

УДК 007:681.512.2

НЕЙРОННЫЕ СЕТИ ДЛЯ ИДЕНТИФИКАЦИИ ПОЛЬЗОВАТЕЛЯ НА ОСНОВЕ АНАЛИЗА ПОСЕЩЕНИЙ НОВОСТНОГО САЙТА

И. Ю. Каширин, д.т.н., профессор кафедры ВПМ РГРТУ, Рязань, Россия;
orcid.org/0000-0003-1694-7410, e-mail: igor-kashirin@mail.ru.

Рассматривается проектирование нейронных сетей со знаниями. Проектируемые сети предназначены для решения задачи классификации пользователей по их психосоматическому типу. Излагается оригинальный метод построения сетей с помощью ICF-онтологий, структурированных в соответствии с причинно-следственной таксономией информационных потребностей пользователя Web-сайта.

Для проектирования онтологической причинно-следственной таксономии используются элементы теории иерархических чисел, позволяющие вычислять семантическую близость концептов онтологии для ее оптимизации. Признаки, используемые в таксономии, применяются в дальнейшем в формировании множества базовых признаков (S-элементов) нейронной сети.

В программной реализации предложенного подхода используется предметная область сайта текущих новостей с множеством разнонаправленных информационных рубрик. Целью применения нейронной сети является классификация пользователей по их психосоматическому типу на основе анализа посещения материалов сайта различными пользователями. В качестве практической части рассматриваются эксперименты, спланированные и поставленные в рамках программного инструментария Python v.3 (Anaconda 3). Полученные результаты позволяют положительно оценить эффективность предложенной технологии.

***Целью работы** является создание оригинального метода построения нейронных сетей для задач классификации с использованием причинно-следственных таксономий в формате ICF-онтологии с применением иерархических чисел.*

***Ключевые слова:** нейронные сети, причинно-следственная таксономия, психосоматическая идентификация пользователей, ICF-онтология, иерархические числа, семантическая близость.*

DOI: 10.21667/1995-4565-2022-82-104-111

Введение

Теоретическая концепция больших данных (Big Data) [1] предполагает массовый сбор и анализ информации в какой-либо актуальной предметной области. Одной из значимых предметных областей является идентификация пользователя информационных ресурсов. Задача идентификации заключается в определении универсума пользователей, занимающихся одним или несколькими видами деятельности с целевыми функциями, характерными для заранее известных или не полностью определенных классов пользователей. При этом пользователи могут быть как клиентами обслуживающих их компаний, так и самостоятельными участниками информационной деятельности, предпринимающими попытки получения актуальных для них сведений.

Анализ деятельности таких участников информационных операций дает возможность выявить классы пользователей по их отношению к определенным кластерам данных. Такой анализ принято чаще всего производить с целью определения информационных потребностей участников информационной деятельности с тем, чтобы в дальнейшем использовать полученные данные в рекламных компаниях для целевых аудиторий.

В то же время существует наиболее общая классификация участников информационной деятельности как личностей определенных психолого-соматических групп, т.е. личностей, идентифицированных по своему психотипу и состоянию здоровья. При этом психотип определяется как система психических признаков, соответствующая обобщенной модели поведе-

ния личности и ее реакции на различные факторы окружающей среды, в том числе информационной. Такие модели личности характерны для больших групп людей, что дает возможность объединять их в классы для идентификации. Личность определенного класса нуждается в создании среди своего повседневного инструментария бытовой и профессиональной деятельности комплекса адекватных своему классу информационных источников. В современной обстановке таким комплексом становятся найденные пользователем Web-сайты, отвечающие его информационным потребностям. Классы пользователей могут быть полностью противоположными или, наоборот, иметь существенную часть теоретико-множественного пересечения интересов. В этом случае можно говорить об иерархии психосоматических типов, причем сопоставление этим типам конкретных наименований может не играть решающей роли для большого множества задач, т.к. такие наименования можно заменить, например, числовыми идентификаторами.

Актуальность задачи классификации или кластеризации [2] пользователей хорошо известна в сфере рекламы, финансового кредитования, выявления потенциальных террористов, поиска специалистов и т.п. В статье предложена концепция решения этой задачи на основе машинного обучения с использованием нейронных сетей и других методов Data Mining [1-3].

