Федеральное государственное бюджетное образовательное
учреждение высшего профессионального образования
«Вятский государственный гуманитарный университет»

Дополнительная подготовка школьников по дисциплине «Информатика и информационные технологии»

Учебный модуль
Информация, ее измерение и кодирование

Н. А. Бушмелева

Киров
2011

СОДЕРЖАНИЕ

1. Понятие информации

1.1. Информация и данные

1.1.1. Определения

Все люди используют в повседневной жизни понятия «информация» и «данные», подразумевая под ними разъяснение, сообщение, изложение, какие-либо сведения. Понятия «данные» и «информация» близки, но не тождественны. Эти понятия часто смешивают.

Бесспорной оказывается важность этих понятий, многие исследователи их важность связывают со значимостью основных физических понятий – «материя» и «энергия». Несмотря на то, что «информация» является центральным понятием в информатике, в настоящее время не существует общепринятого определения этого понятия, в литературе предлагаются самые различные определения.

В обычном, «житейском» смысле информация – это сведения, которые получает некоторый субъект об окружающем мире, о протекающих в нем процессах или явлениях.

В большом энциклопедическом словаре приводится следующее определение информации: 1) информация – это сведения, передаваемые людьми устным, письменным или каким-либо другим способом (с помощью условных знаков, сигналов, технических средств и т.п.); 2) с середины XX века информация – это обмен сведениями между людьми, человеком и автоматом, автоматом и автоматом, обмен сигналами в живом и растительном мире, передача признаков от клетки к клетке, от организма к организму [1].

Информация – это сведения, уменьшающие неопределённость нашего знания об окружающем нас мире, которые являются объектом хранения, преобразования, передачи и использования.

Информация – это мера неоднородности распределения материи и энергии в пространстве и во времени, показатель изменений, которыми сопровождаются все происходящие в мире процессы (В. М. Глушков).

Трудность в построении общего определения информации состоит в том, что существуют разные типы информации.

Понятие «данные» встречается не менее часто, чем информация и также является основным в информатике. В отличие от понятия «информация» есть несколько различных по форме, но равнозначных по сути определений этого понятия. Например, наиболее общее определение: *данные* – это

зарегистрированные сигналы. *Сигнал* – это изменяющийся во времени физический процесс. К регистрации сигналов можно отнести: запись музыки на магнитофон, запись лекции в тетрадь, запись наблюдений в ходе эксперимента в виде чисел, графиков, фотографирование каких-либо объектов, запоминание учеником материала на уроке, нарисованный план, запись данных в память компьютера, на жёсткий диск и т. д.

Существуют определения понятия «данные» через понятие «информация». *Данные* – это информация об объекте или отношениях объектов, выраженная в знаковой форме. Это определение носит прикладной характер и имеет отношение к базам данных.

Некоторые определения сужают общность понятия «данные» до уровня данных, используемых в технике. *Данные* – это информация, представленная в виде, позволяющем запоминать, хранить, передавать или обрабатывать её с помощью технических средств.

Данные и информация взаимосвязаны. Информация не может существовать без данных, без какого-либо носителя: она как-то должна быть представлена с помощью данных. С другой стороны, любые данные всегда несут в себе какую-то информацию.

Данные могут восприниматься человеком или техническим устройством; их можно переводить из одной знаковой системы в другую без потери содержащейся в них информации.

Для выделения информации из данных нужно применить к ним методы обработки, «адекватные» этим данным.

1.1.2. Информационные процессы

Информацию можно:

- создавать;
- передавать;
- воспринимать;
- использовать;
- запоминать;
- принимать;
- копировать;
- формализовать;
- распространять;
- преобразовывать;
- комбинировать;
- обрабатывать;
- делить на части;
- упрощать;
- собирать;
- хранить;
- искать;
- измерять;
- разрушать;

и др.

Все эти процессы, связанные с определенными операциями над информацией, называются *информационными процессами*.

1.2. Свойства информации

1. Информация *достоверна*, если она отражает истинное положение дел. Недостоверная информация может привести к неправильному пониманию или принятию неправильных решений. Достоверная информация со временем может стать недостоверной, так как она обладает свойством устаревать, то есть перестаёт отражать истинное положение дел.

2. Информация *полна*, если её достаточно для понимания и принятия решений. Как неполная, так и избыточная информация сдерживает принятие решений или может повлечь ошибки.

