Методическая разработка

для проведения лекции

Занятие 14. Информация и ее свойства

Учебные вопросы занятия:
1. Информация и избыточность
2. Передача сообщений в каналах связи

Заключительная часть

1. Информация и избыточность

1.1 Понятие информации, свойства информации

Современный этап научно-технического прогресса заключается в глобальной информатизации общества, базирующейся на передовых компьютерных технологиях и телекоммуникационных системах, объединяющих в единое целое совокупность разнородных информационно-вычислительных ресурсов и отдельных потребителей информации. Это объясняется постоянно возрастающими потребностями людей в обмене различного рода сведениями, а также многообразием технических и программных средств их формирования. Информационные процессы отличаются активностью, что вызывает необходимость их исследования и описания в базисе основных понятий в этой области.

В теории информации непосредственно с термином «информация» ассоциируются сведения, являющиеся объектом некоторых операций: передачи, распределения, преобразования, хранения или использования [14].

Информацию следует считать особым видом ресурса, представленного в виде суммы неких знаний относительно материальных предметов либо энергетических, структурных или других характеристик предмета. В отличие от материальных ресурсов информационные ресурсы являются неистощимыми и предполагают принципиально иные методы воспроизведения и обновления, чем материальные ресурсы.

В рамках указанного подхода возникает необходимость акцентировать внимание на следующих свойствах информации:

– запоминаемость – возможность хранения информации;

– передаваемость – способность информации к копированию;

– воспроизводимость – тождественность информации самой себе при копировании;

– преобразуемость – способность к увеличению или уменьшению объема информации в ходе информационных процессов;

– стираемость – способность преобразования информации таким образом, чтобы ее количество становилось равным нулю;

– объективность (субъективность) – свойство, отражающее наличие либо отсутствие зависимости информации от чьего-либо мнения, суждения;

– достоверность (адекватность) – соответствие реальному объекту или явлению;

– полнота – достаточность данных для обоснованного принятия решения;

– доступность – возможность получения информации по конкретной тематике;

– актуальность – степень соответствия информации текущему моменту времени.

В полном объеме указанные свойства проявляются в процессе информационного взаимодействия, состоящем в передаче информации, представленной в определенной форме. Процессы такого вида предполагают наличие источника информации (передатчика) S и ее получателя (интерпретанта) Pr. Описание состояний источника S и получателя информации Pr осуществляется с помощью соответствующих им множеств понятий, мощности которых определяются числом возможных состояний, свойств и целей объектов информационного взаимодействия.

1.2 Сообщение как форма представления информации, виды и модели сообщений

Под сообщением длиной μ следует понимать систему $\langle s_i \rangle_\mu$, $i = 1(1)N$ символов (знаков) источника S, определенных на множестве (алфавите) объемом N, находящихся в определенных отношениях и связях друг с другом и образующих определенную целостность. Источник $S_{\{N\}}$ использует алфавит объемом N.

Таким образом, сообщение в общем смысле есть форма представления информации в виде последовательности длиной μ взаимосвязанных символов $s_i \in S_{\{N\}}$, удобная для передачи на расстояние. Наиболее распространенные виды сообщений представлены на рисунке 1.

