УДК 303.732.4

МЕТРИЗАЦИЯ ИЗМЕРИТЕЛЬНЫХ ШКАЛ РАЗЛИЧНЫХ ТИПОВ И СОВМЕСТНАЯ СОПОСТАВИМАЯ КОЛИЧЕСТВЕННАЯ ОБРАБОТКА РАЗНОРОДНЫХ ФАКТОРОВ В СИСТЕМНО-КОГНИТИВНОМ АНАЛИЗЕ И СИСТЕМЕ «ЭЙДОС»[1]

Луценко Евгений Вениаминович
д.э.н., к.т.н., профессор
Кубанский государственный аграрный университет, Россия, 350044, Краснодар, Калинина, 13,
prof.lutsenko@gmail.com

В статье измерительные шкалы рассматриваются как инструмент создания формальных моделей реальных объектов и инструмент повышения степени формализации этих моделей до уровня, достаточного для их реализации на компьютерах. Описываются различные типы измерительных шкал, позволяющие создавать модели различной степени формализации; приводятся типы преобразований, допустимые при обработке эмпирических данных, полученных с помощью шкал различного типа; ставится задача метризации шкал, т.е. преобразования к наиболее формализованному виду; предлагается 7 способов метризации всех типов шкал, обеспечивающих совместную сопоставимую количественную обработку разнородных факторов, измеряемых в различных единицах измерения за счет преобразования всех шкал к одним универсальным единицам измерения в качестве которых выбраны единицы измерения количества информации. Все эти способы метризации реализованы в системно-когнитивном анализе и интеллектуальной системе «Эйдос»

Ключевые слова: МЕТРИЗАЦИЯ, АНАЛИЗ, ИЗМЕРИТЕЛЬНЫЕ ШКАЛЫ, ФАКТОРЫ, СОПОСТАВИМОСТЬ, СИСТЕМА, СИСТЕМНО-КОГНИТИВНЫЙ, «ЭЙДОС»

UDC 303.732.4

METRIZATION OF MEASURING SCALES OF DIFFERENT TYPES AND JOINT COMPARABLE QUANTITATIVE PROCESSING OF HETEROGENEOUS FACTORS IN SYSTEM-COGNITIVE ANALYSIS AND THE EIDOS SYSTEM

Lutsenko Evgeny Veniaminovich
Dr.Sci.Econ., Cand.Tech.Sci., professor
Kuban State Agrarian University, Krasnodar, Russia

The article considers measuring scales as a tool for creating formal models of real objects and a tool for increasing the degree of formalization of these models to a level sufficient to implement them on computers. It also describes the different types of measuring scales, allowing to create models of varying degrees of formalization; lists the types of transformation valid during the processing of empirical data obtained with scales of different types; develops the task of metrization of the scales, i.e. conversion to the most formalized mind; it proposes 7 ways of metrization of all the types of scales, providing a joint comparable quantitative processing of heterogeneous factors measured in different units of measure due to the conversion of all scales to one universal unit of measurement in which the measurement number of information is selected. All of these methods of metrization have been implemented in the system-cognitive analysis and in the Eidos intellectual system

Keywords: METRIZATION, ANALYSIS, MEASURING SCALES, FACTORS, COMPARABILITY, SYSTEM, SYSTEM-COGNITIVE, EIDOS

"Системы искусственного интеллекта позволяют решать сложнейшие проблемы, которые не возникали, пока этих систем не было"
/Из компьютерного фольклора/

Измерительные шкалы рассматриваются как инструмент создания формальных моделей реальных объектов и инструмент повышения степе-

[1] Работа выполнена при финансовой поддержке РГНФ (проект № 13-02-00440).

ни формализации этих моделей до уровня, достаточного для их реализации на компьютерах.

Различные подходы к классификации измерительных шкал, отраженны в работах [1, 2, 3]. Наиболее строго математически обоснованным являемся подход, предложенный проф. А.И.Орловым в работе [1]. В этой работе описываются различные типы измерительных шкал, позволяющие создавать модели различной степени формализации (таблица 1):

Таблица 1. Основные шкалы измерения по проф. А.И.Орлову [1].

Тип шкалы	Определение шкалы	Примеры	Группа допустимых преобразований $\Phi = \{\varphi\}$
Шкалы качественных признаков			
Наименований	Числа используют для различения объектов	Номера телефонов, паспортов, ИНН, штрих-**коды**	Все взаимно-однозначные преобразования
Порядковая (ранговая)	Числа используют для упорядочения объектов	Оценки экспертов, баллы ветров, отметки в школе, полезность, номера домов	Все строго возрастающие преобразования
Шкалы количественных признаков (описываются началом отсчета и единицей измерения)			
Интервалов	Начало отсчета и единица измерения произвольны	Потенциальная энергия, положение точки, температура по шкалам Цельсия и Фаренгейта	Все линейные преобразования $\varphi(x) = ax + b$, a и b произвольны, $a>0$
Отношений	Начало отсчета задано, единица измерения произвольна	Масса, длина, мощность, напряжение, сопротивление, температура по Кельвину, цены	Все подобные преобразования $\varphi(x) = ax$, a произвольно, $a>0$
Разностей	Начало отсчета произвольно, единица измерения задана	Время	Все преобразования сдвига $\varphi(x) = x + b$, b произвольно
Абсолютная	Начало отсчета и единица измерения заданы	Число людей в данном помещении	Только тождественное преобразование $\varphi(x) = x$

С данными эмпирических измерений, полученными с помощью измерительной шкалы определенного типа, корректно могут быть проведены лишь вполне определенные математические преобразования, допустимые в данной шкале, тогда как другие преобразования над ними являются некорректными и, строго говоря, бессмысленными.

На практике это часто не осознается, особенно руководством, или осознается, но недостаточно четко и на это попросту «закрывают глаза».

Например, оценки в школе или вузе представляют собой порядковые оценки уровня знаний и, хотя внешне выглядят точно как числа, фактиче-

ски числами не являются. Это наглядно демонстрируется тем, что, не смотря на то, что 2+3=5 суммарные знания двоечника и троечника не равны знаниям отличника. Тем более некорректно вычислять некие средние баллы аттестатов или полученные учащимися факультета по результатам государственных экзаменов или защиты дипломных проектов, но это всегда делается.

В таблице 1 шкалы приведены в порядке повышения *степени формализации моделей*, создаваемых с их использованием.

Спрашивается, а зачем повышать степень формализации модели? Дело в том, что чем выше степень формализации модели, тем более развитые и точные математические методы могут быть применены в этих моделях и тем точнее решаются различные задачи в реальной области[2] с использованием этих моделей, в частности тем проще использовать эти модели при проектировании и создании искусственных систем (таблица 2):

Таблица 2. Основные измерительные шкалы
и возможные математические операции с их градациями

Степень формализации шкалы	Тип шкалы	Определение шкалы	Примеры	Допустимые математические операции
1	**Наименований (номинальная)**	Числа используют для различения объектов, т.е. в качестве **кодов**	Номера телефонов, паспортов, ИНН, штрих-коды	Наличие или отсутствие тождества, эквивалентности
2	**Порядковая (ранговая)**	Числа используют для упорядочения объектов	Оценки экспертов, баллы ветров, отметки в школе, полезность, номера домов	Отношения больше, меньше
3	**Интервалов**	Начало отсчета и единица измерения произвольны	Потенциальная энергия, положение точки, температура по шкалам Цельсия и Фаренгейта	Сложение и вычитание
4	**Разностей**	Начало отсчета произвольно, единица измерения задана	Время	Сложение и вычитание
5	**Отношений**	Начало отсчета задано, единица измерения произвольна	Масса, длина, мощность, напряжение, сопротивление, температура по Кельвину, цены	Сложение и вычитание, умножение и деление
6	**Абсолютная**	Начало отсчета и единица измерения заданы	Число людей в данном помещении	Сложение и вычитание, умножение и деление

[2] Прежде всего это задачи идентификации, прогнозирования и принятия решений.

