрении функционирования информационных систем во многих случаях необходимо обеспечивать фильтрацию информации [1, 2]. Тогда обеспечивается выбор таких информационных сообщений, которые характеризуются такой совокупностьюсведений, которая является наиболее вероятной.

Множество объектов позволяют рассматривать всю анализируемую информацию таким образом

$$G_{исх} = \bigcup_{n=1}^{N_{исх}} g_n, \quad (1)$$

в ходе моделирования считается, что $N_{исх}$ рассматривается в виде объем первичной выборки.

Набор показателей связан с каждым из наблюдений:

$$\forall n : \; g_n \to P_n = \left\{P_n^1, P_n^2, ..., P_n^i, ..., P_n^{I_{исх}}\right\} \quad (2)$$

в ходе рассмотрения считается, что $i = \overline{1, I_{исх}}$ будет индексом показателя,

$n = \overline{1, N_{исх}}$ при наблюдении считается порядковым номером.

Требуется, прежде всего, указать нижнюю и верхнюю возможные границы P^i_{min} и P^i_{max}, $(i = \overline{1, I_{исх}})$, которые относятся к каждому показателю P^i $(i = \overline{1, I_{исх}})$. За границы можно выйти лишь вследствие того, что естьошибки визмерениях или в ходе записей показателей [3]. После этого ведется отсев по недостоверным сообщениям. Возникнуть они могут вследствие того, что за некоторые установленные границы будет выходопределенных параметров. Формируется множество

$$G_{гр} = \bigcup_{n=1}^{N_{гр}} g_n. \quad (3)$$

Это позволит включить в него лишь измерения, которые связаны с таким условием:

$$\forall n, \forall i: P^{i}_{min} \le P^{i}_{n} \le P^{i}_{max} \quad (4)$$

После этого в фильтрации выделяется второй этап. При этом обозначается некоторый порог w_0. Оценкой достоверности w_n $(n = \overline{1, N_{гр}})$ по информационным сообщениям должна превышать указанный порог. Если в информации есть лингвистические значения, тогда необходимо осуществить процессы преобразования с привлечением экспертных оценок в численный вид.

Если информация представлена численным способом [4, 5], тогда необходимо, чтобы множество, связанное с информационными сообщениями (3), было отображено к множеству оценок достоверности первичных сообщений

$$W = \bigcup_{n=1}^{N_{гр}} w_n. \quad (5)$$

Реализуется процесс, в котором формируется множество $G \subseteq G_{гр}$ тех сообщений, которые будут отфильтрованными $G_{ф} = \bigcup_{f=1}^{N_{ф}} g_f$ на основе правила

$$\forall n \left(n = \overline{1, N_{гр}}\right) : g_n \to \begin{cases} g_f \mid w_n \ge w_0 \\ 0 \mid w_n < w_0 \end{cases}, \quad (6)$$

В указанном выражении 0 – рассматривается в виде пустого множества.

На практике можно столкнуться с тем, что в разных единицах измерения будут измеряться, различные показатели. Должна быть проведена нормировка по всем показателям, чтобы обеспечивать эффективную работу алгоритма фильтрации информации. Единообразное описаниепо всем признакам позволяет достичь нормировки. Тогда объекты могут быть формальным образом сопоставлены [6, 7]. Мы можем осуществлять какой-то из переходов от первичных значений *x* к нормированным *z*, если реализуется нормировка:

$$z^1 = (x - \bar{x}) / \sigma, \tag{7}$$

$$z^2 = x / \bar{x}, \tag{8}$$

$$z^3 = x / x', \tag{9}$$

$$z^4 = x / x_{max}, \tag{10}$$

$$z^5 = (x - \bar{x}) / (x_{max} - x_{min}), \tag{11}$$

в ходе моделирования $\bar{x}, \sigma$ рассматриваются в виде среднего и среднеквадратического отклонения *x*;

x' рассматривается в виде эталонного значения x;

при анализе x_{max}, x_{min} являются наибольшим и наименьшим значением*x*.

По шкалам, связанным с отношениями и интерваламииспользуются нормировки z^1 и z^2. Другие из них применяются лишь по шкалам отношений.

