Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования

«Амурский государственный университет»

С.Г. Самохвалова

ТЕОРИЯ ИНФОРМАЦИИ

Методические указания к практическим занятиям

для студентов очной формы обучения

Благовещенск

2017

ББК32.811
Т33

Теория информации. Методические указания к практическим занятиям для студентов очной формы обучения. / сост. С.Г. Самохвалова – Благовещенск.: ФГБОУ ВО «АмГУ», 2017 г. – 45 с.

Основу методических указаний составляют краткие теоретические сведения из теории информации, примеры, задания для самостоятельной работы, контрольные вопросы. Практические занятия призваны обеспечить закрепление полученных теоретических знаний по основам теории информации, выработать необходимые навыки вычисления количественных характеристик систем передачи информации, таких как, энтропия, скорость передачи информации, пропускная способность канала связи и т.д.

Методические указания рекомендуется студентам направления подготовки 09.03.01 – Информатика и вычислительная техника и 09.03.02 - Информационные системы и технологии, изучающим дисциплину «Теория информации», а также могут быть полезны для преподавателей и студентов, преподающих и осваивающих эту дисциплину в рамках других направлений подготовки, где она (и практические занятия по ней) предусмотрены учебными планами.

Рецензенты:

Чалкина Н.А. доцент, к.п.н. доцент кафедры общей математики и информатики ФГБОУ ВО АмГУ

СОДЕРЖАНИЕ

Введение

Теория информации является существенной, неотъемлемой частью кибернетики – науки, изучающей общие законы получения, передачи хранения информации.

Объектом исследования кибернетики являются все управляемые системы.

Примеры кибернетических систем: автоматические регуляторы в технике, ЭВМ, мозг человека или животных, биологическая популяция, социум. Часто кибернетику связывают с методами искусственного интеллекта, т.к. она разрабатывает принципы создания систем управления и систем автоматизации умственного труда.

Основными разделами кибернетики считаются: теория информации, теория алгоритмов, теория автоматов, исследование операций, теория оптимального управления и теория распознавания образов.

Родоначальниками кибернетики считаются американские ученые Норберт Винер и Клод Шеннон. Клод Шеннон – основоположник теории информации.

Норберт Винер (26 ноября 1894, Колумбия, штат Миссури, США — 18 марта 1964, Стокгольм, Швеция) — американский учѐный, выдающийся математик и философ, основоположник кибернетики и теории искусственного интеллекта.

В 4 года Винер уже был допущен к родительской библиотеке, а в 7 лет написал свой первый научный трактат по дарвинизму. Норберт никогда по-настоящему не учился в средней школе. Зато 11 лет от роду он поступил в престижный Тафт-колледж, который закончил с отличием уже через три года получив степень бакалавра искусств.

В 18 лет Норберт Винер получил степени доктора философии по математической логике в Корнельском и Гарвардском университетах. В девятна-

дцатилетнем возрасте доктор Винер был приглашѐн на кафедру математики Массачусетского технологического института.

Винер ввел основную категорию кибернетики – управление, показал существенные отличия этой категории от других, например, энергии, описал несколько задач, типичных для кибернетики и привлек всеобщее внимание к особой роли вычислительных машин, считая их индикатором наступления новой НТР. Выделение категории управления позволило Винеру воспользоваться понятием информации, положив в основу кибернетики изучение законов передачи и преобразования информации.

Клод Элвуд Шеннон родился 30 апреля 1916 г. в городе Петоцки, расположенном на берегу озера Мичиган штата Мичиган (США), в семье юриста и преподавателя иностранных языков. Шеннон закончил общеобразовательную среднюю школу в 1932 г. в возрасте шестнадцати лет, одновременно получив дополнительное образование на дому. В 1932 г. Шеннон поступил в Мичиганский университет, который окончил в 1936 г., получив степень бакалавра по двум специальностям: математика и электротехника.

В конце 1936 г. Шеннон поступает в магистратуру, а уже в 1937 г. он пишет реферат диссертации на соискание степени магистра. В 1948 г. он опубликовал свой эпохальный труд «Математическая теория связи» Последние годы жизни Клод Шеннон тяжело болел. Он скончался в феврале 2001 г. в массачусетском доме престарелых от болезни Альцгеймера на 85-м году жизни.

Клод Шеннон оставил богатое прикладное и философское наследие. Им создана общая теория устройств дискретной автоматики и вычислительной техники, технология эффективного использования возможностей канальной среды. Все современные архиваторы, используемые в компьютерном мире, опираются на теорему Шеннона об эффективном кодировании.

Теория информации представляет собой математическую теорию, посвященную измерению информации, ее потока, характеристик канала связи и

т.п., особенно применительно к радио, телевидению и другим средствам связи.

Основные понятия теории информации

Информация происходит от латинского слова informatio, что в переводе означает сведение, разъяснение, ознакомление.

В зависимости от области знания существуют различные подходы к определению понятия информации.

Под информацией понимают: в быту — это сведения об окружающем мире и протекающих в нем процессах, воспринимаемые человеком или специальными устройствами; в технике — это сообщения, передаваемые в форме знаков или сигналов.

В настоящее время нет общепринятого и однозначного понимания термина "информация". Понятие информации имеет много определений, от наиболее широкого (информация есть формализованное отражение реального мира) до практического (это – сведения и данные, являющиеся объектом хранения, передачи, преобразования, восприятия и управления).

Информация – это сведения, снимающие неопределенность об окружающем мире, которые являются объектом хранения, преобразования, передачи и использования.

Сведения – это знания, выраженные в сигналах, сообщениях, известиях, уведомлениях и т.д.

Сигнал – представляет собой любой процесс, несущий информацию.

Данные – это совокупность фактов, результатов наблюдений, измерений о каких-либо объектах, явлениях или процессах материального мира, представленных в формализованном виде: количественном или качественном.

Сообщение – это информация, представленная в определенной форме и предназначенная для передачи.

Различают две формы представления информации – непрерывную (аналоговую) и дискретную.

Дискретные сообщения формируются в результате последовательной выдачи источником сообщений отдельных элементов - знаков.

При этом все множество возможных различных знаков называют **алфавитом сообщения,** а размер множества - **объемом алфавита**.

Непрерывные сообщения в свою очередь не разделены на элементы, а описываются непрерывными сигналами - функциями времени, принимающими значения из непрерывного континуума.

Носителями информации являются сигналы, то в качестве сигналов могут использоваться физические процессы различной природы.

Сигнал называется непрерывным, если его параметр в заданных пределах может принимать любые промежуточные значения.

Сигнал называется дискретным, если его параметр в заданных пределах может принимать отдельные фиксированные значения

Источник информации – это субъект или объект, порождающий информацию и представляющий ее в виде сообщения.

Приемник информации – это субъект или объект, принимающий сообщение и способный правильно его интерпретировать.

Факт приема сообщения еще не означает получение информации. Информация может считаться полученной только в том случае, если приемнику известно правило интерпретации сообщения. Например, слыша речь на незнакомом языке, человек оказывается приемником сообщения, но не приемником информации.

Совокупность технических средств используемых для передачи сообщений от источника к приемнику информации называется **системой связи**. К ним относятся телеграф, телефон, радио и телевидение, компьютерные телекоммуникации и пр.

Качество информации является одним из важнейших параметров для потребителя информации. Оно определяется следующими свойствами:

репрезентативность – правильность отбора информации в целях адекватного отражения источника информации;

достаточность – минимальный, но достаточный состав данных для достижения целей, которые преследует потребитель информации, как неполная, так и избыточная информация снижает эффективность принимаемых пользователем решений;

доступность – простота (или возможность) выполнения процедур получения и преобразования информации;

актуальность – определяется степенью сохранения ценности информации для управления в момент ее использования и зависит от динамики изменения ее характеристик и от интервала времени, прошедшего с момента возникновения данной информации;

своевременность – означает ее поступление не позже заранее назначенного момента времени, согласованного с временем решения поставленной задачи;

точность – степень близости получаемой информации к реальному состоянию объекта, процесса, явления и т.п.;

адекватность – это определенный уровень соответствия создаваемого с помощью полученной информации образа реальному объекту, процессу, явлению и т.п.;

устойчивость – способность информации реагировать на изменения исходных данных без нарушения необходимой точности.

Канал связи — это среда передачи информации, которая характеризуется в первую очередь максимально возможной для неё скоростью передачи данных (ёмкостью канала связи).

Помехи — случайные искажения сигнала в канале связи при передаче информации.

Кодирование — преобразование дискретной информации из одного вида в другой.

Энтропия

Базисным понятием всей теории информации является понятие энтропии. Энтропия – мера неопределенности некоторой ситуации.

Впервые энтропия была предложена Клодом Шенноном в его фундаментальной работе "Математические основы теории связи" опубликованной в 1948г в которой были заложены основы современной теории информации.

Факт получения информации всегда связан с уменьшением разнообразия или неопределенности. Рассмотрим источник информации, который может в каждый момент времени случайным образом принять одно из конечного множества возможных состояний. Такой источник называют дискретным источником информации. Различные состояния реализуются вследствие выбора их источником. Каждому состоянию источника *a* ставится в соответствие условное обозначение в виде знака (в частности, буквы) из алфавита данного источника: $a_1, a_2 \ldots\ldots a_i, \ldots a_n$. Поскольку одни состояния выбираются источником чаще, а другие реже, то в общем случае он характеризуется *ансамблем A*, т.е. полной совокупностью состояний с вероятностями их появления:

$$A = \begin{bmatrix} a_1, a_2 \ldots\ldots a_i, \ldots a_n \\ p_1, p_2 \ldots\ldots, p_i, \ldots p_n \end{bmatrix}$$

Задано множество с N возможными состояниями $a_1, a_2 \ldots\ldots a_i, \ldots a_n$ и заданным на нем распределением вероятностей $p(a_i)$ таким, что для всех $i=1,N$ $p(a_i) \geq 0$, а $\sum p(a_i) = 1$.

