ИНТЕЛЛЕКТУАЛЬНЫЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ

УДК 007:681.512.2

DATA MINING С ИСПОЛЬЗОВАНИЕМ ИЕРАРХИЧЕСКИХ ЧИСЕЛ В РЕТРОСПЕКТИВНОЙ ДИАГНОСТИКЕ

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

Представлена новая концепция проектирования алгоритмов интеллектуального анализа данных (Data Mining) с использованием модели представления знаний в онтологической форме. Для ретроспективного анализа динамики данных в области медицинской диагностики применяется вычисление семантической близости концептов и признаков, использующее прикладную ICF-онтологию. Для анализа семантической близости признаков и концептов используется алгебраическая система иерархических чисел. Программная реализация основана на алгоритмах анализа данных с обучением. Проведенные эксперименты с использованием инструментария Python v.3 (Anaconda 3) показывают эффективность предложенного подхода.

***Целью работы** является создание наукоемкой технологии проектирования алгоритмов Data Mining с обучением для решения задач диагностического характера.*

***Ключевые слова:** интеллектуальный анализ данных, Data Mining, алгоритмы с обучением, медицинская диагностика, ретроспективный анализ, ICF-онтология, иерархические числа, семантическая близость, кластеризация.*

DOI: 10.21667/1995-4565-2022-79-81-88

Введение

Современное состояние методов интеллектуального анализа данных дает возможность использовать все более сложные и затратные по вычислительным ресурсам технологии, базирующиеся на интеллектуальных и статистических алгоритмах поиска закономерностей [1]. Одной из актуальных задач в этой сфере являются задачи диагностирования, в частности медицинского, учитывающего не только текущие признаки наблюдаемого субъекта, но и аналитику динамики набора этих признаков. Эффективное решение таких проблем существует в области обучающихся статистических и интеллектуальных алгоритмов [2].

Теоретическая часть

Сложность ретроспективного анализа признаков заключается в том, что появляются дополнительные задачи:

– определение глубины обучающей выборки;

– выбор способов оптимизации количества и качества признаков;

– структуризация априорных моделей знаний [3] для выявления диагностических родовидовых и причинно-следственных таксономий [4].

В качестве предметной области будем использовать медицинскую диагностику. Задача ставится следующим образом.

Пусть есть ряд измерительных приборов и автоматизированных способов, позволяющих фиксировать следующие классы признаков, относящихся к здоровью человека:

– климатические;
– физиологические;
– психосоматические.

Требуется:
– определить причину негативных отклонений здоровья человека;
– выявить закономерности и взаимозависимости динамики физиологических признаков;
– при необходимости выбрать рекомендации по текущей терапии.

В качестве измерительных приборов обычно выступают наручные браслеты или интеллектуальные часы, способные производить холтеровское мониторирование (измерение пульса, давления, аритмии), снимать кардиограмму, определять факторы внешней среды (температура, влажность), жизненные фазы (сон, бодрствование). В программных приложениях часто могут использоваться более сложные и разнообразные приборы, такие как силомер, эргометр, спирометр, калипер и др.

Перечисленные особенности характеризуют задачу ретроспективной диагностики как более сложную, нежели классические задачи, решаемые многослойными нейронными сетями или алгоритмами интеллектуального анализа данных [4]. Для решения такой задачи необходимо использовать базы знаний и интеллектуальные информационные хранилища.

Для этого предложим использовать ICF-онтологию[5] и теорию иерархических чисел [6]. Приведем фрагмент ICF-таксономии для предметной области медицинской диагностики «Влияние на организм» (рисунок 1). Особенностью такого отношения является смежное наследование [5]. Здесь оно выражается в том, что факторами влияния на организм человека могут быть пища, физическая и психологическая нагрузки, экология, которые могут быть полезными и вредными. И наоборот, полезное и вредное влияния могут проявляться как посредством пищи, так и физической и психологической нагрузок, а также внешней (экологической).