3. *Точность* информации определяется степенью ее близости к реальному состоянию объекта, процесса, явления и т. п.

4. *Ценность* информации зависит от того, насколько она важна для решения задачи, а также от того, насколько в дальнейшем она найдёт применение в каких-либо видах деятельности человека.

5. Только *своевременно* полученная информация может принести ожидаемую пользу. Одинаково нежелательны как преждевременная подача информации (когда она ещё не может быть усвоена), так и её задержка.

6. Если ценная и своевременная информация выражена непонятным образом, она может стать бесполезной.

7. Информация *понятна*, если она выражена языком, на котором говорят те, кому предназначена эта информация.

8. Информация должна преподноситься в *доступной* (по уровню восприятия) форме.

9. Информацию по одному и тому же вопросу можно изложить *кратко* (сжато, без несущественных деталей) или *пространно* (подробно, многословно). Краткость информации необходима в справочниках, энциклопедиях, учебниках, всевозможных инструкциях.

2. Измерение информации

2.1. Комбинаторный подход к определению количества информации

В 1928 году американский инженер Р. Хартли первым предложил методологию *измерения количества информации*. При этом Р. Хартли считал, что *измеряемая информация* – это «… группа физических символов – слов, точек, тире и т. п., имеющих по общему соглашению известный смысл для корреспондирующих сторон». Р. Хартли предложил рассматривать процесс получения информации как выбор одного сообщения из конечного наперёд заданного множества из N *равновероятных* сообщений. Тогда количество информации I, содержащееся в выбранном сообщении, определяет как двоичный логарифм N:

$$\boxed{I = \log_2 N}$$ – формула Р. Хартли

Выбор функции для описания количества информации в физической системе обоснован следующими условиями.

1. Значение функции в 1 должно быть равно 0, т.е.

$$I(1)=0.$$

2. Функция должна быть аддитивной, т. е.

$$I(N_1 \cdot N_2) = I(N_1) + I(N_2).$$

3. Функция должна быть возрастающей, т.е.

если $N_1 > N_2$, то $I(N_1) > I(N_2)$.

Этим условиям в полной мере удовлетворяет *логарифмическая* функция.

1. Если заданное множество имеет одно единственное сообщение (N=1), то количество информации, содержащееся в нем равно нулю:

$$I = I\,(1) = \log_2 1 = 0.$$

2. При наличии двух независимых множеств сообщений с мощностями N_1 и N_2 соответственно количество информации, приходящееся на одно сообщение (взятое из объединения этих множеств), равно сумме количеств информации, которые были бы получены от двух независимых множеств, взятых порознь:

$$I\,(N_1 \cdot N_2) = \log_2(N_1 \cdot N_2) = \log_2(N_1) + \log_2(N_2) = I(N_1) + I(N_2).$$

3. Большее количество информации содержит сообщение, выбранное из большего числа равновероятных сообщений:

если $N_1 > N_2$, то $I(N_1) > I(N_2)$, т.к. $\log_2(N_1) > \log_2(N_2)$.

2.2. Вероятностный подход к определению количества информации

Для определения количества информации не всегда возможно использовать формулу Хартли. Её применяют, когда выбор любого элемента из множества, содержащего N элементов, равнозначен. Или, при алфавитном подходе, все символы алфавита встречаются в сообщениях, записанных с помощью этого алфавита, одинаково часто. Однако в действительности символы алфавитов естественных языков в сообщениях появляются с разной частотой.

В 1948 году американец Клод Шеннон предложил формулу определения количества информации, учитывающую возможную *неодинаковую вероятность* сообщений в наборе:

$$\boldsymbol{I = -\,(\,p_1 \log_2 p_1 + p_2 \log_2 p_2 + \ldots + p_N \log_2 p_N\,),}$$

где p_i – вероятность того, что именно i-ое сообщение выделено в наборе из N сообщений

– формула К.Шеннона

2.3. Алфавитный (объемный) подход к определению количества информации

Алфавитный (объемный) подход применяется в технике, где информацией считается любая хранящаяся, обрабатываемая или передаваемая последовательность знаков, сигналов.