Если изменяются значения показателей, тогда удобно применять нормировку по допустимому диапазону. Необходимо применять следующую формулу в задаче (3):

$$\forall n, \forall i : P_norm_n^i = \frac{P_n^i - P^i_{min}}{P^i_{max} - P^i_{min}}. \tag{12}$$

Применяется концепция типичности, когда определяется степень достоверности в информационных сообщениях в ходе рассмотрения задачи (5)-(6). Будет тем больше значение в достоверности w_n информационных сообщениях, чем оно типично в анализируемой. Численные значения используются для того, чтобы представлять сведения по сообщениям $\{P_n^i\}$. В этой связи можно применят геометрический подход. Он дает возможности для того, чтобы проводить рассмотрение информационных сообщений в виде «созвездий», которые связаны с в i-мерным гиперпространством признаков. То, насколько первичная выборка по сообщениям $G_{гр}$ является «засоренной», будет оказывать влияние на тот метод решения, который используется.

В случае малой «засоренности» выборки $G_{гр}$ можно это представить соответствующим геометрическим образом. Тогда будет группировка по сообщениям g_n вокруг некоторого мнимого центра тяжести. Группировка будет симметричной. Чем меньше расстояние до обобщенного сообщения g_0, тем будет больше значение вероятности. Для такого сообщения есть численный набор сведений $P_0 = \{P_0^1, P_0^2, ..., P_0^i, ..., P_0^{I_{гр}}\}$. Выполняется условие

$$\forall i \ : \ P_0^i = \sum_{n=1}^{N_{гр}} P_norm_n^i / N. \qquad (13)$$

Необходимо найти решение. Оно связано с тем, что вычисляется значение, соответствующее вектору расстояний $S = \{S_1, S_2, ..., S_n, ..., S_{N_{гр}}\}$. между сообщениями $g_n \in G_{гр}$ и обобщенным сообщением g_0. При этом необходимо ориентироваться на некоторую метрику. В качестве примера может быть евклидова метрика:

$$\forall n \ : \ S_n = \left[\sum_{\forall i} (P_norm_n^i - P_norm_0^i)^2 \right]^{1/2} \qquad (14)$$

Для того, чтобы по сообщениям g_n дать оценку степени достоверности, можно опираться на выражение

$$w_n = S_{min} / S_n, \qquad (15)$$

учитывается, что $S_{min} = \min\limits_{\forall n} S_n$.

Может быть случай большей «засоренности» по выборке $G_{гр}$. Например, может анализироваться половина выборки. В таких случаях можно говорить об асимметрии распределения анализируемых данных. Использование обобщенного сообщения g_0 не дает преимуществ в рассмотрении выборки $G_{гр}$. Тогда может помочь использование соответствующей метрики. Она должна быть адекватной с точки зрения сведений метрики. Происходит вычисление между информационными сообщениями значений векторов суммарных расстояний $S = \{S_1, S_2, ..., S_n, ..., S_{N_{гр}}\}$ таким образом

$$s_n = \left[\sum_{m=1}^{N_{гр}} \sum_{i=1}^{I_{исх}} \left(P_n^i - P_m^i \right)^2 \right]^{1/2}, \quad n = \overline{1, N_{гр}} \tag{16}$$

Также как в (15) можно по сообщениям рассмотреть степень достоверности.

В случае сильной «засоренности» выборки $G_{гр}$ можно учесть гипотезу компактности. Она относится другим возможным группировкам. В таком случае можно применять способы, связанные с кластерным анализом. Для того, чтобы определить степеньдостоверности информациимоможно применять соответствующие методики [8]. Происходит организация $M=N_{гр}-2$ итерационных циклов, в которых есть индексы $C = \overline{2, (N_{гр} - 1)}$, в случае, когда в выборке $G_{гр}$ объем будет $N_{гр}$. Выборка $G_{гр}$ классифицируется по С классам для каждого итерационного цикла. Происходит подсчет по всем итерациям при объеме $V_C \geq 2h_n$ общего числа включений каждого сообщения в классы K_C

$$h_n = \sum_{C=2}^{N_{гр}-1} 1 \;\Big|\; g_n \to K_C \,\&\, V_C, \quad n = \overline{1, N_{гр}} \tag{17}$$

