Семантические, аксиоматические и численные основы информационных технологий. Часть 1

Геннадий ХУДЯКОВ,
д. т. н., профессор

Широчайшее распространение информационных технологий и всевозможное их использование требуют фундаментальных оснований. Однако до сих пор ни на семантическом уровне теории информации, ни на формально-математическом пользователи результатов развития информатики таких оснований не имеют.
Автор сделал попытку систематически изложить основания теории информации как на семантическом ее уровне, так и на аксиоматическом.

Введение

Развитие и широкое практическое применение математической теории информации (МТИ) в ряде прикладных научных областей — радиоэлектронике, физике, экспериментальной биологии, инженерной психологии, теоретической лингвистике — за последние полвека дали весьма важные теоретические и полезные практические результаты. Однако эффективное использование МТИ в различных областях знаний существенно затрудняется двумя обстоятельствами.

Во-первых, сама математическая теория информации как раздел теории вероятностей обычно излагается на основе математического аппарата, значительно превосходящего уровень математической подготовки не-математиков: теория меры Лебега и интеграла Стильтьеса, теория групп Абеля и др.

Во-вторых, даже в математической теории связи, на почве которой, собственно говоря, и выросла теория информации, до сих пор не сформулированы строгие и полные формально-математические основания.

К. Шеннон не создал аксиоматических оснований математической теории информации, соответствующих требованиям современной математики. Ближайший последователь К. Шеннона Р. Фано в книге [10] формулирует четыре постулата, относящиеся к взаимной информации в ансамблях случайных точек дискретного пространства, которые должны быть следствием более простых постулатов. Терминология сборника АН СССР [9] требования современной математики также не удовлетворяет. В известной автору другой литературе по теории информации (в том числе [2, 4, 5, 11]) также отсутствует вполне удовлетворительная аксиоматика теории информации.

Основные понятия общей теории информации

Информация — предельно обобщенное (абстрактное) понятие, аналогичное физическому понятию «энергия».

«Прикладной теорией энергии» в механике является термодинамика равновесных и неравновесных макроскопических систем. Законы термодинамики физики-теоретики сделали настолько абстрактными и «неконструктивными», что они, при всей их простоте, плохо усваиваются неспециалистами в области статистической физики. Тем не менее специалистам в различных областях физики и техники необходимо их знать и правильно понимать, чтобы не изобретать разного рода «вечных двигателей».

Аналогичная ситуация наблюдается и в прикладной теории информации. Различные средства формирования, преобразования, хранения и передачи информации имеют дело с конкретными сообщениями, которые могут быть зафиксированы на соответствующих носителях информации различной физической природы.

Задачей общей (семантической) теории информации является определение меры количества семантической (смысловой) информации, содержащейся в сообщениях, которые фиксируются на любых носителях сообщений, и формулировка закономерностей, сопровождающих процессы формирования (выработки), преобразования, хранения и передачи сообщений с помощью разнообразных технических устройств и систем.

В различных информационных технических системах мы имеем дело с конкретными сообщениями, которые обладают различной ценностью для нас и которые зафиксированы на разных носителях: это бумажные книги, фотографии, компакт-диски, магнитофонные и киноленты, компьютерные диски и «флэшки», запоминающие устройства и т. п.

Для практических применений мы должны определить меру количества информации, содержащейся в данном сообщении, и на ее основе разработать способы расчета избыточности источника сообщений, минимальной длины кодовых слов, надежности средств хранения и передачи сообщений, процента потерь информации при хранении и передаче сообщений, быстродействия каналов передачи информации, эффективности предлагаемых способов сохранения информации в условиях воздействия на сообщения различных помех и т. д., и т. п.

Если же мы говорим об общей теории информации, то следует сформировать общий взгляд на окружающую нас действительность и выявить информационные аспекты такого подхода.

Для более четкого и ясного понимания математических основ прикладной теории информации определим общие (базовые) понятия, относящиеся к информации вообще и к информации в технических подсистемах информационных эргатехнических систем в частности [1, 6–8, 10, 12 и 13].

Первое, что мы должны отметить: окружающая нас действительность не представляет собой некоторое аморфное образование, а вполне структурирована и состоит из различного рода систем.

Системой назовем целостную совокупность определенным образом взаимодействующих между собой функционально законченных элементов. Системы бывают материальными (существующие объективно) и идеальными (существующие в нашем сознании).

