ИНТЕЛЛЕКТУАЛЬНЫЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ

УДК 007:681.512.2

НЕЙРОСЕТИ СО ЗНАНИЯМИ ДЛЯ АНАЛИЗА БОЛЬШИХ ДАННЫХ

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 важнейшая задача проектирования обучающихся алгоритмов интеллектуального анализа больших данных, включая нейросетевые. Для повышения их эффективности используются таксономии для моделей знаний искусственного интеллекта. В качестве базисного отношения таксономии предлагается использовать отношение смежного наследования ICF. Такой подход позволяет эффективно проектировать общие и прикладные онтологии для решения задач искусственного интеллекта. В статье используется математический аппарат иерархических чисел, дающий возможность анализировать прикладные ICF-онтологии с помощью вычисления основных характеристик таксономических иерархий. К таким характеристикам относятся, например, мера семантической близости концептов или сложность структурного пути в таксономической иерархии между двумя концептами. В качестве примера, подтверждающего возможность эффективного использования ICF-таксономий в решении реальных задач анализа данных, предлагается решение задач кластеризации и классификации в предметной области «Сокращение оттока и привлечение клиентов телекоммуникационной компании». Для нее показана возможность взаимозовисимого проектирования нейросетей и прикладных онтологий с априорными библиотеками причинно-следственных таксономий, например, в технологии Semantic Web. В экспериментальной части для апробации возможностей информационного поиска в глобальных сетях на основе технологии Semantic Web с помощью ICF-таксономии рассматривается программа-экспериментальный стенд для методов интеллектуального анализа данных (ЭСМИАД v.12.02.2021), позволяющий путем помещения фрагментов текста программы в комментарии или наоборот выполнять эксперименты для аналитических алгоритмов с обучающими выборками. Экспериментальный стенд реализован на языке Python v.3.7 в инструментальной среде Anaconda v.3.0. Назначение этого стенда – анализ эффективности методов интеллектуального анализа данных, использующих модели представления знаний.

***Целью работы** является анализ особенностей проектирования обучающихся алгоритмов анализа данных с применением общей и прикладной ICF-онтологий с помощью аппарата иерархических чисел для эффективного решения актуальных задач кластеризации и классификации.*

***Ключевые слова:** модели знаний, большие данные, искусственный интеллект, нейронные сети, ICF-онтология, онтологические таксономии, иерархические числа, инструментальные программные средства, мера семантической близости концептов, технология Semantic Web, Python, Anaconda, родовидовые и причинно-следственные отношения.*

DOI: 10.21667/1995-4565-2021-75-71-84

Введение

Современные методы интеллектуального анализа данных по своей сути представляют собой математическое решение ряда задач, так или иначе связанных с классификацией и кластеризацией. В эти методы добавляется немного технической алгоритмизации только тогда, когда речь заходит о разбиении матриц входных данных на обучающую и тестирующую

выборки [1]. Таким образом, претендовать на принадлежность искусственному интеллекту эти концепции могут весьма условно, и в большей части интеллектуальность относится лишь к нейросетям, в которых управление связями между нейронами можно отнести к инженерным задачам высоких технологий.

В подтверждение сказанного следует добавить, что множество научных трудов об анализе больших данных использует примеры, искусственно сгенерированные алгоритмами рандомизации [1].

В то же время, реальная экономическая или социальная отдача от анализа больших данных начинается только в результате привязки полученной методологии к реальным задачам конкретных предметных областей [2]. Однако и в этом случае чаще всего данные рассматриваются как мешок непонятных характеристик, которые нужно как-то отсортировать (кластеризовать) и принять решение об их пригодности для решения поставленной задачи (классифицировать).

Цель настоящей статьи показать, как можно использовать современные системы представления знаний теории искусственного интеллекта в анализе больших данных, так, чтобы методы такого анализа назывались интеллектуальными по праву.

Долгое время параллельно с разработкой математических методов кластеризации и классификации инженерами по знаниям разрабатывались модели представления знаний, разнообразие и эффективность которых не уступает упомянутым математическим методам [3]. Однако сейчас применение теории представления знаний главным образом связывается со следующими направлениями:

– теорией формирования и доказательства гипотез;
– теорией унификации для поиска нетривиальных ответов в базе знаний;
– поиском самих знаний (фактов) в глобальных сетях;
– общению с компьютером на естественном языке.

В то же время в результате работ в этих направлениях появились весьма конструктивные понятия, такие как таксономии следующих видов:

– родовидовые деревья;
– причинно-следственные деревья;
– таксономии «часть-целое»;
– ассоциативные таксономии.

Следует также обратить внимание на рожденные в рамах идеологии искусственного интеллекта псевдофизические логики пространства и времени, паранепротиворечивые, многозначные и нечеткие логики как прикладные исчисления первого и высших порядков. Некоторые из этих концепций, например, нечеткие множества, уже работают в современных автоматизированных системах классификации [1].

Постановка задачи

Предлагаемый здесь подход основывается на том, что любые хранилища данных позволяют выделить в них наименование объектов или характеристик объектов и отношений в конкретной предметной области. То есть, либо задача *идентификации* характеристик сразу решена, либо позволяет это сделать с помощью кластеризации или другого подходящего алгоритма. Кроме того, известна также *задача* анализа данных, позволяющая принять на его основе правильное управленческое решение.

В этом случае перед инженером по знаниям можно поставить аналитическую задачу по построению модели предметной области, в которой нужно выделить:

– объекты;
– характеристики объектов;
– отношения;
– характеристики отношений;
– элементы динамики предметной области.

Элементы динамики, например, могут включать причинно-следственные связи и операции принятия решений с их результатами. Объекты и отношения могут и должны включаться в цепочки или даже деревья таксономий [3].

Все перечисленные строительные материалы позволят создать общую и прикладную онтологии предметной области [4-8]. При достаточном опыте инженера по знаниям общие онтологии не нужно будет создавать заново, а просто выбрать одну подходящую из инструментального множества уже имеющихся в онтологической библиотеке. Это должно происходить примерно так же, как выбор готового метода классификации в обширных математических библиотеках инструментария Python (например, k ближайших соседей, метод опорных векторов или наивный байесовский метод и т.п.). Так же, как для инструментария этих библиотек, для онтологий должна существовать возможность их автоматизированной адаптации к предметной области с помощью настройки цепочки упрощений. Поставим задачу определить зависимость между существующими аналитическими алгоритмами и прикладными таксономиями онтологий, показать эффективность их комплексного использования.

Теоретическая часть

При использовании механизма ICF-онтологии [9] возможно накопление библиотеки *элементов* дихотомических отношений, «вырезанных» из ICF-дерева общей онтологии. В результате должна появиться прикладная онтология, простая, но концептуально связанная и достаточно полная для ассоциации ее терминальных вершин с признаками и метками хранилища больших данных. Элементы дерева должны иметь возможность новой композиции их в классифицирующее дерево, подходящее для предметной области, объемные данные которой являются исходными для интеллектуальной аналитики.