В информационном сообщении g_n будет степень достоверности

$$w_n = h_n / \; (N_{гр} - \; 2) \tag{18}$$

Метод «средней связи» может быть использован при классификации выборки сообщений $G_{гр} = \bigcup_{\forall T} g_n$ с учетом соответствующего числа классов С. Тогда происходит построение матрицы S взаимных расстояний среди сообщений. При этом необходимо выбирать метрику, например евклидову

$$s_{mn} = \left[\sum_{i=1}^{I_{исх}} \left(P_m^i - P_n^i \right)^2 \right]^{1/2}, \quad n = \overline{1, N_{гр}} \tag{19}$$

После этого применяется гипотеза компактности, чтобы классифицировать сообщения. Тогда происходит расположение сходных сообщений внутри гиперпространства сведений компактным образом [9]. При этом расстояния среди центров классов будут

 Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International

превышать расстояния среди сообщений, которые лежат в одном классе.

Установленное значение w_0 оказывает влияние на объем отфильтрованных сообщений $G_ф$.

Анализ особенностей алгоритма, позволяющего вести заполнение пробелов

Были рассмотрены возможности формирования алгоритма, который позволяет пропущенные значения соответствующим образом заполнять. В ходе его создания мы опирались на несколько предпосылок.

Первая связана с тем, что когда используются на практике таблицы с данными, то в них наблюдается свойство избыточности. То есть объекты – строки будут похожими. Столбцы характеризуются зависимыми свойствами. Когда в таблице находятся случайные числа, тогда не будет избыточности. В таком случае нет возможностей для того, чтобы один прогноз был предпочтен другому.

Предпосылка вторая связана с тем, что при близости пары объектов близка с точки зрения (n-1) свойств, можно считать, что будет наблюдаться близость относительно n-ого свойства. То есть, выполняется гипотеза аналогичности.

Предпосылка третья базируется на локальном характере избыточности (локальная компетентность). Тогдасуществует подмножество объектов-аналогов по каждому объекту. Также подмножество свойств-аналогов может быть выделено по каждому свойству. При заполнении пробелов необходимо соответствующим образом делать выбор строк и столбцов отдельным образом по каждому предсказываемому элементу [10]. Нормировка по всем показателям должна быть осуществлена, чтобы этот алгоритм был использован. Можно указать три этапа по анализируемому алгоритму.

1. Выделяется множество строк, которые можно использовать и после этого по таким строкам выделяется множество столбцов с точки зрения «соответствия».

2. Проводится проверка на то, чтобы прогнозная ошибка по была минимальной в ходе расчетов.

3. Реализуется процесс прогнозирования интересующих элементов.

Когда рассматривается 1-я строка, то она будет «соответствовать» с точки зрения i-ой строки, если выбирается величина L_{il}. Она должна быть обратно пропорциональной расстоянию между указанными строками. Будет обеспечиваться «соответствие» между k-м столбцами j-м столбцом при выборе величины L_{jk}. Она должна быть пропорциональнойтому расстоянию, которое будетсреди столбцов. После этого осуществляетсяформирование подматрицы. В нее входят от 3-х до 7-и строк и столбцов.

В тех строках и столбцах, которые обеспечивают «соответствие», не должны быть пробелы для j-го столбца (i-ой строке).

Должна быть обеспечена нормировка по всем показателям, чтобы алгоритм функционировал эффективным способом. Тогда будет введена новой единицы измерения, что дает возможности для того, чтобы все признаки были представлены единообразным способом. Тогда сопоставление объектов формальные будет обеспечено формальным образом.