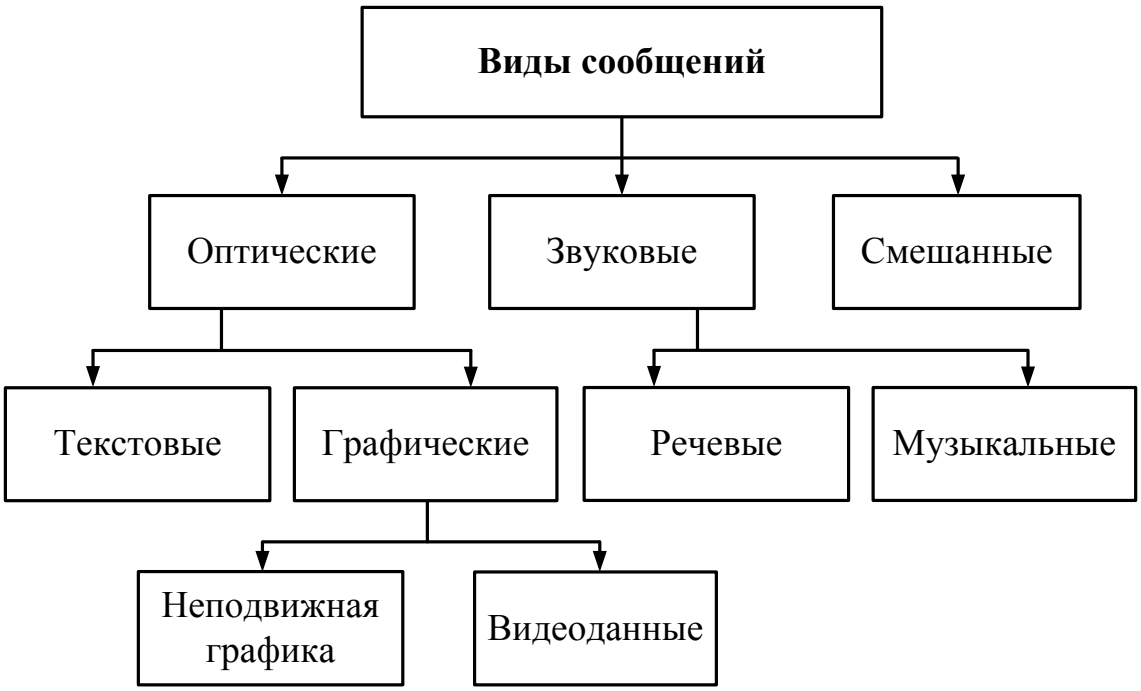

Рисунок 1 – Наиболее распространенные виды сообщений

Модели сообщений различных видов рассматриваются на трех основных уровнях: синтаксическом, семантическом и прагматическом.

Модели синтаксического уровня описывают комбинаторику символов (знаков), не принимая во внимание значение и ту ценность, которую представляют эти символы и их сочетания как для лица или устройства, передающего информацию (передатчика), так и для потребителя информации (интерпретанта). Формирование модели на этом уровне подразумевает анализ внутренних статистических свойств сообщений без учета их местоположения, иерархии и взаимосвязей в структуре генерирующего источника. Непосредственно в качестве предмета анализа на этом уровне выступают отношения между символами s_i, $i = 1(1)N$ системы $\langle s_i \rangle_\mu$.

Проблемы обработки на синтаксическом уровне в целом ориентированы на создание теоретических основ построения систем приема, основные показатели функционирования которых были бы близки к предельно возможным, а также на совершенствование существующих систем с целью повышения эффективности их применения. Это чисто технические проблемы совершенствования методов обработки сообщений и их материального воплощения – сигналов.

Иначе говоря, на этом уровне особо актуальны проблемы обработки сообщений на приемной стороне как совокупности взаимосвязанных знаков при полной абстракции от их смыслового и прагматического содержания. Результаты решения ряда проблем именно такого уровня составляют основу синтаксической теории информации. Она опирается на понятие «количество

информации», которое является мерой частоты употребления знаков и их комбинаций.

Модель семантического уровня учитывает отношения между символами и их содержанием (в терминах семиотики – между означающим и означаемым), игнорируя при этом состояния как источника, так и приемника (интерпретанта) этих символов. Предметом исследования и описания на этом уровне являются отношения между сочетаниями элементов принятого сообщения $\langle s_i \rangle_\mu$ и понятиями, которые образуются в процессе обратной интерпретации сообщения $\langle s_i \rangle_\mu$. Ряд предложений по решению задач описания отношений такого рода представлены в работах Бар-Хиллела. Выдвигаемые в них Бар-Хиллелом идеи измерения семантической информации при описании текстовых сообщений опираются на предложенную Карнапом теорию логических вероятностей. В качестве семантического аналога информации здесь выступает некоторая функция числа всех возможных условий, при которых сообщение было бы истинным. Для реализации этой идеи авторы вводят понятие «возможные условия» с помощью предложенного Виттгенштейном понятия «описание». При этом число элементов множества описаний зависит от того языка, с помощью которого осуществляется это описание.