Чем больше величина *N*, тем больше неопределенность выбора конкретного элемента ансамбля.

Попробуем ввести количественную меру неопределенности.

В качестве такой меры нужно использовать непрерывную функцию, зависящую от числа состояний источника N – $f(N)$. Она должна удовлетворять следующим требованиям:

1. $f(1)=0$, так как при $N=1$ исход опыта не является случайным и неопределенность отсутствует;

2. условию монотонного возрастания при увеличении числа возможных состояний источника N;

3. условию аддитивности, которое формулируется следующим образом: если два независимых источника с числом равновероятных состояний N и M, соответственно, рассматривать как один источник, одновременно реализующий пары состояний $n_i \cdot m_j$, то неопределенность объединенного источника должна равняться сумме неопределенности исходных источников

$$f(NM) = f(N) + f(M)$$

Данное соотношение выполняется, если в качестве меры неопределенности источника с равновероятными состояниями принять функцию:

$$H(A) = \log N .$$

Для этой функции выполняются все три выдвинутые требования:

1. $\log(1)=0$;

2. функция монотонно возрастает с возрастанием N;

3. выполняется условие аддитивности, так как $\log(N \cdot M)=\log(N)+\log(M)$.

Указанная мера была предложена американским ученым Р.Хартли в 1928 г.

Основание логарифма не имеет принципиального значения и определяет только масштаб или единицу измерения неопределенности. Так как современная информационная техника базируется на элементах, имеющих два устойчивых состояния, то самым удобным основанием логарифма оказывается число 2. Единица измерения неопределенности при двух возможных равновероятных состояниях источника называется бит (от английского binary digit "двоичный разряд" или "двоичная единица"). Единицы при других осно-

ваниях логарифма: трит (основание логарифма равно три), нат (натуральный логарифм), дит (основание логарифма равно десяти).

Предложенная мера позволяет решать определенные практические задачи. Но она не получила широкого применения, поскольку была рассчитана на слишком грубую модель источника информации, приписывающую всем его возможным состояниям одинаковую вероятность.

Попробуем обобщить формулу на ситуацию, когда состояния **не равновероятны**, например, $p(A_1)$ и $p(A_2)$. Тогда:

$$H_1 = -p(A_1) \cdot \log_2 p(A_1) \quad и \quad H_2 = -p(A_2) \cdot \log_2 p(A_2)$$

$$H_0 = H_1 + H_2 = -p(A_1) \cdot \log_2 p(A_1) - p(A_2) \cdot \log_2 p(A_2)$$

Обобщая это выражение на n **неравновероятных** состояний, получим:

$$H(A) = -\sum_{i=1}^{M} p(a_i) \log p(a_i) \qquad (1)$$

Введенная таким образом величина получила название ***энтропии дискретного источника информации*** и представляет собой неопределенность появления на выходе источника сообщений буквы первичного алфавита.

Мера неопределенности выбора дискретным источником состояния из ансамбля A была предложена американским ученым К. Шенноном.

Ансамбль сообщений на выходе приемника будем называть ансамблем приемника сообщений и обозначать буквой B. Для того чтобы отличить переданные и принятые сигналы, абстрактный алфавит в котором представлен ансамбль приемника сообщений, обозначается $\{b_1, b_2, \ldots, b_j, \ldots, b_n\}$, а соответствующие вероятности - $p(b_1), p(b_2), \ldots, p(b_i), \ldots p(b_n)$.

Энтропия приемника сообщений

$$H(B) = -\sum_{j=1}^{N} p(b_j) \log p(b_j)$$

и представляет собой неопределенность появления на входе приемника буквы после ее появления на выходе источника сообщений. Если в канале связи

не происходит потерь информации, то всегда буква a_1 соответствует букве b_1, $a_2 - b_2$ и т.д. При этом $H(A)=H(B)$.

Понятие энтропии используется не только при передаче сообщений. Энтропия широко применяется для описания состояния механических и термодинамических систем, для изучения свойств алфавитов различных языков, при исследовании экономических систем.

Свойства энтропии.

1. Энтропия является вещественной и неотрицательной величиной, так как для любого i ($1<i<N$) p_i изменяется в интервале от *0* до *1*, $\log p_i$ отрицателен и, следовательно, $-p_i \log p_i$ положительна.

2. Энтропия – величина ограниченная.

3. Энтропия обращается в нуль лишь в том случае, если вероятность одного из состояний равна единице; тогда вероятности всех остальных состояний, равны нулю.

4. Энтропия максимальна, когда все состояния источника равновероятны.

$$H_{\max} = \log N$$

5. Энтропия объединения нескольких статистически независимых источников информации равна сумме энтропий исходных источников.

В случае статистической независимости источников информации *а* и *в* запишем $p(a_i,в_j)=p(a_i)p(в_j)$

тогда

$$H(a,b) = -\sum_{i=1}^{N}\sum_{j=1}^{K} p(a_i)p(b_j)\log p(a_i)p(b_j) =$$

$$= -\sum_{i=1}^{N} p(a_i)\log p(a_i)\sum_{j=1}^{K} p(b_j) - \sum_{j=1}^{K} p(b_j)\log p(b_j)\sum_{i=1}^{N} p(a_i)$$

Учитывая, что

$$\sum_{i=1}^{N} p(a_i) = 1 \quad \text{и} \quad \sum_{i=1}^{N} p(b_j) = 1$$

получим

$$H(A,B)=H(A)+H(B)=H(B,A)$$

Пример 1. Источник сообщений выдает символы алфавита $A=\{a_i\}, i=\overline{1,4}$ с вероятностями $p_1=0{,}2,\ p_2=0{,}3,\ p_3=0{,}4,\ p_4=0{,}1$. Найти энтропию источника.

Решение. Вычислим энтропию по формуле (1)

$$H(A)=-(0{,}2\log 0{,}2+0{,}3\log 0{,}3+0{,}4\log 0{,}4+0{,}1\log 0{,}1)=1{,}86 \text{ } бит$$

Пример 2. Имеются два ящика, в каждом из которых лежит по 12 шаров. В первом – 3 белых, 3 черных и 6 красных; во втором – каждого цвета по 4. Опыты состоят в вытаскивании по одному шару из каждого ящика. Что можно сказать относительно неопределенностей исходов этих опытов?

Решение. Согласно (1) находим энтропии обоих опытов:

$$H(A)=-(\frac{3}{12}\log\frac{3}{12}+\frac{3}{12}\log\frac{3}{12}+\frac{6}{12}\log\frac{6}{12})=1{,}5 \text{ } бит$$

$$H(B)=-(\frac{4}{12}\log\frac{4}{12}+\frac{4}{12}\log\frac{4}{12}+\frac{4}{12}\log\frac{4}{12})=1{,}58 \text{ } бит$$

Поскольку $H(B) > H(A)$, неопределенность исхода в опыте B выше и, следовательно, предсказать его можно с меньшей долей уверенности, чем исход опыта A.

Пример 3. Определить энтропию сообщения из пяти букв, если число букв в алфавите равно 32 и все сообщения равновероятные.

Решение: Общее число пятибуквенных сообщений равно: $N = m^n = 32$

Энтропия для равновероятных сообщений равна:

$$H(A)=\log N=5\log 32=25$$

Контрольные вопросы

1.Дать определение энтропии.

2.Запишите формулу Шеннона и формулу Хартли.

3.Перечислите основные свойства энтропии.

4. Что является единицей измерения энтропии?

5. В каких случаях энтропия равна нулю?

6. При каких условиях энтропия принимает максимальное значение?

7. В чем состоит правило сложения энтропий для независимых источников?

Задачи для самостоятельного решения

Задача 1. Имеются две урны, содержащие по 20 шаров - 10 белых, 5 черных и 5 красных в первой и 8 белых, 8 чёрных и 4 красных во второй. Из каждой урны вытаскивают по одному шару. Исход, какого из этих двух опытов следует считать более неопределенным?

Задача 2. Известно, что каждое из *к* возможных равновероятных двоичных сообщений несет 4 бита информации. Чему равно *к?*

Задача 3. Опытный индивидуальный предприниматель знает, что 25% всех его документов составляют налоговые декларации. Для неопытного предпринимателя появление любого типа документа - равновероятно. Определите, какое количество информации получит опытный и неопытный предприниматели при получении налоговой декларации?

Задача 4. Определить энтропию источника сообщений, если статистика распределения вероятностей появления символов на выходе источника сообщений представлена следующей схемой:

$$A = \begin{vmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 & a_9 & a_{10} \\ 0,35 & 0,035 & 0,07 & 0,15 & 0,07 & 0,07 & 0,14 & 0,035 & 0,01 & 0,07 \end{vmatrix}$$

Задача 5. Определить энтропию системы, состоящей из двух подсистем. Первая подсистема состоит из трех элементов, каждый из которых может находиться в двух состояниях с вероятностями p_1=0,6; p_2=0,4. Вторая

подсистема состоит из двух элементов, каждый из которых может находиться в трех состояниях с вероятностями $P1=0,1;\ p_2=0,4;\ P3=0,5$.