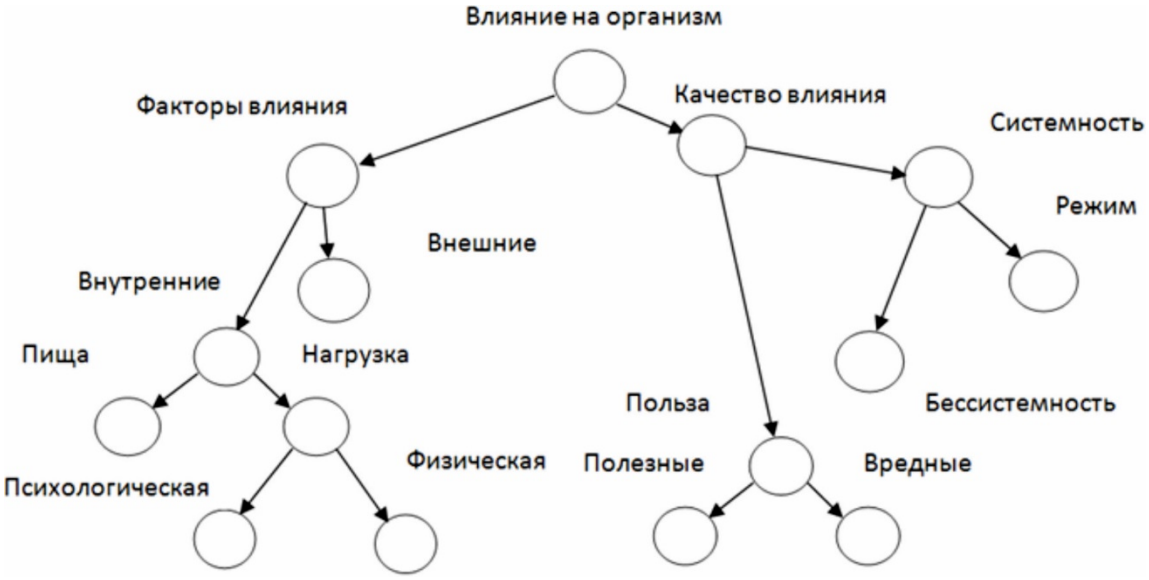

Рисунок 1 – ICF-отношение «Влияние на организм»
Figure 1 – ICF-relation «Influence on the body»

ICF-таксономия, заданная соответствующей иерархической схемой, отражает тот факт, что в любой вершине иерархии отражаются как составляющие все остальные вершины, расположенные как сверху по наследованию, так и снизу. Таким образом, можно «взяться за какую-либо вершину и вытянуть ее наверх иерархии», вся иерархия будет внутренней структурой концепта, соответствующего этой вершине. В то же время составляющие концепта могут быть более сильными и более слабыми. Например, вершины «Польза» и «Вредные» явно ближе друг другу, чем «Польза» и «Психологическая». Это свойство ICF-таксономии с оче-

видностью прослеживается на ее схеме. Математически точно степень смысловой близости концептов выражается при использовании теории иерархических чисел [6].

Кратко идея иерархических чисел выглядит так. Пусть N – множество целых положительных чисел с элементами $n_i \in N$, пусть также есть выделенный символ «.».

Множество A = N ∪ «.» ∪ ʌ определяется как алфавит с целыми числами n, где «∪» – операция объединения множеств, а ʌ – пустой символ. Тогда грамматика:

$\hat{h} \rightarrow$ ʌ, $\hat{h} \rightarrow$ h,

h → < n >, h → < n > . <h>

описывает множество иерархических чисел $\hat{H}$ с элементами $\hat{h}$.

Пример иерархического числа: 0.1.24.8 .

Алгебраическая система иерархических чисел определяется системой множеств:

$\hat{H}$ = < $\hat{H}$, { -, ∩ }, {>, <, =} >,

в которой { -, ∩ } – множество операций, а {>, <, =} – множество отношений на иерархических числах. Семантика операций дана в [6]. Операция «$\hat{h}_i$ – $\hat{h}j$» вычисляет минимальное количество вершин в некотором дереве, находящихся на пути между $\hat{h}_i$, $\hat{h}_j$, если такой путь существует. Эта разность задает расстояние между иерархическими числами. Если деревом является дерево онтологической модели знаний, то такое расстояние является семантической близостью двух концептов $\hat{h}_i$ и $\hat{h}_j$, соответствующих двум разным вершинам этого дерева.

Отношения {>, <, =} определяют, находится ли одна вершина выше или ниже другой или же является одной и той же вершиной.

В задаче диагностики нужно разделить признаки на полезные и вредные, удвоив таким образом их количество, а также на регулярные и бессистемные, что еще раз приведет к удвоению признаков. Такие удвоения, безусловно, оправданы, но между ними будет потеряна смысловая зависимость. Эта структура не дает возможности вычислить меры смысловой близости между концептами и, следовательно, существенно затрудняет задачу поиска закономерностей и взаимозависимостей динамики признаков, сформулированную в начале настоящей статьи.