Из этого ясно, что при эмпирических исследованиях:

– необходимо четко отдавать себе отчет о том, какого типа измерительные шкалы в нем используются;

– надо стремиться к использованию измерительных шкал наиболее высокой степени формализации.

Но раз так, то почему же тогда абсолютные шкалы или хотя бы шкалы отношений не применяются всегда, а в ряде случаев на практике используются номинальные, порядковые и интервальные шкалы, а также шкала разностей, имеющие ограничения на возможные математические операции с эмпирическими данными, полученными с помощью этих шкал? Иногда этого и не требуется по условиям задачи, но чаще всего просто потому, что отсутствуют[3] соответствующие измерительные системы[4] с необходимыми для этого возможностями, т.е. способные *сразу, т.е. непосредственно в процессе измерений, представить измеряемые величины в абсолютной шкале или шкале отношений*.

Но оказывается это возможно сделать и *после* завершения самого процесса измерения, т.е. уже после прекращения контакта измерительной системы с измеряемым объектом. Иначе говоря, *возможно провести такую математическую обработку данных, полученных в результате измерений с помощью измерительной шкалы определенной степени формализации, которая бы повысила эту степень формализации*.

Из таблиц 1 и 2 видно, что *для этого необходимо обоснованно ввести на исходной шкале отношения порядка по степени выраженности свойства, измеряемого шкалой, начало отсчета и единицу измерения*. Эта идея, по-видимому, впервые была четко сформулирована в 1958 году датским математиком Г. Рашем (*Georg Rasch*) [2][5] и им же была поставлена и

[3] Или где-то существуют, но на практике исследователям недоступны

[4] Т.е. измерительные инструменты, методики и технологии, включая датчики измерений, каналы связи между датчиками и системой обработки, а также методы математической обработки

[5] См. так называемую «Модель Раша».

решена соответствующая «задача метризации шкал», т.е. задача преобразования шкалы к наиболее формализованному виду. Это название связано с понятием ***метрики, под которой в физике понимается способ измерения расстояний*** между градациями (значениями) шкалы. Иначе говоря, метризация шкалы проводится с целью повышения степени ее формализации и осуществляется путем ввода метрики, т.е. единицы измерения на этой шкале. В современном понимании ***метризация шкалы*** *предполагает не только введение единицы измерения, но также и отношений порядка и начала отсчета на ней*.

Модель Г.Раша математически тесно связана с моделью логитов, предложенной в 1944 году Джозефом Берксоном (*Joseph Berkson*)[6] и здесь мы ее не приводим, т.к. она подробно описана в литературе. Модель Г.Раша (с учетом ее модификаций) является чуть ли не единственной широко известной в настоящее время моделью метризации измерительных шкал.

Однако в системно-когнитивном анализе (СК-анализ) и его программном инструментарии: интеллектуальной системе «Эйдос» [4] предлагается еще 7 способов метризации *всех* типов шкал[7], обеспечивающих, кроме того еще и корректную *совместную* сопоставимую количественную обработку *разнородных* по своей природе факторов[8], измеряемых в различных единицах измерения.

В СК-анализе факторы формально описываются шкалами, а значения факторов – градациями шкал. Существует три основных группы факторов: физические, социально-экономические и психологические (субъективные) и в каждой из этих групп есть много различных видов факторов, т.е. есть много различных физических факторов, много социально-экономических и

[6] http://www.machinelearning.ru/wiki/index.php?title=Функция%20Логит
[7] даже шкалы отношений и абсолютной шкалы
[8] физических, социальных и субъективных, и в каждой из этих групп факторов есть много различных видов факторов

много психологических, но в СК-анализе все они рассматриваются ***с одной единственной точки зрения****: сколько информации содержится в их значениях о переходе объекта, на который они действуют, в определенное состояние, и при этом сила и направление влияния всех значений факторов на объект измеряется в одних общих для всех факторов единицах измерения:* ***единицах количества информации***. Именно по этой причине вполне корректно складывать силу и направление влияния всех действующих на объект значений факторов, независимо от их природы, и определять результат *совместного* влияния на объект *системы* значений факторов. При этом в общем случае объект является *нелинейным и* факторы внутри него взаимодействуют друг с другом, т.е. для них не выполняется принцип суперпозиции [5].

Если же разные факторы измеряются в различных единицах измерения, то результаты сравнения объектов будут зависеть от этих единиц измерения, что совершенно недопустимо из теоретических соображений.

Представим себе, что мы сравниваем студентов по их росту и весу, причем рост выражен в сантиметрах, а вес в килограммах (таблица 3):

Таблица 3. Сравнение студентов по их росту и весу, измеряемым в их обычных единицах измерения

	1-й студент	2-й студент	3-й студент	Сумма
Рост (см)	178	173	173	351
Вес (кг)	75	65	75	140
Сумма	253	238	248	491

Для сравнения студентов мы просто складываем рост и вес для каждого студента, и потом сравниваем эти числа, например, находим модуль их разности: |253-238|=15 и считаем, что она отражает сходство-различие студентов по этим параметрам. Проверим корректность этого метода путем сравнения 3-го студента с ростом как у 2-го студента 173 сантиметра и

весом как 1-го студента 75 килограммов. Спрашивается, на какого студента он больше похож: на 1-го или 2-го? Очевидно, что он должен иметь одинаковое сходство и различие с обоими этими студентами, т.к. у него в равной степени представлены признаки их обоих. Однако, для 3-го студента сумма роста и веса равна: 173+75=248 и его отличие от 1-го составляет |253-248}=5, а от 2-го: |238-248|=10, т.е. получается, что третий студент в отличатся от 2-го больше, чем от 1-го. Этот результат является некорректным и связан с тем, что рост 1-го и 2-го студентов отличается на 5 сантиметров, а вес на 10 килограммов. Конечно, сложение и вычитание величин, измеряемых в разных единицах измерения, некорректно само по себе. Но особенно хорошо это заметно, когда мы меняем единицы измерения. Так если рост измерять не в сантиметрах, а в миллиметрах, то его числовое выражение возрастет в 10 раз как и его влияние на сходство-различие студентов, а роль веса при этом сравнении соответственно снизится. И наоборот, если рост оставить в сантиметрах, а вес начать измерять не в килограммах, а в граммах, то тогда сходство-различие студентов в основном будет определять уже их вес, т.к. его количественное выражение и влияние на результаты сравнения возрастет в 1000 раз.

В СК-анализе и системе предложено кардинальное решение проблем сравнения объектов, описанных в измерительных шкалах различных типов и размерностей [6]. Продолжим пример со студентами. В соответствии с методологией СК-анализа и методикой применения системы «Эйдос» *для сравнения студентов используем не их рост и вес в обычных единицах измерения, а количество информации о том, что перед нами тот или иной студент, которое содержится в его росте и весе*. Можно сравнить 3-го студента с первыми двумя по суммарному количеству информации в его признаках о сходстве с 1-м и 2-м студентами. Это будет вполне корректно и результат такого сравнения вообще не будет зависеть от исходных еди-

ниц измерения роста и веса, т.е. будет *инвариантным* относительно единиц измерения исходных признаков, как и должно быть.

Рассмотрим численный пример, демонстрирующий, что выбор единиц измерения никак не влияет на модель и результат сравнения с ее применением.

