РОССИЙСКАЯ АКАДЕМИЯ НАУК
ОРДЕНА ЛЕНИНА ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
ИМЕНИ М.В. КЕЛДЫША

Н.Л. Брошкова, С.В. Попов

О ПРОЕКТИРОВАНИИ ИНФОРМАЦИОННЫХ СИСТЕМ

Москва, 2005г.

УДК 519.72 (075)

Брошкова Н.Л., Попов С.В. О проектировании информационных систем. Препринт Института прикладной математики им. М.В. Келдыша РАН. Москва, 2005г.

На содержательном уровне обосновывается возможность использования аппарата теории информации при разработке информационных моделей физических систем. При этом под физическими понимаются детерминированные системы, задаваемые конечными дискретными отображениями.

Broshkova N.L., Popov S.V. About designing information systems. Preprint of the Keldysh Institute of Applied Mathematics of RAS. Moscow, 2005.

At a substantial level the opportunity of use the theory of the information proves by development of information models for physical systems. Under physical there are understood the determined systems set by limited discrete functions.

На дне колодца знаний залегает истина.
Выпив всю воду, ты легко обнаружишь искомое.
Авессалом .Подводный.

Введение. Основополагающей работой теории информации служит статья К.Шеннона [1]. Теория информации, главным образом, исследует количественные закономерности передачи и обработки информации в различных каналах связи. Были предприняты разнообразные попытки использования аппарата теории информации в физике, химии, биологии и даже психологии.

В настоящей работе авторы показывают, что понятия и методы теории информации могут применяться и при разработке информационных систем (ИС) для различных предметных областей (ПО). При этом мы исходим из содержательного представления об информации как это принято в теории информации: информация о некоторой явлении, факте, событии положительна, если *a priory* не известное состояние затем становится известным. И информации тем больше, чем менее ожидаемый результат получен. В контексте передачи информации, когда рассматриваются три компонента: источник информации, передающий канал и приемник, и исследуются выходные сигналы канала связи, такое представление вполне естественно. В контексте проектирования ИС, используя по существу то же содержательное представление, иначе выглядит область информационных процессов. Хотя, как показано далее, и такой подход является аналогом традиционного, описывающего передачу информации от источника к приемнику.

Пусть имеется физическая систем, ее формальная модель и наблюдатель, сравнивающий выход формальной модели с теми событиями физической си-

стемы, которые представляются формальной моделью. Формальная модель адекватно отображает физическую систему, иными словами, является информативной, если расхождение между ее реакцией на входные воздействия и соответствующими событиями физической системы находятся в допустимых пределах. В этом случае модель обладает положительной информацией о физической системе. В противном случае она мало информативна, чтобы считать ее аналогом физической системы.

***Пример* 1**. Пусть ИС создана для управления некоторым объектом управления (например, производством). Тогда она представляет интерес лишь в случае адекватного отображения поведения объекта управления. Действительно, если в ответ на запрос к ее банку данных (БД) о числе отпускников на предприятии в августе ИС выдаст число заболевших сотрудников в январе, то ее пригодность для целей управления предприятием выглядит сомнительной.

Мы полагаем, что в ответ на запрос о характеристиках физической системы формальная модель выдает ответ мгновенно, не используя для этого существенных ресурсов, связанных, например, с перебором данных. Естественно, что это идеализация, но мы ее используем, так как исследуем информационные процессы, возникающие при анализе физических систем и отображении их в моделях.

Таким образом, формальная модель более информативна, когда в ответ на заранее определенные воздействия ее (мгновенная) реакция мало отличается от ожидаемой реакцией физической системы. В любом случае можно

ожидать некоторого рассогласования поведения физической системы и предсказывающей это поведение формальной модели. Но этой разницей пренебрегают в силу ее незначительности. Естественно возникает вопрос о количественном измерении полученной информации. Как видно, ситуация с измерением информации совпадает с традиционными представлениями теории информации: мы обладаем тем большей информацией о физической системе, чем точнее реакция формальной модели соответствует ее поведению.

Можно сформулировать некоторые аналогии между понятиями теории информации и введенными представлениями.

Передатчику, расположенному на передающем конце канала связи, соответствует исследуемая физическая система (объект управления), а сигналу передатчика - ее поведение. Формальная модель физической системы соответствует каналу связи, выходной сигнал которого соответствует поведению формальной модели. Наблюдатель в обоих случаях выполняет одну и ту же роль: анализирует сигналы (в одном случае на выходе канала связи, в другом – наблюдаемые сигналы формальной моделью) и сравнивает их с сигналами передатчика или поведением физической системы. Введенные аналогии подчеркивают сходство и различие задач, информационных процессов при передаче информации по каналам связи и при проектировании ИС.

1. *Физическая система.* Чтобы сравнивать информационные процессы физических систем и формальных моделей введем их уточнение. Начнем с *физической системы*. В данном контексте нас не интересуют ее конструктивные особенности, так как наблюдатель анализирует лишь ее поведение, игнорируя механизм, который позволяет его получить. Поведение всякой

(детерминированной) системы можно описать как отображение: множество входных воздействий отображаются в поведенческих характеристиках. При изменении входных воздействий реакция системы меняется, но всякий раз она определена однозначно.