Этот подход основан на *подсчете числа символов в сообщении,* т. е. связан только с длиной сообщения и не учитывает его содержания. Но длина сообщения зависит не только от содержащейся в нем информации. На нее влияет мощность алфавита используемого языка. Множество используемых в тексте символов называется *алфавитом***.** Полное количество символов алфавита называется *мощностью алфавита***.**

В вычислительной технике наименьшей единицей измерения информации является **1 бит** (binary digit). Один бит соответствует одному знаку двоичного алфавита, т.е. 0 или 1. Таким образом, **1 бит = 0 или 1.**

Единицы измерения информации

Для удобства помимо бита применяются более крупные единицы измерения информации:

1 байт = 8 бит,
1 Кб (килобайт) = 1024 байт,
1 Мб (мегабайт) = 1024 Кб,
1 Гб (гигабайт) = 1024 Мб,
1 Тб (терабайт) = 1024 Гб.

Для того чтобы подсчитать количество информации в сообщении необходимо умножить количество информации, которое несет 1 символ, на длину сообщения.

Информационный объем сообщения (информационная емкость сообщения) – количество информации в сообщении, измеренное в битах, байтах или производных единицах (Кбайтах, Мбайтах и т.д.).

2.4. Закон аддитивности информации

Количество информации H(x1, x2), необходимое для установления пары (x1, x2), равно сумме количеств информации H(x1) и H(x2), необходимых для независимого установления элементов x1 , x2:

$$H(x1, x2) = H(x1) + H(x2).$$

Пример 1. Вычислить, какой объем памяти компьютера потребуется для хранения одной страницы текста на английском языке, содержащей 2400 символов.

Решение. Мощность английского алфавита, включая разделительные знаки, $N = 32$. Тогда для хранения такой страницы текста в компьютере понадобится $2400 \cdot \log_2 32$ бит $= 2400 \cdot 5 = 12000$ бит $= 1500$ байт.

Пример 2. В течение 5 секунд было передано сообщение, объем которого составил 375 байт. Каков размер алфавита, с помощью которого записано сообщение, если скорость передачи составила 200 символов в секунду?

Решение.

1) 375 байт / 5 с = 75 байт/с – скорость передачи,
2) Так как 75 байт/с соответствуют 200 симв./с., то в одном символе содержится 75 байт / 200 = 0,375 байт = 3 бита.
3) $\log_2 N = 3$ бита, следовательно, $N = 2^3 = 8$ символов.

3. Кодирование информации

3.1. Основные понятия теории кодирования

Теория кодирования информации является одним из разделов теоретической информатики. Она охватывает широкий круг проблем и стимулируется развитием средств связи. Теория кодирования решает задачи передачи информации (согласование параметров передаваемой информации с особенностями канала связи; разработка приемов, обеспечивающих надежность передачи информации по каналам связи) и задачи обработки и хранения информации (разработка принципов наиболее экономичного кодирования информации).

Информация передаётся в виде сообщений. Передаваемое *сообщение* (комбинация символов первичного алфавита) представляется *кодовым словом* или *кодом* (комбинацией символов вторичного алфавита).

Процесс перевода исходного сообщения в соответствующий код называется *кодированием*. Если вторичный алфавит состоит из двух символов, то кодирование называется *двоичным кодированием*.

Необходимость кодирования возникает, прежде всего, из потребности приспособить форму сообщения к данному «каналу связи» или какому-либо устройству, предназначенному для преобразования или хранения информации.

Декодирование – операция, *обратная* кодированию, т. е. восстановление информации в первичном алфавите по полученному коду.

Операции кодирования и декодирования называются *обратимыми,* если их последовательное применение обеспечивает возврат к исходной информации без каких-либо ее потерь.

3.2. Требования, предъявляемые к кодированию

1) *Взаимно однозначное соответствие кодового слова и исходного сообщения.* Это требование недопустимости одинаковых кодовых слов для разных сообщений и недопустимости представления одного сообщения различными кодовыми словами.

2) *Наибольшая экономность.* Это требование заключается в том, чтобы передаваемые кодовые слова были как можно короче и, благодаря этому, требовали бы наименьшее время на передачу сообщения.

3) *Префиксность*, означающая, что ни одно кодовое слово нельзя было получить из другого, более короткого, путем дополнительных символов.

Другими словами ни одно более короткое кодовое слово не является началом ни одного более длинного кодового слова.

Пример. Пусть $A = \{a_1, a_2, a_3\}$, $B = \{0, 1\}$. Ниже представлены некоторые возможные коды для букв алфавита А.