Структура и движение (развитие) материальных систем в пространстве и во времени определяется их составом и типом взаимодействия их элементов.

Энергетическое взаимодействие основано на четырех видах элементарного (непосредственного) физического взаимодействия (слабого, сильного, гравитационного и электромагнитного) и присуще всем материальным системам. Интенсивность (сила) этого взаимодействия характеризуется энергией. Как видим, говорить о том, что Вселенная «пронизана потоками энергии», по крайней мере некорректно, так как энергия — не сущность энергетического взаимодействия, а его свойство.

Информационное взаимодействие (первичное понятие, то есть категория) — специфическая форма взаимодействия биологических элементов (организмов) друг с другом и с окружающей средой, отличная от энергетического взаимодействия физических и биологических элементов различных зоологических систем. Оно характеризуется тем, что реакции биологических организмов на внешние физико-химические воздействия (стимулы, раздражители) опосредованы рецепторами («датчиками») и органами формирования, хранения и переработки информации. Эти реакции осуществляются не непосредственно вслед за воздействиями (стимулами), как при энергетических взаимодействиях, но в зависимости от внешней обстановки и внутреннего состояния, а также потребностей биоорганизма — путем подачи «органом управления» (например, центральной нервной системой) соответствующих сигналов на специализированные «исполнительные органы» биоорганизма.

Если физическое взаимодействие осуществляется за счет обмена энергией (потенциальной и кинетической), то информационное взаимодействие осуществляется за счет взаимного обмена информацией между биоорганизмами, а при ее формировании — за счет энергетического взаимодействия биоорганизма с физическими и биологическими объектами.

В биосфере Земли имеется пять основных уровней информационного взаимодействия биоорганизмов (особей) с окружающей средой и друг с другом: элементарная раздражимость; безусловно-рефлекторное взаимодействие животных организмов (в том числе человеческих особей) с окружающим миром; условно-рефлекторное взаимодействие; сигнальное взаимодействие в сообществах высших животных и человека; знаковое (семиотическое) взаимодействие, присущее только человеку.

Сигнальное взаимодействие осуществляется в таксонах: стаях, стадах, прайдах и т. п. У каждого вида высших животных имеется совокупность звуковых сигналов, поз, жестов и мимики (этограмма), посредством которых в таксоне осуществляется сигнальное взаимодействие.

Знаковое взаимодействие — гораздо сложнее и качественно иное. Оно характерно только для человеческого сообщества.

Сигналы — физико-химические процессы или явления, воздействующие на рецепторы (датчики информации), протекающие в центральной нервной системе или подающиеся на исполнительные органы высших животных и являющиеся конкретным воплощением сообщения о каком-либо событии, явлении, состоянии объекта внешнего или внутреннего мира, либо физико-химические процессы выдачи команд управления, оповещения и т. п. Сигналы являются динамическими материальными носителями информации и представляют собой процессы соответствующего изменения в пространстве и (или) во времени некоторой физико-химической (измеримой) величины.

В подсистемах формирования, сбора и хранения информации сигналы кодируются состоянием элементов некоторого статического или динамического материального носителя информации («запоминающего устройства») — с последующим считыванием (извлечением, предоставлением) информации.

Знак — материальный чувственно воспринимаемый предмет, явление, событие или действие, выступающее при информационном взаимодействии людей в качестве элементарного представителя некоторого сложного объекта или явления. Смысл (семантика) знака (его отношение к обозначаемому предмету или явлению) должен быть понятным и правильно воспринимаемым всеми субъектами информационного взаимодействия.

Символ — некоторый материальный предмет, явление, событие или действие, служащее для условного обозначения (по взаимной договоренности между субъектами информационного взаимодействия) какого-либо объекта материального или духовного мира. Обычно символы — это вторичные знаки; они характерны только для человеческого сообщества.

Знаковая система — совокупность четко различаемых знаков (например, алфавит), обладающая определенной структурой (синтаксисом) и предназначенная для определенного вида информационного взаимодействия в человеческих сообществах. Знаковая система есть материальный посредник для информационного взаимодействия.

Высшей формой знаковой системы является человеческий язык.

Различные знаковые системы являются предметом изучения науки семиотики, состоящей из синтактики («структуроведения»), семантики («смысловедения») и прагматики («ценностноведения»).

Помеха — нежелательное изменение сигнала или знака.

Шум — случайная помеха в динамических технических информационных системах.