***Пример* 2**. Пусть физическая система представляет собой объект управления, поведение которого описывается функцией $y = \varphi(x)$, где $\boldsymbol{x}$ – это входные параметры, а $\boldsymbol{y}$ – демонстрируемое поведение. Допустим, что физическая система представлена БД, который выступает ее формальной моделью. Тогда отнесение кортежа к тому или иному отношению определяется его атрибутами, так как всякая таблица БД обладает фиксированным набором атрибутов. В этом случае можно описать функцию, которая по известным параметрам объекта относит его к соответствующему отношению. И такая БД может служить формальной моделью исходной физической системы.

Таким образом, говоря о физической системе, будем иметь в виду произвольную дискретную функцию $\varphi: A \to B$, где A есть множество входных воздействий, а B – выходные значения. Не уменьшая общности условимся, что аргумент и значение этой функции суть скалярные величины.

Полагаем, что входы и выходы физической системы представляют собой (в общем случае бесконечные) двоичные слова. Поэтому аргумент $\boldsymbol{x}$ отображения $\varphi(\boldsymbol{x})$ представляется в виде конечной или бесконечной последовательности $x_0, x_1, \ldots, x_i, \ldots$ двоичных переменных. Аналогично выходное значение $\boldsymbol{y} = \varphi(\boldsymbol{x})$ также представляется в виде конечной или бесконечной последовательности $y_0, y_1, \ldots, y_i, \ldots$ двоичных переменных. Если область из-

менения аргумента (или значения) есть множество двоичных чисел, то младшие разряды будем располагать левее и значения входа и выхода отображения представляется двоичными последовательностями с бесконечной нулевой правой частью. Если область определения аргумента (или значения) есть слова в некотором алфавите, то представляющие его двоичные последовательности могут иметь ограниченную длину.

Введем следующую договоренность. Так как нас интересуют информационные процессы реальных физических систем, которые описываются формальными моделями, то ограничимся рассмотрением лишь физическими системами, области определения которых суть ограниченные множества. При этом на мощность этих множеств мы не накладываем никаких ограничений, лишь считая их ограниченными некоторой известной величиной. Поэтому длины векторов, которые суть двоичные значения входа и выхода всякой физической системы конечны. Это не уменьшает общности наших рассуждений, так как для всякой реальной физической системы всегда можно указать такую величину.

Говорим, что σ_x есть ***x***-набор, если $\sigma_x = \sigma_{i_0}\sigma_{i_1}...\sigma_{i_k}\ldots$ представляет собой (бесконечную) двоичную последовательность, означивающую переменные соответственно $x_{i_0}x_{i_i}... \ x_{i_k}\ \ldots$. Мы говорим о *начальном* ***x***-наборе, если $\sigma_x = \sigma_0\sigma_1...\sigma_k$ представляет собой конечную двоичную последовательность, означивающую переменные соответственно $x_0x_1... \ x_k$. Аналогичные определения вводятся для означиваний двоичных переменных $y_0y_1... \ y_k\ldots$ значения ***y***.

2. Неопределенность физических систем. Сформулируем задачу, которую мы решаем при изучении физической системы. Основной вопрос, ко-

торый нас интересует, когда мы анализируем поведение физической системы φ, состоит в следующем: для всякой пары $<\sigma_x, \sigma_y>$ соответственно ***x***-набора и ***y***-набора установить выполняется ли соотношение $\varphi(\sigma_x) = \sigma_y$. Как правило, ответ на него не известен заранее, он характеризуется неоднозначностью (наподобие события, наступление которого происходит с некоторой вероятностью) и поэтому обладает неопределенностью, которая зависит от сложности физической системы и наших знаний о ней. (Сейчас мы не уточняем термина «сложность физической системы», понимая его содержательно.) Действительно, когда физическая система сложна, и мы не до конца ее представляем, демонстрируемое ею поведение по большей части мало предсказуемо. В лучшем случае мы можем говорить о некоторой вероятностной модели ее поведения, когда на основе имеющихся знаний с определенной достоверностью можем предсказать реакцию системы в ответ на входные воздействия.

По мере исследования знания о физической системе увеличиваются, и соответственно уменьшается неопределенность ее поведения. Поэтому задача будет состоять в том, чтобы оценить в какой степени возрастают знания и уменьшается неопределенность при исследовании физической системы и какие ресурсы для этого используются. Понятно, что в общем случае лишь поверхностное знакомство с системой не дает возможности понять ее характерные особенности. Последнее достигается за счет определенных ресурсов. Содержательно в данном контексте под ресурсами мы понимаем совершенную при анализе ПО работу, что отображается в сложности построенной

формальной модели. Использованный ресурс тем больше, чем сложнее модель, возникающая по мере изучения физической системы.

Очевидно, что на каждом этапе исследования неопределенность знаний о физической системе в общем случае уменьшается, но этот процесс не строго монотонный, так как некоторые шаги при исследования физической системы, могут оказаться неэффективными – они не увеличивают наших знаний о ее поведении. Это совпадает с содержательными представлениями, если рассматривать процесс получения знаний о некоторой области: как правило, по мере изучения какого-либо не простого предмета некоторые шаги дают увеличение наших знаний, в то время, как другие не приводят к такому результату.

Исходя из конечности физических систем, введем следующие определения.