Теоретическая часть

Для инициации решения задачи машинного обучения [4] необходимо сформировать входной набор данных большого объема, позволяющий как можно более полно охватить все возможные сочетания признаков, используемых для идентификации пользователя. Такое формирование является само по себе нетривиальной и актуальной задачей. Эта задача может быть решена наиболее эффективно, если у дата-сайентиста (инженера по знаниям и большим данным) есть возможность обслуживать несколько сайтов различной тематики. В противном случае, при анализе работы пользователей с одним сайтом, можно исследовать общие характеристики работы, к которым можно отнести:

- частоту посещения отдельных рубрик Web-страницы;
- длительность присутствия в рамках каждой из рубрик;
- быстроту перехода от одной информационной рубрики к другой;
- дни недели (будни, праздники, выходные) посещения Web-страницы;
- географическое положение пользователя.

Перечисленные признаки дают возможность как минимум выявить базовый психотип пользователя информационного ресурса [5]:

- мыслительный;
- сенсорный;
- эмоциональный;
- интуитивный;
- интравертный;
- экстравертный;
- психологический (флегматик, сангвиник, холерик, меланхолик).

В любом случае, если решать задачу идентификации пользователя, на администрируемый сайт целесообразно добавить новостной раздел, возможно, использующий RSS-ссылки, такие как news-rss.

Анализ собранной хотя бы за несколько месяцев информации даст возможность воспользоваться одним из перечисленных далее методов Data Mining [6, 7].

1. Кластеризация записей (строк) входного набора данных, т.е. выделение множества записей, схожих по информационным отличительным признакам (Features). Решение этой задачи дает возможность косвенно определить класс клиентов, потенциально заинтересованных в продолжении использования Web-сайта. Кроме того, появляется возможность группирования Web-сайтов различной тематической направленности и/или востребованности (по-

пулярности). Это, в свою очередь, может в дальнейшем способствовать решению следующих задач классификации и регрессионного анализа.

2. Классификация записей (строк) входного набора данных, т.е. определение высокой вероятности, как правило, бинарного выбора дальнейших действий клиента: будет ли пользователь продолжать пользоваться Web-ресурсом или нет.

3. Регрессионный анализ данных входного набора, т.е. определение различных не бинарных значений вероятностей продолжения клиентом работы с Web-ресурсом.

В соответствии с концепцией использования семантики признаков [8] можно выделить основную таксономию причинно-следственных связей между классами признаков. В соответствии с этой концепцией необходимо сформулировать вопросы к экспертам предметных областей Web-сайта о возможных причинно-следственных связях в каждой из этих областей.

Приведем примерные разделы причинно-следственных таксономий [8] для идентификации пользователей (рисунок 1).

- По данным о дате и времени посещения Web-сайта
 - ⇒ Тип психики клиента (жаворонок, сова) утро, день, вечер, ночь
 - ⇒ По выходным и праздничным дням – работающий или пенсионер
 - ⇒ По пропускам дней – болезнь или отпуск
 - ⇒ Длительность пребывания на сайте
- По пользовательскому агенту
 - ⇒ Мобильный телефон или компьютер
 - ⇒ Примерный возраст пользователя (современный браузер, часто обновляется)
 - ⇒ IT-специалист, продвинутый пользователь или нет
 - ⇒ Программа-робот или интеллектуальный агент
- По тематике
 - ⇒ Интересы пользователя
 - ⇒ Возраст
 - ⇒ Болезни
 - ⇒ Политическая ориентация
 - ⇒ Пол
- По IP адресу
 - ⇒ Город пользователя
 - ⇒ Погода на нужную дату (давление, ненастье, благоприятная для отдыха)
 - ⇒ Список IP одного пользователя (много => продвинутый пользователь)

Рисунок 1 – Примерные разделы причинно-следственных таксономий для идентификации пользователей Web-ресурсов

Figure 1 – Sample sections of causal taxonomies for identifying users of Web resources

Свойства пользователей являются одновременно признаками тематических рубрик, которые они посещают. Если несколько пользователей посещают одни и те же тематические рубрики, то можно выделить несколько заключений:

- эти рубрики примерно одинаковой тематики (прямая задача);
- эти пользователи относятся к одному и тому же классу по интересам (обратная задача);
- можно оценить популярность тематики рубрик;
- можно оценить продвинутость пользователя по аналогии с лучшими пользователями.

Безусловно, получение достаточно полной информации для использования представленных причинно-следственных зависимостей требует привлечения дополнительных внешних источников данных и достаточно трудоемкой разработки соответствующего программного инструментария.