В теории Карнапа-Бар-Хиллела рассматривается и описывается семантико-информационное содержание лишь простых суждений (декларативных предложений) вне всякой связи с прагматическими вопросами полезности, ценности и эвристичности информации для ее потребителя. Иными словами, семантические вопросы связи (для речевой коммуникации это вопросы лингвистической истинности) остаются за пределами семантической теории Карнапа-Бар-Хиллела. Если учесть также, что практически невозможно определить весь ряд альтернативных логических описаний и распределить на него меру вероятности, становится ясным, что перспективы применения рассмотренной семантической теории для исследования значений в сообщениях семантических форматов данных недостаточно определенны. На современном уровне при семантическом подходе информация рассматривается как с точки зрения формы, так и содержания. При этом информацию связывают с тезаурусом, т.е. полнотой систематизированного набора данных о предмете информации, что не исключает количественного анализа.

При моделировании свойств сообщения на *прагматическом уровне* анализируется его смысловое содержание, отношение к источнику информации. При этом в качестве предмета анализа рассматриваются отношения между понятиями внутри некоторой системы, либо между понятиями, принадлежащими различным системам.

Одна из этих систем образуется в процессе обратной интерпретации принятого сообщения $\langle s_i \rangle_\mu$, вторая система отражает знания интерпретанта

сообщения. В этом случае прагматика исследует символы сообщения $\langle s_i \rangle_\mu$ с точки зрения их ценности для интерпретанта, а иногда и для источника информации. Именно на этом уровне осуществляется во всей полноте реальный процесс обмена информацией между людьми, а символы сообщений рассматриваются в определенных обстоятельствах при определенном окружении с привлечением проблем их ценности, полезности, узнавания и интерпретации. Проблемы обработки сообщений на прагматическом уровне связаны с формализацией смысла передаваемой информации, например, введением количественных оценок близости информации к истине, т.е. оценок ее качества. Эти проблемы чрезвычайно сложны, так как смысловое содержание информации больше зависит от получателя, чем от сообщения, представляемого в каком-либо виде.

Информация заложена в сообщении, но проявляется только при взаимодействии с получателем, так как может быть зашифрована. Кроме того, если получатель – человек, то и незашифрованное сообщение может быть понято по-разному. Объясняется это тем, что различное понимание того или иного фрагмента может сильно изменить смысл сообщения. Восприятие человеком информации зависит также и от его эмоционального состояния, жизненного опыта и других факторов. Рассмотрение задач такого рода в процессе несанкционированного анализа осложняется возможностью наличия дезинформации относительно состояний и свойств взаимодействующих объектов.

При обработке сообщений получателем информации их свойства на прагматическом уровне проявляются во взаимосвязях обрабатываемого сообщения с ранее сформированной базой знаний конкретного потребителя информации. В рамках обобщенной модели сообщения они формируют дополнительные ограничения на множество совокупностей символов, разрешенных с точки зрения свойств предыдущих уровней.

Следует отметить, что модели сообщений на прагматическом уровне предполагают наличие вероятностных оценок тех ситуаций, в которых может оказаться получатель информации, принявший и правильно понявший сообщение (для определения ценности сообщения с точки зрения источника информации необходимо соответственно оценить ситуации, в которых он окажется после передачи сообщения). Сложность этой задачи очевидна, и поэтому не случайно, что на современном этапе развития техники передачи информации приемлемая работающая процедура, с помощью которой можно было бы получать количественные оценки параметров модели прагматического уровня, отсутствует.

В рамках обобщенной модели сообщений, учитывающей их свойства на всех рассмотренных уровнях, наиболее широкое использование получил статистический подход к описанию на основе измерения количества информации, развитый Шенноном. Несмотря на то, что информационные оценки Шеннона, ориентированные на синтаксический аспект, не могут быть использованы для непосредственного измерения семантики и прагматики

сообщений различных видов, следует иметь в виду, что синтактика, семантика и прагматика сообщений не просто находятся в тесном взаимодействии, но определенным образом коррелированны и взаимообусловлены. Каждое сообщение по теории Шеннона выбирается из некоторого множества возможных. Возможность же появления этих сообщений определяется, как правило, соотношением, существующим между способом общения и сообщением, с одной стороны, и объективной реальностью – с другой. Что же касается возможности перенесения результатов статических экспериментов на прагматику сообщений, то ее предпосылкой служит соотнесенность «языка-кода» определенного вида и порождаемых им сообщений с коллективным опытом носителей данного «языка» и коллективной оценкой ими пользы и эвристичности их содержания.