Объемом физической системы φ назовем множество

$$\Sigma_{\varphi} = \{<\boldsymbol{x}, \boldsymbol{y}>: \boldsymbol{x} \in \mathrm{Def}\ \varphi, \boldsymbol{y} = \varphi(\boldsymbol{x})\}.$$

Неопределенностью физической системы φ называется $l\log_2 |\Sigma_{\varphi}|$. (В дальнейшем будем использовать логарифмы только по основанию 2, поэтому знак основания логарифма будем опускать.) Как видно, в данном случае неопределенность физической системы выступает и как характеристика ее сложности.

Приведем следующие содержательные рассуждения, которые касаются процесса исследования физической системы с целью построения ее формальной модели. Представим себе, что нам полностью известно влияние множества $x' \subset x$ входных параметров физической системы: при всяком их

допустимом означивании известен тип демонстрируемого поведения. В итоге можно считать, что система разбивается на определенную совокупность разделяемых по типу поведения подсистем, но поведение каждой из них не известно. Понятно, что каждая из результирующих подсистем обладают меньшей неопределенностью, чем вся исходная.

В результате система φ разделяется на подсистемы $\varphi_1,\varphi_2,...,\varphi_q$, перечисляются все допустимые x'-означивания и указываются тип подсистемы из $\varphi_1,\varphi_2,...,\varphi_q$, к которой приводит каждое из них. В общем случае, результирующих подсистем меньше, чем допустимых x'-означиваний, так как некоторые различные входные воздействия x' приводят к одинаковому поведению физической системы при всех остальных входах. Теперь первоначальный вопрос: касающийся произвольной пары $<\sigma_x,\sigma_y>$ соответственно ***x***-набора и ***y***-набора относится уже только к той части ***x***-набора σ_x, которая касается означивания компонентов ***x*** - x'. Содержательно понятно, что такой вопрос обладает меньшей неопределенностью, нежели исходный.

Следовательно, если мы обладаем полным знанием о поведении системы в зависимости от входных параметров x', и эти знания представлены в промежуточной формальной модели, то можно говорить, что исходный вопрос превращается в два следующих.

Первый касается уже спроектированной части модели и выглядит так: какой тип поведения из известных $\varphi_1,\varphi_2,...,\varphi_q$ будет демонстрировать исходная физическая система при известном входном векторе $\sigma_{x'}$? Этот вопрос

можно переформулировать следующим образом: какое событие из совокупности $\varphi_1,\varphi_2,...,\varphi_q$ наступит при произвольном входном воздействии $\sigma_{x'}$?

Второй относится к исследованию собственно каждой из результирующих подсистем $\varphi_1,\varphi_2,...,\varphi_q$, поведение которых нам заранее не известно. В данном случае мы исследуем каждую из подсистем $\varphi_1,\varphi_2,...,\varphi_q$ в точности так же, как делали это с исходной системой φ.

Таким образом, неопределенность наших знаний на этом этапе исследования физической системы складывается из двух компонентов: первый определяется видом построенной части формальной модели и второй – это оставшаяся неопределенность тех подсистем исходной физической системы, которые остаются не исследованными. В итоге задача сводится к количественной оценке этих величин и демонстрированию их связи с исходной неопределенностью физической системы.

По поводу первого пункта приведем следующие рассуждения.

Пусть для произвольного события A, происходящего с определенной частотой, $\gamma(A)$ есть показатель его *неожиданности*, зависящий от частоты p_A, с которой данное событие происходит: $\gamma(A) = -l\log p_A$. Чем ближе p_A к единице, тем меньше неожиданность события и показатель $\gamma(A)$. При приближении p_A к нулю, показатель неожиданности возрастает. Тем самым функция $\gamma(A)$ соответствует содержательному пониманию неожиданности наступления события. Поэтому $\gamma(A)$ назовем *частной неопределенностью* H_A события A.

Пусть теперь имеются q различных проявлений $\varphi_1,\varphi_2,...,\varphi_q$ одной случайной величины, каждое из которых возникает с определенной частотой соот-

ветственно p_1, p_2, …, p_q. Тогда формула для усредненной неопределенности принимает вид

$$H = -\sum_{i=1}^{q} p_i \, \gamma(\varphi_i)$$

и обозначает усреднение показателей неожиданности возможных проявлений одной случайной величины. Если рассматривать φ_i, как возможные значения случайной величины A, имеющей распределение p_i, $i = 1, 2, ..., q$, то последнюю формулу можно представить так: $H = \mathbf{M}[\gamma(A)] = \mathbf{M}[H_A]$. Неопределенность в таком случае представляет собой математическое ожидание неопределенности H_A случайной величины A.

Приведенные содержательные рассуждения позволяют ввести следующие, важные для последующего изложения определения.

Пусть $\sigma_{x'}$ есть некоторый ***x***-набор, означивающий подмножество x' двоичных переменных из совокупности $x_0x_1... \, x_k \, \ldots$. Подставим эти значения вместо соответствующих переменных. Назовем означенные переменные *неподвижными*, а оставшиеся – *изменяемыми*. Допустим теперь, что при всевозможных означиваниях оставшихся двоичных переменных из ***x*** так, что результирующие ***x***-наборы будут принадлежать множеству Def φ, определенные двоичные переменные из множества $y_0y_1...y_k \ldots$ принимают фиксированные значения, которые не меняются в результате любого такого означивания. Так же как в случае совокупности переменных аргумента назовем такие переменные – *неподвижными* для σ_x, а остальные – *изменяемыми*.