Для большей строгости изложения заметим, что объективно отражающая реальность структура признаков многократно сложнее структуры, представленной на предложенных рисунках. Например, логично было бы дополнить ее признаками «Интенсивность нагрузки», «Аэробика (как свежий воздух)» и «Апартаменты (как присутствие в доме, квартире)». Из сказанного несложно понять, в какую часть ICF-иерархии можно вставить каждый из этих признаков и какие дополнительные вершины для этого понадобятся.

В целях прозрачности изложения материала из приведенной структуры для области здравоохранения выпущены такие важные вершины, как «Ретроспективные признаки» и «Текущая ситуация». Однако понятно, что два этих фактора весьма существенны в анализе данных. Текущие данные и их предыстория являются важными для каждого из концептов, представленных терминальными вершинами рисунков 1 и 2. Мало того, очень вероятна необходимость дальнейшего уточнения этих концептов такими денотатами, как: «Пульс», «Давление», «Интенсивность нагрузки», «Сон», «Бодрствование», «Белки», «Жиры», «Углеводы», «Жидкость», «Овощи», «Фрукты», «Стресс», «Положительные эмоции» и т.п. Каждый из этих признаков может использоваться в качестве исходных данных для нейронных сетей или обучающихся интеллектуальных алгоритмов анализа. Эти данные могут быть включены в прикладную онтологическую модель данных [7], не претендующих на ICF-основу. Однако для каждого из этих данных текущее значение характеристик и их динамика очень важны.

Обладание каким-либо свойством для концептов каждой из вершин таксономии является главным признаком существования этого свойства в ICF-онтологии для выбранной предметной области. В этом случае онтологию нужно расширить в сторону большей абстракции, т.е. графически добавить более высокие вершины (рисунок 2).

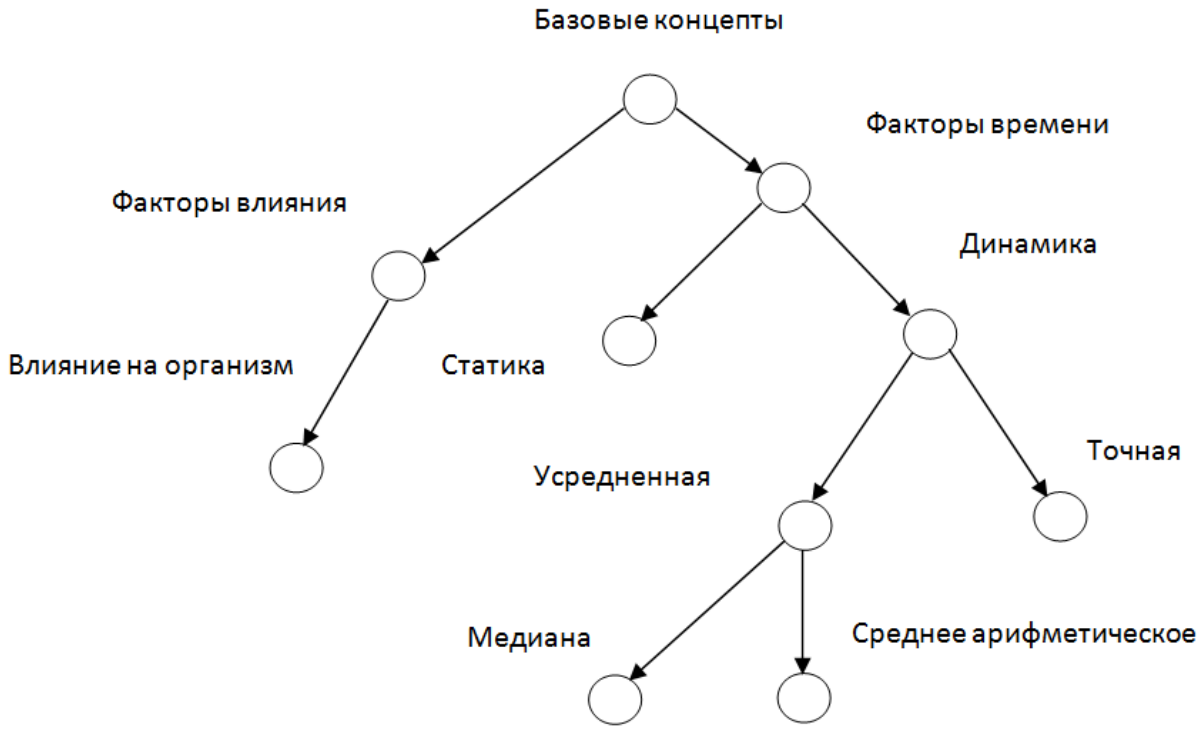

Рисунок 2 – Расширение таксономии факторами динамики и статики
Figure 2 – Expansion of taxonomy by factors of dynamics and statics