По допустимому диапазону, в котором изменяются значения показателей, можно обеспечивать процессы нормировки. В анализируемой задаче следующую формулу можно применять на практике:

$$\forall n,\forall i : y_norm_n^i = \frac{y_n^i - y_{min}^i}{y_{max}^i - y_{min}^i} \quad (20)$$

Происходит выработка «подсказок» b(k) = F(X(k)) с привлечением уравнений регрессии, когдапредсказывается значение пробела на основе зависимостей среди j-ым и другими (k-ыми) столбцами. Используется вес, который пропорциональны компетентности соответствующего столбца при у усреднения q подсказок, когда внутри под матрицы анализировался q+1 столбец. Тогда,

вследствие избыточности, которая есть внутри столбцовбудет формироваться прогнозная величина b(j):

$$b(j) = \sum_{k=1}^{q} \left(b(k) \cdot L(jk)^{\alpha} \right) / \sum_{k=1}^{q} L(jk)^{\alpha} \qquad (21)$$

В ходе моделирования используется α – коэффициент, который позволяет вести регулирование влияние соответствия на результат предсказания. Чем больше значение используется для α, тем больше будет влияние со стороны соответствия столбцов. Разные значения α используются в ходе предсказания всех известных элементов в j-м столбце. После этого обеспечение минимальной ошибки по прогнозу обеспечивается при выборе соответствующего значения α. Осуществляется процесс прогноза b(j) величины того элемента, который пропущен, при выбранном значении α на основе формулы (21).

Используется зависимость среди i-ой строкой и s остальными (1-ыми) строками (1, 2,..l...s), если заполняется пробел. Тогда можно опираться на следующую формулу

$$b(i) = \sum_{l=1}^{s} \left(b(l) \cdot L(il)^{\alpha} \right) / \sum_{l=1}^{s} L(il)^{\alpha} \qquad (22)$$

Исходим из того, что достигается минимальная ошибка в предсказании всех известных элементов вi-ой строке, когда осуществляется процесс выбора α. Усреднение b(i) и b(j) позволяет прогнозировать y'_{ij} значение в пропущенном элементе y_{ij}.

В «соответствующей» подматрице проводится вычисление по дисперсии величин подсказок b(k) и b(l), которые могут быть получены относительно всех k столбцов и l строк. Тогда определяется ожидаемая ошибкав предсказании. Не будет устойчивая закономерная связь среди элемента y_{ij} и остальными элементами в подматрице, если наблюдаются большие значения в дисперсии.

На рис. 1 дана иллюстрация схемы алгоритма, позволяющего реализовывать информационную фильтрацию.