X-набор $\sigma_{x'}$ следующим образом определяет функция $\varphi|_{\sigma_{x'}}(\boldsymbol{u})$:

1. Переменные ***u*** – это изменяемые переменные аргумента, перечисленные в том же порядке, в котором они встречаются в ***x***.
2. Область определения Def $\varphi|_{\sigma_{x'}}$ получается в результате ограничения области определения Def φ теми двоичными словами, соответствующие компоненты которых совпадает с $\sigma_{x'}$.
3. Значения функции $\varphi|_{\sigma_{x'}}(\boldsymbol{u})$ представляются векторами, образованными из двоичного вектора $y_0 y_1 \ldots y_k \ldots$ игнорированием неподвижных для σ_x переменных. Можно полагать, что на места неподвижных компонентов подставлено единственное значение «_», понимаемое, как «не существенное». Изменяемые компоненты этого бинарного вектора рассматриваются в том же порядке, в котором они встречаются в ***y***.

Пример **3**. Пусть $\varphi(\boldsymbol{x}) = \boldsymbol{x} + 1$ и $\sigma_x = 01011 = 26_{10}$. Тогда $\varphi|_{\sigma_x}(\boldsymbol{u}) = 2^5\boldsymbol{u}$, так как неподвижные компоненты значений функции суть компоненты с нулевого по четвертый включительно, а остальные – изменяемые и отсутствует перенос в пятый разряд. Все значения функции $\varphi(\boldsymbol{x})$ имеют своим префиксом 11011, если ограничить область ее определения двоичными наборами с префиксом $\sigma_x = 01011$.

Говорим, что два ***x***-набора σ'_x и σ''_x по функции φ определяют одну функцию, если функции $\varphi|_{\sigma'_x}$ и $\varphi|_{\sigma''_x}$ совпадают. При этом учитывается не только порядок изменяемых переменных, но также их положение в векторе значений, т.е. каждый изменяемый компонент ассоциируется с тем местом,

которое он занимает в исходном векторе ***y***. Такие ***x***-наборы σ'_x и σ''_x назовем *эквивалентными* (для φ) и обозначим $\sigma'_x \approx \sigma''_x$.

***Пример* 4**. Пусть $\varphi(\boldsymbol{x}) = \boldsymbol{x} + 1$ и σ'_x = 01011, σ''_x= 11111, σ'''_x= 11011. Показано, что $\varphi|_{\sigma'_x}(\boldsymbol{u}) = 2^5\boldsymbol{u}$, легко проверить, что $\varphi|_{\sigma'''_x}(\boldsymbol{u}) = 2^5\boldsymbol{u}$ так как неподвижные переменные вектора значений функции также суть компоненты с нулевого по четвертый включительно. Но $\varphi|_{\sigma''_x}(\boldsymbol{u}) = 2^5\boldsymbol{u} + 1$. Значения функции $\varphi(\boldsymbol{x})$ имеют своим префиксом:

- 11011, если область ее определения ограничена двоичными наборами с префиксом σ'_x,

- 00000, если область ее определения ограничена двоичными наборами с префиксом σ''_x,

- 00111, если область ее определения ограничена двоичными наборами с префиксом σ'''_x.

Но во втором случае возникает единица переноса в пятый разряд двоичного вектора значений. В первом и третьем случаях переноса нет. Именно этот перенос делает функцию $\varphi|_{\sigma''_x}(\boldsymbol{u})$ отличной от $\varphi|_{\sigma'_x}(\boldsymbol{u})$ и $\varphi|_{\sigma'''_x}(\boldsymbol{u})$. Поэтому первый и третий наборы эквивалентны, а второй – не эквивалентен им.

Отношение эквивалентности ***x***-наборов определяет семейство функций – по одной для каждого класса эквивалентности. Значения этих функций зависят только от изменяемых компонентов. Поэтому, выделив такое семейство функций мы сводим задачу исследования исходной физической системы к описанию поведения совокупности (можно полагать независимых) физических систем. Для каждой из них нам известны означивания аргумента,

которые преобразуют в нее исходную физическую систему, но поведение ни одной частичной системы нам не известно.

Таким образом, эквивалентные означивания аргумента задают в определенном смысле совпадающие функции. Содержательно понятно, чем больше возникает классов эквивалентности, тем большее число подсистем можно выделить в исходной системе, а поэтому функция φ сложнее зависит от своих аргументов.

В теории информации неопределенность служит мерой *свободы* системы: чем больше у системы степеней свободы, т.е. чем меньше на нее наложено ограничений, тем больше ее неопределенность. Поэтому неопределенность максимальна при одинаковой доле наблюдаемых событий, а всякое отклонение от него приводит к ее уменьшению. В пределе, когда доля одного события равна 1, неопределенность равна нулю. Применительно к построенной части модели, под степенями свободы можно мыслить относительные доли классов эквивалентности означиваний аргумента. Чем меньше классов, тем больше определенность того, какая подсистема определяет поведение исходной при заданном частичном означивании аргументов и, следовательно, тем меньше неопределенность этого события. И наоборот, чем больше классов эквивалентности и чем однороднее доли p_1, p_2, …, p_q, тем больше возможностей для выбора того или иного промежуточной подсистемы и, следовательно, больше неопределенность.