Неявное знание (tacit knowing) — индивидуальный опыт у высших животных и человека. Это знание, которое не может быть полностью формализовано и передается через непосредственное обучение (показ и подражание). Неявное знание относится к психике, а не к сознанию.

Знание (knowledge) — отражение в индивидуальном сознании человека сведений об окружающем мире в форме, в которой эти сведения могут быть переданы другим индивидам. Материальной формой существования знаний является язык (слово). Посредством языка происходит переход от восприятия (образа) к понятиям, включение понятий в систему знаний, передача своих знаний другим людям и т. д.

Мышление — процесс (целенаправленного или «свободного») оперирования понятиями, в результате которого формируется элемент системы знаний — мысль, которая либо включается в систему знаний, либо материализуется в виде сообщения. Мышление и язык — неразрывны, как форма и содержание.

Мысль — это суждение, являющееся идеальным отражением какого-либо материального или идеального явления окружающего субъект мира и выраженное в словесной форме данного языка как знаковой системы. Отношение данной мысли к отражаемому явлению есть содержание этого явления, или его смысл (со-мысль).

Значение — есть то, что данный знак (слово, символ, сигнал) обозначает, то есть к чему и каким образом он относится к обозначаемому предмету. Это характерно как для человека, так и для высших животных.

Смысл (со-мысль) — внутреннее (сущностное, глубинное) содержание знания или сообщения, понятное другим субъектам знакового информационного взаимодействия.

Сознание (со-знание: совместное знание группы людей) — способность человека идеального (психического) отражения действительности. Содержание сознания — субъективный образ объективного и субъективного мира. Сознание выступает в двух взаимосвязанных формах: индивидуальной (личностной) и общественной.

Индивидуальное сознание как процесс накопления личного опыта есть познание. Результат этого процесса — индивидуальное знание (явное и неявное).

Общественное сознание — это совокупность науки, философии, искусства, идеологии, морали, религии, предрассудков и т. п. данного социума.

Сообщение (со-общение) — материальная форма представления информации для ее хранения, передачи, распределения, обработки, преобразования или непосредственного использования (письмо, текст, устная речь, изображение, цифровые данные, электрические сигналы и т. п.), а информация есть идеальное содержание сообщений любой при-

роды. И так же, как количество энергии характеризует интенсивность взаимодействия физических систем, так и количество информации, содержащееся в данном сообщении, характеризует эффективность процесса информационного взаимодействия между собой субъектов общения.

Тезáурус (Большая советская энциклопедия, 1976, тт. 23 и 25: от греческого слова «сокровище») — «множество смысловыражающих единиц некоторого языка с заданной на нем системой семантических отношений. Тезаурус фактически определяет семантику языка».

«В широком смысле тезаурус интерпретируют как описание системы знаний о действительности, которыми располагает индивидуальный носитель информации или группа носителей. Этот носитель может выполнять функции приемника дополнительной информации, вследствие чего изменяется и его тезаурус. Исходный тезаурус определяет при этом возможности приемника при получении им семантической информации».

Тезаурус включает в себя явные знания и систему их взаимосвязей (субъективный мир), оценки их соотношений с образами внешнего мира (семантику), оценки степени достоверности этих знаний об объективном мире (истинности элементов тезауруса), оценки ценности этих знаний (прагматику), а также оценки степени новизны и возможности (вероятности) реализации смысловыражающих единиц тезауруса при приеме субъектом некоторого сообщения.

«Семантическую информацию сообщений любой природы можно оценивать как степень изменения системы знаний (тезауруса) адресата в результате восприятия данного сообщения».

Из последнего заключения следует очень важный, для общей (семантической) теории информации, вывод: количество семантической информации, содержащейся в данном сообщении, измеряется степенью изменения тезауруса данного получателя информации в совокупности с изменением системы оценок новизны или вероятности реализации смысловыражающих единиц тезауруса.

Только усвоив сформулированные выше основные понятия общей (семантической) теории информации, можно достаточно глубоко (для практических приложений) понять аксиоматическое построение математической теории информации.

Основные положения математической теории информации

Из основных понятий общей (семантической) теории информации следует, что в человеческом сообществе имеется несколько уровней информационного взаимодействия человеческих особей друг с другом и с окружающей средой. На высшем уровне информационного взаимодействия — знаковом — основным видом информационного взаимодействия субъектов между собой является язык.