Условная энтропия и энтропия объединения

Условная энтропия

Если состояния элементов системы не зависят друг от друга, если состояние одной системы не зависит от состояния другой системы, то неопределенность того, что некоторый элемент системы будет находиться в одном из k возможных состояний полностью определялась бы вероятностными характеристиками отдельных элементов системы, либо вероятностными характеристиками состояний самих систем. При этом подразумевается, что символы сообщения взаимонезависимы, т.е. с приходом одного символа распределение вероятностей последующих символов не изменяется.

На практике же чаще всего встречаются взаимозависимые символы и сообщения. Если передавать не просто отдельные буквы алфавита, а смысловые сообщения, то можно убедиться, что существует взаимозависимость передаваемых символов. Одни буквы встречаются чаще, другие реже, одни буквы и слова часто следуют за другими, другие редко.

Понятие условной энтропии широко используется для определения информационных потерь при передаче информации.

Если элементы источника сообщений принимают состояния $a_1,\ a_2,\ \ldots..a_n$ с вероятностями соответственно $p(a_1),\ p(a_2),\ldots.p(a_n),$ а элементы адресата – состояния $в_1,\ в_2,\ldots..в_m,$ с вероятностями соответственно $p(в_1),\ p(в_2),\ldots..p(в_m),$ то понятие условной энтропии $H(в_j/a_i)$ выражает неопределенность того, что отправив a_i, мы получим $в_j$. Если в канале связи присутствуют помехи, то с различной степенью вероятности может быть принят любой из сигналов b_j, и наоборот, принятый сигнал $в_j$ может появится в результате отправления любого из сигналов a_i. Если в канале связи помехи отсутствуют, то всегда посланному сигналу a соответствует принятый сигнал $в$ и т.д. При этом энтропия источника $H(A)$ равна энтропии приемника $H(B)$. Если в кана-

ле связи присутствуют помехи, то они уничтожают часть передаваемой информации.

Информационные потери полностью описываются через частную и общую условную энтропию. Вычисление частных и общей условной энтропии удобно производить при помощи канальных матриц. **Если канал связи описывается со стороны источника сообщений (т.е. известен посланный сигнал)**, то вероятность того, что при передаче сигнала а по каналу связи с помехами мы получим сигнал в, обозначается как условная вероятность $p(b_j / a_i)$, а канальная матрица имеет вид:

B A	b_1	b_2	...	b_j	.....	b_m
a_1	$p(b_1/a_1)$,	$p(b_2/a_1)$,	..,	$p(b_j/a_1)$,	..,	$p(b_m/a_1)$
a_2	$p(b_1/a_2)$,	$p(b_2/a_2)$,	..,	$p(b_j/a_2)$,	..,	$p(b_m/a_2)$
	.....					
a_i	$p(b_1/a_i)$,	$p(b_2/a_i)$,	..,	$p(b_j/a_i)$,	..,	$p(b_m/a_i)$
	.....					
a_n	$p(b_1/a_n)$,	$p(b_2/a_n)$,	..,	$p(b_j/a_n)$,	..,	$p(b_m/a_n)$

Вероятности, которые расположены по диагонали, определяют вероятности правильного приема, остальные – ложного. Значения цифр, заполняющих колонки канальной матрицы, обычно уменьшаются по мере удаления от главной диагонали и при полном отсутствии помех все, кроме цифр, расположенных на главной диагонали, равны нулю.

Прохождение данного вида сигнала со стороны источника сообщений в данном канале связи описывается распределением условных вероятностей вида $p(b_j / a_i)$. Например, для сигнала a_1 распределением вида

$$p(b_1/a_1)+p(b_2/a_1)+\ldots+p(b_j/a_i)+\ldots+p(b_m/a_1)=1$$

Потери информации, приходящиеся на долю сигнала a_i описываются при помощи частной условной энтропии. Например, для сигнала a_1

$$H(b_j / a_1) = -\sum p(b_j / a_1) \log p(b_j / a_1)$$

Суммирование производится по j, так как i-е состояние остается постоянным.

Потери при передаче всех сигналов по данному каналу связи описываются при помощи общей условной энтропии. Для ее вычисления следует просуммировать все частные условные энтропии, т.е. произвести двойное суммирование по i и по j. При этом, в случае равновероятных появлений сигналов на выходе источника сообщений

$$H(B / A) = -\frac{1}{N} \sum_j \sum_i p(b_j / a_i) \log p(b_j / a_i)$$

В случае неравновероятного появления символов источника сообщений следует учесть вероятность появления каждого символа, умножив на нее соответствующую частную условную энтропию. При этом общая условная энтропия

$$H(B / A) = -\sum_i \sum_j p(a_i) p(b_j / a_i) \log p(b_j / a_i)$$

Если исследовать канал связи ***со стороны приемника сообщений*** (т.е. известен принятый сигнал), то с получением сигнала $в_j$ предполагаем, что был послан какой-то из сигналов a_i. При этом канальная матрица будет иметь вид:

B A	b_1	b_2	...	b_j	.....	b_m
a_1	$p(a_1 / b_2)$,	$p(a_2 / b_1)$, ..,		$p(a_1 / b_j)$,	..,	$p(a_1 / b_m)$
a_2	$p(a_1 / b_2)$,	$p(a_2 / b_2)$, ..,		$p(a_2 / b_j)$,	...,	$p(a_1 / b_m)$
	.....					
a_i	$p(a_i / b_1)$,	$p(a_i / b_2)$, ...,		$p(a_i / b_j)$,	..,	$p(a_i / b_m)$
	.....					
a_n	$p(a_n / b_1)$,	$p(a_n / b_2)$,	...,	$p(a_n / b_j)$, ...,		$p(a_n / b_m)$

В этом случае единице должны равняться суммы условных вероятностей не по строкам, а по столбцам канальной матрицы

$$p(a_1 / b_j) + p(a_2 / b_j) + \ldots + p(a_i / b_j) + \ldots + p(a_n / b_j) = 1$$

Частная условная энтропия

$$H(a_i / b_j) = -\sum_{i=1}^{N} p(a_i / b_j) \log p(a_i / b_j)$$

Общая условная энтропия

$$H(A / B) = -\sum_{j} \sum_{i} p(b_j) p(a_i / b_j) \log p(a_i / b_j).$$

Понятие условной энтропии в теории информации используется при определении взаимозависимости между символами кодируемого алфавита, для определения потерь при передаче информации по каналам связи, при вычислении энтропии объединения.

Во всех случаях при вычислении условной энтропии в том или ином виде используются условные вероятности.

Если в канале связи помехи отсутствуют, то все элементы канальной матрицы, кроме элементов, расположенных на главной диагонали, равны нулю. Вероятность получения правильного сигнала станет безусловной, а условная энтропия будет равна нулю. Канальная матрица будет иметь вид

$$p(a/b) = \begin{vmatrix} 1\,0\,0\ldots.0 \\ 0\,1\,0\ldots.0 \\ \ldots\ldots\ldots\ldots \\ 0\,0\,0\ldots.1 \end{vmatrix}$$

В этом случае условная энтропия будет равна нулю.

Энтропия объединения

Взаимная энтропия, или как ее часто называют, *энтропия объединения* используется для вычисления энтропии совместного появления статистических зависимых сообщений.

Пусть $(a_1,a_2,...,a_i,...,a_n)$ есть выборочное пространство A, характеризующее источник сообщений, а $(b_1,b_2,...,b_j,...,b_m)$ есть выборочное пространство B, характеризующее приемник сообщений. При этом a есть сигнал на входе шумящего канала, а $в$ – сигнал на его выходе. В этом случае взаимная энтропия представляет собой информацию о переданном сигнале a_i, содержащегося в принятом сигнале b_j. Взаимосвязь переданных и принятых сигналов описывается вероятностями совместных событий вида $p(a_i,b_j)$, а взаимосвязь выборочных пространств A и B описывается матрицей объединения вида:

$$p(a_i,b_j)=\begin{vmatrix} p(a_1,b_1) & p(a_1,b_2) & \dots & p(a_1,b_m) \\ p(a_2,b_1) & p(a_2,b_2) & \dots & p(a_2,b_m) \\ \dots\dots & \dots\dots & \dots & \dots\dots \\ p(a_n,b_1) & p(a_n,b_2) & \dots & p(a_n,b_m) \end{vmatrix}$$

Если матрица описывает канал связи, то число строк матрицы равно числу столбцов, $m=n$, и пределы суммирования по i и по j одинаковы.

Независимо от равенства или неравенства числа строк числу столбцов матрица объединения обладает следующими свойствами

1) $\sum_i p(a_i,b_j)=p(b_j)$ Сумма вероятностей по столбцам равна вероятности приёмника.

2) $\sum_j p(a_i,b_j)=p(a_i)$ Сумма вероятностей по строкам рана вероятности источника.

3) $\sum_i p(a_i) = \sum_j p(b_j) = 1$, т. е. $\sum_i \sum_j p(a_i, b_j) = 1$Сумма всех элементов равна 1.