В представленной ICF-таксономии допущен ряд упрощений, например «Статика» и «Динамика» в общих онтологиях располагаются значительно выше по уровню абстрагирования, поскольку близки к базовым категориям «Пространства» и «Времени». Пример такой базовой онтологии можно найти в [5].

Иерархические числа, рассмотренные ранее, позволяют вычислить структурную семантическую близость концептов (характеристик). Семантическая близость вершин таксономии является величиной, которая при ее минимализме соответствует концептам, механизмы определения которых, по сути дела, одинаковы. Например, если концепт «Пища» разделить далее на «Белки», «Жиры» и «Углеводы», то механизм воздействия этих факторов на организм будет более схожим, чем, например, «Утренняя зарядка». Однако, если сравнивать «Пищу», «Физическую нагрузку» и «Внешнюю среду», то их семантическая близость не столь очевидна и может быть уточнена только ретроспективным анализом больших данных. Эти соображения предопределяют взаимозависимость алгоритмов интеллектуального анализа данных с моделями знаний соответствующей предметной области [8]. Точное знание значений семантической близости дает возможность упростить и сделать более эффективными алгоритмы анализа данных с обучением, так как позволяют априори определить признаки, подлежащие ретроспективному анализу, и признаки, которые можно рассматривать в их текущих значениях.

Практическая часть

Обратимся к задаче медицинской диагностики. Приведем фрагмент примерной таблицы набора данных для обучения алгоритмов и анализа данных (таблица 1).

Для поставленной в начале настоящей статьи задачи необходимо в инструментарии Python исследовать все возможные характеристики на взаимовлияние, использовав такой фрагмент программы:

```
import pandas.rpy.common as corr_health
import seaborn as seaborn_helth
%matplotlib inline
df = corr_health.load_data('Helth_Data')
```

```
corr = df.corr()
seaborn_helth.heatmap(corr,
        xticklabels=corr.columns,
        yticklabels=corr.columns)
```

Таблица 1 – Выборка из исходного набора данных (начало)
Table 1 – Sample from original dataset (beginning)

t, C	Humidity	Pressure	Dream Continuous	Dream Total	SYS pressure	DIA pressure	Pulse	SYS after
-4,2	28	758	7,0	7,0	165	94	56	121
-2,7	26	755	6	7	171	106	59	119
-3	32	756	7	7	176	92	55	122
-2	25	746	4	7	157	95	58	131
0	28	756	6,5	7	180	107	56	121
0	27	761	5	7,5	164	87	53	123
0	35	744	9	9	146	96	71	118
-3	30	742	6,5	7	169	97	56	132
-17	20	756	7	7	176	104	55	120
-21	26	761	7,5	7,5	155	93	57	123
-22	26	762	7	7,5	155	90	58	164
-15	24	761	4,5	8	144	80	56	124
-10	24	760	6	7	160	93	57	127
-15	30	752	8	8	169	94	56	133
-15	31	763	5	8	164	101	61	135
-8	28	770	4	8	171	97	58	134
-10	29	756	6	8	172	92	52	131
-9	27	756	7	7	189	98	53	130
-6	25	760	6	8	188	110	60	139

Таблица 1 – Выборка из исходного набора данных (окончание)
Table 1 – Sample from original data set (end)

Food yesterday	Food healthy	Food Amount	Regime	Walk Total	Walk Amount	Weight
0,5	17	5	1	99	2	104,5
0,7	14	5	1	60	1	104,75
0,3	16	5	1	65	1	105
0,3	14	5	1	87	2	105
0,5	12	4,5	1	100	2	105,1
0,7	12	5	1	115	2	104,5
0,6	12	4,5	1	111	2	104,5
0,74	11	5	0	89	2	105
0,8	11	5	1	48	1	105,1
0,7	12	5	1	52	1	105
0,7	15	5	0	72	2	105
0,5	12	4	1	70	1	105
0,74	12	4	1	60	1	106
0,7	12	5	1	60	2	105,8
0,7	10	4	0	108	2	106

0,8	11	4	0	90	2	106
0,65	10	4	0	93	2	106,4
0,7	11	4	0	99	2	106
0,7	12	5	0		2	106,1

Для этого набора данных приняты следующие обозначения (таблица 2).