В силу указанных аспектов результаты статистических экспериментов могут быть положены в основу формирования модели синтаксического, семантического и прагматического уровней представления информации в сообщениях.

1.3 Количество информации, основные подходы к измерению количества информации, энтропия и избыточность

Рассуждая о количестве, содержании и ценности информации в сообщениях, следует исходить из возможностей соответствующего анализа знаковых структур. Изучение связей между символами сообщения может быть реализовано путем измерения количества информации, содержащейся в сообщениях. В решении проблемы измерения количества информации существуют два основных подхода, появившихся почти одновременно. Вероятностный подход основывается на понятии «энтропии» как меры неопределенности случайной ситуации, а работы по созданию ПЭВМ привели к «объемному» подходу.

В рамках *вероятностного подхода* в качестве основной характеристики сообщения выступает величина, называемая количеством информации. Это понятие не затрагивает смысла и важности передаваемого сообщения, а связано со степенью его неопределенности в пространстве возможных сообщений. Для источника $S_{\{N\}}$ такое пространство определяется объемом алфавита N и длиной сообщения μ, а количество информации определяется величиной

$$I(S_{\{N\}}) = \log N^{\mu} = \mu \log N . \qquad (1)$$

Указанная мера была предложена американским ученым Хартли в 1928 г. Шеннон обобщил понятие количественной меры информации на общий случай неравновероятных исходов.

В общем случае каждый из знаков s_i, $i = 1(1)N$ появляется в сообщении $\langle s_i \rangle_{\mu}$ с различной вероятностью.

Пусть на основании статистического анализа известно, что в сообщении длиной μ знак s_i появляется μ_i раз. Тогда вероятность появления знака определяется выражением

$$p(s_i) = \frac{\mu_i}{\mu},\ i = 1(1)N. \tag{2}$$

Все знаки алфавита определяют полную группу случайных событий

$$\sum_{i=1}^{N} p(s_i) = 1. \tag{3}$$

С учетом указанных статистических характеристик для описания источника сообщений в целом используется математическое ожидание количества информации, называемое энтропией и обозначаемое $H(S_{\{N\}})$,

$$H(S_{\{N\}}) = M[\log \frac{1}{p(s_i)}]. \tag{4}$$

Энтропия характеризует меру неопределенности сообщения.

Основные свойства энтропии состоят в следующем.

1. Энтропия неотрицательна. Она равна нулю только для «вырожденного» ансамбля, когда одно сообщение передается с вероятностью равной единице, а остальные имеют нулевую вероятность.

2. Энтропия аддитивна. Это означает, что, если рассматривать последовательность из *n* сообщений как одно «укрупненное» сообщение, то энтропия источника таких укрупненных сообщений будет в *n* раз больше исходного источника.

3. Для источника $S_{\{N\}}$ выполняется неравенство вида

$$H(S_{\{N\}}) \le \log N, \tag{5}$$

причем равенство имеет место только для источника независимых и равновероятных сообщений.

Переходя к вероятностям и логарифмам с произвольным основанием, могут быть получены формулы Шеннона для количества информации и энтропии:

$$I(S_{\{N\}}) = -\mu \sum_{i=1}^{N} p(s_i) \log\ p(s_i); \tag{6}$$

$$H(S_{\{N\}}) = -\sum_{i=1}^{N} p(s_i) \log\ p(s_i). \tag{7}$$

Для источника равновероятных сообщений при $p(s_i) = \frac{1}{N},\ i = 1(1)N$ формула Шеннона переходит в формулу Хартли.

В рамках проблемы распознавания образов распространение получила мера информации, введенная Кульбаком, выступающая в качестве меры

неопределенности распределения вероятностей $P_1(s)$ символов источника относительно распределения $P_2(s)$

$$H_K = \sum_s P_1(s) \log \frac{P_1(s)}{P_2(s)} . \quad (8)$$

К задачам экспериментальных исследований приближается мера информации Фишера, представляющая собой значение количества информации по Кульбаку в частном случае двух близких гипотез о значении параметра распределения, т.е. в случае, когда в определении количества информации по Кульбаку

$$P_1(s) = P(s, \theta),\ P_2(s) = P(s, \theta + \Delta\theta), \quad (9)$$

где θ – многомерный параметр,

$\theta + \Delta\theta$ – точка, соседняя к θ.