Условные вероятности при помощи матрицы объединения находятся следующим образом

$$p(a_i / b_j) = \frac{p(a_i, b_j)}{\sum_i p(a_i, b_j)} = \frac{p(a_i, b_j)}{p(b_j)}.$$

$$p(b_j / a_i) = \frac{p(a_i, b_j)}{\sum_j p(a_i, b_j)} = \frac{p(a_i, b_j)}{p(a_i)}.$$

Взаимная энтропия ансамблей *A* и *B* при помощи матрицы объединения вычисляется путем последовательного суммирования по строкам или по столбцам всех вероятностей вида *p(a,b)*, умноженных на логарифм этих же вероятностей

$$H(A,B) = -\sum_i \sum_j p(a_i, b_j) \log p(a_i, b_j) \quad \text{бит/два символа}$$

Размерность «бит/два символа» объясняется тем, что взаимная энтропия представляет собой неопределенность возникновения пары символов, то есть неопределенность на два символа.

Взаимная энтропия передаваемого ансамбля *A* и принимаемого ансамбля *B* равна сумме безусловной энтропии *H(A)* и условной энтропии *H(B/A)*

$$H(A,B) = H(A) + H(B / A)$$

H(B/A) в данном случае представляет ту добавочную информацию, которую дает сообщение *B* после того как стала известна информация, содержащаяся в сообщении *A*.

Таким образом, условная энтропия представляет собой неопределенность того, что при приеме *в* было послано *а*, а взаимная энтропия отражает неопределенность возникновения пары вида *ав*.

Так как взаимная энтропия есть неопределенность относительно пары символов, сигналов, состояний, в общем случае, относительно пары элементов взаимосвязанных выборочных пространств *A* и *B*, то не имеет значения

имеет ли эта пара вид *ав* или *ва,* так как неопределенность возникновения такого сочетания – одинакова. *Взаимная энтропия обладает свойством симметрии.*

$$H(A,B) = H(B,A)$$

Если построена матрица вероятностей *p(а,в),* описывающая взаимосвязь двух произвольных выборочных пространств, в частности взаимосвязь входа и выхода шумящего канала связи, то остальные информационные характеристики могут не задаваться, так как матрица объединения обладает **информационной полнотой**.

Определение. Набор информационных характеристик произвольного канала связи считается информационно полным, если с помощью этого набора, путем алгебраических преобразований, можно получить любую другую информационную характеристику того же канала связи.

Пример: Определить общую условную энтропию дискретного канала связи, если задана матрица объединения

$$p(a,b) = \begin{pmatrix} 0{,}1 & 0 & 0{,}1 \\ 0{,}1 & 0{,}3 & 0 \\ 0 & 0{,}1 & 0{,}3 \end{pmatrix}$$

Решение: Вычисляем вероятности появления символов на входе источника $p(a_i)$ и вероятности появления символов на входе приемника $p(b_j)$.

$$\begin{aligned} p(a_1) &= 0.2 \qquad & p(b_1) &= 0.2 \\ p(a_2) &= 0.4 \qquad & p(b_2) &= 0.4 \\ p(b_3) &= 0.4 \qquad & p(b_3) &= 0.4 \end{aligned}$$

Определяем условные вероятности и строим канальные матрицы

$$p(a_i / b_j) = \frac{p(a_i, b_j)}{p(b_j)} \qquad\qquad p(b_j / a_i) = \frac{p(a_i, b_j)}{p(a_i)}$$

$$p(a_i / b_j) = \begin{pmatrix} 0.5 & 0 & 0.25 \\ 0.5 & 0.75 & 0 \\ 0 & 0.25 & 0.75 \end{pmatrix} \qquad p(b_j / a_i) = \begin{pmatrix} 0.5 & 0 & 0.5 \\ 0.25 & 0.75 & 0 \\ 0 & 0.25 & 0.75 \end{pmatrix}$$

Находим общую условную энтропию $H(A/B)$ и общую условную энтропию $H(B/A)$

$$H(A/B) = -\sum_{i=1}^{3}\sum_{j=1}^{3} p(b_j)p(a_i/b_j)\log p(a_i/b_j) = -[0.2(0.5\log 0.5 + 0.5\log 0.5) +$$
$$+ 0.4(0.75\log 0.75 + 0.25\log 0.25) + 0.4(0.25\log 0.25 + 0.75\log 0.75)] \approx 0.85$$

$$H(B/A) = -\sum_{i=1}^{3}\sum_{j=1}^{3} p(a_i)p(b_j/a_i)\log p(b_j/a_i) = -[0.2(0.5\log 0.5 + 0.5\log 0.5) +$$
$$+ 0.4(0.25\log 0.25 + 0.75\log 0.75) + 0.4(0.75\log 0.75 + 0.25\log 0.25)] \approx 0.85$$

Контрольные вопросы

1. Дать определение условной энтропии.
2. Какие формулы используются для расчета условной энтропии?
3. Какие формулы используются для расчета взаимной информации?
4. Как определяется полная средняя взаимная информация?
5. Почему вводится понятие условной энтропии?
6. Приведите выражение для энтропии двух взаимно связанных ансамблей.

Задачи для самостоятельного решения

Задача 1. Чему равна условная энтропия и энтропия объединения, если канальная матрица имеет вид

$$p(b/a) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Задача 2. Канал связи, в котором передаются сигналы a_1, a_2, a_3, a_4, описан следующей канальной матрицей:

$$p(a,b) = \begin{pmatrix} 0{,}01 & 0{,}1 & 0{,}11 & 0{,}02 \\ 0{,}02 & 0{,}02 & 0{,}05 & 0{,}07 \\ 0{,}2 & 0{,}08 & 0{,}07 & 0{,}03 \\ 0{,}02 & 0{,}03 & 0{,}06 & 0{,}01 \end{pmatrix}$$

Найти долю информационных потерь, которые припадают на сигнал a_2 при передаче сигналов a_1 - a_4 по данному каналу связи.

Задача 3. Определить частную условную энтропию относительно каждого символа источника сообщений при передаче по каналу связи, описанному следующей канальной матрицей:

$$p(a,b) = \begin{pmatrix} 0{,}2 & 0{,}1 & 0{,}07 \\ 0{,}1 & 0{,}3 & 0{,}1 \\ 0 & 0{,}1 & 0{,}03 \end{pmatrix}$$

Задача 4. Взаимодействие двух систем *A* и *B* описывается следующей матрицей:

$$p(a,b) = \begin{pmatrix} 0{,}2 & 0{,}1 & 0 \\ 0 & 0{,}3 & 0{,}1 \\ 0 & 0{,}1 & 0{,}2 \end{pmatrix}$$

Определить безусловную энтропию системы *A* и системы *B*, и *H(A/B)*.

Задача 5. Определить все возможные информационные характеристики канала связи, в котором взаимосвязь источника с приемником может быть описана матрицей вида:

$$p(a,b) = \begin{pmatrix} 0{,}2 & 0{,}05 & 0 \\ 0{,}1 & 0{,}2 & 0{,}05 \\ 0 & 0{,}1 & 0{,}3 \end{pmatrix}$$

Задача 6. В результате статистических испытаний канала связи получены следующие условные вероятности перехода одного сигнала в другой: $p(b_1/a_1)=0.85$, $p(b_2/a_1)=0.1$, $p(b_3/a_1)=0.05$, $p(b_1/a_2)=0.09$, $p(b_2/a_2)=0.91$, $p(b_3/a_2)=0$, $p(b_1/a_3)=0.08$, $p(b_2/a_3)=0$ $p(b_3/a_3)=0.92$. Построить канальную матрицу и определить общую условную и взаимную энтропию сообщений, передаваемых по данному каналу связи, если на выходе источника сигналы появились с равной вероятностью.

Задача 7. Обладают ли информационной полнотой следующие информационные характеристики:

$p(a_1)=0.3,\ p(a_2)=0.2,\ p(a_3)=0.5 \quad p(b/a)=\begin{vmatrix}1&0&0\\0&1&0\\0&0&1\end{vmatrix}$

Задача 8. Вероятность появления сигналов на входе приемника сообщений равна соответственно $p(в_1)=0.2,\ p(в_2)=0.3,\ p(в_3)=0.5$. Канал связи описан следующей канальной матрицей:

$$p(a/b)=\begin{vmatrix}0{,}97 & 0 & 0{,}01\\0{,}02 & 0{,}98 & 0{,}01\\0{,}01 & 0{,}02 & 0{,}98\end{vmatrix}$$

Определить энтропию источника сообщений.

Задача 9. Вероятность появления сигналов на выходе источника сообщений равна соответственно $p(a_1)=0.4,\ p(a_2)=0.3,\ p(a_3)=0.3$. Канал связи описан следующей канальной матрицей:

$$p(b/a)=\begin{vmatrix}0{,}9 & 0 & 0{,}1\\0{,}2 & 0{,}8 & 0\\0{,}1 & 0{,}2 & 0{,}7\end{vmatrix}$$

Определить информационные потери которые приходятся на сигнал b_1.