Таблица 2 – Выборка признаков исходного набора данных (начало)
Table 2 – Selection of features of original data set (beginning)

№ п/п	Обозначение признака	Наименование признака
1	**t, C**	Температура воздуха
2	**Humidity**	Влажность в помещении
3	**Pressure**	Атмосферное давление
4	**Pulse**	Пульс утром
5	**SYS pressure**	Давление систолическое
6	**DIA pressure**	Давление диастолическое

Таблица 2 – Выборка признаков исходного набора данных (окончание)
Table 2 – Selection of features of original data set (end)

№ п/п	Обозначение признака	Наименование признака
7	**Weight**	Вес
8	**Dream Total**	Продолжительность сна
9	**DreamContinuous**	Качество сна (количество пробуждений)
10	**Food Amount**	Количество приемов пищи
11	**Food healthy**	Оценка разнообразия полезных продуктов
12	**Food yesterday**	Оценка качества питания накануне
13	**Regime**	Поддержание выбранного режима жизни
14	**Walk Total**	Длительность прогулок на свежем воздухе
15	**Walk Amount**	Количество прогулок на свежем воздухе

Однако прямая корреляция признаков не дает возможности решить задачу поиска причины нежелательных отклонений показателей здоровья, поскольку наиболее значимым параметром является предыстория динамики признаков. Для такого анализа необходимо задействовать аппарат моделей знаний на основе таксономий (рисунок 1).

Определение множества признаков для ретроспективного анализа вычисляется по формуле допустимой семантической близости следующим образом. Пусть b – таксономическая вершина концепта «Польза», I (b) – иерархическое число, соответствующее b, U – универсум всех оцениваемых признаков, включенных в таксономию прикладной задачи о здоровье, а F – вычисляемое множество признаков, чье исследование нуждается в ретроспективном анализе. Обозначим через B порог допустимой семантической близости для выбора анализируемых признаков. Тогда критерием отбора таких признаков является следующая формула:

$$\forall x,\ x \in U,\ |\ I(x) - I(\ b)\ | < B \rightarrow x \in F,$$

где | I(x) – I(b) | – значение семантической близости для концептов (признаков) x и b; «→» – операция логического следования (импликации); «<» – арифметическое отношение «меньше».

После выбора множества признаков, которые должны исследоваться в их динамике, эксперименту подвергаются способы исследования динамики. Это могут быть такие методы, как:

– среднее арифметическое или геометрическое всех ретроспективных значений;
– среднее арифметическое или геометрическое выборочных ретроспективных значений;
– медианное вычисление значений;
– вычисление ограниченной ретроспективной выборки (например, за последнюю неделю, месяц);
– учет всех значений ограниченной ретроспективной выборки.

Экспериментальные исследования

Испытательный стенд (ЭСМИАД v.10.02.2022), реализованный на языке Python v.3.7 [4] в среде Anaconda 3, дает возможность определить семантическую близость «Факторы влияния» и «Польза» как наиболее перспективную. В результате экспериментов становится понятным, что признаки 4 – 10 из таблицы 2 должны быть подвергнуты ретроспективному анализу. Практические эксперименты с небольшой группой испытуемых (14 человек) выявили эффективность предложенной концепции интеллектуального ретроспективного анализа и показали возможность кластеризации множества пациентов по индивидуальным особенностям их организмов. Так, например, если в одной группе оказались испытуемые, у которых при незначительном повышении веса (1 – 2 %) уменьшалось кровяное давление, в то время как у второй группы испытуемых давление падало при таком же уменьшении веса. Кроме того, у первой группы ухудшалось самочувствие при регулярных прогулках перед сном, тогда как у второй – наоборот, основные показатели здоровья приходили в норму.

Заключение

В статье предложена наукоемкая технология использования иерархических чисел и ICF-онтологии для проектирования алгоритмов интеллектуального анализа данных в ретроспективной диагностике на примере предметной области медицинской диагностики. Технология дает возможность разрабатывать автоматизированные интеллектуальные диагностические системы с оптимизацией моделями знаний.