Язык выступает в двух основных формах: речь и письмо. Если речевое общение (словесное информационное взаимодействие субъектов) происходит на больших расстояниях, то основной задачей проектирования технического средства речевого общения (системы телефонной или радиосвязи) является обеспечение на выходе технической информационной подсистемы наименее искаженного воспроизведения звуковых колебаний воздуха, которые несут речевое сообщение источника сообщений. Эта информационная эрготехническая («человеко-машинная») система (ЭТИС) является системой с динамическим каналом передачи сообщений (каналом связи). Качество воспроизведения речи количественно измеряется ее разборчивостью и узнаваемостью голоса источника сообщений (верностью). Это является задачей оптимальной фильтрации речевых сообщений на фоне шумов, которая относится к статистической теории радиосистем [10].

Письменные сообщения состоят из отдельных знаков: букв, соответствующих (но не всегда!) отдельным фонемам речевых сообщений, пробелов между буквами и словами, цифр, знаков препинания и т. п. Эти знаки фиксируются в определенной физической форме на некотором носителе (например, на телетайпной перфоленте или в ОЗУ), а носители знаковых сообщений передаются (в пространстве и/или во времени) получателю информации. Такими статическими носителями информации могут являться: письма, книги, магнитофонные ленты, дискеты, перфоленты, компакт-диски, «флэшки», каменные плиты, берестяные, кожаные (пергаментные) или металлические пластинки и т. п. Здесь главным свойством носителей информации является обеспечение длительной сохранности знаковых сообщений в условиях воздействия на них различных физико-химических факторов среды хранения и передачи — помех.

Каким же образом можно объективно оценить количество передаваемой информации (чтобы рассчитать необходимую пропускную способность канала передачи и/или информационную емкость носителей информации) и процент ее потерь при разрушающем воздействии на носители информации различного рода помех? Это — одна из фундаментальных проблем математической теории информации.

Вспомним, что информация, материализованная в какой-либо из знаковых систем, имеет три главные стороны, или уровня (им соответствуют три раздела теории знаковых систем — семиотики [6–8]). Она имеет определенную внутреннюю структуру (совокупность знаков и их взаимосвязь — синтактический, или формальный уровень семиотики), определенную систему связей (через первую или вторую сигнальные системы человека) с материальными или идеальными объектами, отражаемыми в сознании человека (семантический уровень — объективная сторона информации), а также систему связей с «внутренним миром» индивида и его потребностями (систему ценностей информации: прагматический уровень — субъективная сторона информации).

Мы уже выяснили, что количество семантической информации, получаемое субъектом-получателем информации (субъектом-ПИ) при приеме им данного знакового сообщения, измеряется степенью изменения тезауруса получателя информации.

Пользуясь системой ЭТИС (например, телефоном), субъекты информационного взаимодействия «интуитивно» применяют ряд способов сохранения качества функционирования системы ЭТИС при различной помеховой обстановке в канале связи системы: снижение скорости передачи сообщений (сужение необходимой полосы пропускания слухового аппарата субъекта-ПИ, что повышает отношение «сигнал/шум»), повышение мощности передаваемых сигналов (переход на крик), дублирование (повторение) сообщений, избыточное (канальное, помехоустойчивое) кодирование элементарных сообщений (фонем) при передаче «по буквам» и ряд других. При этом субъекты общения «машинально» оценивают качество канала связи исходя из допустимой степени потерь информации — и дают себе или собеседнику команды на применение тех или иных способов повышения качества передачи информации за счет психофизиологических средств субъектов общения. Так или иначе, но общающиеся субъекты добиваются, по возможности, почти безошибочной передачи информации даже при очень большом уровне помех в канале связи системы ЭТИС.

В современных системах ЭТИС типа цифровых сотовых радиотелефонных систем подвижной электросвязи все эти способы обеспечения качества передачи информации (и некоторые другие) используются и осуществляются автоматически — техническими средствами самой системы ЭТИС. Эти способы можно подразделить на энергетические (повышение мощности передатчика, чувствительности приемника, сужение полосы частот и т. п.) и информационные (дублирование сообщений, перезапрос, помехоустойчивое кодирование и т. п.). Энергетические способы обеспечения заданного качества передачи информации с помощью систем ЭТИС входят в компетенцию статистической теории радиосистем [10].