Вместе с тем реализация рассмотренных задач даст возможность не только повысить качество ML-моделей для предсказания дальнейших действий клиента, но и решить следующие дополнительные задачи.

1. Построить психосоматическую модель клиента, позволяющую идентифицировать пользователя по его поведению в рамках Web-ресурсов.

2. Построить классификацию клиентов, позволяющую идентифицировать класс пользователя по его поведению в рамках Web-ресурсов.

3. Получить классификацию Web-ресурсов по поведению пользователей различных классов в рамках этих Web-ресурсов.

4. Выработать рекомендации для компаний-владельцев Web-ресурсов по изменению их содержательной и дизайнерской частей.

Кроме того, решение задачи идентификации дает возможность выработать рекомендации для решения таких задач, как:

SVD (Singular Value Decomposition – сингулярное разложение матрицы оценок для формирования рекомендаций),

Timing (выбор подходящего момента или интервала времени для рекламных действий),

Clustering (кластеризация как определение классов объектов без заранее заданного критерия),

Customer Lifetime Value LTV (прогнозирование длительности заинтересованности потребителя).

Для продвижения перспективных проектов в области идентификации пользователей в качестве рабочего примера предлагается следующая тематическая таксономия, которая может выглядеть следующим образом (рисунок 2).

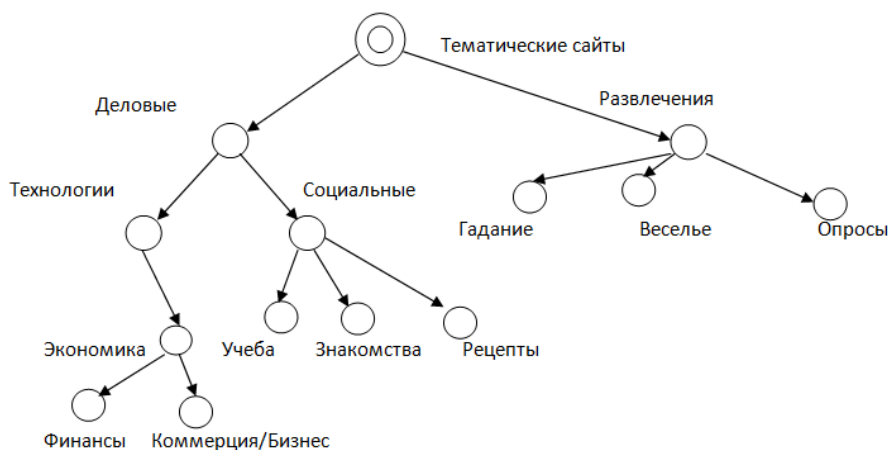


Рисунок 2 – Пример родовидовой тематической таксономии Web-сайтов
Figure 2 – An example of generic thematic taxonomy of Web sites

Приведенный фрагмент таксономии представляет собой иерархию и сложен для реализации нейронной сети, где вершины используются в качестве признаков (S-элементов).

В этом случае для графического представления таксономии предлагается использовать ICF-онтологию [1-3]. В рамках этого формализма терминальные вершины таксономии могут быть упорядочены в соответствии с величинами семантической близости соответствующих им концептов (понятий, признаков). Такая технология предполагает применение правила минимальности пути смежного наследования концептов:

$$\left| \bigcup_{i=1}^m \bigcup_{j=n-1}^1 Is - A(x_i^j, x_i^{j+1}) \right| = \tau.$$

$$[\exists \mu, \mu < \tau],$$

$$\left| \bigcup_{i=1}^m \bigcup_{j=k-1}^1 Is - A(x_i^j, x_i^{j+1}) \right| = \mu,$$

где μ и τ – меры семантической близости концептов, m – количество терминальных вершин таксономии, x_i^1 – корневая вершина, являющаяся предком всех терминальных вершин, x_i^n – терминальные вершины таксономии, $\exists$ – квантор существования. Числителем и знаменате-

лем здесь являются мощности множества всех путей от терминальных вершин до корневого предка как элементов бинарного родовидового отношения Is-A соответственно до применения упорядочивания признаков и после него. По сути дела такое упорядочивание состоит в трансформации онтологии, уменьшающей мощность исходного множества путей.