Одним из положительных эффектов описанной технологии является возможность внесения изменений в модули библиотеки онтологий на основании статистического анализа данных существующими методами аналитики данных, главным образом, деревьями решений и нейронными сетями. Например, это может касаться перемещения концептов в родовидовых таксономиях, как в сторону большей абстракции, так и в направлении конкретизации.

Перечислим преимущества использования моделей знаний в алгоритмах анализа больших данных.

1. Автоматическое упорядочивание таксономий.
2. Сокращение перебора при обучении нейросети на этапе формирования межуровневых связей.
3. Выявление недостающих информационных полей (нейронов).
4. Идентификация (выявление) сложных стратифицированных информационных полей. Выявляется принадлежность признаков одной группе или даже одной характеристике.
5. Выявление априорной оценки меры взаимозависимости признаков до начала обучения нейросети на основе концепции иерархических чисел [7].
6. Автоматизированное формирование гипотез о неинформативных признаках.
7. Формирование логической схемы гипотезы о существующей закономерности для поставленной цели обучения нейронной сети.

Для практического подтверждения изложенного приведем пример из предметной области «Сокращение оттока и привлечение клиентов телекоммуникационной компании»,

взятого из международного репозитория Kaggle [1]. Kaggle – система организации конкурсов по исследованию данных, а также социальная сеть специалистов по обработке данных и машинному обучению Этот интернет ресурс с марта 2017 года принадлежит корпорации Google.

Основная задача выбранной предметной области как цель обучения нейросети или другого инструмента классификации: определить причины оттока заказчиков телекоммуникационных услуг. Смежная задача: определить условия привлечения новых заказчиков.

Существующий сбор данных позволил предоставить в распоряжение инженера по знаниям следующие информационные поля:

– ID пользователя;
– пол;
– работник;
– пенсионер;
– иждивенец;
– прошедшее время использования услуг;
– пользовался ли телефонным сервисом;
– пользовался ли многоканальностью;
– пользовался ли интернетом;
– использовал ли средства электронной защиты;
– использовал ли резервное копирование;
– использовал ли защиту устройства связи;
– пользовался ли техподдержкой;
– пользовался ли потоковым телевидением;
– смотрел ли потоковые фильмы;
– длительность и вид контракта;
– использовал ли безбумажный биллинг;
– способ оплаты;
– объем ежемесячных платежей;
– всего заплачено;
– хочет ли пользователь покинуть компанию.

Для проектирования обучающихся алгоритмов, в том числе нейросети, предназначенных для анализа данных из представленного репозитория будем использовать основы алгебры иерархических чисел [7] и концепцию прикладной ICF-онтологии [6].

Алгебраическую систему иерархических чисел можно описать так. Пусть N – множество целых положительных чисел с элементами $n_i \in N$, пусть также есть выделенный символ «.». Множество $A = N \cup \text{«.»} \cup \Lambda$ определим как алфавит с целыми числами n, где «∪» – операция объединения множеств, а ʌ – пустой символ. Тогда грамматика:

$$\hat{h} \rightarrow \Lambda, \quad \hat{h} \rightarrow h,$$
$$h \rightarrow < n >, \quad h \rightarrow \ < n > . <h>$$

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

Пример иерархического числа: 0.1.12.2 .

Алгебраическая система иерархических чисел определяется системой множеств:

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

в которой { -, ⌢ } – сигнатура, или множество операций, а {>, <, =} – множество отношений на иерархических числах. Семантика операций дана в [9]. Упрощенно 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. Если пути между $\hat{h}_i$ и $\hat{h}_j$ не существует, результатом является 0. Операция «$\hat{h}_i \frown \hat{h}j$» вычисляет вершину, являющуюся общим предком двух вершин $\hat{h}_i$ и $\hat{h}_j$. Если такого предка нет, результатом считается 0.

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

В настоящей статье нам понадобится только понимание операции вычисления семантической близости концептов в родовидовой таксономии прикладной онтологической модели знаний.

Главная задача анализа данных в выбранном примере – обеспечение потребности пользователя.

Пользователь, как и поставщик, являются разновидностями участника отношений, а следовательно, имеют собственные свойства и используют в своих отношениях какие-либо услуги или товары. Далее будет рассмотрено поэтапное проектирование структурной прикладной ICF-онтологии для предметной области «Сокращение оттока и привлечение клиентов телекоммуникационной компании». Основой родовидовой таксономии прикладной онтологии является следующая иерархическая схема (рисунок 1).

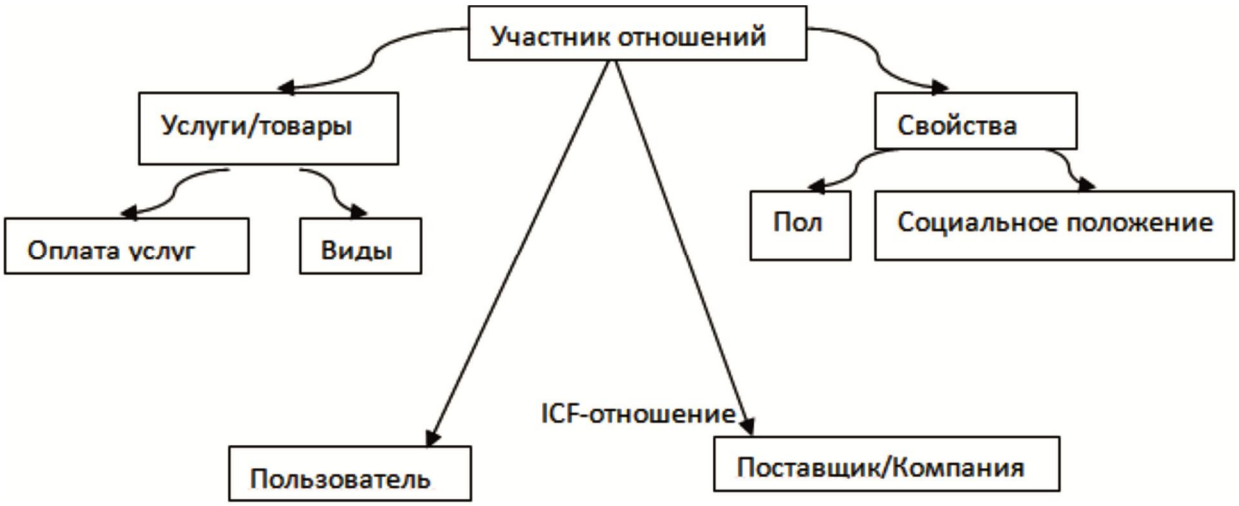

Рисунок 1– Основой родовидовой таксономии предметной области «Оказание услуг»
Figure 1 – The basis of generic taxonomy of «Service delivery» subject area

Принцип использования ICF-отношений говорит, что, следуя онтологической структуре, представленной рисунком 1, «Пользователь» как «Участник отношений» имеет в ICF-связке «Услуги/товары», а также «Свойства». То же самое можно сказать и про «Поставщика».