Очевидно, что для описания небольшого числа различных моделей поведения требуется меньшее число параметров (в нашем случае – компонентов аргумента) и наоборот, чем более разнообразны проявления системы,

тем больше параметров необходимо для ее описания. При прочих равных условиях, более сложная система наблюдателю кажется менее определенной и наоборот, чем больше определенности у наблюдателя, тем проще зависимость демонстрируемого поведения системы от входных параметров. В контексте этой работы под сложностью зависимости системы от аргументов понимается число классов эквивалентности. Это согласуется с определением неопределенности функции от аргументов, как сложности зависимости от ее аргументов. Чем меньше классов эквивалентности, тем, с одной стороны, меньше средний показатель неожиданности, а с другой, - тем проще выражается зависимость функции от аргументов. Если же классов эквивалентности много, то больше средний показатель неожиданности и сложнее зависимость функция от аргументов.

Для этих содержательных рассуждений в последующем мы сформулируем и докажем их формальные уточнения.

Приведем несколько примеров, которые демонстрируют наши содержательные рассуждения и введенные определения. В последующем мы введем все необходимые определения и докажем ряд утверждений о свойствах указанных неопределенностей.

***Пример* 5.** Пусть некоторая ПО описывается БД, для простоты состоящим из одной таблицы T, первые k атрибутов которой образуют первичный ключ. Допустим далее, что размерность таблицы T, т.е. число ее кортежей равно N (обозначим это $|T| = N$). Неопределенность H_0 данной ПО полагаем равной $log\ N$, считая все кортежи таблицы равновероятными. Эта величина

характеризует неопределенность всей ПО, при отсутствии каких-либо знаний о ней.

Представим каждый кортеж $t \in T$ следующим образом: $t = t't''t'''$, где t' - первые ключевые l ($l < k$) атрибутов, t'' - остальные k - l ключевых атрибутов и t''' - оставшиеся (не ключевые) атрибуты.

Допустим теперь, что всевозможные допустимые означивания первых l атрибутов разбивают таблицу T на не пересекающиеся подмножества следующим образом: два кортежа t_1' и t_2' относятся к одному подмножеству (обозначим это $t_1' \approx t_2'$) тогда и только тогда, когда для каждого продолжения $t_1'' t_1'''$ имеется совпадающее с ним продолжение $t_2'' t_2'''$ и наоборот. Иными словами, если ключевые атрибуты t_1' определяет некоторое множество объектов, которые характеризуются определенными значениями оставшихся атрибутов, то атрибуты t_2' определяют то же самое множество продолжений. Содержательно это обозначает, что при всяком продолжении атрибутов t_1' означиванием оставшихся k - l ключевых атрибутов, если такое означивание приводит к кортежу из T, то имеется аналогичное продолжение кортежа t_2' и наоборот. Такое разбиение таблицы T на непересекающиеся подмножества определяет эквивалентность $\approx$ кортежей.

Понятно, что по мере увеличения l и приближения его значения к k число классов эквивалентности уменьшается, при $l = k$ каждый класс содержит в точности один кортеж.

Пусть в нашем случае $T = T_1 \cup T_2 \cup \ldots \cup T_q$ и $p_i = |T_i| / |T|$, $i = 1, 2, \ldots, q$.

Допустим, что все продолжения означиваний кортежей из класса T_i образуют множество T_i', $i = 1,2, \ldots, q$. Имеет место равенство

$$|T| = \sum_i |T_i \times T_i'|,$$

так как классы T_i, $i = 1,2, \ldots, q$ не пересекаются.

Полагаем, что на некотором этапе исследования физической системы, нам известны все классы эквивалентности T_i и их доли p_i, $i = 1, 2, \ldots, q$. Эту часть наших знаний, которая хранится в некотором легко доступном месте, можно считать фрагментом формальной системы, представляющей исследованную часть физической системы.

Теперь введем две величины, первая из них описывает неопределенность уже построенной части формальной модели, а вторая – неопределенность оставшейся части ПО. Неопределенность уже построенного фрагмента формальной системы определим, исходя из основного вопроса и основываясь на следующих содержательных рассуждениях относительно поведения физической системы.

Так как исходный вопрос превратился в два, как было показано выше, то каждый из них обладает некоторой неопределенностью. Очевидно, что ответ на вопрос связанный с построенным фрагментом формальной модели (здесь построение классов эквивалентности T_i, $i = 1, 2, \ldots, q$) зависит от доли каждого класса среди остальных и не зависит от его мощности. Это вытекает из того, что при разделении кортежей из l ключевых атрибутов мы решаем задачу отнесения каждого из них к тому или иному типу. Можно полагать, что отнесение кортежа к определенному типу есть событие, которое наступает с вероятностью, совпадающей с долей этого типа среди остальных. Но тогда неопределенность, ассоциированная с уже построенным фрагментом модели тем больше, чем труднее решается задача установления к какому классу эк-

вивалентности относится означивающий кортеж. А неопределенность такой задачи, как следует из приведенных ранее рассуждений, представляет собой неопределенность системы из q независимых событий, обладающей распределением вероятностей $p_1, p_2, \ldots, p_q$.