Таблица 4. Исходные данные

Источник данных	Классификационная шкала	Описательные шкалы	
	Студент	Рост (см)	Вес (кг)
1-й студент	1-й	178	75
2-й студент	2-й	173	65

С помощью программного интерфейса системы «Эйдос-Х++» (рисунок 1) данные из таблицы 4 вводятся в систему.

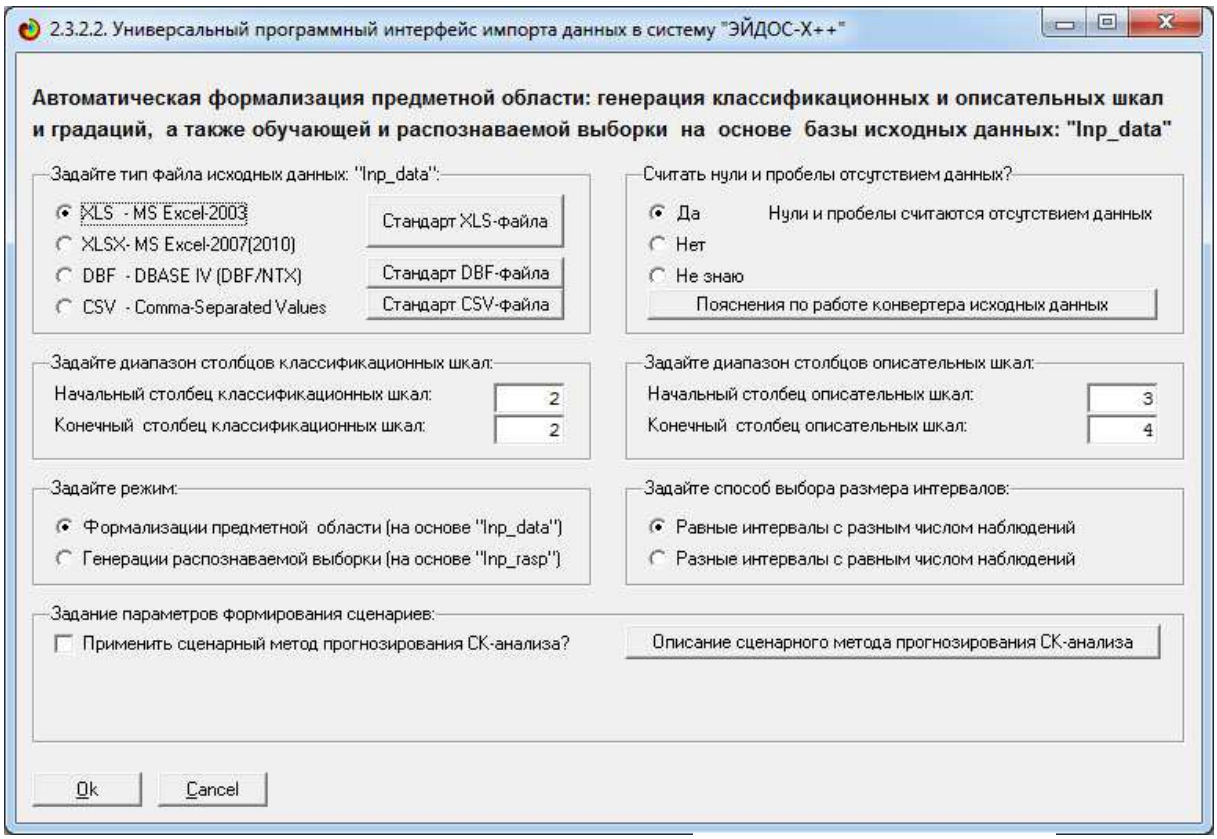

Рисунок 1. Начальная экранная форма программного интерфейса системы «Эйдос-Х++» с внешними базами данных

В первой экранной форме задается диапазон столбцов таблицы исходных данных 4 классификационными шкалами и диапазон столбцов с описательными шкалами. В экранной форме, представленной на рисунке 2,

задается количество интервалов в числовых классификационных и описательных шкалах, если они есть.

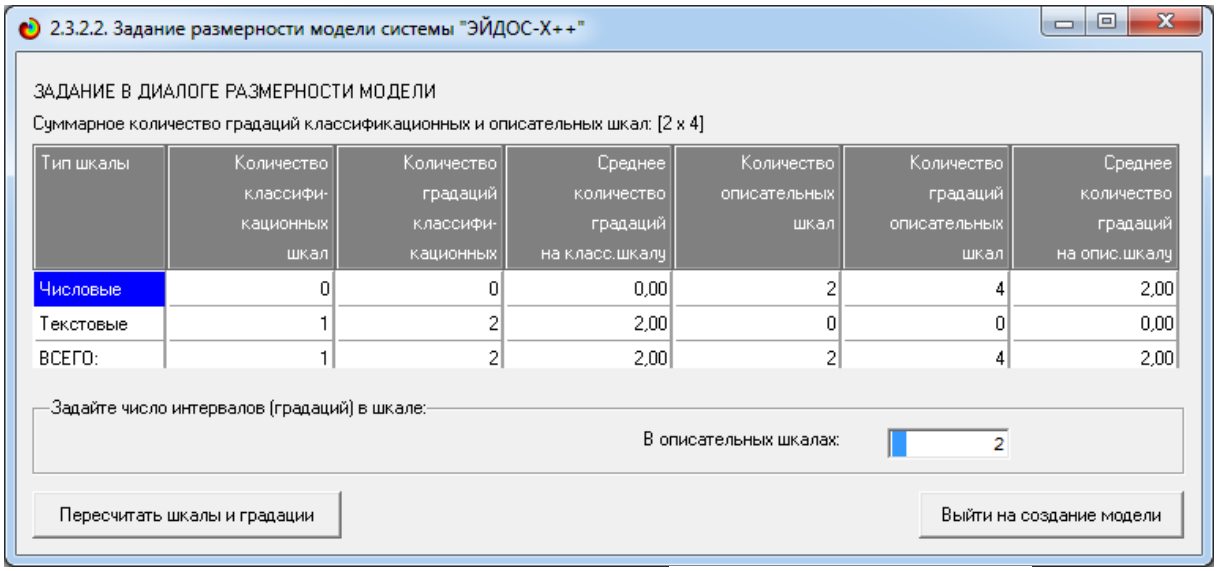

Рисунок 2. Вторая экранная форма программного интерфейса системы «Эйдос-X++» с внешними базами данных

В текущей версии системы «Эйдос-X++» суммарное количество классификационных и описательных шкал не должно превышать 1500, а суммарное количество градаций в них ограничено только размерами дисковой памяти[9].

При этом программным интерфейсом создаются справочники классификационных и описательных шкал и градаций и с их использованием кодируются исходные данные и формируется обучающая выборка (таблицы 5, 6, 7):

Таблица 5. Справочники классификационных шкал и градаций

Код класса	Наименование класса
1	СТУДЕНТ-1-й
2	СТУДЕНТ-2-й

Классы представляют собой градации классификационных шкал.

[9] Проводились численные эксперименты до 100000 градаций классификационных шкал и 100000 градаций описательных шкал. Программный интерфейс испытывался на вводе в систему «Эйдос-X++» данных и Excel-файла с 880000 строк, это заняло 7 минут.

Таблица 6. Справочники описательных шкал и градаций

Код признака	Наименование признака
1	РОСТ (СМ)-1/2-{173.0000000, 175.5000000}
2	РОСТ (СМ)-2/2-{175.5000000, 178.0000000}
3	ВЕС (КГ)-1/2-{65.0000000, 70.0000000}
4	ВЕС (КГ)-2/2-{70.0000000, 75.0000000}

Признаки представляют собой градации описательных шкал.