Если же концепт «Участник отношений» приподнять в пространстве над плоскостью рисунка, то эффект множественного наследования в ICF [6] можно представить рисунком 2.

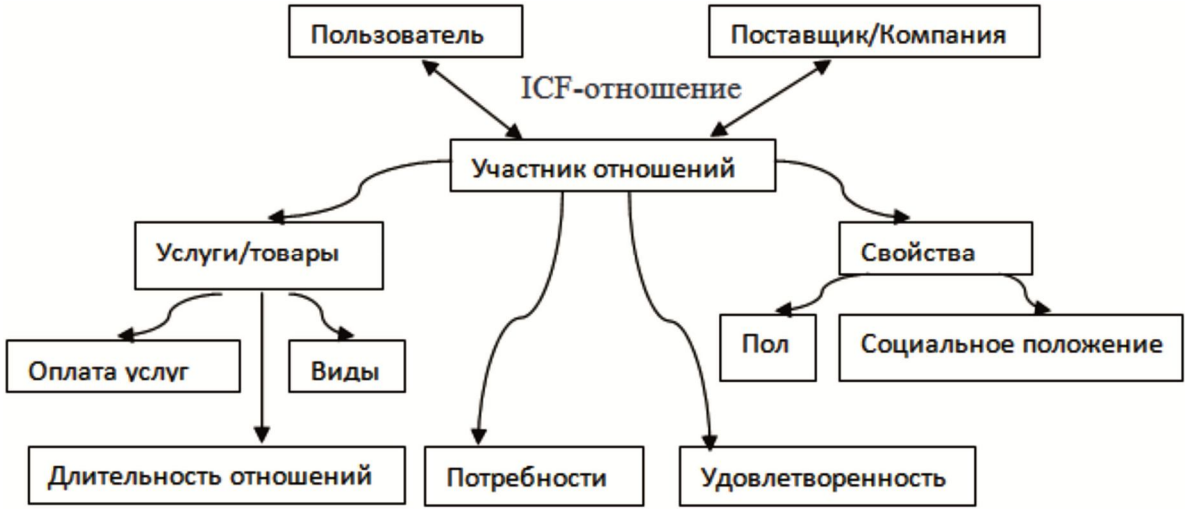

Рисунок 2 – Множественное наследование в ICF-онтологии
Figure 2 – Multiple inheritance in ICF-ontology

Более подробная онтология, претендующая на общую онтологию сферы оказания услуг вообще, представлена рисунком 3. На рисунке помещены также иерархические числа как индексы элементов онтологии. При этом символ переменной x свидетельствующий о существовании более абстрактных концептов в полной онтологии, в прикладной онтологии для вычисления семантической близости может быть принят за пустой символ «л».

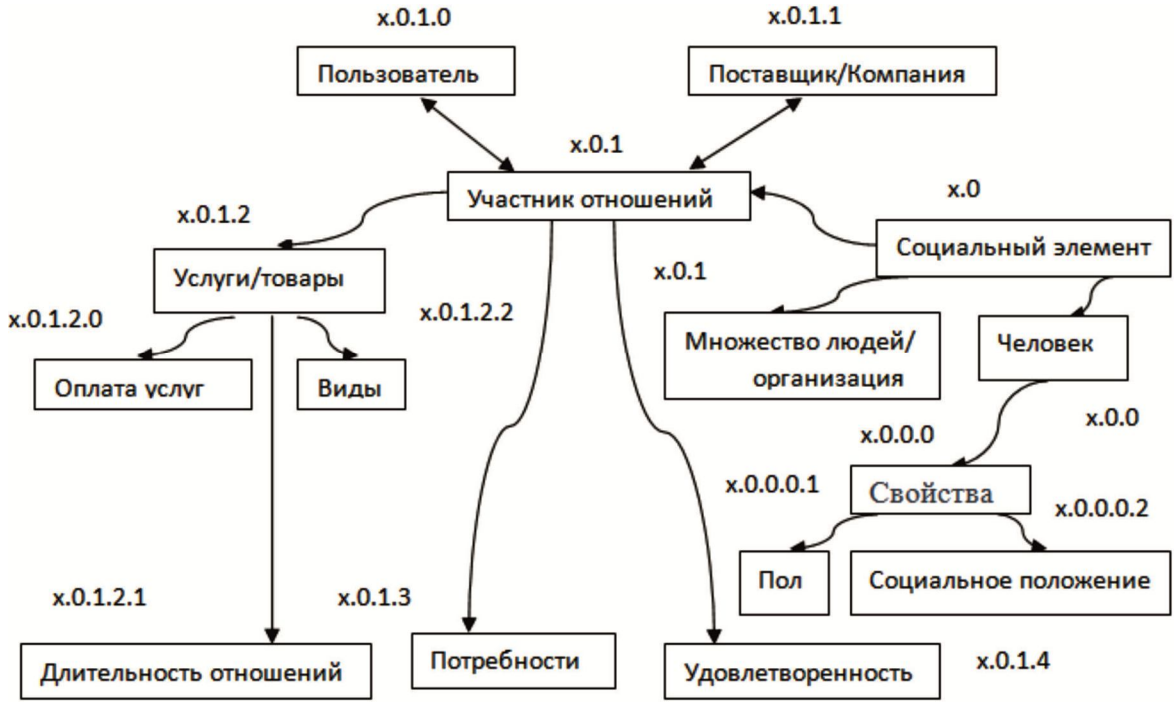

Рисунок 3 – Фрагмент общей онтологии сферы оказания услуг
Figure 3 – Fragment of general ontology of service delivery sector

Следуя рисункам 2 и 3, «Участник отношений», будь то «Пользователь» или «Поставщик», имеет множество естественных «Свойств», например, «Пол» и «Социальное положение». Кроме того, «Участник отношений» связан, например, с «Услугами», которые оплачиваются и имеют свои разновидности. «Участник отношений» может иметь «Потребности» и быть удовлетворенным собой, другими участниками отношений, своими «Свойствами», а также «Услугами», их «Видами» и «Оплатой». До сих пор, представленную на рисунке 3 онтологию можно считать общей онтологией [2].