Задача 10. Определить энтропию источника сообщений, передаваемых по каналу связи, и состоящих из равновероятных символов и информационные потери при передачи информации, если влияние помех в канале описывается матрицей:

$$p(b/a)=\begin{vmatrix}0{,}9 & 0 & 0{,}1 & 0\\0{,}1 & 0{,}6 & 0{,}2 & 0{,}1\\0{,}1 & 0{,}1 & 0{,}7 & 0{,}1\\0 & 0{,}2 & 0{,}3 & 0{,}5\end{vmatrix}$$

Задача 11. При передачи информации по каналу связи с помехами статистические испытания дали следующие результаты:

100 раз передавали цифру <1> приняли цифру <1> 93 раз
100 раз передавали цифру <1> приняли цифру <2> 5 раза
100 раз передавали цифру <1> приняли цифру <3> 2 раза
100 раз передавали цифру <2> приняли цифру <2> 97 раз
100 раз передавали цифру <2> приняли цифру <3> 1 раз

100 раз передавали цифру <2> приняли цифру <1> 1 раз
100 раз передавали цифру <2> приняли цифру <4> 1 раз
100 раз передавали цифру <3> приняли цифру <3> 98 раз
100 раз передавали цифру <3> приняли цифру <1> 1 раз
100 раз передавали цифру <3> приняли цифру <4> 1 раз
100 раз передавали цифру <4> приняли цифру <4> 100 раз

Определить информационные потери при передаче каждой отдельной цифры по данному каналу связи, если символы этого алфавита появляются на выходе источника сообщений с вероятностями: *p(1)=0.3, p(2)=0.3, p(3)=0.3, p(4)=0.1*.

Количество информации и избыточность

Количественная оценка информации

Для того чтобы оценить и измерять количество информации применяются различные подходы и методы. Среди них выделяются статистический, семантический, прагматический и структурный. Исторически наибольшее развитие получил статистический подход.

Статистический подход. Он изучается в обширном разделе кибернетики, называемой теорией информации. Основоположником этого подхода считается К.Шеннон, опубликовавший в 1948 г. свою математическую теорию связи. Большой вклад в теорию информации до него внесли ученые Г. Найквист и Р. Хартли, которые соответственно в 1924 и 1928 гг. напечатали работы по теории телеграфии и передачи информации. Признаны во всем мире исследования по теории информации российских ученых А.Н.Колмогорова, А.Я.Хинчина, В.А. Котельникова, А.А.Харкевича и т.д.

К.Шенноном было введено понятие количества информации как меры неопределенности состояния системы, снимаемой при получении информации. Количественно выраженная неопределенность состояния получила название энтропии по аналогии с подобным понятием в статистической механике. При получении информации уменьшается неопределенность, т.е. энтропия системы. Чем больше информации получает наблюдатель, тем боль-

ше снижается неопределенность, и энтропия системы уменьшается. При энтропии, равной нулю, о системе имеется полная информация, и наблюдателю она представляется целиком упорядоченной.

Найдем среднее количество информации, содержащееся в любом принятом элементе сообщения относительно переданного источником. До получения конкретного элемента сообщения средняя неопределенность, имеющаяся у адресата, относительно реализации источником любого элемента сообщения равна энтропии источника. Ее называют *априорной энтропией* источника.

Средняя неопределенность относительно любого состояния источника, остающаяся у адресата после получения конкретного элемента сообщения b_j, характеризуется частной условной энтропией $H(a_i / b_j)$:

$$H(a_i / b_j) = -\sum_{i=1}^{N} p(a_i / b_j) \log p(a_i / b_j)$$

Это случайная величина, зависящая от того, какой конкретно элемент сообщения принят.

Средняя неопределенность по всему ансамблю принимаемых элементов сообщений равна условной энтропии $H(A/B)$:

$$H(A/B) = -\sum_{i=1}^{N} \sum_{j=1}^{N} p(b_j) p(a_i / b_j) \log p(a_i / b_j)$$

Эту условную энтропию называют апостериорной энтропией источника информации.

При наличии помех среднее количество информации, содержащееся в каждом принятом элементе сообщения, относительно любого переданного равно разности априорной и апостериорной энтропий:

$$I(A,B) = H(A) - H(A/B) = H(B) - H(B/A).$$

Если частный характер количества информации специально не оговаривается, то мы имеем дело с количеством информации, приходящимся в среднем на один элемент сообщения.

Основные свойства количества информации

Свойство 1. Количество информации величина положительная.

Свойство 2. При отсутствии статистической связи между случайными величинами A и B, $H(A/B) = H(A)$ следовательно, в этом случае $I(A,B) = 0$

Принятые элементы сообщения не несут никакой информации относительно переданных.

Свойство 3. Количество информации в B относительно A равно количеству информации в A относительно B.

$$I(A,B) = I(B,A)$$

Свойство 4. При взаимно однозначном соответствии между множествами передаваемых и принимаемых элементов сообщений, что имеет место в отсутствии помех, апостериорная энтропия равна нулю и количество информации численно совпадает с энтропией источника

$$I(A,B) = H(A).$$

Это максимальное количество информации о состоянии дискретного источника. Для непрерывного источника оно равно бесконечности.

Семантический подход. Этот подход является наиболее трудно формализуемым и до сих пор окончательно неопределившимся.

Наибольшее признание для измерения смыслового содержания информации получила тезаурусная мера, предложенная Ю.И. Шнейдером. Идеи тезаурусного метода были сформулированы ещё основоположником кибернетики Н. Винером. Для понимания и использования информации её получатель должен обладать определенным запасом знаний.

Если индивидуальный тезаурус получателя информации близок к нулю, $S_П \approx 0$, то в этом случае и количество воспринятой информации равно нулю: $I_C = 0$.

Иными словами, получатель не понимает принятого сообщения, и, как следствие, для него количество воспринятой информации равно нулю. Такая ситуация эквивалентна прослушиванию сообщения на неизвестном ино-

странном языке. Несомненно, сообщение не лишено смысла, однако оно непонятно, а значит, не имеет информативности.

Количество семантической информации I_C в сообщении также будет равно нулю, если пользователь информации абсолютно все знает о предмете, т.е. его тезаурус $S_П$, и сообщение не дает ему ничего нового.

Тезаурусный метод подтверждает тезис о том, что информация обладает свойством относительности и имеет, таким образом, относительную, субъективную ценность.

Прагматический подход. Он определяет количество информации как меру, способствующую достижению поставленной цели. Одной из первых работ, реализующих этот подход, явилась статья А.А. Харкевича. В ней он предлагал принять за меру ценности информации количество информации, необходимое для достижения поставленной цели. Этот подход базируется на статической теории Шеннона и рассматривает количество информации как приращение вероятности достижения цели. Так, если принять вероятность достижение цели до получения информации равной p_0, а после её получения - p_1, то прагматическое количество информации $I_П$ определяется как

$$I_П = \log \frac{p_1}{p_0}.$$

Если основание логарифма сделать равным двум, то $I_П$ будет измеряться в битах, как и при статистическом подходе.

Структурный подход. Он связан с проблемами хранения, реорганизации и извлечения информации и по мере увеличения объемов накапливаемой в компьютерах информации приобретает все большее значение.

При структурном подходе абстрагируются от субъективности, относительно ценности информации и рассматривают логические и физические структуры организации информации.

Избыточность информации

Для нахождения максимальной пропускной способности системы связи необходимо уметь определять максимальное количество информации, которое может быть передано при помощи символов данного алфавита за единицу времени.

Максимальное количество информации на символ сообщения можно получить только в случае равновероятных и независимых символов. Реальные коды редко полностью удовлетворяет этому условию, поэтому информационная нагрузка на каждый элемент обычно меньше той, которую они могли бы переносить. Энтропия сообщений представленная такими кодами меньше максимальной.

Раз элементы кода, представляющие сообщения недогружены, то само сообщение обладает информационной *избыточностью*. Понятие избыточности в теории информации введено для количественного описания информационного резерва кода, из которого составлено сообщение. Поставка такой задачи стала возможной именно потому, что информация является измеримой величиной, каков бы не был частный вид рассматриваемого сообщения.

Различают избыточность: *естественную и искусственную*. *Естественная избыточность* характерна для первичных алфавитов, а *искусственная* – для вторичных.

Естественная избыточность может быть подразделена на семантическую и статистическую.

Семантическая избыточность в том, что мысль, высказанная в сообщении может быть выражена короче. Если сообщение можно сократить без изменения смысла, а затем восстановить содержание, то оно обладает семантической избыточностью.

Статистическая избыточность обуславливается неравновероятностным распределениям качественных признаков первичного алфавита и их взаимнозависимостью.

При учёте частоты появление букв в текстах, следование букв в различных сочетаниях и слов в различных сообщениях, передаваемую инфор-

мацию можно значительно сжать, сократить. Отношение $H/\log N = \mu$ называют коэффициентом сжатия, а для характеристики величины, на которую у удлиняются сообщения на данном языке по сравнению с минимальной длинной, необходимой для передачи той же информации, вводят специальный параметр D–избыточность:

$$D = 1 - \frac{H}{H_{\max}} = 1 - \frac{H}{\log N}$$

где N – число различных букв используемого алфавита;

H – энтропия, приходящаяся на одну букву смыслового текста при учете всех k-буквенных сочетаний;

$H_{max} = \log N$ – максимальная энтропия, приходящаяся на букву, когда буквы независимы и равновероятны.

Энтропия может быть определена как информационная нагрузка на символ сообщения. Избыточность определяет недогруженность символов. Если $H=\log N$, то недогруженности не существует.

Кроме общего понятия статистическая избыточность существуют различные частные понятия, основными из которых являются следующие: избыточность D_s, вызванная статистической связью между символами сообщения, и избыточность D_p, обусловленная неравновероятностным распределением символов сообщения.

Избыточность от округления находится по формуле:

$$D_0 = \frac{k - \varphi}{k}$$

где $\varphi = \frac{\log_2 m_1}{\log_2 m_2}$, k - округленное до ближайшего целого значение φ.