После произведенной оптимизирующей трансформации необходимо использовать алгебру иерархических чисел [3], дающую возможность представить терминальные вершины приведенного графа сложными числами, которые можно трансформировать в битовые последовательности с дальнейшей нормализацией в числовой отрезок [0,1]. Графически такой отрезок может представляться с помощью оси семантического пространства из n признаков, которой можно условно дать наименование оси делового уровня (уровня деловой активности пользователя).

Качественные значения оси делового уровня можно представить следующей упорядоченной цепочкой элементов: «финансы (1) – коммерция (0.9) – учеба (0.7) – рецепты (0.6) – знакомства (0.5) – опросы (0.3) – гадание (0.2) – веселье(0)». При необходимости сокращения размерности семантического пространства цепочку признаков можно упростить: «Экономика/Технологии (1) – Социум (0.5) – Развлечения (0)».

Аналогичным образом на уровне S-элементов нейронной сети можно представить причинно-следственную таксономию рисунка 1.

Полученные числовые характеристики являются числовыми характеристиками их семантической близости и поэтому могут быть использованы при оптимизации набора признаков нейронных сетей.

Экспериментальные исследования

Фрагменты программного кода, реализующего работу нейронной сети и оценку точности классификации, приведены далее.

```
# Функция построения модели нейросети (исходная версия)
def build_model():
    model = Sequential()
    model.add(Dense(200, input_dim=55, activation="relu"))
    model.add(Dense(800, activation="relu"))
    model.add(Dense(600, activation="relu"))
    model.add(Dense(400, activation="relu"))
    model.add(Dense(1, activation='linear'))
    model.compile(optimizer='adamax', loss='mse', metrics=[r2, 'accuracy'])
    return model

print("ОТОБРАЖЕНИЕ ROC-КРИВОЙ")
fpr, tpr, thresholds = roc_curve(y_test, y_pred)
def plot_roc_curve(fpr, tpr, label=None):
    plt.plot(fpr, tpr, linewidth=2, label=label)
    plt.plot([0, 1], [0, 1], 'k--') # dashed diagonal
    plt.xlabel('False Positive Rate, FPR (1 - specificity)')
    plt.ylabel('True Positive Rate, TPR (Recall)')
    plt.title('ROC curve')
    plt.savefig("ROC.png")
    return 0

plot_roc_curve(fpr, tpr)
plt.show()

from sklearn.metrics import confusion_matrix
from sklearn.metrics import mean_squared_error
from sklearn.metrics import precision_score

print('\nВариант 1:\n ')
print('Actual Negative [[True Negative][False Positive]')
print('Actual Positiv  [[Fase Negative][True Positive]\n')
print(confusion_matrix(y_test, y_pred)) # Возвращает матрицу ошибок
plot_confusion_matrix(SVC, X_test, y_test, cmap='gray_r')
# Нужно выделить y_true из тестового набора
```

```

print('\nВариант 2:')
# Нормализует по Positive значениям
print(confusion_matrix(y_test, y_pred, normalize='true')) #print(model.score(X_
test, y_test))
# Возвращает точность, полноту и F1-score
print(classification_report(y_test, y_pred)) print('\nВариант 3:')
# Нормализует по предсказанным значениям
print(confusion_matrix(y_test, y_pred, normalize='pred')) #print(lr.score(X_tes
t, y_test))
print(classification_report(y_test, y_pred))
from sklearn.metrics import recall_score
print ("Стратегии усреднения (averagigng strategies)")
print("Weighted average: ", recall_score(y_test, y_pred, average="weighted")) #
Вычисление полноты с учетом дисбаланса классов
print("Macro average: ", recall_score(y_test, y_pred, average="macro")) #
Без учета дисбаланса

```

Качественные характеристики нейронной сети приведены на рисунке 3.