Таблица 7. Обучающая выборка

Код объекта	Наименование объекта	Классы	Признаки	
		CLS1	ATR1	ATR2
1	1-й студент	1	2	4
2	2-й студент	2	1	3

В результате синтеза и верификации моделей в режиме 3.5 системы «Эйдос-Х++» создаются матрица абсолютных частот (таблица 8) и матрица информативностей (таблица 9):

Таблица 8. Матрица абсолютных частот

Код признака	Наименование описательной шкалы и градации	Классы	
		1-й студент	2-й студент
1	РОСТ (СМ)-1/2-{173.0000000, 175.5000000}	0	1
2	РОСТ (СМ)-2/2-{175.5000000, 178.0000000}	1	0
3	ВЕС (КГ)-1/2-{65.0000000, 70.0000000}	0	1
4	ВЕС (КГ)-2/2-{70.0000000, 75.0000000}	1	0

Таблица 9. Матрица информативностей

Код признака	Наименование описательной шкалы и градации	Классы	
		1-й студент	2-й студент
1	РОСТ (СМ)-1/2-{173.0000000, 175.5000000}	0,0000000	0,5000000
2	РОСТ (СМ)-2/2-{175.5000000, 178.0000000}	0,5000000	0,0000000
3	ВЕС (КГ)-1/2-{65.0000000, 70.0000000}	0,0000000	0,5000000
4	ВЕС (КГ)-2/2-{70.0000000, 75.0000000}	0,5000000	0,0000000

Из таблицы 9 видно, что каждому интервальному значению роста и веса соответствует 0.5 бит информации о принадлежности студента с этим признаком к тому или иному классу. Ясно, что если в таблицах 6, 8 и 9 одинаково переставить десятичную запятую в интервальных значениях роста и веса, то коды в обучающей выборке (таблица 7), а значит и на абсо-

лютные частоты их наблюдения по классам и количество информации, рассчитываемое на их основе, это никак не повлияет.

Рассмотрим *этапы* последовательного повышения степени формализации модели путем преобразования исходных данных в информацию, а ее в знания, применяемые в автоматизированном системно-когнитивном анализе и системе «Эйдос-Х++» [7] (рисунок 3).

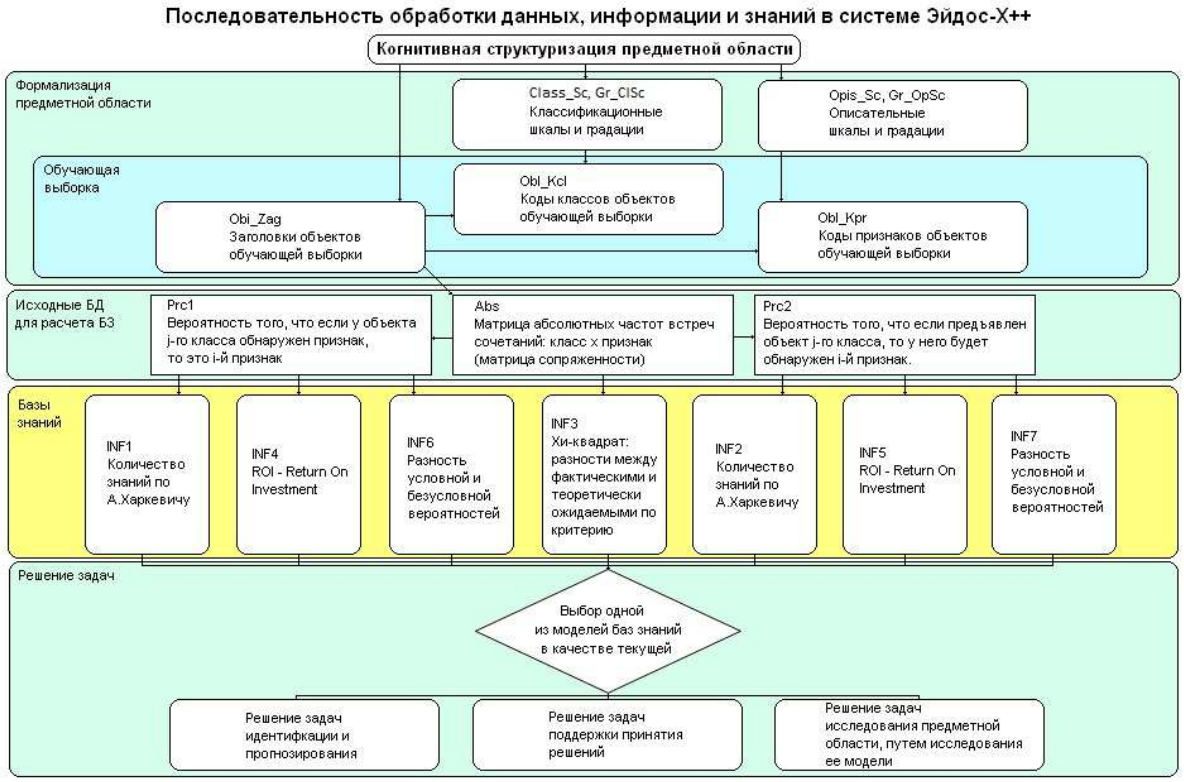

Рисунок 3. Этапы последовательного повышения степени формализации модели путем преобразования исходных данных в информацию, а ее в знания, применяемые в автоматизированном системно-когнитивном анализе и системе «Эйдос-Х++»

Прежде всего, кратко рассмотрим соотношение содержания понятий: «данные», «информация» и «знания».

Данные – это информация, рассматриваемая безотносительно к ее смысловому содержанию, находящаяся на носителях или в каналах связи и представленная в определенной системе кодирования или на определенном языке (т.е. в формализованном виде).

Информация – это *осмысленные* данные. Смысл, семантика, содержание (согласно концепции смысла Шенка-Абельсона [4, 10]) – это знание причинно-следственных зависимостей.

Знания – это информация, *полезная* для достижения целей (рисунок 4).