Для получения прикладной онтологии [2] требуется добавить конкретные виды услуг, формы и тарифы оплаты. Можно также конкретизировать виды или уровни социального положения (рисунок 4). На рисунке приведены также онтологические индексы, которые нетрудно самостоятельно продолжить в нижней части онтологической схемы.

Эта онтология содержит большинство семантических категорий (концептов и их свойств), которые имеются в репозитории предметной области выбранного примера. Фрагмент базовой таблицы репозитория представлен таблицей 1. Этот факт позволяет сделать вывод о пригодности прикладной онтологии на рисунках 3 и 4 для ее использования в качестве одного из элементов конструктивного инструментария для интеллектуального анализа данных об услугах телекоммуникационной компании.

Можно рассмотреть причинно-следственную таксономию, формируемую из аксиоматических (априорных) взаимосвязей характеристик обобщенной предметной области «Сфера услуг». Таксономия состоит из описания взаимосвязей, задаваемых между анализируемыми характеристиками данных с помощью отношения зависимости «=►» и взаимозависимости «◄ = ►». Стрелками «↓» и «↑» определяется понижение числового значения характеристики и его повышение соответственно. Для этой таксономии тоже можно использовать графическое представление (рисунок 5).

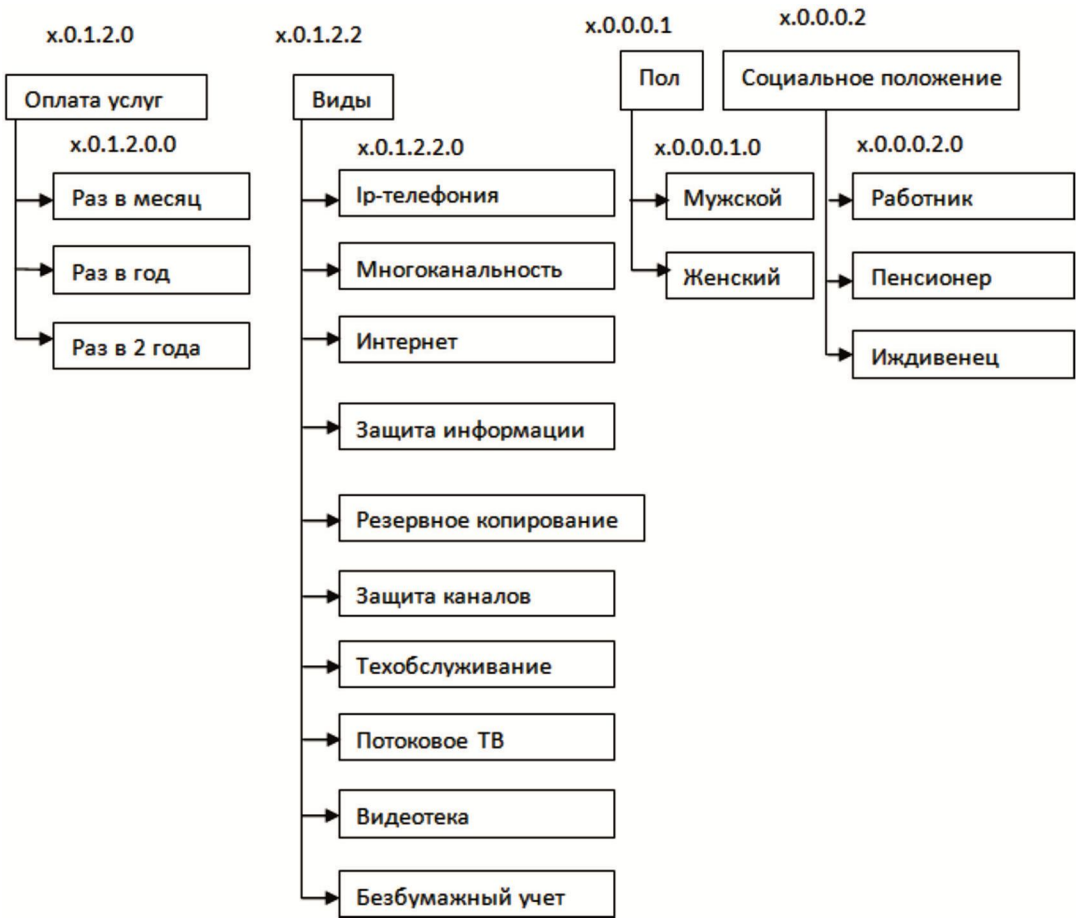

Рисунок 4 – Дополнение общей онтологии сферы услуг до прикладной онтологии «Услуги телекоммуникационной компании»
Figure 4 – Addition of general ontology of service sector to applied ontology «Telecommunication company services»

Приведем далее фрагмент библиотеки причинно-следственной таксономии.

Цена товара ↓ ◄=► Объем продаж ↑
Качество товара ↑ =► Объем продаж ↑
Ассортимент товара ↑ =► Объем продаж ↑
Удовлетворенность потребителя ↑ =► Время отношений с поставщиком ↑
Качество товара ↑ =► Цена товара ↑
Доход потребителя ↑ =► Покупательная способность ↑
Покупательная способность ↑ =► Объем продаж ↑
Качество товара ↑ =► Время отношений с поставщиком ↑
Качество товара ↑ =► Удовлетворенность потребителя ↑

Гипотеза: чем меньше иерархическое число как расстояние между потребностью и другими концептами, тем более числовые характеристики этих концептов влияют на удержание пользователя в списке потребителей.

Близко расположены, например: оплата услуг, уровень доходов, длительность потребления (надежность).

В то же время, в общем случае анализирующий алгоритм может иметь большое число характеристик концептов. Среди них, как правило, можно выделить характеристики, тривиально (или очевидно) влияющие на результат анализа больших данных. Безусловно, высказанную для нашего примера гипотезу можно будет переформулировать так: «чем меньше иерархическое число как расстояние между тривиально влияющей на результат характеристикой и другими концептами, тем более числовые характеристики этих концептов влияют на заключительный результат анализа.

Таблица 1 – Данные из репозитория клиентов телекоммуникационной компании
Table 1 – Data from customer repository of telecommunications company

Customer ID	Gender	Tenure	Phone Service	Multiple Lines	...	Internet Service	Payment Method	Monthly Charges	Total Charges	Churn
7590-VHVEG	Female	1	No	No phone service	...	DSL	Electronic check	29,85	29,85	No
5575-GNVDE	Male	34	Yes	No	...	DSL	Mailed check	56,95	1889,5	No
3668-QPYBK	Male	2	Yes	No	...	DSL	Mailed check	53,85	108,15	Yes
7795-CFOCW	Male	45	No	No phone service	...	DSL	Bank transfer (automatic)	42,3	1840,75	No
9237-HQITU	Female	2	Yes	No	...	Fiber optic	Electronic check	70,7	151,65	Yes
9305-CDSKC	Female	8	Yes	Yes	...	Fiber optic	Electronic check	99,65	820,5	Yes
1452-KIOVK	Male	22	Yes	Yes	...	Fiber optic	Credit card (automatic)	89,1	1949,4	No
6713-OKOMC	Female	10	No	No phone service	...	DSL	Mailed check	29,75	301,9	No
7892-POOKP	Female	28	Yes	Yes	...	Fiber optic	Electronic check	104,8	3046,05	Yes
6388-TABGU	Male	62	Yes	No	...	DSL	Bank transfer (automatic)	56,15	3487,95	No
9763-GRSKD	Male	13	Yes	No	...	DSL	Mailed check	49,95	587,45	No
...	...	...	...	...	...	...	...	...	...	...