Энтропия ансамбля характеризует среднее количество полной информации, содержащейся в сообщении. Часто возникает задача определения количества информации, содержащейся в одном ансамбле относительно другого, например, в принятом сигнале относительно переданного сообщения. Для этого необходимо рассмотреть объединение двух зависимых дискретных множеств – $S_{\{N\}}$ и $Z_{\{N\}}$. Это объединение можно интерпретировать как пару множеств сообщений либо как множества сообщения и сигнала, с помощью которых сообщение передается.

Пусть $p(s_i, z_j)$ – совместная вероятность элементов $s_i, i = 1(1)N$ и $z_j,\ j = 1(1)N$. Тогда совместная энтропия множеств $S_{\{N\}}$ и $Z_{\{N\}}$ может быть представлена следующим выражением:

$$H(S_{\{N\}}, Z_{\{N\}}) = M[\log \frac{1}{p(s_i, z_j)}]. \quad (10)$$

Для условной энтропии имеет место аналогичное выражение

$$H(S_{\{N\}} \left| Z_{\{N\}}\right.) = M[\log \frac{1}{p(s_i \left| z_j\right.)}], \quad (11)$$

где $p(s_i \left| z_j\right.)$ – условная вероятность s_i, если имеет место z_j.

При этом математические ожидания рассчитываются по объединенному ансамблю $S_{\{N\}} Z_{\{N\}}$. Например, для источников без памяти справедливо следующее уравнение

$$H(S_{\{N\}} \left| Z_{\{N\}}\right.) = \sum_{i=1}^{N} \sum_{j=1}^{N} p(s_i, z_j) \log \frac{1}{p(s_i \left| z_j\right.)}. \quad (12)$$

Для условной энтропии справедливо неравенство

$$0 \le H(S_{\{N\}} \left| Z_{\{N\}}\right.) \le H(S_{\{N\}}). \quad (13)$$

При этом равенство вида

$$H(S_{\{N\}}\big|Z_{\{N\}}) = 0 \qquad (14)$$

имеет место в случае наличия строгой функциональной зависимости между элементами z_j и s_i. Другими словами, $Z_{\{N\}}$ содержит полную информацию о $S_{\{N\}}$.

В случае отсутствия зависимости между элементами множеств $S_{\{N\}}$ и $Z_{\{N\}}$ условные вероятности $p(s_i\big|z_j)$ определяются значениями $p(s_i)$ и имеет место другое равенство

$$H(S_{\{N\}}\big|Z_{\{N\}}) = H(S_{\{N\}}). \qquad (15)$$

В этом случае знание значений из множества $Z_{\{N\}}$ не уменьшает неопределенности относительно элементов множества $S_{\{N\}}$.

В общем случае условная энтропия $H(S_{\{N\}}\big|Z_{\{N\}})$ меньше безусловной $H(S_{\{N\}})$ и измеренные значения из $Z_{\{N\}}$ снижают в среднем первоначальную неопределенность относительно $S_{\{N\}}$. Таким образом, разность $H(S_{\{N\}}) - H(S_{\{N\}}\big|Z_{\{N\}})$ представляет собой количество информации, содержащейся в $Z_{\{N\}}$ относительно $S_{\{N\}}$. Указанная разность называется взаимной информацией между $S_{\{N\}}$ и $Z_{\{N\}}$

$$I(S_{\{N\}}, Z_{\{N\}}) = H(S_{\{N\}}) - H(S_{\{N\}}\big|Z_{\{N\}}). \qquad (16)$$

Взаимная информация измеряется в тех же единицах, что и энтропия. Основные свойства взаимной информации состоят в следующем.

1. $I(S_{\{N\}}, Z_{\{N\}}) \geq 0$,

причем равенство имеет место только в том случае, когда элементы множеств $S_{\{N\}}$ и $Z_{\{N\}}$ не зависят друг от друга.

2. $I(S_{\{N\}}, Z_{\{N\}}) = I(Z_{\{N\}}, S_{\{N\}})$,

т.е. $Z_{\{N\}}$ содержит столько же информации относительно $S_{\{N\}}$, сколько $S_{\{N\}}$ содержит относительно $Z_{\{N\}}$.