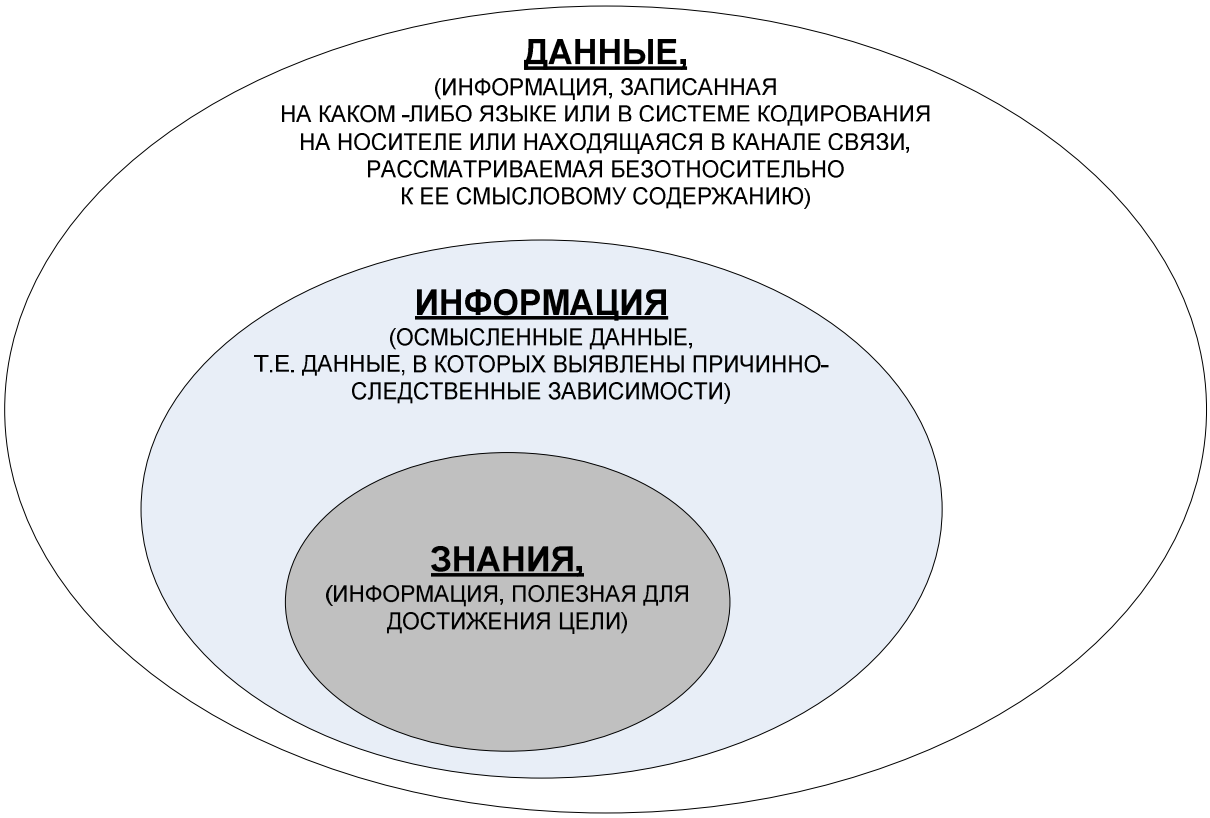

Рисунок 4. Соотношение содержания понятий:
«данные», «информация», «знания»

Знания могут быть представлены в различных формах, характеризующихся различной *степенью формализации*:

– вообще неформализованные знания, т.е. знания в своей собственной форме, ноу-хау (мышление без вербализации есть медитация);

– знания, формализованные на естественном вербальном языке;

– знания, формализованные в виде различных методик, схем, алгоритмов, планов, таблиц и отношений между ними;

– знания в форме технологий, организационных производственных, социально-экономических и политических структур;

– знания, формализованные в виде математических моделей и методов представления знаний в автоматизированных интеллектуальных системах (логическая, фреймовая, сетевая, продукционная, нейросетевая, нечеткая и другие).

Таким образом, для решения задачи метризации шкал в АСК-анализе необходимо осознанно и целенаправленно ***последовательно повышать степень формализации*** исходных данных до уровня, который позволяет ввести исходные данные в интеллектуальную систему, а затем:

– преобразовать исходные данные в информацию;

– преобразовать информацию в знания;

– использовать знания для решения задач прогнозирования, принятия решений и исследования предметной области.

Для этого в АСК-анализе предусмотрены следующие этапы [4]:

1. Когнитивная структуризация предметной области, при которой определяется, что мы хотим прогнозировать и на основе чего (конструирование классификационных и описательных шкал).

2. Формализация предметной области [8]:

– разработка градаций классификационных и описательных шкал (номинального, порядкового и числового типа);

– использование разработанных на предыдущих этапах классификационных и описательных шкал и градаций для формального описания (кодирования) исследуемой выборки.

3. Синтез и верификация (оценка степени адекватности) модели [9].

4. *Если модель адекватна*, то ее использование для решения задач идентификации, прогнозирования и принятия решений, а также для исследования моделируемой предметной области [4].

Для синтеза моделей в АСК-анализе в настоящее время используется 7 частных критериев знаний (таблица 10), а для верификации моделей 2 интегральных критерия:

Таблица 10. Частные критерии знаний, используемые в настоящее время в СК-анализе и системе «Эйдос-Х++»

Наименование модели знаний и частный критерий	Выражение для частного критерия	
	через относительные частоты	через абсолютные частоты
INF1, частный критерий: количество знаний по А.Харкевичу, 1-й вариант расчета относительных частот: *Nj* – суммарное количество признаков по *j*-му классу. Относительная частота того, что если у объекта j-го класса обнаружен признак, то это i-й признак	$I_{ij} = \Psi \times Log_2 \frac{P_{ij}}{P_i}$	$I_{ij} = \Psi \times Log_2 \frac{N_{ij} N}{N_i N_j}$
INF2, частный критерий: количество знаний по А.Харкевичу, 2-й вариант расчета относительных частот: *Nj* – суммарное количество объектов по *j*-му классу. Относительная частота того, что если предъявлен объект j-го класса, то у него будет обнаружен i-й признак.	$I_{ij} = \Psi \times Log_2 \frac{P_{ij}}{P_i}$	$I_{ij} = \Psi \times Log_2 \frac{N_{ij} N}{N_i N_j}$
INF3, частный критерий: Хи-квадрат: разности между фактическими и теоретически ожидаемыми абсолютными частотами	---	$I_{ij} = N_{ij} - \frac{N_i N_j}{N}$
INF4, частный критерий: ROI - Return On Investment, 1-й вариант расчета относительных частот: *Nj* – суммарное количество признаков по *j*-му классу[10]	$I_{ij} = \frac{P_{ij}}{P_i} - 1 = \frac{P_{ij} - P_i}{P_i}$	$I_{ij} = \frac{N_{ij} N}{N_i N_j} - 1$
INF5, частный критерий: ROI - Return On Investment, 2-й вариант расчета относительных частот: *Nj* – суммарное количество объектов по *j*-му классу	$I_{ij} = \frac{P_{ij}}{P_i} - 1 = \frac{P_{ij} - P_i}{P_i}$	$I_{ij} = \frac{N_{ij} N}{N_i N_j} - 1$
INF6, частный критерий: разность условной и безусловной относительных частот, 1-й вариант расчета относительных частот: *Nj* – суммарное количество признаков по *j*-му классу	$I_{ij} = P_{ij} - P_i$	$I_{ij} = \frac{N_{ij}}{N_j} - \frac{N_i}{N}$
INF7, частный критерий: разность условной и безусловной относительных частот, 2-й вариант расчета относительных частот: *Nj* – суммарное количество объектов по *j*-му классу	$I_{ij} = P_{ij} - P_i$	$I_{ij} = \frac{N_{ij}}{N_j} - \frac{N_i}{N}$

Обозначения:

i – значение прошлого параметра;

j - значение будущего параметра;

N_{ij} *– количество встреч j-го значения будущего параметра при i-м значении прошлого параметра;*

M – суммарное число значений всех прошлых параметров;

W - суммарное число значений всех будущих параметров.

N_i *– количество встреч i-м значения прошлого параметра по всей выборке;*

N_j *– количество встреч j-го значения будущего параметра по всей выборке;*

N – количество встреч j-го значения будущего параметра при i-м значении прошлого параметра по всей выборке.

I_{ij} *– частный критерий знаний: количество знаний в факте наблюдения i-го значения прошлого параметра о том, что объект перейдет в состояние, соответствующее j-му значению будущего параметра;*

Ψ – нормировочный коэффициент (Е.В.Луценко, 2002), преобразующий количество информации в формуле А.Харкевича в биты и обеспечивающий для нее соблюдение принципа соответствия с формулой Р.Хартли;

P_i *– безусловная относительная частота встречи i-го значения прошлого параметра в обучающей выборке;*

P_{ij} *– условная относительная частота встречи i-го значения прошлого параметра при j-м значении будущего параметра .*

[10] Применение предложено Л.О. Макаревич

Все эти способы метризации с применением 7 частных критериев знаний (таблица 10) реализованы в системно-когнитивном анализе и интеллектуальной системе «Эйдос» и обеспечивают сопоставление градациям всех видов шкал числовых значений, имеющих смысл количества информации в градации о принадлежности объекта к классу. Поэтому является корректным применение интегральных критериев, включающих операции умножения и суммирования, для обработки числовых значений, соответствующих градациям шкал. Это позволяет единообразно и сопоставимо обрабатывать эмпирические данные, полученные с помощью любых типов шкал, применяя при этом все математические операции [8].