Для построения информационной теории систем электросвязи как приложения математической теории информации необходимо определить меру количества информации в системах ЭТИС не на семантическом,

а на синтактическом уровне семиотики. Для этого нужно, исходя из семантического определения количества информации как степени изменения тезауруса получателя информации, сформулировать постулаты, которые одинаково соответствовали бы как семантическому уровню семиотики, так и синтактическому.

Если в качестве сообщения используется текст, то он состоит из ряда предложений, которые представляют собой некоторые последовательности $S_i^{(n)} = (u_{i1}, u_{i2}, \ldots, u_{ik}, \ldots, u_{in})$ знаков (первичных символов) из множества $\{u_j\}_1^N$ — алфавита знаковой системы U. Здесь n — длина (количество знаков) i-го сообщения $S_i^{(n)}$.

При этом на k-м месте последовательности $S_i^{(n)}$ знак u_j может появиться с вероятностью $P(u_j) = P_j$ независимо от того, какие знаки предшествовали ему — в простейшем случае, либо (как в тексте реального языка) — в зависимости от реализации в последовательности $S_i^{(n)}$ ряда знаков из множества $\{u_j\}_1^N$, стоящих до появления очередного знака u_{ik}. В последнем случае следует применять вероятностные модели типа цепей Маркова. В первом (простейшем) случае ситуация определяется только совокупностью $\mathbf{P} = \{P_j\}_1^N$ априорных вероятностей появления знаков $\{u_j\}_1^N$ в любой из последовательностей $S_i^{(n)}$; $i, n = 1, 2, \ldots$

Можно утверждать, что количество информации, содержащейся в некотором знаковом сообщении, непосредственно связано с априорной вероятностью реального события, о котором информирует субъекта-ПИ данное сообщение.

Если субъект-ПИ получил два никак не связанных между собой сообщения, то количество информации, содержащейся в этих двух сообщениях вместе, должно равняться сумме количеств информации, содержащейся в каждом из них в отдельности: свойство аддитивности количества информации.

Кроме того, количество семантической информации — величина положительная (неотрицательная), так как бессмысленное предложение или предложение на непонятном субъекту-ПИ языке воспринимается им только как факт наличия языкового сообщения, недоступного для его понимания. А дезинформация в тезаурусе субъекта-ПИ содержит позитивную информацию, если он не знаком с происхождением этого сообщения.

Следовательно, на семантическом уровне семиотики количество информации, содержащейся в принятом субъектом-ПИ сообщении, определяется «априорной вероятностью» P_r, с которой оценивает субъект-ПИ в своем тезаурусе данное сообщение из бесконечного множества ему подобных. Это количество информации — величина положительная и может изменяться от нуля (при $P_r = 1$) до произвольно большого значения.

Если же субъект-ПИ получил последовательность независимых (никак не связанных между собой) сообщений, то общее количество полученной им информации равно сумме количеств информации, содержащейся в каждом отдельном сообщении.

Для математического анализа технических информационных подсистем систем ЭТИС мы можем формально (аксиоматически) перенести количественные свойства информации, выявленные на семантическом уровне семиотики, на синтактический уровень, ибо техническая подсистема информационной эрготехнической системы не может оперировать со смыслами передаваемых сообщений и объективно оценивать субъективные ожидания субъекта-ПИ.

Для технических информационных систем — как подсистем ЭТИС — в качестве элементарных событий должны выступать элементарные сообщения u_j ($j = 1, 2, \ldots, N$), выдаваемые источником дискретных (знаковых) сообщений — ДИС, а в качестве их априорных вероятностей P_j — *частотности* этих сообщений в длинных последовательностях сообщений $S_i^{(n)} = (u_{i1}, u_{i2}, \ldots, u_{ik}, \ldots, u_{in})$, то есть при $n \to \infty$.

Лучше всего элементарным сообщением считать отдельный языковый знак: букву, цифру, знак препинания, пробел и т. д. Всего таких «возможных элементарных сообщений» будет менее сотни. Эти языковые элементы назовем элементарными сообщениями (синтактического уровня), или знаками, или же первичными символами.

Будем считать, что поток элементарных сообщений (сложное сообщение — текст $S_i^{(n)}$), поступающий в статическую подсистему связи ЭТИС, представляет собой последовательность знаков (первичных символов). Частотность появления этих знаков в текстах будет различной. Поэтому информационная модель источника ДИС должна состоять, во-первых, из множества знаков (алфавита) $\mathbf{U} = \{u_j\}_1^N$ и, во-вторых, из соответствующего ему множества $\mathbf{P} = \{P_j\}_1^N$ априорных вероятностей появления знаков в различных текстах. При этом вероятности P_j можно оценить статистически для данного типа длинных текстов:

$$P_j = \lim_{n\to\infty}(n_{ij}/n),$$

где n_{ij} — количество знаков u_j, содержащихся в любом i-м сообщении $S_i^{(n)}$.