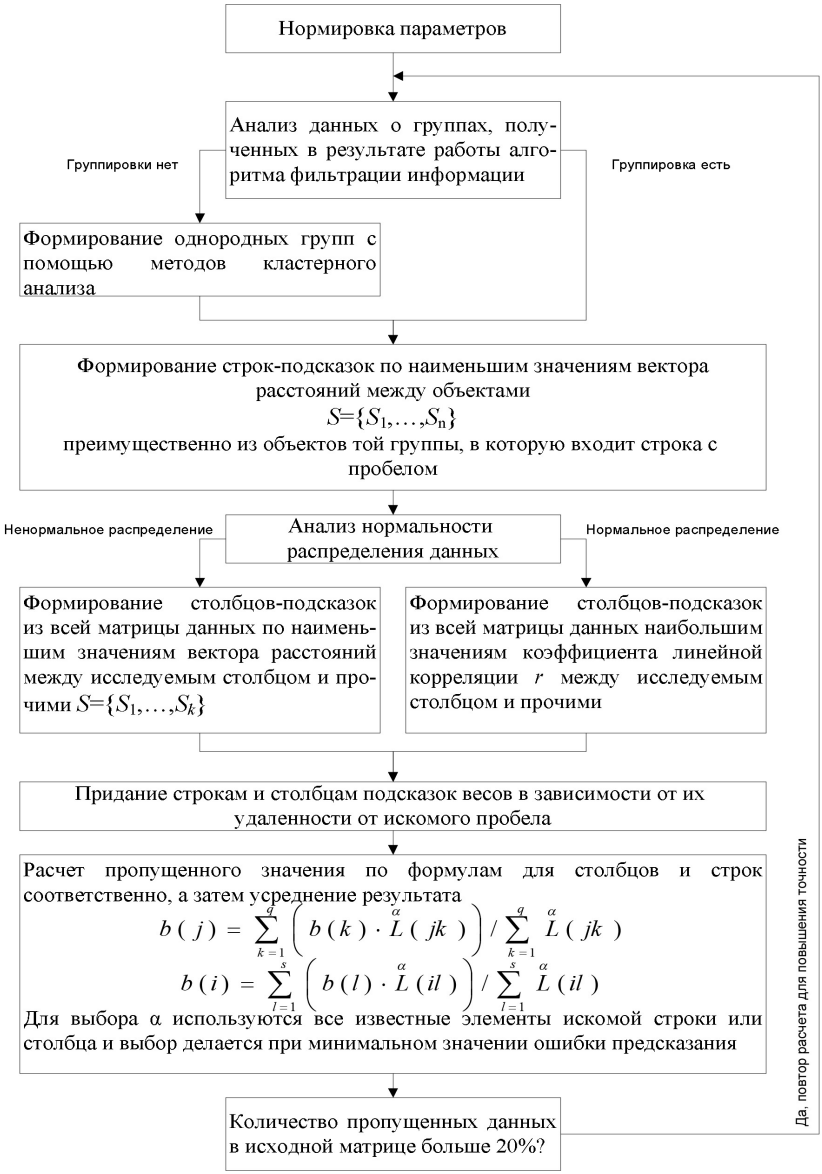

Рис. 1. Иллюстрация схемы алгоритма, который позволяет реализовать заполнение пробелов

Выводы

Когда строятся классификационно-прогностические модели, тогда необходимо чтобы была использовались лингвистические переменные, если информация будет преобразована к численному виду. Качество исходных данных оказывает влияние на точность применяемых моделей. Информация должна быть предварительным образом обработана. Предложен алгоритм заполнения пробелов в информации. Его эффективность может быть повышена за счет того, используются уравнения регрессии высших порядков.

Список литературы

1. Львович Я. Е., Преображенский Ю. П., Ружицкий Е. Анализ особенностей приема и передачи сигналов в компьютерных сетях // Вестник Воронежского института высоких технологий. 2022. № 1 (40). С. 75–78.
2. Львович Я. Е., Преображенский Ю. П., Ружицкий Е. Исследование характеристик рубрикации текстов // Вестник Воронежского института высоких технологий. 2022. № 2 (41). С. 105–108.
3. Чупринская Ю.Л., Линкина А.В. Краткий обзор современных технологических трендов в контексте цифровой трансформации // Вестник Воронежского института высоких технологий. 2022. № 1 (40). С. 107–109.
4. Львович К. И. Управление эффективностью деятельности персонала в условиях цифровой трансформации организационных систем // Моделирование, оптимизация и информационные технологии. 2020. Т. 8. № 3 (30). https://doi.org/10.26102/2310-6018/2020.30.3.039
5. Rykova V.V. The northern sea route: a scientometric analysis of information arrays in Web of Science and Scopus databases // International Journal of Advanced Studies. 2021. Т. 11. № 1. С. 77–86. https://doi.org/10.12731/2227-930X-2021-11-1-77-86
6. Львович Я. Е., Преображенский Ю. П., Ружицкий Е. О проблемах использования современных технологий в «умных городах»

// Вестник Воронежского института высоких технологий. 2022. № 2 (41). С. 116–119.
7. Львович Я. Е., Преображенский Ю. П., Ружицкий Е. Исследование характеристик алгоритмов классификации // Вестник Воронежского института высоких технологий. 2022. № 2 (41). С. 96–99.
8. Plotnikova T.A. Computer aided schematic design: how to manage large data in the early stages of urban design projects // International Journal of Advanced Studies. 2021. Т. 11. № 2. С. 52–64. https://doi.org/10.12731/2227-930X-2021-11-2-52-64
9. Rykova V.V. The arctic regional security: a scientometric analysis of information arrays in databases Web of Science and Scholar Sibirica // International Journal of Advanced Studies. 2020. Т. 10. № 3. С. 18–29. https://doi.org/10.12731/2227-930X-2020-3-18-29
10. Линкина А.В., Лихачев С. А. Применение корреляционного и регрессионного анализа в геоинформационных системах // Вестник Воронежского института высоких технологий. 2022. № 2 (41). С. 80–82.