Рассмотрим интегральные критерии знаний, используемые в настоящее время в СК-анализе и системе «Эйдос-X++» для верификации моделей и решения задач идентификации и прогнозирования.

1-й интегральный критерий «Сумма знаний» представляет собой суммарное количество знаний, содержащееся в системе факторов различной природы, характеризующих сам объект управления, управляющие факторы и окружающую среду, о переходе объекта в будущие целевые или нежелательные состояния.

Интегральный критерий представляет собой аддитивную функцию от частных критериев знаний и имеет вид::

$$I_j = (\vec{I}_{ij}, \vec{L}_i).$$

В выражении круглыми скобками обозначено скалярное произведение. В координатной форме это выражение имеет вид:

$$I_j = \sum_{i=1}^{M} I_{ij} L_i,$$

где: M – количество градаций описательных шкал (признаков);

$\vec{I}_{ij} = \{I_{ij}\}$ *– вектор состояния j–го класса;*

$\vec{L}_i = \{L_i\}$ *– вектор состояния распознаваемого объекта, включающий все виды факторов, характеризующих сам объект, управляющие воздействия и окружающую среду (массив–локатор), т.е.:*

$$\vec{L}_i = \begin{cases} 1, \textit{ если } i-\textit{й фактор действует}; \\ n, \textit{ где}: n > 0, \textit{ если } i-\textit{й фактор действует с истинностью } n; \\ 0, \textit{ если } i-\textit{й фактор не действует}. \end{cases}$$

В текущей версии системы «Эйдос-Х++» значения координат вектора состояния распознаваемого объекта принимались равными либо 0, если признака нет, или n, если он присутствует у объекта с интенсивностью n, т.е. представлен n раз (например, буква «о» в слове «молоко» представлена 3 раза, а буква «м» - один раз).

2-й интегральный критерий «Семантический резонанс знаний» представляет собой *нормированное* суммарное количество знаний, содержащееся в системе факторов различной природы, характеризующих сам объект управления, управляющие факторы и окружающую среду, о переходе объекта в будущие целевые или нежелательные состояния.

Интегральный критерий представляет собой аддитивную функцию от частных критериев знаний и имеет вид:

$$I_j = \frac{1}{\sigma_j \sigma_l M} \sum_{i=1}^{M} \left(I_{ij} - \bar{I}_j \right) \left(L_i - \bar{L} \right),$$

где:

M *– количество градаций описательных шкал (признаков);*

$\bar{I}_j$ *– средняя информативность по вектору класса;*

$\bar{L}$ *– среднее по вектору объекта;*

σ_j *– среднеквадратичное отклонение частных критериев знаний вектора класса;*

σ_l *– среднеквадратичное отклонение по вектору распознаваемого объекта.*

Приведенное выражение для интегрального критерия «Семантический резонанс знаний» получается непосредственно из выражения для критерия «Сумма знаний» после замены координат перемножаемых векторов их стандартизированными значениями:

$$I_{ij} \to \frac{I_{ij} - \bar{I}_j}{\sigma_j}, \quad L_i \to \frac{L_i - \bar{L}}{\sigma_l}.$$

Свое наименование интегральный критерий сходства «Семантический резонанс знаний» получил потому, что по своей математической форме является корреляцией двух векторов: состояния j–го класса и состояния распознаваемого объекта.

Таким образом, в АСК-анализе:

1. Рассматривается ряд объектов (фактов), представляющих в совокупности исследуемую выборку.

2. Каждый из объектов исследуемой выборки представляет собой систему, имеющую сложную многоуровневую структуру признаков (экстенсионально описание).

3. Для каждого из объектов исследуемой выборки известно, к каким обобщенным категориям (классам) он относится (интенсионально описание).

4. Необходимо сформировать модель, обеспечивающую идентификацию объектов по их признакам, т.е. определение их принадлежности к обобщенным классам.

Если признаки и классы относятся к одному времени, то имеет место задача идентификации (распознавания). Если же признаки (факторы, причины) относятся к прошлому, а классы, характеризующие состояния объектов, – к будущему, то это задача прогнозирования. Математически эти задачи не отличаются.

Совокупность экстенсионального и интенсинального описания каждого объекта обучающей выборки, по сути, представляет собой его *определение* через подведение под более общее понятие и выделение специфических признаков. Иначе говоря каждый объект обучающей выборки описывается принадлежностью к более общей категории (классу) и наличием у него ряда признаков. Например, так определяется понятие «млекопитаю-

щее»: это животное (более общее понятие), выкармливающее своих детей молоком (специфический признак). ***На основе ряда определений конкретных объектов путем их обобщения можно получить определения классов.*** Если привести в качестве примеров исследуемой выборки множество различных животных, как млекопитающих, так и других, каждый из таких примеров определить множеством признаков и построить модель, то окажется, что наиболее характерным признаком млекопитающих является не наличие шерсти или когтей, а именно вскармливание детенышей молоком.

Процедура преобразования исходных данных в информацию – это ***анализ*** данных, состоящий из трех шагов:

– разработка справочников фактов и событий;

– выявление в исходных данных *фактов* или *событий* и их кодирование;

– выявление причинно-следственных связей (зависимостей) между этими событиями.

Фактически для преобразования исходных данных в информацию необходимо:

1. Разработать классификационные и описательные шкалы и градации.

2. С использованием классификационных и описательных шкал и градаций ***закодировать*** *исходные данные*, в результате чего получится обучающая выборка, состоящая из *фактов*, представляющих собой примеры в единстве экстенсионального и интенсинального описания.

3. Произвести расчет матриц абсолютных частот, условных и безусловных процентных распределений и матрицы информативностей, отражающей причинно-следственные связи между значениями факторов и принадлежностью объектов к классам.

Таким образом, информация по задаче – это исходные данные плюс классификационные и описательные шкалы и градации, обучающая выборка, а также матрицы частот, процентных распределений и информативностей.

Процедура преобразования информации в знания – это оценка полезности информации для достижения *цели*.

Значит знания по задаче – это информация плюс цель и оценка степени полезности информации для достижения этой цели.

Знания получаются из информации, когда мы классифицируем будущие состояния объекта управления как желательные (целевые) и нежелательные.

Банк данных – это базы данных плюс система *управления базами данных* (СУБД) (стандартные термины). СУБД – это, по сути, *система управления данными*.

Информационный банк – это информационные базы плюс информационные системы (предлагается стандартизировать эти термины). Информационная система – это, по сути, *система управления информацией*.

Банк знаний – это базы знаний плюс интеллектуальные системы (стандартные термины). Интеллектуальная система – это, по сути, *система управления знаниями*.

Существует очевидная параллель между терминами и понятиями, связанными с данными, информацией и знаниями, наглядно представленная в таблице 11.

Таблица 11. Параллель между понятиями и терминами, касающимися данных, информации и знаний

Наполнение	**Объект**	**Субъект**	**Система**
Данные	База данных (БД)	Система управления базами данных (СУБД)	Банк данных=БД+СУБД
Информация	Информационная база (ИБ)	Информационная система (система управления информационными базами – СУИБ)	Информационный банк=ИБ+СУИБ
Знания	База знаний (БЗ)	Интеллектуальная система (система управления базами знаний – СУБЗ)	Банк знаний=БЗ+СУБЗ

Сформулируем ***требования*** к форме представления данных, информации и знаний, позволяющие оценить *степень их пригодности* для решения задач прогнозирования и принятия решений, а также исследования предметной области (например, кластерного анализа).