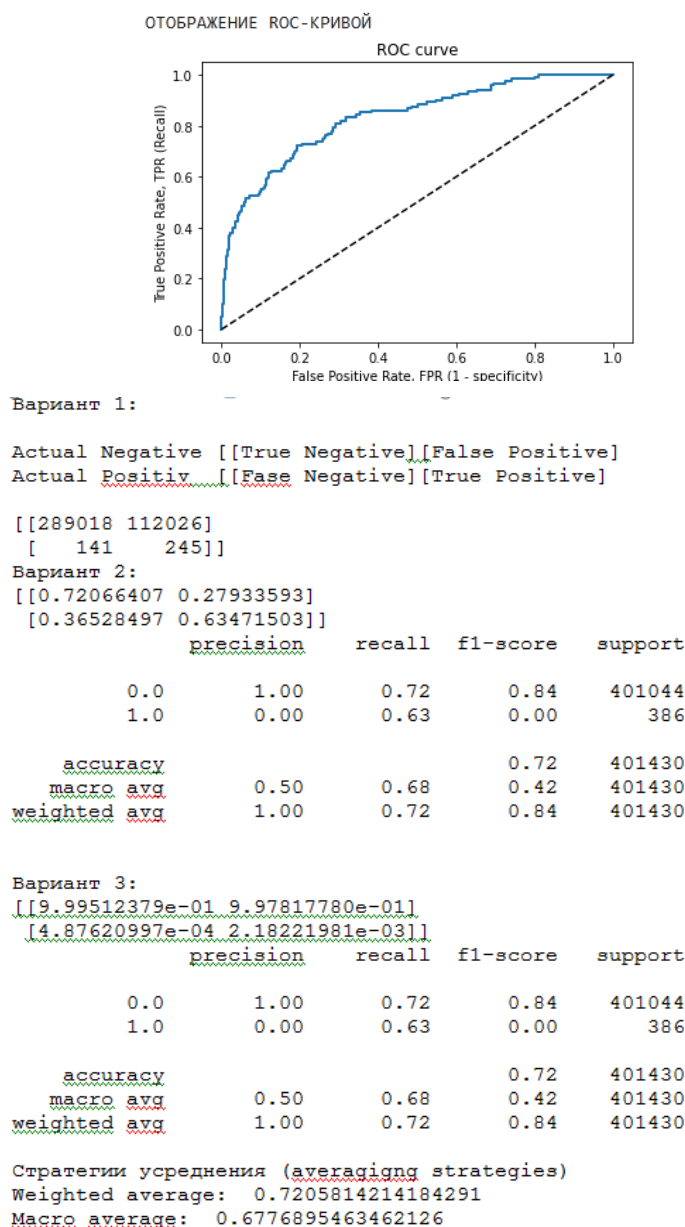


Рисунок 3 – Характеристики качества обучения нейронных сетей
Figure 3 – Characteristics of neural networks training quality

Оптимизация набора входных S-элементов нейронной сети на основе предложенной технологии повышает на 7-10 % качество обучения. Это подтверждается результатами ряда экспериментов (общим числом более 70) для предметной области «Анализ пользовательской активности в новостном сайте».

Эксперименты показали, что испытательный стенд (ЭСМИАД v.14.10.2022), реализованный на языке Python v.3.7 [3] в среде Anaconda 3, дает возможность повысить эффективность применения правил структуризации признаков для входного набора данных в различных предметных областях.

В качестве вывода можно констатировать, что для проведенных экспериментов характеристика macro-average ROC AUC score (площадь под ROC кривой, receiver operating characteristic, рабочая характеристика приёмника) в среднем возросла от 0,67 до 0,72 соответственно до использования правил структуризации и после их использования.

Библиографический список

1. **Каширин И. Ю.** Data Mining с использованием иерархических чисел в ретроспективной диагностике // Вестник Рязанского государственного радиотехнического университета. 2022. № 79. С. 118-126.
2. **Каширин И. Ю.** Интеллектуальный анализ данных в ретроспективной диагностике // V Международный научно-технический форум СТНО-2022: сборник трудов. Том 4. Рязань: Book Jet, 2022. С. 11-17.
3. **Каширин И. Ю.** Иерархические числа в проектировании общих онтологий. V Международный научно-технический форум СТНО-2022: сборник трудов. Том 4. Рязань: Book Jet, 2022. С.17-24.
4. **Каширин И. Ю.** Применение теории иерархических чисел в проектировании ICF-таксономии для оптимизации нейронных сетей // Вестник Рязанского государственного радиотехнического университета. 2022. № 80. С. 36-43.
5. **Утлик Э. П.** Психология личности: учебник / Э. П. Утлик. М.: Academia, 2018. 448 с.
6. **Kashirin I. Yu., Filatov I. Yu.** Formalized Description Of Intuitive Perception Of Spatial Situations // 8th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 2019, pp. 1-4.
7. **Каширин И. Ю., Каширина О. И.** Обзор концепций формального исследования инструментальных программных средств // Вестник Рязанского государственного радиотехнического университета. 2015. № 2. С. 74-83.
8. **Каширин И. Ю.** Нейронные сети, использующие модели знаний. IV Международный научно-технический форум СТНО-2021: сборник трудов. Том 4, Рязань: Book Jet, 2021. С.9-13.
9. **Каширин И. Ю.** Иерархические числа для проектирования ICF-таксономий искусственного интеллекта // Вестник Рязанского государственного радиотехнического университета. 2020. № 71. С. 71-82.
10. **Каширин И. Ю.** Нейронные сети, использующие модели знаний // Вестник Рязанского государственного радиотехнического университета. 2021. № 75. С.71-84.