Библиографический список

1. **Каширин И. Ю., Маликова Л. В., Маркова В. В.** Основы формальных систем / под ред. И.Ю. Каширина – М.: Минобразования России, НИЦПрИС, 1999. 80с.

2. **Каширин И. Ю.** Нейронные сети, использующие модели знаний // Вестник Рязанского государственного радиотехнического университета. 2021. № 75. С.71-84.

3. **Каширин Д. И., Каширин И. Ю**. Модели представления знаний в системах искусственного интеллекта // Вестник Рязанского государственного радиотехнического университета. № 31. 2010. С. 36-43.

4. **Каширин И. Ю.** Нейронные сети, использующие модели знаний. IV Международный научно-технический форум СТНО-2021 // сб. тр. Т. 4. Рязань, Book Jet, 2021. С. 9-13.

5. **Каширин Д. И.** Формализм ICF-онтологий для представления знаний в глобальной сети нового поколения Semantic // Вестник Тамбовского государственного технического университета. 2007. – Т. 13. № 1. С. 91-103.

6. **Каширин И. Ю.** Иерархические числа для проектирования ICF-таксономий искусственного интеллекта // Вестник Рязанского государственного радиотехнического университета. 2020. № 71. С. 71-82.

7. **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.

8. **Каширин И. Ю., Каширина О. И.** Обзор концепций формального исследования инструментальных программных средств // Вестник Рязанского государственного радиотехнического университета. 2015. № 52. С. 74-83.

UDC 007:681.512.2

DATA MINING USING HIERARCHICAL NUMBERS IN RETROSPECTIVE DIAGNOSIS

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

A new concept of designing data mining algorithms (Data Mining) using a knowledge representation model in an ontological form is presented. For retrospective analysis of data dynamics in the field of medical diagnostics, the calculation of semantic similarity of concepts and features is used using applied ICF ontology. An algebraic system of hierarchical numbers is used to analyze the semantic proximity of features and concepts. Software implementation is based on learning data analysis algorithms. The experiments performed using Python v.3 (Anaconda 3) tools show the effectiveness of the proposed approach.

***The aim of the work** is to create a science-intensive technology for designing Data Mining algorithms with training to solve the problems of diagnostic nature.*

***Key words:** Daia Mining, supervised algorithms, medical diagnostics, retrospective analysis, ICF ontology, hierarchical numbers, semantic proximity, clustering.*

DOI: 10.21667/1995-4565-2022-79-81-88

References

1. **Kashirin I. Yu, Malikova L. V. Markova V. V.** *Osnovy formalnyh system*. Pod red. I. Yu. Kashirina Moscow: MINOBRAZOVANYA Roaaii, NICPrIS, 1999. 80 p. (in Russian).

2. **Kashirin I. Yu.** Neuronnye seti, ispolzuyuschie modeli znsnij. *Vestnik Rjazanskogo gosudarstvennogo radiotehnicheskogo universiteta*. 20210, no. 75, pp. 71-84 (in Russian).

3. **Kashirin D. I., Kashirin I. Yu.** Modeli predstavlenya znanij v systemah iskusstvennogo intellekta. *Vestnik Rjazanskogo gosudarstvennogo radiotehnicheskogo universiteta*. 2010, no. 31, pp. 36-43. (in Russian).

4. **Kashirin I. Yu.** IV *Mejdunarodny nauchno-technicheskij forum STNO-2021. Sbornik trudov*, vol. 4. Ryazan, Book Jet. 2021, pp. 9-13.

5. **Kashirin D. I.** Formalizm ICF-ontologij dlya predstavleniya znanij v globalnoj seti novogo pokolenya Semantic Web. *Vestik Tambovskogo gosudarstvennogo tehnicheskogo universiteta*. 2007. vol. 13, no. 1, pp. 91-103. (in Russian).

6. **Kashirin I. Yu.** Ierarhicheskie chisla dlya proectirovaniya ICF-taksonomij iskusstvennogo intellekta. *Vestnik Rjazanskogo gosudarstvennogo radiotehnicheskogo universiteta*. 2010. no. 71, pp. 71-82 (in Russian).

7. **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.

8. **Kashirin I. Yu, Kashirina O. I.** Obzor koncepcij formalnogo issledovanya instrumentalnyh programmnyh sredstv. *Vestnik Rjazanskogo gosudarstvennogo radiotehnicheskogo universiteta*. 2015, no. 2, pp. 74-83. (in Russian).