Отметим, что множество $\mathbf{U}$ — не упорядочено (бессмысленно говорить, что буква «а» больше или меньше буквы «я»). Введение на множестве $\mathbf{U}$ информационной меры $\mathbf{I} = \{I_j\}_1^N$ делает его (линейно) упорядоченным, а значит — измеримым. То есть про любые два различных элемента $u_j \in \mathbf{U}$ и $u_l \in \mathbf{U}$ можно будет сказать, что один из них (скажем, u_j) содержит не меньше информации, чем другой (например, u_l): $I(u_j) \geq I(u_l)$, или $I_j \geq I_l$. При этом элемент u_j содержит больше информации, чем u_l, на $[I(u_j)-I(u_l)]$ единиц информации.

В простейшем (идеализированном) случае будем считать, что появление данного знака $u_j \in \mathbf{U}$ в данном месте текста $S_i^{(n)}$ не зависит от реализации предыдущей части текста. Тогда простейшая знаковая (семиотическая, «дискретная») информационная модель источника ДИС $\{\mathbf{U}, \mathbf{P}, \mathbf{I}, \mathbf{S}\}$ будет состоять из:

- конечного множества (алфавита) элементарных сообщений (знаков, первичных символов) $\mathbf{U} = \{u_1, u_2, \ldots, u_j, \ldots, u_N\}$;
- множества $\mathbf{P} = \{P_j\}_1^N$ априорных вероятностей (частотностей) их появления в разных местах различных текстов;
- множества $\mathbf{S}$ последовательностей $S_i^{(n)}$ ($i = 1, 2, \ldots, N^n$) элементарных сообщений ($S_i^{(n)} \in \mathbf{S}$) вида $S_i^{(n)} = (u_{i1}, u_{i2}, \ldots, u_{ik}, \ldots, u_{in})$, то есть множества всех возможных сообщений длины $n = 1, 2, \ldots$;
- информационного множества $\mathbf{I} = \{I_1, I_2, \ldots, I_j, \ldots, I_N\}$, где I_j — количество информации, заключающейся в элементарном сообщении (знаке) $u_j \in \mathbf{U}$,
- а также содержать правила вычисления количества информации $I(S_i^{(n)})$, заключающейся в любом из сообщений $S_i^{(n)} \in \mathbf{S}$.

Определим величину количества синтактической информации $I(u)$ для элементарного «дискретного» сообщения (знака) $u \in \mathbf{U}$, имеющего априорную вероятность, или частотность P, исходя из следующих постулатов.

1. Величина $I(u)$ является некоторой функцией $I(u) = f(P)$ от априорной вероятности P элементарного сообщения $u \in \mathbf{U}$. Эта функция должна обладать следующими необходимыми свойствами.
2. Функция $f(P)$ неотрицательна, то есть количество информации, содержащейся в любом элементарном сообщении $u \in \mathbf{U}$, больше или равно нулю; $f(P) = 0$ только при $P = 1$; при $P = 0$ значение функции $f(P)$ не определено, то есть $f(0) = +\infty$.
3. Если имеется совокупность из двух элементарных сообщений, то есть $u_j \in \mathbf{U}$ и $u_l \in \mathbf{U}$, с априорными вероятностями $P_j = P(u_j)$ и $P_l = P(u_l)$, и если $P_l > P_j$, то $f(P_l) < f(P_j)$, то есть функция $f(P)$ — строго убывающая.

   Таких функций $f(P)$ существует бесконечное множество. Однако, исходя из семантического принципа аддитивности информации, для однозначного определения вида функции $f(P)$ достаточно ввести еще один (четвертый) постулат.
4. Должны выполняться следующие равенства:

$$f(S_i^{(2)}) = f[P(u_j, u_l)] = f(P_j \times P_l) = \\ = f(P_j) + f(P_l),$$

   то есть количество информации, содержащейся в любой i-й последовательности $S_i^{(2)} = (u_{ij}, u_{il})$ из двух элементарных *независимых* сообщений $u_j \in \mathbf{U}$ и $u_l \in \mathbf{U}$, равно сумме количеств информации, содержащейся в каждом из них в отдельности (аддитивность информационной меры).