UDC 007:681.512.2

NEURAL NETWORKS FOR USER IDENTIFICATION BASED ON NEWS SITE VISITS ANALYSIS

I. Yu. Kashirin, Dr. Sc. (Tech.), full professor, RSREU, Ryazan, Russia;
orcid.org/0000-0003-1694-7410, e-mail: igor-kashirin@mail.ru

The article deals with the design of neural networks with knowledge. The designed networks are intended to solve the problem of classifying users according to their psychosomatic type. An original method of building networks using ICF-ontologies, structured in accordance with the causal taxonomy of the information needs of the Web site user, is presented.

To design an ontological causal taxonomy, elements of the theory of hierarchical numbers are used, which allow calculating the semantic similarity of ontology concepts for its optimization. The features used in the taxonomy are applied later in the formation of a set of basic features (S-elements) of neural network.

Software implementation of the approach proposed uses the subject area of current news site with many multidirectional information headings. The purpose of using a neural network is to classify users according to their psychosomatic type based on the analysis of visits to site materials by various users. As a practical part, experiments planned and delivered within the framework of Python v.3 software toolkit (Anaconda 3) are considered. The results obtained allow us to positively assess the effectiveness of the technology proposed.

The aim of the work is to create an original method for constructing neural networks for classification problems using causal taxonomies in ICF ontology format using hierarchical numbers.

Key words: neural networks, causal taxonomy, psychosomatic user identification, ICF ontology, hierarchical numbers, semantic proximity.

DOI: 10.21667/1995-4565-2022-82-104-111

References

1. **Kashirin I. Yu.** Data Mining s ispol'zovaniyem iyerarkhicheskikh chisel v retrospektivnoy diagnostike. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta*. Ryazan', 2022, no. 79, pp.118-126.
2. **Kashirin I. Yu.** Intellektual'nyy analiz dannykh v retrospektivnoy diagnostike. *V Mezhdunarodnyy nauchno-tekhnicheskiiy forum STNO-2022. Sbornik trudov*, vol. 4, Ryazan', Book Jet, 2022, pp. 11-17.
3. **Kashirin I. Yu.** Iyerarkhicheskiye chisla v proyektirovanii obshchikh ontologiy. *V Mezhdunarodnyy nauchno-tekhnicheskiiy forum STNO-2022. Sbornik trudov*, vol. 4, Ryazan', Book Jet, 2022, pp. 17-24.
4. **Kashirin I. Yu.** Primeneniye teorii iyerarkhicheskikh chisel v proyektirovanii ICF-taksonomii dlya optimizatsii neyronnykh setey. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta*. Ryazan', 2022, no. 80, pp. 36-43.
5. **Utlik, E. P.** *Psikhologiya lichnosti: Uchebnik* E.P. Utlik. Moscow: Akademiya, 2018. 448 p.
6. **Kashirin I. Yu., Filatov I. Yu.** Formalized Description Of Intuitive Perception Of Spatial Situations. *2019 8th Mediterranean Conference on Embedded Computing (MECO)*, Budva, Montenegro, 2019, pp. 1-4.
7. **Kashirin I. Yu., Kashirina O. I.** Obzor koncepcij formal'nogo issledovaniya instrumental'nyh programmnykh sredstv. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta*. 2015, no. 2, pp. 74-83. (in Russian).
8. **Kashirin I. Yu.** Neyronnye seti, ispol'zuyushchie modeli znsnij. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta*. 2021, no. 75, pp. 71-84 (in Russian).
9. **Kashirin I. Yu.** Iyerarkhicheskiye chisla dlya proyektirovaniya ICF-taksonomiy iskusstvennogo intellekta. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta*. 2020, no. 71, pp. 71-82.
10. **Kashirin I. Yu.** Neyronnye seti, ispol'zuyushchiye modeli znaniy. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta*. 2021, no. 75, pp. 71-84.