Поэтому неопределенность, ассоциированную с уже построенным фрагментом формальной системы мы будем характеризовать величиной

$$H_1 = -\sum_{i=1}^{q} p_i \log p_i .$$

После определения принадлежности кортежа из l ключевых атрибутов каждому классу T_i, $i = 1,2, \ldots, q$, его продолжение определяется содержанием класса T_i', вид которого заранее не известен. Поэтому его неопределенность равна $\log |T_i'|$, а неопределенность всей системы, представленной независимыми подмножествами $T_1', T_2', ..., T_q'$ есть величина

$$H_2 = \sum_{i=1}^{q} p_i \log |T_i'| .$$

Эта формула усредненной неопределенности, так как все множества T_i' независимы, а принадлежность кортежа множеству T_i' определяется только известной его частью из l ключевых атрибутов.

Таким образом, на этапе проектирования формальной модели, когда известна зависимость поведения от $l < k$ ключевых атрибутов, суммарная неопределенность наших знаний об исходной ПО равна сумме $H_1 + H_2$. Рассмотрим две крайние точки этапа проектирования формальной модели. Первая – это начало проектирования, т.е., когда $l = 0$. В этом случае первая сумма равна 0, а вторая - H_0 так как имеется лишь один класс эквивалентности, состоящий из всей таблицы T. Содержательно понятно, что в этом случае не-

определенность знаний максимальна, что и показывают равенства $H_1 = 0$, $H_2 = H_0$. С другой стороны, когда формальная система спроектирована полностью ($l = k$), оба слагаемых равны нулю, так как вновь имеется лишь один класс эквивалентности, мощность которого равна 1. В этом случае неопределенность наших знаний о ПО равна нулю, ПО описана полностью и это описание представлено формальной моделью. Интерес представляют промежуточные состояния, связанные с проектированием формальной модели (в данном контексте $0 < l < k$). Именно в этом случае возникает не нулевое значение неопределенности H_1, связанной с уже построенным фрагментом модели.

В последующем будет показано, что выполняется соотношение

$$H_1 + H_2 \le H_0.$$

В итоге на промежуточном этапе построения формальной системы наши знания, по сравнению с исходными, увеличились на величину

$$H_0 - (H_1 + H_2).$$

Ее можно считать информацией, которую мы приобрели в результате совершенной работы (затраченных ресурсов).

***Пример* 6**. Пусть $\pi(x)$ есть программа в некотором базисе, входом которой служат бинарные слова длины n. Можно полагать, что эта программа задана своей блок-схемой, в которой, как обычно, имеются два типа поименованных узлов: логические и арифметические. Тогда мы можем говорить об определенной траектории вычисления, которая задается входным значением σ_x. Под *траекторией вычисления* мы понимаем последовательность арифметических и логических операторов, которые выполняются при заданном

входе программы. В результате вычисления выходная переменная ***y*** принимает определенное значение. Неопределенность H_0 этой программы зависит от множества T всех возможных траекторий вычислений и равна $l\log|T|$.

Введем эквивалентность бинарных последовательностей длины $l \le n$ следующим образом. Две такие последовательности σ_1 и σ_2 эквивалентны, обозначается $\sigma_1 \approx \sigma_2$, если всякий вход этой программы, имеющий соответствующие входные переменные x', означенными наборами σ_1 и σ_2, приводит вычисления к одному и тому же состоянию (можно считать, что эти вычисления приводят к выполнению одной и той же подпрограммы, которая есть часть исходной программы).

В итоге, множество вычислений разбивается на классы, каждый из которых определяется эквивалентными означиваниями входных переменных. В каждый такой класс входят траектории вычислений, определяемые эквивалентными означиваниями входа программы.

Если полагать, что программа $\pi(x)$ есть физическая система, которую мы наблюдаем как сторонние наблюдатели, то после некоторого исследования возможных вычислений оказывается, что наши знания о программе увеличились, хотя неопределенность еще остается. Она полностью исчезает лишь, когда известны все траектории вычислений, определенные всеми возможными входными значениями.

Традиционно мы исследуем задачу о выполнении для всех пар $<\sigma_x,\sigma_y>$ свойства $\pi(\sigma_x)=\sigma_y$. Можно полагать, что после исследования всех вычислений, которые задаются начальными входами длины l, мы обладаем фрагментом формальной модели, описывающей эти вычисления, и в ответ на наш

запрос мы мгновенно получаем имя подпрограммы, которая должна выполняться. Пусть множество, выделенных таким образом подпрограмм, имеет вид: $\pi_1, \pi_2, \ldots, \pi_q$, множество траекторий вычислений, ведущих в одну подпрограмму π_i, имеет мощность t_i и $p_i = t_i/t$ есть доля этого множества среди всех таких траекторий, $i = 1, 2, \ldots, q$. Тем самым, мы располагаем фрагментом модели программы, которая несет определенную информацию о самой программе, хотя и сама, как показано выше, является источником неопределенности. Неопределенность этой части модели связана с вопросом: какое событие из совокупности $\pi_1, \pi_2, \ldots, \pi_q$ наступит при произвольном входном воздействии $\sigma_{x'}$, которое получается в результате означивания части x' входных переменных. Как и в предыдущем примере, эта неопределенность H_1 зависит лишь от долей p_i, не зависит от мощностей t_i, $i = 1, 2, \ldots, q$ и ее значение равно

$$-\sum_{i=1}^{q} p_i \log p_i .$$

Оставшаяся не исследованной часть исходной программы представляет собой совокупность $\pi_1, \pi_2, \ldots, \pi_q$ программ, обладающих собственными множествами траекторий соответственно $T_1', T_2', \ldots, T_q'$, где $|T_i'| = t_i'$, $i = 1, 2, \ldots, q$. Для окончательного построения формальной модели мы должны исследовать каждую из полученных подпрограмм в точности так, как делали это с исходной программой. Это приводит к средней неопределенности H_2, имеющей значение

$$\sum_{i=1}^{q} p_i \log t_i' .$$

Таким образом, неопределенность наших знаний на этом этапе исследования включает две составляющие: первая определяется видом полученного фрагмента формальной модели и вторая – это неопределенность совокупности подпрограмм, поведение которых еще не известно.