Эти четыре постулата позволяют однозначно определить меру синтактической информации, содержащейся в элементарном сообщении $u \in \mathbf{U}$. К. Шеннон показал, а А. Я. Хинчин (1894–1959) математически строго доказал, что единственной функцией, соответствующей всем четырем постулатам (1–4), является логарифмическая функция: количество синтактической информации $I(u)$, содержащейся в некотором знаке $u \in \mathbf{U}$, который имеет априорную вероятность появления в любом месте i-й последовательности $S_i^{(n)} = (u_{i1}, u_{i2}, \ldots, u_{ik}, \ldots, u_{in})$ независимых элементарных сообщений (знаков) длины n, определяется формулой:

$$I(u) = -K\log_a P(u),$$

где K — некоторая положительная константа.

Поскольку величина P оценивается статистически, то мера $I(u)$ называется статистической информационной мерой Шеннона.

С помощью построенной информационной модели $\{\mathbf{U}, \mathbf{P}, \mathbf{I}, \mathbf{S}\}$ источника знаковых сообщений можно проводить различные информационные расчеты, связанные с этим источником. Остается договориться о единицах измерения количества синтактической информации $I_j = -K\log_a P_j$ в знаковой системе $\mathbf{U} = \{u_j\}_1^N$.

Для простоты вычислений, как это принято в физических дисциплинах, лучше всего положить $K = 1$. Если в качестве основания a логарифмов выбрать число «2», то такие единицы называются битами (bit = BInary uniT — двоичная единица, или BInary digiT — двоичная цифра); если трансцендентное число $e \approx 2{,}7183$ — то натами (натуральные единицы); если $a = 10$ — то дитами (десятичными единицами, или хартами — в честь Р. Хартли (1888–1970), который одним из первых пытался обосновать логарифмическую меру информации).

За последние полвека наибольшее распространение в качестве единиц измерения количества информации получили биты. «Эталоном» процесса выработки сообщений, содержащих количество информации в 1 бит, является бросание идеальной монеты, поскольку в сообщении u, имеющем априорную вероятность появления в тексте $P(u) = 0{,}5$, содержится $I_0(1/2) = -\log_2(1/2) = 1$ бит.

Пятый постулат теории информации связан с величиной потери информации при условии воздействия на ее носитель различного рода помех и будет рассмотрен во второй части статьи.

Заключение

В этой части статьи была предпринята попытка сформулировать семантические и аксиоматические основания теории информации, пригодные для широкого применения в информационных технологиях.

В следующей части будут рассмотрены информационные характеристики источников сообщений и каналов связи, а также численный метод расчета информационных характеристик различных каналов электросвязи. ■

Литература

1. Адаменко А. Н., Ашеров А. Т., Бердников И. Л. и др. Информационно-управляющие человеко-машинные системы: исследование, проектирование, испытания. Справочник. М.: Машиностроение, 1993.
2. Галлагер Р. Теория информации и надежная связь. М.: Советское радио, 1974.
3. Колесник В. Д., Полтырев Г. Ш. Курс теории информации. М.: Наука, 1982.
4. Колмогоров А. Н. Теория передачи информации. М.: Изд-во АН СССР, 1956.
5. Пирс Дж. Символы, сигналы, шумы. Закономерности и процессы передачи информации. М.: Мир, 1967. (Оригинал: 1962 г.)
6. Пирс Ч. Логические основания теории знаков. СПб.: СПб ГУ, Алетейя, 2000.
7. Семиотика: Антология/Сост. Ю. С. Степанов. М.: Академический проект; Екатеринбург: Деловая книга, 2001.
8. Степанов Ю. С. Семиотика / Ю. С. Степанов. М.: Наука, 1971.
9. Теория передачи информации. Терминология. М.: Наука, 1984.
10. Фано Р. Передача информации. Статистическая теория связи / Р. Фано. — М.: Изд-во Мир, 1965.
11. Хэмминг Р. В. Теория кодирования и теория информации / Р. В. Хэмминг. М.: Радио и связь, 1983.
12. Шеннон К. Работы по теории информации и кибернетике. Пер. с англ. под ред. Р. Л. Добрушина и О. Б. Лупанова / К. Шеннон. М.: Изд-во ИЛ, 1963.
13. www.ru.wikipedia.org; www.en.wikipedia.org