A	a_1	a_2	a_3
a)	1	0	01
b)	01	10	11
c)	0	10	111

Вариант а) не является однозначно декодируемым кодом, так как код a_2 является начальной частью комбинации a_3. Для доказательства этого достаточно рассмотреть, например, двоичную последовательность 0101. Она может быть декодирована одним из сообщений: $a_2a_1a_2a_1$; a_3a_3; $a_2a_1a_3$; $a_3a_2a_1$.

Вариант b) декодируется однозначно, поскольку здесь представлено равномерное кодирование – все кодовые слова этого кода имеют равные длины и различны.

Вариант с) также однозначно декодируемый, поскольку никакое кодовое слово не является началом (префиксом) другого более длинного кодового слова.

Существует большое количество методов кодирования, которые можно классифицировать различным образом. Вот некоторые из них. Методы кодирования делятся на следующие:

1) алфавитные и групповые;
2) равномерные и неравномерные;
3) эффективные (оптимальные) и неэффективные;
4) адаптивные, полуадаптивные и неадаптивные;
5) симметричные и несимметричные.

Различные методы кодирования могут сравниваться по значению их цены.

Цена кодирования – это среднее число двоичных символов приходящихся на один символ исходного сообщения.

3.3. Методы кодирования

3.3.1. Метод Морзе

Этот метод кодирования был изобретён в 1838 году американцем Сэмюэлем Финли Бриз Морзе для передачи сообщений по телеграфным линиям.

Морзе продемонстрировал свою систему кодирования 24 мая 1844-го года в первом в истории США сеансе телеграфной связи, который проводился между городами Балтимор (штат Мэриленд) и Вашингтон (Округ Колумбия). Он послал сообщение «What hath God wrought!» («Чудесно творение господне!»).

Метод кодирования является двоичным (вторичный алфавит содержит два символа – точку и тире) и базируется на следующем принципе: наиболее часто употребляемым буквам ставились в соответствие наиболее короткие последовательности из точек и тире, что существенно сокращало длину сообщения.

Вот как выглядит знаменитое телеграфное сообщение – сигнал бедствия «SOS» (***S**ave **O**ur **S**ouls* – спасите наши души) в коде азбуки Морзе, применяемом к английскому алфавиту:

• • • – – – • • •

Три точки (буква S), три тире (буква O), три точки (буква S). Две паузы отделяют буквы друг от друга.

При реальной передаче данных сигнал для тире в 3 раза превосходит по длительности сигнал для точки. Сигналы точек и тире в совокупностях, которыми обозначаются буквы, разделяются интервалами, длительность которых равна длительности сигнала точки. «Пробел» между буквами, формирующими то или иное слово, обозначается интервалом, который по длительности равен утроенной длительности сигнала точки (иными словами, длительность этого интервала равна длительности сигнала для тире). Пробел между словами обозначается интервалом, по длительности равным ушестерённой длительности сигнала точки (иными словами, длительность этого интервала равна удвоенной длительности сигнала для тире).

Интересно, что Морзе подсчитывал частоту использования букв не путём изучения текстов, а путём подсчёта литер каждого типа в типографском наборе. Результатом его поистине каторжного труда стал высокоэффективный метод кодирования, который с некоторыми изменениями используется до сих пор, хотя с момента его изобретения прошло уже более 160 лет.

3.3.2. Метод Бодо

В 1874-м году примитивный печатающий телеграфный аппарат («телетайп») запатентовал во Франции Жан Морис Эмиль Бодо. Изобретатель использовал новую 5-битную систему кодирования символов. Вот ее фрагмент:

Двоичное значение	Буквы	Знаки
00011	A	1
11001	B	?
01110	C	:
01001	D	$
00001	E	3
01101	F	!
11010	G	&
…	…	…

Этот способ кодирования подразумевал использование 5 бит, однако 32-х позиций ($2^5 = 32$) недостаточно для представления символов латинского алфавита вместе с арабскими цифрами и знаками препинания. Поэтому в коде Бодо используется специальная схема «сдвига» для переключения между двумя «внутренними» таблицами символов, по 32 символа в каждой, которую можно сравнить с системой «сдвига», используемой в пишущих машинках для переключения на верхний регистр. Технически это реализовывалось с помощью использования управляющих (в данном случае работой телетайпов) двоичных последовательностей.

Бодо был вынужден ограничить свой метод кодирования использованием двоичных последовательностей длиною в 5 бит из-за аппаратных ограничений.