А суммарная неопределенность знаний о программе равна сумме $H_1 + H_2$. Рассмотрим две крайние точки этапа проектирования. Первая – это начало проектирования, т.е., когда $l = 0$. В этом случае первая сумма равна 0, а вторая равна H_0. Содержательно понятно, что в этом случае неопределенность максимальна, что и показывают равенства. С другой стороны, когда формальная система спроектирована полностью, т.е. описывает все траектории вычислений, оба слагаемых равны нулю. В этом случае неопределенность наших знаний о программе равна нулю, описание всех ее вычислений содержится в формальной модели . Интерес представляют промежуточные состояния, связанные с проектированием формальной модели (в данном контексте $0 < l < n$).

Как и в предыдущем примере выполняется соотношение

$$H_1 + H_2 \le H_0.$$

В итоге на промежуточном этапе построения формальной системы наши знания по сравнению с исходными увеличились на величину

$$H_0 - (H_1 + H_2).$$

Эту разность уместно считать той информацией, которую мы приобрели в результате исследования программы.

2. ***Формальная модель физической системы***. Теперь необходимо уточнить, что мы понимаем под формальной моделью физической системы.

Для этого расширим булевский язык над базисом $\{\vee, \wedge, \neg\}$ за счет введения обобщений дизъюнкции и конъюнкции на не более, чем счетное множество аргументов. Для обобщенных дизъюнкции и конъюнкции будем использовать выражения соответственно $\cup_{i=k,\ h}$ и $\cap_{i=k,\ h}$, где i называется *индексом* этой связки, k- *нижней границей*, а h – *верхней*. k и h – это либо константы, либо символ бесконечности - ω, либо функции, зависящие от других индексов.

Следует отметить, что такое расширение булевского языка обладает весьма большими выразительными возможностями. В последующем мы покажем, что для наших целей исследования физических систем достаточно ограничиться его некоторым подмножеством, обладающим рядом интересных свойств, облегчающих анализ систем. Основное свойство функций из этого класса – это их локальность. Содержательно локальность понимается как ограниченность числа классов эквивалентности означивающих наборов некоторой константой, не зависящей от вида этих наборов.

Будем говорить, что логическая формула $F(\boldsymbol{x}, \boldsymbol{y})$, где $\boldsymbol{x} = x_0x_1 \ldots x_i\ldots$, $\boldsymbol{y} = y_0y_1 \ldots y_i\ldots$ суть совокупности логических переменных, *представляет* двоичную функцию $\varphi(\boldsymbol{x})$, если при означивании переменных $\boldsymbol{x}$ бинарным вектором σ_x таким, что $\sigma_x \in \mathrm{Def}\ \varphi$, формула $F(\sigma_x, \sigma_y)$ истинна тогда и только тогда, когда $\varphi(\sigma_x) = \sigma_y$. Если логическая функция представляет физическую систему, то назовем ее *формальной моделью* последней.

***Пример* 7.** Обозначим $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{u}$ наборы двоичных переменных соответственно: $x_0, x_1, \ldots, x_n,\ldots$; $y_0, y_1, \ldots, y_n,\ldots$ и $u_0, u_1, \ldots, u_n,\ldots$. Нетрудно увидеть, что формула

$$\text{Сл}_1(\boldsymbol{x}, \boldsymbol{y})\text{: } (y_0 = \overline{x}_0) \wedge \cap_{i=1,\,\omega}\,(y_i = x_i + \cap_{j=0,\,i-1}\,x_i)$$

представляет функцию прибавления единицы двоичной арифметики. Подформула $\cap_{j=0,\,i-1}\,x_i$ представляет перенос в i-ый разряд, он равен 1 тогда и только тогда, когда все разряды до $(i - 1)$-го включительно равны 1.

В точности так же и формула

$$\text{Сл}_1(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{y})\text{: } u_0 = x_0 \wedge \cap_{i=0,\omega}\,(u_{i+1} = u_i \wedge x_i) \wedge (y_0 = \overline{x}_0) \wedge \cap_{i=1,\omega}\,(y_i = x_i + u_i)$$

представляет функцию прибавления единицы. Легко увидеть, что новая переменная u_i определяет перенос в i-ый разряд, который вычисляется после прибавления 1 к вектору с разрядами от нулевого до $(i - 1)$-го включительно. Перенос в нулевой разряд совпадает с нулевым разрядом самого числа.

Если $F(\boldsymbol{x}, \boldsymbol{y})$ есть формальная модель для конечной физической системы $\varphi(\boldsymbol{x})$, то объем последней совпадает с множеством Σ_F единичных означиваний функции F. Поэтому исходная неопределенность физической системы совпадает с величиной $l\text{og}\ |\Sigma_F|$, которую назовем *неопределенностью* ее формальной модели.