Искусственная избыточность необходима для повышения помехоустойчивости кодов и её вводят в виде добавочных символов $n_к$.

Если в коде всего n разрядов из них n_u – несут информационную нагрузку, то $n_к = n - n_u$ и характеризует *абсолютную корректирующую избыточность*, а величина

$$D_k = \frac{n - n_u}{n_u}$$

характеризует *относительную корректирующую избыточность*.

Информационная избыточность – семантическая или статистическая - явление естественное и заложена такая избыточность в первичном алфавите.

Искусственная избыточность – явление искусственное и заложена она во вторичном алфавите.

Уменьшая избыточность сообщений, можно увеличить скорость его передачи. Увеличивая избыточность сообщения, можно уменьшить вероятность его искажения под действием помех.

Пример 1. Вычислить энтропию источника, выдающего два символа 0 и 1 с вероятностями *p(0) = p(1) = 1/N* и определить его избыточность.

Решение: Энтропия для случая независимых, равновероятных элементов равна: $H(x) = \log_2 N = \log_2 2 = 1$ *[дв. ед/симв.]*

При этом $H(x) = H_{max}(x)$ и избыточность равна *0*.

Пример 2. Алфавит состоит из 6 букв, вероятности появления которых равны $p(a_1) = \frac{4}{14}, p(a_2) = \frac{4}{14}, p(a_3) = \frac{2}{14}, p(a_4) = \frac{2}{14}, p(a_5) = \frac{1}{14}, p(a_6) = \frac{1}{14}$

Найти избыточность источника сообщений при Вычислить энтропию источника независимых сообщений, выдающего два символа 0 и 1 с вероятностями *p(0) = 3/4, p(1) = 1/4*.

Решение: Энтропия для случая независимых, не равновероятных элементов равна:

$$H(X) = -\sum_{i=1}^{m} p(x_i)\log_2 p(x_i) = -[p(0)\log_2 p(0) + p(1)\log_2 p(1)] =$$

$$= -[\frac{3}{4}\log_2 \frac{3}{4} + \frac{1}{4}\log_2 \frac{1}{4}] = 0.815\,[дв.ед./симв].$$

При этом избыточность равна *D = 1–0,815=0,18*

Пример 3. По каналу связи передается сообщение «многоногое». Найти: а) избыточность D_1 источника сообщений при статистической независимости букв; б) избыточность D_2 с учетом зависимости между буквами.

Решение. Для передаваемого сообщения, таблица распределения вероятностей появления символов имеет вид

a_i	Кол-во появлений букв в слове	Вероятность появления букв в слове
м	1	$p(м)=0,1$
н	2	$p(н)=0,2$
о	4	$p(о)=0,4$
г	2	$p(г)=0,2$
е	1	$p(е)=0,1$
	$\sum=10$	$\sum=1$

а) Алфавит состоит из 5 символов, значит $N=5$. $H_{max} = \log N=\log 5=2,322$ *бит*

$$H=-\sum_{i=1}^{5} p(a_i)\log p(a_i)$$

$$H=-(0,1\log 0,1+0,2\log 0,2+0,4\log 0,4+0,2\log 0,2+0,1\log 0,1)=2,122\,бит$$

По определению избыточности имеем

$$D=1-\frac{H}{\log N}=1-\frac{2,122}{2.322}=0,086$$

б) При учете статистической зависимости между буквами заполняем таблицу частот появления двухбуквенных сочетаний

а \ в	м	н	о	г	е	$n(a)$
м		1				1
н			2			2
о		1		2	1	4
г			2			2
е	1					1
$n(b)$	1	2	4	2	1	$\sum=10$

Таблица распределения вероятностей имеет следующий вид

а \ в	м	н	о	г	е	$n(a)$
м		1/10				1/10
н			2/10			2/10
о		1/10		2/10	1/10	4/10

г			2/10			2/10
е	1/10					1/10
n(b)	1/10	2/10	4/10	2/10	1/10	$\sum=1$

Избыточность с учетом зависимости между буквами вычисляется по формуле $D=1-\frac{H}{\log N}$, где H – энтропия на букву при учете двухбуквенных сочетаний.

Энтропия двухбуквенного текста

$$H(A,B)=-\sum_{i=1}^{5}\sum_{j=1}^{5}p(a_i,b_j)\log p(a_i,b_j)=-(4*(0{,}1\log 0{,}1)+3*(0{,}2\log 0{,}2))=2{,}39.$$

Следовательно $H=\frac{H(A,B)}{2}=1{,}195$ *бит*.

$$D=1-\frac{H}{\log N}=1-\frac{1{,}195}{2{,}322}\approx 0{,}485 \quad \textit{бит}.$$

Контрольные вопросы

1. Что необходимо учитывать при выборе способа измерения количества информации?

2. В каких единицах измеряется количество информации?

3. Как связаны между собой понятия количества информации и энтропии?

4. Основные свойства количества информации

5. Что такое избыточность?

6. Каковы последствия от наличия избыточности сообщений?

Задачи для самостоятельного решения

Задача 1. При передачи информации по каналу связи с помехами статистические испытания дали следующие результаты:

100 раз передавали цифру <1> приняли цифру <1> 95 раз
100 раз передавали цифру <1> приняли цифру <2> 3 раза
100 раз передавали цифру <1> приняли цифру <3> 2 раза
100 раз передавали цифру <2> приняли цифру <2> 97 раз
100 раз передавали цифру <2> приняли цифру <3> 1 раз
100 раз передавали цифру <2> приняли цифру <1> 1 раз

100 раз передавали цифру <2> приняли цифру <4> 1 раз
100 раз передавали цифру <3> приняли цифру <3> 99 раз
100 раз передавали цифру <3> приняли цифру <4> 1 раз
100 раз передавали цифру <4> приняли цифру <4> 100 раз

Определить количество информации, которое теряется в данном канале связи при передаче сообщений, составленных из алфавита 1,2,3,4, если символы этого алфавита появляются на выходе источника сообщений с вероятностями: *p(1)=0.3, p(2)=0.3, p(3)=0.3, p(4)=0.1.*

Задача 2. Определить количество информации при передаче сообщений по каналу связи, описанному следующей канальной матрицей:

$$p(b/a) = \begin{pmatrix} 0{,}25 & 0{,}25 & 0{,}25 & 0{,}25 \\ 0{,}25 & 0{,}25 & 0{,}25 & 0{,}25 \\ 0{,}25 & 0{,}25 & 0{,}25 & 0{,}25 \\ 0{,}25 & 0{,}25 & 0{,}25 & 0{,}25 \end{pmatrix}$$

если на выходе источника сообщений символы встречаются с вероятностями $p(a_1)=0.1,\ p(a_2)=0.1,\ p(a_3)=p(a_4)=0.4$.

Задача 3. Определить количество информации на сообщение, передаваемое по каналу связи, описанному матрицей

$$p(a,b) = \begin{vmatrix} 0{,}1 & 0 & 0 \\ 0{,}1 & 0{,}3 & 0 \\ 0 & 0{,}1 & 0{,}4 \end{vmatrix}$$

В сообщении, составленном из пяти качественных признаков, последние используются с разной частотой, т.е. вероятности их различны и равны соответственно $p_1=0.8,\ p_2=0.15,\ p_3=0.03,\ p_4=0.015$ *и* $p_5=0.005$. Определить количества информации.

Задача 4. Влияние помех в канале связи описывается следующим распределением условных вероятностей:

$$p(b/a) = \begin{vmatrix} 0{,}98 & 0{,}01 & 0{,}01 \\ 0{,}15 & 0{,}75 & 0{,}01 \\ 0{,}3 & 0{,}2 & 0{,}5 \end{vmatrix}$$

Вычислить количество информации, которое переносится одним символом сообщения при равновероятном появлении символов в сообщении.

Задача 5.

Влияние помех в канале связи описывается следующей матрицей

$$p(a/b) = \begin{vmatrix} 0,9 & 0,05 & 0,2 \\ 0,1 & 0,75 & 0,3 \\ 0 & 0,2 & 0,5 \end{vmatrix}$$

Вычислить количество информации, если символы на входе приемника появляются при вероятностях $p(в_1)=0.7;\ p(в_2)=0.2;\ p(в_3)=0.1$.

Информационные характеристики источника сообщений и канала связи

Опираясь на введенную меру количества информации, рассмотрим информационные характеристики источника дискретных сообщений и дискретный канал связи, позволяющие установить пути повышения эффективности систем передачи информации, и в частности, определить условия, при которых можно достигнуть максимальной скорости передачи сообщений по каналу связи, как в отсутствие, так и при наличии помех.

Источник дискретных сообщений формирует дискретные последовательности из ограниченного числа элементарных сообщений.

Под *каналом связи* подразумевают совокупность устройств и физических сред, обеспечивающих передачу сообщений из одного места в другое (или от одного момента времени до другого). Если канал используется для передачи дискретных сообщений, он называется *дискретным* каналом.

Если временным действием помех в канале можно пренебречь, то для анализа используется модель в виде идеализированного канала, называемого каналом без помех. В идеальном канале сообщению на входе однозначно соответствует определенное сообщение на выходе и наоборот.

Когда требования к достоверности велики и пренебрежение неоднозначностью связи между сообщениями A и B недопустимо, используется более сложная модель - канал с помехами.