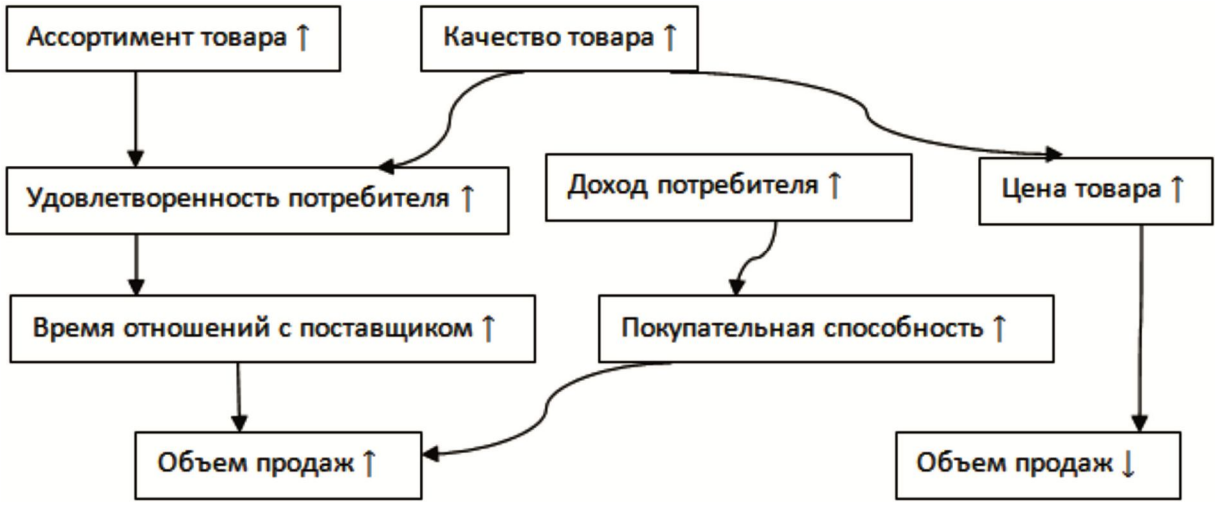

Рисунок 5 – Иерархическое представление причинно-следственной таксономии
Figure 5 – Hierarchical representation of causal taxonomy

Однако, предложенное правило удобно использовать главным образом в алгоритмах теории доказательств или в правилах вывода формальных систем [10, 11]. Очевидность взаимосвязей в алгоритмах анализа с обучением, например, в нейросетях делает близкие к тривиально зависимым характеристики практически бесполезными, поскольку их можно отождествить с самими тривиально зависимыми характеристиками. Следовательно, гораздо большую пользу в практическом случае принесут характеристики, далеко отстоящие от тривиально зависимых. Дело в том, что такие характеристики не просто имеют большое иерархическое число, как рас-

стояние между концептами в прикладной онтологии, но они объективно зависят от большого числа других факторов (характеристик, свойств, динамики предметной области).

Например, то, что человек является пенсионером, свидетельствует о его пожилом возрасте и/или наличии серьезного заболевания. Это правило является фактическим определением пенсионера и по своей значимости является почти бесполезным. Другой факт, говорящий о том, что пенсионер, как правило, имеет меньший доход, чем здоровый человек средних лет, безусловно неочевиден и не всегда справедлив, но именно он более подходит для включения его в нейросеть в качестве одной из связей нейронов.

Таким образом, можно сделать вывод: «характеристики концептов, имеющие большие иерархические числа в качестве оценки семантического расстояния между ними, более весомы в алгоритмах анализа данных с обучением». Еще один важный обратный вывод: «если два концепта настолько близки по своим свойствам друг другу, что один из них может быть исключен из характеристик при анализе данных, то эти концепты расположены в прикладной онтологии рядом, т.е. имеют относительно малые иерархические числа в качестве оценки семантического различия».

Оба этих вывода можно принять в качестве структурных правил при проектировании нейросетей и прикладных онтологий. Взаимозависимость этих правил, даже в случае несправедливости в реальном мире высказанных ранее гипотез, дает уверенность, что построение онтологий и аналитических алгоритмов, основанных на больших данных, в соответствии с изложенными правилами сделает аналитические алгоритмы более эффективными. Это должно проявиться особенно четко на развитых этапах получения онтологической модели данных, когда она содержит большое количество подтвержденных фактов зависимости концептов.

Для иерархических чисел в рассмотренных правилах должен существовать «порог семантической зависимости (близости)», превышение которого делает использование какого-либо из двух концептов бесполезным и свидетельствует о неинформативности его характеристик. Существование такого порога предполагает его вычисление отдельно для каждой из общих и прикладных онтологий, поскольку отдельные реализации онтологий могут существенно отличаться уровнями подробности рассмотрения таксономии базовых концептов. В то же время можно сделать весьма правдоподобное утверждение, что в абсолютно объективной онтологии, полностью адекватной реальному миру, «порог семантической зависимости (близости)» является неизменной константой. Асимптотическое приближение к такой «абсолютной истине» может происходить по мере накопления разработанных человечеством онтологических библиотек и разработки формальных и инструментальных средств их непротиворечивого комбинирования (комплексирования).

Экспериментальные исследования

Приведем программу на языке Python v.3.7 в среде Anaconda3 [1], с помощью которой будем исследовать подключение новых характеристик, рекомендованных для исследования прикладной онтологией, рассмотренной ранее. При этом используем методы: К ближайших соседей, опорных векторов, дерева решений и нейронной сети.

Далее приводится полный текст программы-экспериментального стенда для методов интеллектуального анализа данных (ЭСМИАД v.12.02.2021), позволяющего путем помещения фрагментов текста программы в комментарии или наоборот, раскомментирования фрагментов программы, выполнять эксперименты для аналитических алгоритмов с обучающими выборками [13]. В программе реализована предварительная обработка данных: обработка пропусков данных, поиск неявных дубликатов, обнаружение выбросов, числовое кодирование, нормализация, разделение на подвыборки. Далее алгоритмы предложенных методов анализа обучаются, производят анализ данных и оцениваются на эффективность.