Прежде всего, результаты решения вышеперечисленных задач должны быть ***инвариантны*** относительно:

– *единиц измерения* градаций факторов (признаков);

– *типов шкал*, используемых для формализации классов и факторов (номинальные, порядковые и числовые);

– различных *статистических характеристик исходной выборки:* частотных распределений объектов по классам (обобщенным категориям), частотных распределений градаций факторов, различий в количестве признаков в описаниях объектов исследуемой выборки, различий в суммарном количестве признаков по классам.

Кроме того, форма представления должна обеспечивать решение вышеперечисленных задач с минимальными дополнительными затратами ручного труда, а это значит, что *вся предварительная обработка должна быть максимально автоматизирована*.

Эти требования можно рассматривать и как *критерии* выбора наиболее подходящей для решения вышеперечисленных задач формы представления данных, информации и знаний.

Рассмотрим **влияние единиц измерения в исходной выборке на результаты решения задач** прогнозирования и принятия решений, а также исследования предметной области (например, кластерного анализа).

Если в исходных данных какие-то значения выражены в больших единицах измерения, то их числовые значения будут малыми, и наоборот, если единицы измерения мелкие, то числовые значения – большие. Большие значения оказывают большее влияние на результаты математической

обработки, чем малые, и *это приводит к возникновению зависимости результатов решения задач идентификации, прогнозирования и принятия решений, а также кластерного анализа, от выбранных* ***размерностей*** *исходных данных, что, на взгляд автора, совершенно неприемлемо и указывает на то, что такое решение нельзя признать корректным и даже вообще решением*. По этой же причине некорректно совместно обрабатывать сами исходные данные, представленные в ***различных*** единицах измерения (натуральных или ценовых), например, складывать расстояния, представленные в километрах и в метрах, а затем прибавлять к ним тонны и килограммы, а затем еще и безразмерные величины. Вроде это очевидно, но, как это ни удивительно, но как показывает опыт на практике это довольно часто делается, а потом еще на основе подобного «анализа» делаются и выводы. Очень странно, что обычно на это *не обращают никакого внимания* при использовании исходных данных, представленных в различных единицах измерения. Например, даже в таких популярных (причем, совершенно заслуженно) системах, как SPSS и Статистика, в подсистеме кластерного анализа приводятся примеры кластерного анализа над исходными данными, представленными в различных единицах измерения.

Для решения поставленной задачи в АСК-анализе проводится *последовательное повышение степени формализации исходных данных до уровня, обеспечивающего их обработку на компьютере в программной системе.* После выполнения когнитивной структуризации и формализации предметной области осуществляется синтез модели [7].

Пример метризованной номинальной шкалы, созданной при решении задачи из работы [7], приведен н рисунке 5:

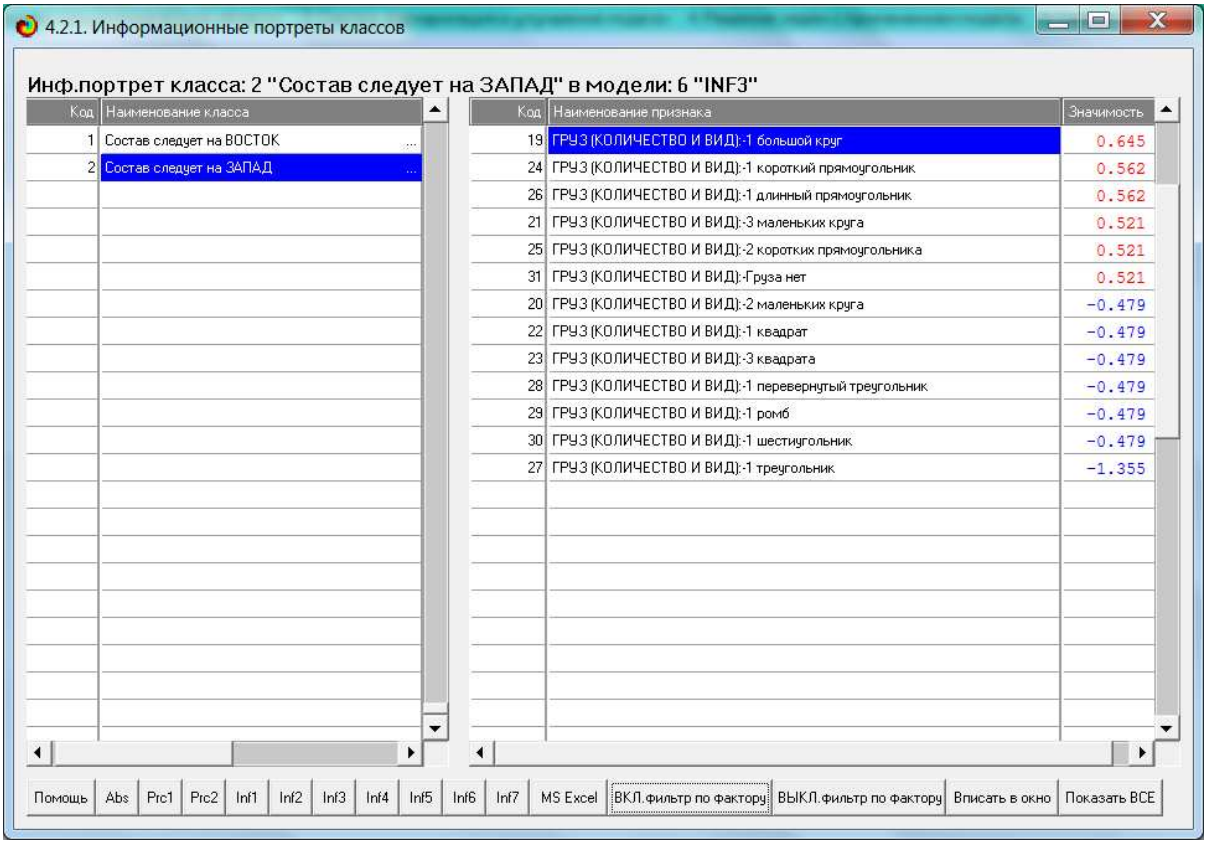

Рисунок 5. Пример метризованной номинальной шкалы «Груз (количество и вид)»

Выводы.

Отображение реальных объектов в формальных шкалах – это и есть измерение. Получается, что система «Эйдос» представляет собой средство для построения и применения измерительных инструментов в различных предметных областях, причем в ней реализованы разнообразные технологии метризации, позволяющие любые свойства объектов, как количественные, так и качественные, исследовать в наиболее сильных абсолютных шкалах знаний.

Материалы статьи могут быть использованы при проведении лекционных и лабораторных занятий по дисциплинам: «Интеллектуальные информационные системы», «Представление знаний в интеллектуальных системах», «Управление знаниями», «Эмпирические социально-экономические и психологические исследования», «Измерения в социаль-

но-экономических и психологических исследованиях», «Эконометрика», «Управление знаниями» и других.