Канал считается заданным, если известны данные о сообщениях на его входе и выходе и ограничения, накладываемые на входные сообщения физическими характеристиками канала.

Модели источника дискретных сообщений

Математической моделью множества возможных реализаций источника является дискретная или непрерывная случайная величина.

На практике, однако, нас чаще всего интересует не одно конкретное состояние источника, а дискретные или непрерывные последовательности состояний, реализуемых источником за длительный промежуток времени, например, телеграммы, видеосюжеты. Для описания таких сообщений используются математические модели в виде дискретных и непрерывных случайных процессов.

Для построения модели необходимо знать объем L алфавита знаков $(a_1 \ldots a_l)$, из которых источником формируются сообщения, и вероятности создания им отдельных знаков с учетом возможной связи между ними.

При доказательстве основных положений теории информации Шенноном использовалась модель, называемая *эргодическим источником сообщений*. Предполагается, что создаваемые им сообщения математически можно представить в виде эргодической случайной последовательности. Такая последовательность, удовлетворяет условиям стационарности и эргодичности. Первое означает, что вероятности отдельных знаков и их сочетаний не зависят от расположения последних по длине сообщения. Из второго следует, что статистические закономерности, полученные при исследовании одного достаточно длинного сообщения с вероятностью близкой к единице, справедливы для всех сообщений, создаваемых источником.

Стационарный источник сообщений, выбирающий знак формируемой последовательности независимо от других знаков, всегда является эргодическим. Его так же называют источником без памяти.

На практике, чаще встречаются источники, у которых вероятность выбора одного знака сообщения зависит от того, какие знаки были выбраны источником до этого (источник с памятью).

Производительность источника дискретных сообщений

Под производительностью источника сообщений подразумевают количество информации, вырабатываемое источником в единицу времени. Эту характеристику источника называют так же скоростью создания сообщений или потоком входной информации.

Длительность выдачи знаков источником в каждом из состояний в общем случае может быть различной.

Если длительность выдачи знака не зависит от состояния источника, для всех знаков одинакова и равна τ, то $\tau_u = \tau$.

Производительность источника $I(A)$ можно выразить формулой

$$I(A) = H(A)/\tau$$

Наибольшая производительность источника в этом случае достигается при максимальной энтропии.

Модели дискретных каналов

Информационная модель канала с помехами задается множеством символов на его входе и выходе и описанием вероятностных свойств передачи отдельных символов. В общем случае канал может иметь множество состояний и переходить из одного состояния в другое как с течением времени, так и в зависимости от последовательности передаваемых символов.

В состоянии канал характеризуется матрицей условных вероятностей $p(b_i / a_j)$ того, что переданный символ a_i будет воспринят на выходе как символ b_j. Значение вероятностей в реальных каналах зависит от многих различных факторов: свойств сигналов, являющихся физическими носителя-

ми символов (энергия, вид модуляции и т.д.), характера и интенсивности воздействующих на канал помех, способа определения сигнала на приемной стороне.

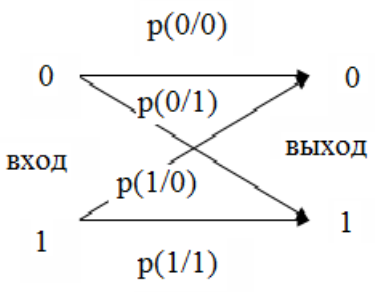
При наличии зависимости переходных вероятностей канала от времени, что характерно практически для всех реальных каналов, он называется нестационарным каналом связи. Если эта зависимость несущественна, используется модель в виде стационарного канала, переходные вероятности которого не зависят от времени.

Канал называется с "памятью" если переходные вероятности в данном состоянии канала зависят от его предыдущих состояний. Если переходные вероятности постоянны, т.е. канал имеет только одно состояние, он называется стационарным каналом без памяти.

Стационарный дискретный двоичный канал без памяти однозначно определяется четырьмя условными вероятностями:

$$p(1/0), p(0/1), p(1/1), p(0/0).$$

Такую модель канала принято изображать в виде графика:



вероятность неискаженной передачи символа $\begin{cases} p(0/0) \\ p(1/1) \end{cases}$

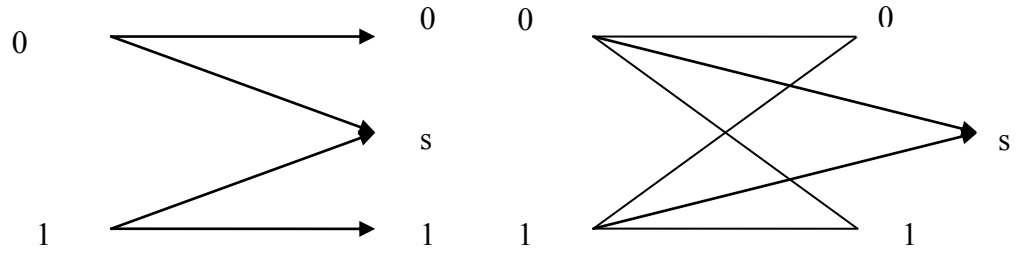
вероятность искаженной передачи символа $\begin{cases} p(0/1) \\ p(1/0) \end{cases}$

Если вероятности искажения можно принять равными, то такой канал называется двоичным симметричным каналом.

Важнейшие результаты, полученные для двоичного симметрического канала, распространены на более широкие классы каналов.

В последнее время приобретает все большее значение дискретный канал со стиранием. Для него характерно, что алфавит выходных символов отличается от алфавита входных символов. На входе, как и ранее, символы *0* и *1*, а на выходе канала фиксируются состояния, при которых сигнал с равным основанием может быть отнесен как к *1* так и к *0*. На месте такого символа не ставится ни нуль, ни единица: состояние отмечается дополнительным символом стирания *S*. При декодировании значительно легче исправить такие символы, чем ошибочно определенные.

На рисунке приведены модели стирающего канала при отсутствии помех и при наличии искажения символов



Скорость передачи информации по дискретному каналу

Характеризуя дискретный канал связи, используют два понятия скорости передачи: технической и информационной.

Под *технической* скоростью передачи подразумевают число элементов сообщения (символов) передаваемых по каналу в единицу времени.

$$V_\tau = \frac{1}{\tau_{ср}}$$

где $\tau_{ср}$ – средние значение длительности передачи одного символа.

Техническая скорость зависит от свойств линии связи и быстродействия аппаратура канала, с учётом возможных различий в длительности символов.

Единицей измерения технической скорости служит *бод* – скорость, при которой за одну секунду передается один символ.

Информационная скорость определяется средним количеством информации, которое передаётся по каналам в единицу времени.

Эта скорость зависит от характеристики данного канала связи (объём алфавита используемых символов, техническая скорость их передачи, статистические свойства помех в линии связи) и от вероятностей поступающих на вход символов и их статистической взаимосвязи.

Информационная скорость задается следующим соотношением:

$$\tilde{\tilde{I}}(A,B) = V_\tau I(A,B)$$

где $I(A,B)$ – среднее количество информации, переносимое одним символом.

Пропускная способность дискретного канала

Предельные возможности канала по передачи информации характеризуется его пропускной способностью. Пропускная способность равна той максимальной скорости передачи информации по данному каналу, которой можно достигнуть при самых совершённых способах передачи и приёма.

$$C_d = \max \tilde{\tilde{I}}(A,B) = \max V_\tau I(A,B)$$

где $I(A,B)$ – среднее количество информации, переносимое одним символом.

Пропускная способность канала как и скорость передачи информации измеряется числом двоичных единиц информации в секунду дв.ед/с

Для дискретного канала без помех:

$$C_d = \max V_\tau I(A,B) = V_\tau (H(A) - H(A/B) = V_\tau \max H(A) = V_\tau \log N$$

При наличии помех пропускная способность канала определяет наибольшее количество информации в единицу времени, которая может быть передана со сколь угодно малой вероятностью ошибки.

Предельные возможности канала никогда не используются полностью степень его загрузки, характеризуется коэффициентом использования канала:

$$\alpha = \frac{I(A)}{C_d}$$

где $I(A)$ – производительность источника сообщений;

C_d – пропускная способность канала связи.

Поскольку нормальное функционирование канала возможно, при изменении производительности источника в пределах $0 < I(A) < C_d$, α теоретически может изменяться в пределах от 0 до 1.

Пример 1. На вход дискретного симметричного канала без памяти поступают двоичные символы $a_1 = 0$ и $a_2 = 1$ с априорными вероятностями $p(a_1) = 0{,}85$ и $p(a_2) = 0{,}15$. Переходные вероятности $p(b_j / a_i)$ в таком канале задаются соотношением

$$p(b_j / a_i) = \begin{cases} p, & i \neq j \\ 1 - p & i = j \end{cases}$$

где *p=0,05* – вероятность ошибки. Определить все апостериорные (условные) вероятности.