Как показывает эксперимент, характеристики эффективности классификации подтверждают продуктивность добавления в анализируемые данные информативных с точки зрения прикладной онтологии признаков. Далее приводится соответствующий скриншот.

```
0.7203690560681334
0.7459190915542938
[[1015  394]
 [   0    0]]
              precision    recall  f1-score   support

           0       0.88      0.79      0.83      1121
           1       0.41      0.56      0.48       288

    accuracy                           0.75      1409
   macro avg       0.64      0.68      0.65      1409
weighted avg       0.78      0.75      0.76      1409

++++++++++++++Score of Tree++++++++++++++++++
[0.81560284 0.77304965 0.72340426 0.80851064 0.74468085
0.76595745
 0.78723404 0.77304965 0.78014184 0.75      ]
++++++++++++++Class Names++++++++++++++++++
[0, 1]
++++++++++++++Result++++++++++++++++++
[1 0 1 ... 0 1 1]
0.7721788502484032
```

Добавление информационного поля gender (пола) и «Признака пенсионера» делает менее эффективными все рассмотренные методы классификации (более всего ухудшает метод K ближайших соседей), кроме дерева решений (рисунок 6). Общие характеристики как возможное разумное количество информативных классов предварительного анализа видны также из дендрограммы кластеризации (рисунок 7).