3. $I(S_{\{N\}}, Z_{\{N\}}) \leq H(S_{\{N\}})$,

причем в том случае, когда по измеренным значениям z_j можно установить значения s_i, имеет место строгое равенство. По аналогии можно записать

$$I(Z_{\{N\}}, S_{\{N\}}) \leq H(Z_{\{N\}}).$$

4. $I(S_{\{N\}}, S_{\{N\}}) = H(S_{\{N\}})$.

Это позволяет интерпретировать энтропию источника как его собственную информацию, т.е. информацию, содержащуюся во множестве $S_{\{N\}}$ о самом себе.

Объемный подход

В 1965 году проведен анализ внутренней сложности описания строки двоичных символов. Если указанную строку рассматривать как последовательность независимых и одинаково распределенных случайных величин, то для ее описания в среднем необходимо число двоичных символов, равное энтропии последовательности.

Вместе с тем, указанная строка может быть сформирована в соответствии с некоторым фиксированным правилом или набором правил, что позволяет получить ее с помощью простой компьютерной программы.

В силу указанных причин Колмогоров определил сложность двоичной строки в виде длины кратчайшей программы для универсального компьютера, способной воспроизвести эту строку. При этом внутренняя сложность описания любого объекта не зависит от компьютера или лица, описывающего данный объект. Можно доказать, что сложность по Колмогорову с учетом ряда допущений аналогична энтропии по Шеннону. Другими словами, в среднем длина кратчайшей компьютерной программы, способной воспроизвести случайный объект, равна энтропии вероятностного распределения, из которого этот объект был извлечен. Сложность по Колмогорову предполагает наличие единого подхода к проблемам сжатия сообщений. Кроме того, она служит основой теории статистических выводов и тесно связана с теорией вычислимости.

Между вероятностным и объемным количеством информации соотношение неоднозначное. Не каждое текстовое сообщение, представленное в виде последовательности двоичных символов, предполагает измерение объема информации в вероятностном (кибернетическом) смысле, но заведомо допускает его измерение в объемном.

Кроме того, если некоторое сообщение допускает измеримость количества информации в обоих смыслах, то это количество не обязательно совпадает, причем кибернетическое количество информации не может быть больше объемного.

В прикладной информатике количество информации в большинстве случаев понимается в объемном смысле.

2. Передача сообщений в каналах связи

Процесс информационного взаимодействия различных объектов с помощью телекоммуникационных систем предусматривает применение последовательности преобразований формы представления информации для передачи в каналах связи. Одним из этапов этой последовательности является кодирование сообщений, цель которого состоит в согласовании

параметров сигналов источника сообщений и канала связи, а также в сжатии передаваемых сообщений. При этом различают кодирование источника (статистическое) и помехоустойчивое кодирование.

Кодирование источника позволяет повысить скорость передачи информации и приблизить ее к пропускной способности каналов передачи. Теоретической основой построения статистических кодов служит первая теорема кодирования Шеннона, состоящая в следующем.

Для канала без помех всегда можно создать систему статистического кодирования дискретных сообщений, у которой среднее число двоичных кодовых сигналов на один символ сообщения будет приближаться как угодно близко к энтропии источника сообщений.

Помехоустойчивое кодирование позволяет повысить достоверность передачи информации путем обнаружения и исправления ошибок. Теоретическую основу помехоустойчивого кодирования составляет вторая теорема кодирования Шеннона, которая применительно к дискретному источнику формулируется следующим образом.

Если производительность источника сообщений V меньше пропускной способности канала C, то существует такой способ кодирования и декодирования, при котором вероятность ошибочного декодирования и ненадежность могут быть сколь угодно малы. Если же указанное условие не выполняется, то таких способов не существует

$$V < C,\ C = \Delta F \log(1 + P_c / P_ш).$$

Теорема кодирования Шеннона справедлива и при передаче дискретных сообщений по непрерывным каналам. Указанная теорема определяет только возможность неискаженной передачи информации по каналу с помехами, не указывая, каким образом это можно сделать.

Заключительная часть.

Подвожу итоги занятия, анализирую степень достижения цели.

Рекомендованная литература:

1. Шеннон К. Работы по теории информации и кибернетике: пер. с англ. – М.: ИЛ, 1963.

2. Харкевич А.А. О ценности информации // ПК, 1960, вып. 4. – С.36–54.