Литература

1. Орлов А.И. Теория измерений как часть методов анализа данных: размышления над переводом статьи П.Ф. Веллемана и Л. Уилкинсона // Социология: методология, методы, математическое моделирование. 2012. № 35. С. 155-174.
2. Дубина И.Н. Математические основы эмпирических социально-экономических исследований: учебное пособие. – Барнаул: Изд-во Алт. ун-та, 2006. – 263 с.
3. ТСиСА. Вопрос №20. Электронный ресурс, режим доступа: http://e-educ.ru/tsisa20.html
4. Луценко Е.В. Автоматизированный системно-когнитивный анализ в управлении активными объектами (системная теория информации и ее применение в исследовании экономических, социально-психологических, технологических и организационно-технических систем): Монография (научное издание). – Краснодар: КубГАУ. 2002. – 605 с.[11]
5. Луценко Е.В. Моделирование сложных многофакторных нелинейных объектов управления на основе фрагментированных зашумленных эмпирических данных большой размерности в системно-когнитивном анализе и интеллектуальной системе «Эйдос-X++» / Е.В. Луценко, В.Е. Коржаков // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета (Научный журнал КубГАУ) [Электронный ресурс]. – Краснодар: КубГАУ, 2013. – №07(091). С. 164 – 188. – IDA [article ID]: 0911307012. – Режим доступа: http://ej.kubagro.ru/2013/07/pdf/12.pdf, 1,562 у.п.л.
6. Луценко Е.В. Метод когнитивной кластеризации или кластеризация на основе знаний (кластеризация в системно-когнитивном анализе и интеллектуальной системе «Эйдос») / Е.В. Луценко, В.Е. Коржаков // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета (Научный журнал КубГАУ) [Электронный ресурс]. – Краснодар: КубГАУ, 2011. – №07(071). С. 528 – 576. – Шифр Информрегистра: 0421100012\0253, IDA [article ID]: 0711107040. – Режим доступа: http://ej.kubagro.ru/2011/07/pdf/40.pdf, 3,062 у.п.л.
7. Луценко Е.В. Методологические аспекты выявления, представления и использования знаний в АСК-анализе и интеллектуальной системе «Эйдос» / Е.В. Луценко // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета (Научный журнал КубГАУ) [Электронный ресурс]. – Краснодар: КубГАУ, 2011. – №06(070). С. 233 – 280. – Шифр Информрегистра: 0421100012\0197, IDA [article ID]: 0701106018. – Режим доступа: http://ej.kubagro.ru/2011/06/pdf/18.pdf , 3 у.п.л.
8. Луценко Е.В. Типовая методика и инструментарий когнитивной структуризации и формализации задач в СК-анализе / Е.В. Луценко // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета (Научный журнал КубГАУ) [Электронный ресурс]. – Краснодар: КубГАУ, 2004. – №01(003). С. 388 – 414. – IDA [article ID]: 0030401016. – Режим доступа: http://ej.kubagro.ru/2004/01/pdf/16.pdf, 1,688 у.п.л.
9. Луценко Е.В. Математический метод СК-анализа в свете идей интервальной бутстрепной робастной статистики объектов нечисловой природы / Е.В. Луценко // Политематический сетевой электронный научный журнал Кубанского государственного

[11] Для удобства читателей эта и другие работы автора размещены на личном сайте: http://lc.kubagro.ru/

аграрного университета (Научный журнал КубГАУ) [Электронный ресурс]. – Краснодар: КубГАУ, 2004. – №01(003). С. 312 – 340. – IDA [article ID]: 0030401013. – Режим доступа: http://ej.kubagro.ru/2004/01/pdf/13.pdf, 1,812 у.п.л.

10. Васильев, Л. Г. Три парадигмы понимания: анализ литературы вопроса Электронный ресурс. / Л. Г. Васильев. — Режим доступа : http://konf-csu.narod.ru/ze/lib/vasilyev.html

Literatura

1. Orlov A.I. Teorija izmerenij kak chast' metodov analiza dannyh: razmyshlenija nad perevodom stat'i P.F. Vellemana i L. Uilkinsona // Sociologija: metodologija, metody, matematicheskoe modelirovanie. 2012. № 35. S. 155-174.

2. Dubina I.N. Matematicheskie osnovy jempiricheskih social'no-jekonomicheskih issledovanij: uchebnoe posobie. – Barnaul: Izd-vo Alt. un-ta, 2006. – 263 s.

3. TSiSA. Vopros №20. Jelektronnyj resurs, rezhim dostupa: http://e-educ.ru/tsisa20.html

4. Lucenko E.V. Avtomatizirovannyj sistemno-kognitivnyj analiz v upravlenii aktivnymi ob#ektami (sistemnaja teorija informacii i ee primenenie v issledovanii jekonomicheskih, social'no-psihologicheskih, tehnologicheskih i organizacionno-tehnicheskih sistem): Monografija (nauchnoe izdanie). – Krasnodar: KubGAU. 2002. – 605 s.

5. Lucenko E.V. Modelirovanie slozhnyh mnogofaktornyh nelinejnyh ob#ektov upravlenija na osnove fragmentirovannyh zashumlennyh jempiricheskih dannyh bol'shoj razmernosti v sistemno-kognitivnom analize i intellektual'noj sisteme «Jejdos-H++» / E.V. Lucenko, V.E. Korzhakov // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta (Nauchnyj zhurnal KubGAU) [Jelektronnyj resurs]. – Krasnodar: KubGAU, 2013. – №07(091). S. 164 – 188. – IDA [article ID]: 0911307012. – Rezhim dostupa: http://ej.kubagro.ru/2013/07/pdf/12.pdf, 1,562 u.p.l.

6. Lucenko E.V. Metod kognitivnoj klasterizacii ili klasterizacija na osnove znanij (klasterizacija v sistemno-kognitivnom analize i intellektual'noj sisteme «Jejdos») / E.V. Lucenko, V.E. Korzhakov // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta (Nauchnyj zhurnal KubGAU) [Jelektronnyj resurs]. – Krasnodar: KubGAU, 2011. – №07(071). S. 528 – 576. – Shifr Informregistra: 0421100012\0253, IDA [article ID]: 0711107040. – Rezhim dostupa: http://ej.kubagro.ru/2011/07/pdf/40.pdf, 3,062 u.p.l.

7. Lucenko E.V. Metodologicheskie aspekty vyjavlenija, predstavlenija i ispol'zovanija znanij v ASK-analize i intellektual'noj sisteme «Jejdos» / E.V. Lucenko // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta (Nauchnyj zhurnal KubGAU) [Jelektronnyj resurs]. – Krasnodar: KubGAU, 2011. – №06(070). S. 233 – 280. – Shifr Informregistra: 0421100012\0197, IDA [article ID]: 0701106018. – Rezhim dostupa: http://ej.kubagro.ru/2011/06/pdf/18.pdf , 3 u.p.l.

8. Lucenko E.V. Tipovaja metodika i instrumentarij kognitivnoj strukturizacii i formalizacii zadach v SK-analize / E.V. Lucenko // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta (Nauchnyj zhurnal KubGAU) [Jelektronnyj resurs]. – Krasnodar: KubGAU, 2004. – №01(003). S. 388 – 414. – IDA [article ID]: 0030401016. – Rezhim dostupa: http://ej.kubagro.ru/2004/01/pdf/16.pdf, 1,688 u.p.l.

9. Lucenko E.V. Matematicheskij metod SK-analiza v svete idej interval'noj butstrepnoj robastnoj statistiki ob#ektov nechislovoj prirody / E.V. Lucenko // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta (Nauchnyj zhurnal KubGAU) [Jelektronnyj resurs]. – Krasnodar: KubGAU, 2004. – №01(003). S. 312 – 340. – IDA [article ID]: 0030401013. – Rezhim dostupa: http://ej.kubagro.ru/2004/01/pdf/13.pdf, 1,812 u.p.l.

10. Vasil'ev, L. G. Tri paradigmy ponimanija: analiz literatury voprosa Jelek-tronnyj re-surs. / L. G. Vasil'ev. — Rezhim dostupa : http://konf-csu.narod.ru/ze/lib/vasilyev.html