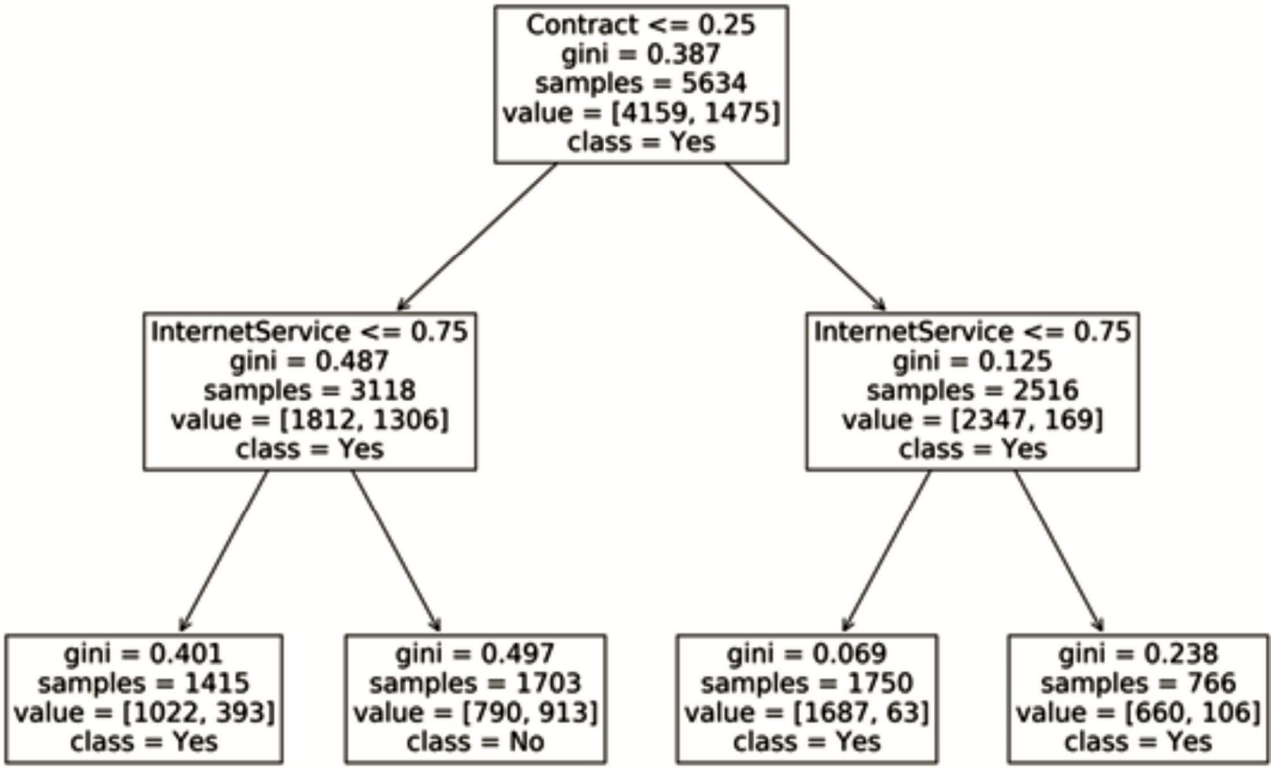

Рисунок 6 – Простейшее дерево решений
Figure 6 – Simplest decision tree

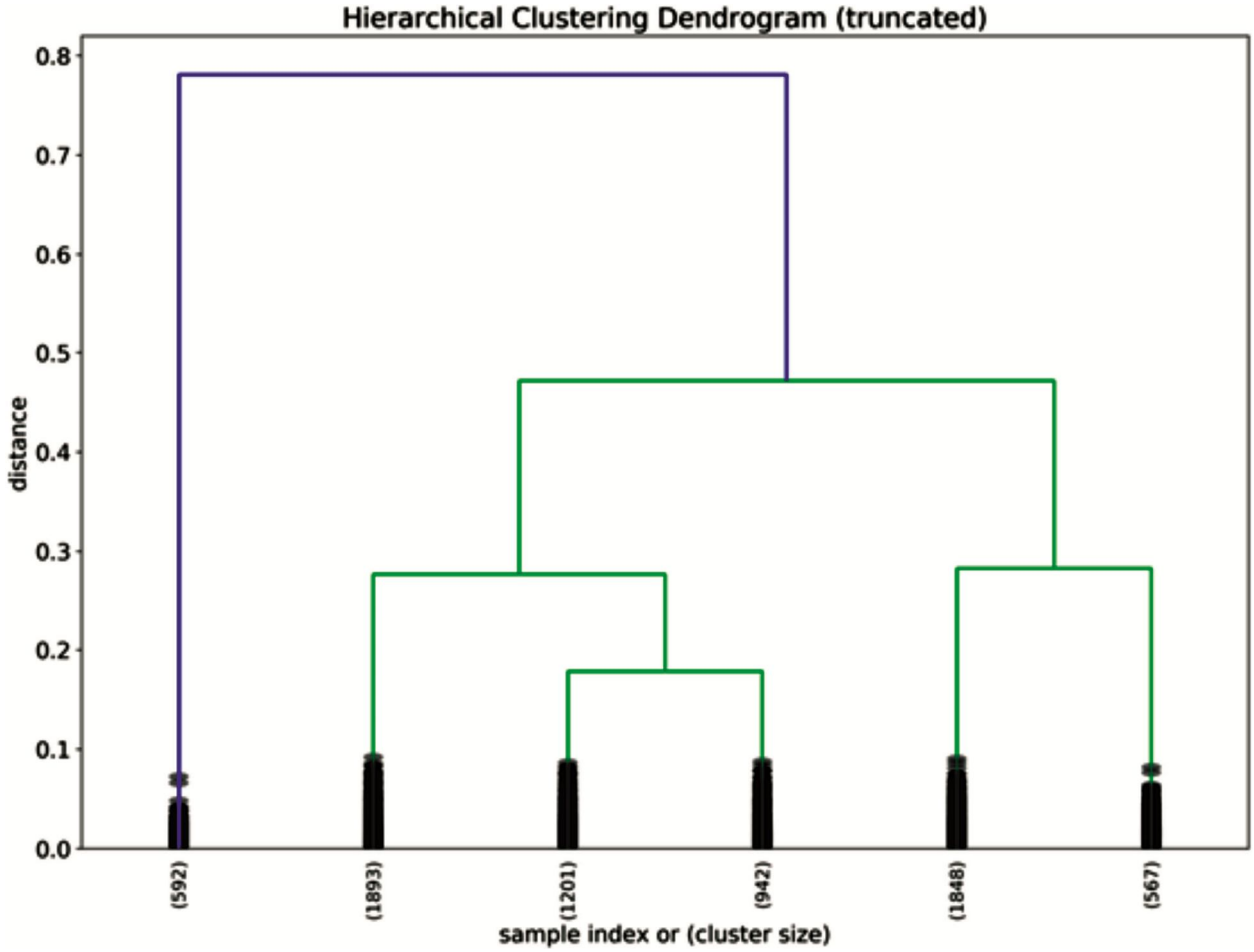

Рисунок 7 – Дендрограмма кластеризации для рассмотренного примера
Figure 7 – Correction of ICF-ontology inheritance

Из следующего скриншота можно сделать вывод о характеристиках эффективности дерева решений.

```
0.7203690560681334
0.7459190915542938
[[1015  394]
 [   0    0]]
              precision    recall  f1-score   support

           0       0.88      0.79      0.83      1121
           1       0.41      0.56      0.48       288

    accuracy                           0.75      1409
   macro avg       0.64      0.68      0.65      1409
weighted avg       0.78      0.75      0.76      1409

++++++++++++++Score of Tree++++++++++++++++++
[0.79432624 0.76595745 0.70921986 0.80851064 0.74468085
0.76595745
 0.76595745 0.81560284 0.75886525 0.71428571]
++++++++++++++Class Names++++++++++++++++++
[0, 1]
++++++++++++++Result++++++++++++++++++
[1 0 1 ... 0 1 1]
0.7771469127040455
```

Добавление информационного поля tenure (длительности общения) дает более эффективными почти все рассмотренные методы классификации, включая нейросеть.

```
++++++++++++++Score of Tree++++++++++++++++++
[0.78014184 0.75886525 0.75177305 0.78723404
0.73758865 0.81560284
 0.79432624 0.78723404 0.85106383 0.75714286]
++++++++++++++Class Names++++++++++++++++++
[0, 1]
++++++++++++++Score of Neural network++++++++++++++
+++
[1 0 1 ... 0 0 0]
0.7899219304471257
```

Полученные результаты в соответствии с ранее сформулированными гипотезами дают возможность обоснованно изменить архитектуру прикладной онтологии для рассмотренного примера, увеличив смысловую близость (приблизив в родовидовой таксономии) концепт «Длительность отношений» к концепту «Участник отношений», а следовательно и к концепту «Удовлетворенность» (рисунок 7).

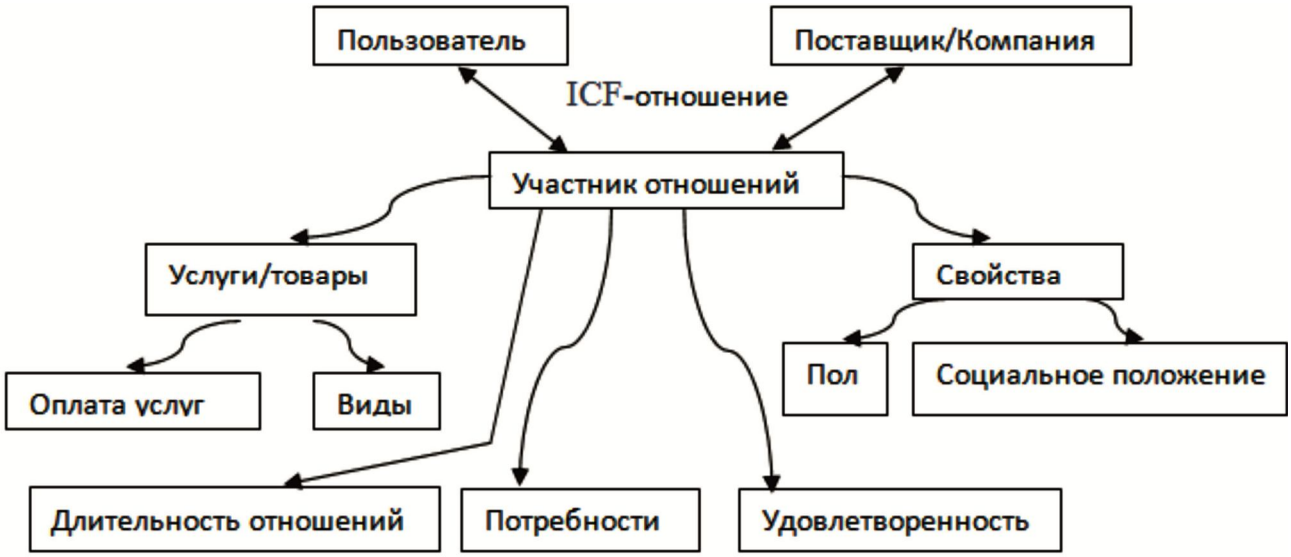

Рисунок 8 – Исправление наследования в ICF-онтологии
Figure 8 – Correction of inheritance in ICF-ontology

Кроме того, к результатам анализа можно отнести подтверждение правильности априорного причинно-следственного отношения «Удовлетворенность потребителя ↑ = ► Время отношений с поставщиком ↑», входящего в библиотеку причинно-следственных таксономий.

Заключение

Результаты работы свидетельствуют об эффективности и продуктивности комплексного использования технологий интеллектуального анализа данных с параллельным проектированием прикладных онтологий одной и той же предметной области. Такой подход позволяет формировать библиотеки родовидовых и причинно-следственных таксономий, а также совершенствовать с их помощью алгоритмы анализа больших данных. Одним из перспективных пунктов приложения результатов, рассмотренных в статье, является анализ данных глобальной сети Internet [14, 15].