Решение. Так как *p=0,05,* то вероятность правильного приема $q = 1 - 0{,}05$. Канальная матрица условных вероятностей описывающих канал связи со стороны источника равна

$$p(b_j / a_i) = \begin{vmatrix} 0{,}95 & 0{,}05 \\ 0{,}05 & 0{,}95 \end{vmatrix}$$

Условные вероятности, описывающие канал связи со стороны приемника найдем по формуле

$$p(a_i / b_j) = \frac{p(a_i, b_j)}{p(b_j)}.$$

Что бы найти вероятности приемника необходимо построить матрицу объединения по формуле

$$p(a_i, b_j) = p(a_i) * p(b_j / a_i) \qquad p(a_i, b_j) = \begin{vmatrix} 0{,}81 & 0{,}04 \\ 0{,}008 & 0{,}142 \end{vmatrix}$$

Вероятность появления символов на входе приемника находим по формуле

$$\sum_i p(a_i, b_j) = p(b_j) \qquad p(b_1) = 0{,}818 \qquad p(b_2) = 0{,}182.$$

Находим апостериорные вероятности

$$p(a_i / b_j) = \begin{vmatrix} 0{,}99 & 0{,}22 \\ 0{,}01 & 0{,}78 \end{vmatrix}$$

Пример 2. В информационном канале используется алфавит, содержащий 8 символов. Длительности всех символов одинаковы и равны $\tau = 2 мкс$. Определить пропускную способность канала при отсутствии шумов.

Решение. Пропускная способность канала при отсутствии шумов вычисляется по формуле $C_d = V_\tau \log N$. Подставляем в формулу N=8 и получаем.

$$C_d = \frac{\log N}{\tau} = \frac{\log 8}{2 \cdot 10^{-6}} = 2 \cdot 10^6 \ \ бит/сек.$$

Контрольные вопросы

1. Что называется технической скоростью?
2. Что называется информационной скоростью?
3. Дайте определение эргодическому источнику.
4. Какой канал называется каналом без помех?
5. Запишите выражения для пропускной способности дискретного канала без помех и с помехами, сравните их.
6. Приведите информационную модель канала связи.
7. Сформулируйте необходимые и достаточные условия неискаженной передачи сигнала по каналу связи.

Задачи для самостоятельного решения

Задача 1. Определить пропускную способность дискретного канала связи, описанного матрицей

$$p(a,b) = \begin{vmatrix} 0{,}1 & 0 & 0 \\ 0{,}1 & 0{,}3 & 0 \\ 0 & 0{,}1 & 0{,}4 \end{vmatrix}$$

Сообщения передаются с длительностью символов равной 1 мс.

Задача 2. Вероятность появления сигналов на входе приемника сообщений равна соответственно $p(b_1) = 0{,}2,\ p(b_2) = 0{,}3,\ p(b_3) = 0{,}5$. Канал связи описан следующей канальной матрицей:

$$p(a/b)=\begin{vmatrix}0{,}97 & 0 & 0{,}01\\ 0{,}02 & 0{,}98 & 0{,}01\\ 0{,}01 & 0{,}02 & 0{,}98\end{vmatrix}$$

Определить информационную скорость, если время передачи одного символа первичного алфавита 0,1 мс.

Задача 3. Двоичный симметричный канал без памяти задан канальной матрицей

$$p(b/a)=\begin{vmatrix}1-p & p\\ p & 1-p\end{vmatrix}$$

и вероятностями элементов на входе $p(a_1)=p(a_2)=\frac{1}{2}$. Построить граф канала, найти скорость создания информации *H(A)*и скорость передачи информации.

Задача 4. По каналу связи передается сообщение из ансамбля

$$X=\begin{pmatrix}x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8\\ 0{,}09 & 0{,}1 & 0{,}22 & 0{,}07 & 0{,}15 & 0{,}17 & 0{,}02 & 0{,}18\end{pmatrix}$$

Средняя длительность передачи одного элемента сообщения в канале 0,44 мс. Шум в канале отсутствует. Определить пропускную способность канала и скорость передачи информации.

Задача 5. Источник вырабатывает три сообщения с вероятностями: $p_1=0{,}1,\ p_2=0{,}2,\ p_3=0{,}7.$ Сообщения передаются с длительностью символов равной 1мс. Определить скорость передачи информации по каналу связи без помех.

Задача 6. По каналу связи передаются сообщения, вероятности которых равны: $p(a_1)=0{,}1,\ p(a_2)=0{,}2,\ p(a_3)=0{,}3,\ p(a_4)=0{,}4$. Канальная матрица, определяющая потери информации в канале связи

$$p(b/a)=\begin{vmatrix}0{,}99 & 0{,}01 & 0 & 0\\ 0{,}01 & 0{,}97 & 0{,}02 & 0\\ 0 & 0{,}01 & 0{,}98 & 0{,}01\\ 0 & 0 & 0{,}01 & 0{,}99\end{vmatrix}.$$ Определить пропускную способность

канала связи, если время передачи одного символа первичного алфавита 0,01 мс.

Задача 7. Определить скорость передачи по двоичному симметричному каналу связи при $\tau = 0{,}001\,c$, если шумы в канале вносят ошибки таким образом, что в среднем четыре символа из *100* принимаются неверно (т.е. *1* вместо *0* и наоборот).

Список использованной литературы

1. Решетников М.Т. Методические указания к практическим занятиям по дисциплине «Теория информации» для студентов специальности 230102 – «Автоматизированные системы обработки информации и управления». Томск: ТУСУР, 2012. – 25 с.

2. Зверева Е.Н., Лебедько Е.Г. Сборник примеров и задач по основам теории информации и кодирования сообщений. – СПб: НИУ ИТМО, 2014. – 76 с.

3. Чикрин Д.Е. Теория информации и кодирования: курс лекций / Д.Е.Чикрин.- Казань: Казанский университет, 2013.-116с.

4. Фурсов В. А. Теория информации: учеб. / В.А. Фурсов. - Самара: Изд-во Самар, гос. аэрокосм, ун-та, 2011. - 128 с.

5. Думачев В.Н. Теория информации и кодирования - Воронеж: Воронежский институт МВД России, 2012.–200с.

6. Кавчук С.В. Сборник примеров и задач по теории информации. Руководство для практических занятий на базе Mathcad 6.0 Plus. Таганрог: Изд-во ТРТУ, 2002. 64 с.

7. Мисюткин, В. И. Элементы теории информации: пособие по одноим. дисциплине для слушателей специальности 1-40 01 73 «Программное обеспечение информационных систем» заоч. формы обучения / В. И. Мисюткин. – Гомель : ГГТУ им. П. О. Сухого, 2015. – 87 с.

8. Дмитриев В.И. Прикладная теория информации: учебник для студентов вузов по специальности "Автоматизированные системы обработки информации и управления". – М.: Высшая школа, 1989. – 320 с.

9. Кузьмин, И.В. Основы теории информации и кодирования / И.В. Кузьмин, В.А. Кедрус; 2-е изд., перераб. и доп. - Киев: Вища школа, 1986.- 238 с.

10. Колесник В.Д. Курс теории информации / В.Д. Колесник, Г.Ш. Полтырев. – М.: Наука. Главная редакция физико-математической литературы. 1982. – 416 с.

Приложение

Для упрощения вычислений при решении задач приведена таблица значений величин $-p \log_2 p$ и таблица двоичных логарифмов целых чисел.

p	$-p \log_2 p$	p	$-p \log_2 p$	p	$-p \log_2 p$
0,01	0,0664	0,36	0,5306	0,71	0,3508
0,02	0,1129	0,37	0,5307	0,72	0,3412
0,03	0,1518	0,38	0,5305	0,73	0,3314
0,04	0,1858	0,39	0,5298	0,74	0,3215
0,05	0,2161	0,4	0,5288	0,75	0,3113
0,06	0,2435	0,41	0,5274	0,76	0,3009
0,07	0,2686	0,42	0,5256	0,77	0,2903
0,08	0,2915	0,43	0,5236	0,78	0,2796
0,09	0,3127	0,44	0,5211	0,79	0,2687
0,1	0,3322	0,45	0,5184	0,8	0,2575
0,11	0,3503	0,46	0,5153	0,81	0,2462
0,12	0,3671	0,47	0,5120	0,82	0,2348
0,13	0,3826	0,48	0,5083	0,83	0,2231
0,14	0,3971	0,49	0,5043	0,84	0,2113
0,15	0,4105	0,5	0,5000	0,85	0,1993
0,16	0,4230	0,51	0,4954	0,86	0,1871
0,17	0,4346	0,52	0,4906	0,87	0,1748
0,18	0,4453	0,53	0,4854	0,88	0,1623
0,19	0,4552	0,54	0,4800	0,89	0,1496
0,2	0,4644	0,55	0,4744	0,9	0,1368
0,21	0,4728	0,56	0,4684	0,91	0,1238
0,22	0,4806	0,57	0,4623	0,92	0,1107
0,23	0,4877	0,58	0,4558	0,93	0,0974
0,24	0,4941	0,59	0,4491	0,94	0,0839
0,25	0,5000	0,6	0,4422	0,95	0,0703
0,26	0,5053	0,61	0,4350	0,96	0,0565
0,27	0,5100	0,62	0,4276	0,97	0,0426

0,28	0,5142	0,63	0,4199	0,98	0,0286
0,29	0,5179	0,64	0,4121	0,99	0,0144
0,3	0,5211	0,65	0,4040		
0,31	0,5238	0,66	0,3956		
0,32	0,5260	0,67	0,3871		
0,33	0,5278	0,68	0,3783		
0,34	0,5292	0,69	0,3694		
0,35	0,5301	0,7	0,3602		

n	***log₂n***	***n***	***log₂n***	***n***	***log₂n***	***n***	***log₂n***	***n***	***log₂n***
1	0,0000	3	1,5850	5	2,3219	7	2,8074	9	3,1699
2	1,0000	4	2,0000	6	2,5850	8	3,0000	10	3,3219