Понятно, что если логические функции $F(\boldsymbol{x}, \boldsymbol{y})$ и $G(\boldsymbol{y}, \boldsymbol{z})$ представляют двоичные функции соответственно $\varphi(\boldsymbol{x}) = \boldsymbol{y}$ и $\phi(\boldsymbol{y}) = \boldsymbol{z}$, то суперпозиция $\phi(\varphi(\boldsymbol{x})) = \boldsymbol{z}$ представляется конъюнкцией $F(\boldsymbol{x}, \boldsymbol{y}) \wedge G(\boldsymbol{y}, \boldsymbol{z})$.

Справедливо следующее утверждение.

***Теорема* 1.** Пусть логическая функция $F(\boldsymbol{x}, \boldsymbol{y})$ представляет вычислимую двоичную $y = \varphi(x)$ и σ_x есть $\boldsymbol{x}$-набор. Тогда $F(\sigma_x, \boldsymbol{y})$ представляет вычислимую функцию $\varphi|_{\sigma_x}$.

Доказательство. Ограничив Def φ двоичными словами, неподвижные компоненты которых заданы набором σ_x, получим область Def $\varphi|_{\sigma_x}$ определения функции $\varphi|_{\sigma_x}$. Все значения функции $\varphi|_{\sigma_x}$ характеризуются фиксированными значениями неподвижных компонент.

Пусть функция $F(\sigma_x, \boldsymbol{y})$ получается из $F(\boldsymbol{x}, \boldsymbol{y})$ означиванием σ_x некоторых переменных из $\boldsymbol{x}$. Для некоторых логических переменных из $\boldsymbol{y}$ существуют фиксированные значения, которые не изменяются при варьировании остальных значений переменных $\boldsymbol{x}$, не означенных набором σ_x. Как и прежде, назовем эти компоненты – *неподвижными,* оставшиеся - *изменяемыми*. Покажем, что означивание σ_x для функции $F(\boldsymbol{x}, \boldsymbol{y})$ определяет в точности то же множество неподвижных переменных, что и для функции φ.

Допустим противное, то есть при некотором $\boldsymbol{x}$-наборе, являющемся расширением σ_x, существуют переменные из $\boldsymbol{y}$, которые неподвижные для φ и изменяемые для $F(\boldsymbol{x}, \boldsymbol{y})$. Но это противоречит тому, что $F(\boldsymbol{x}, \boldsymbol{y})$ представляет функцию φ, так как некоторая изменяемая переменная будет принимать оба значения 0 и 1 в то время, как для функции $\varphi|_{\sigma_x}$ ее значение единственное либо 0 либо 1. Доказательство в обратную сторону, когда некоторая переменная для набора σ_x и формулы $F(\boldsymbol{x}, \boldsymbol{y})$ неподвижна, но изменяемая для φ, подобно.

Разобраны все случаи. Теорема доказана.

Следствие. Пусть для двоичной вычислимой функции φ $\boldsymbol{x}$-наборы σ'_x и σ''_x эквивалентны. Тогда $F(\sigma'_x, \boldsymbol{y})$ и $F(\sigma''_x, \boldsymbol{y})$ представляют одну вычислимую функцию $\varphi|_{\sigma'_x} = \varphi|_{\sigma''_x}$.

***Пример* 8.** Рассмотрим означивания переменных $\boldsymbol{x}$ функции

$$\text{Сл}_1(\boldsymbol{x}, \boldsymbol{y}): (y_0 = \bar{x}_0) \wedge \cap_{i=1,\,\omega} (y_i = x_i + \cap_{j=0,\,i-1} x_i),$$

представляющей прибавление единицы двоичной арифметики начальными $\boldsymbol{x}$-наборами $\sigma'_x = 01011$, $\sigma''_x = 11111$, $\sigma'''_x = 11011$, как в Примере 4. В результате означивания первым и третьим наборами функция $\text{Сл}_1(\boldsymbol{x}, \boldsymbol{y})$ превращается в функцию

$$\cap_{i=5,\,\omega} (y_i = x_i),$$

а в результате означивания вторым – в функцию

$$(y_5 = \bar{x}_5) \wedge \cap_{i=6,\,\omega} (y_i = x_i + \cap_{j=5,\,i-1} x_i).$$

В следующих работах будет введено понятие логической энтропии, и описаны свойства неопределенности формальных моделей, а затем показано, как установленные закономерности переносятся на физические системы.

Литература

1. Shannon C.E. The mathematical theory of communication. Bell System Techn. J., 27 (1948) №3, 379-423, 27 (1948) №4 625-656.
2. Шеннон К. Работы по теории информации и кибернетике, М.: ИЛ. 1963, - 830 с.
3. Колмогоров А.Н. Теория информации и теория алгоритмов, М.: Наука, 1987, - 303 с.
4. Файнстейн А. Основы теории информации. М.: ИЛ, 1960. – 140 с.
5. Мишулина О.А. Основные понятия статистической теории информации. М.: МИФИ, 2000. - 92 с.
6. Бриллюэн Л. Наука и теория информации. М.: Физматгиз, 1969. – 392 с.

7. Яглом А.М., Яглом И.М. Вероятность и информация. М.: Наука, 1973. – 512 с.
8. Урсул А.Д. Природа информации. М.: Сов. Радио, 1975. – 424 с.