Библиографический список

1. Международный репозиторий для анализа данных и оригинальных технологических решений. [Электронный ресурс]. 2020. Дата обновления: 05.10.2020. URL: https://www.kaggle.com/ (дата обращения: 05.10.2020).

2. **Каширин Д. И., Каширин И. Ю., Пылькин А. Н.** Полиморфическое представление знаний в Semsntic Web. М.: Горячая линия-Телеком, 2009. 136 с.

3. **Каширин Д. И., Каширин И. Ю.** Модели представления знаний в системах искусственного интеллекта. Вестник Рязанского государственного радиотехнического университета 2010. № 35. С. 36-43.

4. **Sergey A. Krylov.** 2018, Ontological Taxonomy As A Means To Inventory The Elements Of The Semantic Metalanguage Of Cognitive Analysis (based on E. V. Rakhilina's monograph Cognitive Analysis of Object Names (Semantics and Collocations), Slovo.ru: baltijskij accent. 2018, vol. 9, no. 3, pp. 25-54. DOI: 10.5922/2225-5346-2018-3-2.

5. **Каширин И. Ю., Минашкин С. А.** Онтологии для представления знаний в интерактивных сервисах информационных сетей // Вестник Рязанского государственного радиотехнического университета. 2020. № 39-2. С. 72-86.

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

7. **Baclawski K., Bennett M., Berg-Cross G., Fritzsche D., Schneider T., Sharma R., Westerinen A.** (2018). Ontology Summit 2017 Communique: AI, Learning, Reasoning and Ontologies // Applied Ontology, 13(1), pp. 3-18.

8. **Davies J., Studer R., Warren P.** Semantic Web technologies: trends and research in ontology-based systems / Davies, J. (N. John), Chichester. 2006. 312 c.

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

11. **Каширин И. Ю.** Формальное исследование интеллектуальных программных систем. М.: Радио и связь, 1997. 160 с.

12. **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.

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

14. **Манцивода А. В., Малых А. А.** Представление и обработка знаний в Интернете. Иркутский государственный университет. 2005. 103 с.

UDC 007:681.512.2

NEURAL NETWORKS WITH KNOWLEDGE FOR BIG DATA ANALYSIS

I. Yu. Kashirin, Dr. in technical sciences, full professor, RSREU, Ryazan, Russia;
orcid.org/0000-0003-1694-7410, e-mail: igor-kashirin@mail.ru.

The most important task to design learning algorithms for mining big data, including neural network, is considered. To improve their efficiency, taxonomies for artificial intelligence knowledge models are used. It is proposed to use ICF contiguous inheritance relation as basic taxonomy relationship. This approach makes it possible to effectively design general and applied ontologies to solve artificial intelligence problems. The article uses mathematical apparatus of hierarchical numbers allowing to analyze applied ICF ontologies by calculating the main characteristics of taxonomic hierarchies. Such characteristics include, for example, a measure of concepts semantic similarity or the complexity of structural path in taxonomic hierarchy between two concepts. As an example that confirms the possibility of effective use of ICF taxonomies in solving real problems of data analysis, we propose the solution of clustering and classification problems in «Reducing churn and attracting customers of telecommunication company» subject area. For this problem the possibility of interdependent design of neural networks and applied ontologies with a priori libraries of causal taxonomies, for example, in Semantic Web technology, is shown. In the experimental part, to test the capabilities of information retrieval in global networks based on Semantic Web technology using ICF taxonomy, an

experimental bench program for data mining methods (ESMIAD v.12.02.2021) is considered, which allows by placing fragments of program text in comments or vice versa perform experiments for analytical algorithms with training samples. The experimental stand is implemented in Python v.3.7 in Anaconda v.3.0 development environment. The purpose of this stand is to analyze the effectiveness of data mining methods using knowledge representation models.

***The aim of the work** is to analyze design features of learning algorithms for data analysis using general and applied ICF ontologies with the apparatus of hierarchical numbers to efficiently solve urgent problems of clustering and classification.*

***Key words:** knowledge models, big data, artificial intelligence, neural networks, ICF ontology, ontological taxonomies, hierarchical numbers, software tools, measure of concept semantic similarity, Semantic Web technology, Python, Anaconda, generic and causal relationships.*

DOI: 10.21667/1995-4565-2021-75-71-84

References

1. International repository for data analysis and original technological solutions. [Electronic resource]. 2020. Date of update: 05.10.2020. URL: https://www.kaggle.com/ (date accessed: 05.10.2020).

2. **Kashirin D. I.**, **Kashirin I. Yu., Pylkin A. N.** *Polimorficheskoe predstavlenie znanij v Semsntic Web*. Moscow.: Goryachaya liniya-Telekom, 2009. 136 p. (in Russian).

3. **Kashirin D. I., Kashirin I. Yu.** Modeli predstavlenya znanij v systemah iskusstvennogo intellekta. *Vestnik Rjazanskogo gosudarstvennogo radiotehnicheskogo universiteta*, no. 1 (vypusk) Ryazan, 2010, pp. 36-43. (in Russian).

4. **Sergey A. Krylov.** *Ontological Taxonomy As A Means To Inventory The Elements Of The Semantic Metalanguage Of Cognitive Analysis* (based on E. V. Rakhilina's monograph Cognitive Analysis of Object Names (Semantics and Collocations), Slovo.ru: baltijskij accent. 2018, vol. 9, no. 3, pp. 25-54. DOI: 10.5922/2225-5346-2018-3-2.

5. **Kashirin I. Yu, Minashkin S. A.** Ontologii dlya predstavleniya znanij v interactivhyh servisah informacionnyh setej. *Vestnik Rjazanskogo gosudarstvennogo radiotehnicheskogo universiteta*. 2012, no. 1. pp. 72-86 (in Russian).

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

7. **Baclawski K., Bennett M., Berg-Cross G., Fritzsche D., Schneider T., Sharma R., Westerinen A.** (2018). *Ontology Summit 2017 Communique: AI, Learning, Reasoning and Ontologies*. Applied Ontology, 13(1), pp. 3-18.

8. **Davies J., Studer R., Warren P.** *Semantic Web technologies: trends and research in ontology-based systems* / Davies, J. (N. John), Chichester, 2006. 312 c.

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

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

11. **Kashirin I. Yu.** *Formalnoe issledovanie intellektualhyh programmnyh system.* M.: Radio I Svyaz, 1997. 160 p. (in Russian).

12. **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.

13. **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).

14. **Mancivoda A. V., Malyh A. A**. *Predstavlenie I obrabotka znanij v Internete.* Irkutsky gosudarstvenny universistet, 2005. 103 p. (in Russian).