References

1. L'vovich YA.E., Preobrazhenskij YU.P., Ruzhickij E. *Vestnik Voronezhskogo instituta vysokih tekhnologij,* 2022, no. 1 (40), pp. 75–78.
2. L'vovich YA.E., Preobrazhenskij YU.P., Ruzhickij E. *Vestnik Voronezhskogo instituta vysokih tekhnologij,* 2022, no. 2 (41), pp. 105-108.
3. CHuprinskaya YU.L., Linkina A.V. *Vestnik Voronezhskogo instituta vysokih tekhnologij,* 2022, no. 1 (40), pp. 107-109.
4. L'vovich K.I. *Modelirovanie, optimizaciya i informacionnye tekhnologii,* 2020, vol. 8, no. 3 (30). https://doi.org/10.26102/2310-6018/2020.30.3.039
5. Rykova V.V. The northern sea route: a scientometric analysis of information arrays in Web of Science and Scopus databases. *International Journal of Advanced Studies,* 2021, vol. 11, no. 1, pp. 77-86. https://doi.org/10.12731/2227-930X-2021-11-1-77-86
6. L'vovich YA.E., Preobrazhenskij YU.P., Ruzhickij E. *Vestnik Voronezhskogo instituta vysokih tekhnologij,* 2022, no. 2 (41), pp. 116-119.

7. L'vovich YA.E., Preobrazhenskij YU.P., Ruzhickij E. *Vestnik Voronezhskogo instituta vysokih tekhnologij,* 2022, no. 2 (41), pp. 96-99.
8. Plotnikova T.A. Computer aided schematic design: how to manage large data in the early stages of urban design projects. International Journal of Advanced Studies. 2021. T. 11. № 2. S. 52-64. https://doi.org/10.12731/2227-930X-2021-11-2-52-64
9. Rykova V.V. The arctic regional security: a scientometric analysis of information arrays in databases Web of Science and Scholar Sibirica. *International Journal of Advanced Studies,* 2020, vol. 10, no. 3, pp. 18-29. https://doi.org/10.12731/2227-930X-2020-3-18-29
10. Linkina A.V., Lihachev S.A. *Vestnik Voronezhskogo instituta vysokih tekhnologij,* 2022, no. 2 (41), pp. 80–82.

ДАННЫЕ ОБ АВТОРАХ

Львович Яков Евсеевич, профессор, доктор технических наук, профессор
Воронежский государственный технический университет
ул. 20 лет Октября, 84, г. Воронеж, 394006, Российская Федерация
Komkovvivt@yandex.ru

Преображенский Андрей Петрович, профессор, доктор технических наук, доцент
Воронежский институт высоких технологий
ул. Ленина, 73а, г. Воронеж, 394043, Российская Федерация
Komkovvivt@yandex.ru

Аветисян Татьяна Владимировна, старший преподаватель
Воронежский институт высоких технологий
ул. Ленина, 73а, г. Воронеж, 394043, Российская Федерация
Komkovvivt@yandex.ru

DATA ABOUT THE AUTHORS

Yakov E. Lvovich, Doctor of Technical Sciences, Professor
Voronezh State Technical University

84, 20 years of October Str., Voronezh, 394006, Russian Federation
Komkovvivt@yandex.ru
ORCID: https://orcid.org/0000-0002-7051-3763

Andrey P. Preobrazhenskiy, Professor, Doctor of Technical Sciences, Associate Professor
Voronezh Institute of High Technologies
73a, Lenin Str., Voronezh, 394043, Russian Federation
Komkovvivt@yandex.ru
ORCID: https://orcid.org/0000-0002-6911-8053

Tatiana V. Avetisyan, Senior Lecturer
Voronezh Institute of High Technologies
73a, Lenin Str., Voronezh, 394043, Russian Federation
Komkovvivt@yandex.ru
ORCID: https://orcid.org/0000-0003-3559-6070

Поступила 27.10.2022 Received 27.10.2022
После рецензирования 15.11.2022 Revised 15.11.2022
Принята 20.11.2022 Accepted 20.11.2022