Метод кодирования Бодо стал первым в мире методом кодирования текстовых данных с помощью двоичных последовательностей. Сообщения, для передачи которых использовалась система кодирования Бодо, распечатывались операторами на узкие ленты с помощью специальных 5-клавишных клавиатур. Причем имелась возможность одновременной работы до 6-ти операторов благодаря применению системы временн*о*го распределения. Это позволило значительно увеличить пропускную способность телеграфной линии. Предложенная Бодо аппаратура

зарекомендовала себя весьма положительно и оставалась в широком применении в XX-м веке.

3.3.3. Метод Шеннона–Фано

Метод алфавитного, неравномерного, эффективного, адаптивного кодирования был предложен в 1948–49 гг. независимо друг от друга Р. Фано и К. Шенноном.

1) Вычислить вероятности появления каждого из символов исходного сообщения.

2) Упорядочить символы исходного сообщения по убыванию значений их вероятностей появления в сообщении.

3) Разбить полученное упорядоченное множество на два подмножества таким образом, чтобы суммарные вероятности каждого из них были по возможности одинаковыми и близкими к 0,5. При этом сообщениям из одного подмножества в качестве первого символа кодового слова присвоить нуль, а сообщениям из другого – единицу.

4) Каждое из двух подмножеств, полученных на предыдущем шаге, рассмотреть как новое множество и подвергнуть аналогичному разбиению, в результате чего будет получен второй символ кодового слова для каждого символа.

Действия 3) и 4) продолжать шагом до тех пор, пока в рассматриваемых множествах не останется по одному элементу.

Таким образом, основу построения кода Шеннона–Фано составляет процедура дихотомии, т. е. последовательного разбиения множества символов на две части.

Пример. Пусть в текстовом сообщении значения вероятностей появления каждого символа вычислены и представлены в таблице в порядке убывания.

X	$p(x)$	Итерации					Код
		I	II	III	IV	V	
x_1	0,50	0					0
x_2	0,20	1	0	0			100
x_3	0,15	1	0	1			101
x_4	0,05	1	1	0			110
x_5	0,04	1	1	1	0	0	11100
x_6	0,03	1	1	1	1	0	11110
x_7	0,02	1	1	1	1	1	11111
x_8	0,01	1	1	1	0	1	11101

Видно, что процесс построения кодовых слов заканчивается после 5 итераций, причем варианты разбиений можно проследить по жирным разграничительным линиям и полутоновому фону.

На первом шаге сообщение x_1 сразу оказывается единственным элементом одного из подмножеств, поскольку его вероятность равна 0,5. Поэтому с присвоением x_1 кодового символа 0 кодирование этого сообщения завершается. Естественно, первый символ всех кодовых слов, отображающих сообщения из второго подмножества, полагается равным 1. Разумеется, конкретное соответствие между упомянутыми символами и подмножествами сообщений несущественно, и с равным успехом можно приписать x_1 символ 1, а остальным сообщениям – 0. Второе разбиение приводит к образованию двух подмножеств с равными суммарными вероятностями, первое из которых включает сообщения x_2 и x_3, а второе – все оставшиеся, т. е. $x_4 \div x_8$. При этом в качестве второго кодового символа нуль приписывается словам первого из подмножеств, тогда как единица – словам второго. Дальнейшие действия ясны из таблицы и не нуждаются в комментарии.

Алгоритм Шеннона–Фано гарантирует соблюдение требования префиксности, так как каждое разбиение заканчивается присвоением разным подмножествам противоположных символов.

Цена кодирования:

$$c=0,5\cdot 1+0,2\cdot 3+0,05\cdot 3+0,15\cdot 3+0,04\cdot 4+0,03\cdot 5+0,02\cdot 5+0,01\cdot 5=2,06$$ *двоичных символов/символ исходного сообщения*

Алгоритм Шеннона–Фано не гарантирует построения наиболее

экономного кода, уступая в этом смысле обсуждаемому ниже алгоритму Хаффмена.

3.4. Оптимальное кодирование методом Хаффмана

Это метод двоичного, неравномерного, адаптивного, оптимального кодирования.

1) Вычислить вероятности появления каждого из символов первичного алфавита $A = \{a_1, a_2, \ldots, a_k\}$ исходного сообщения – $p_1, p_2, \ldots, p_k$.

2) Упорядочить символы исходного сообщения по убыванию значений их вероятностей появления в сообщении.

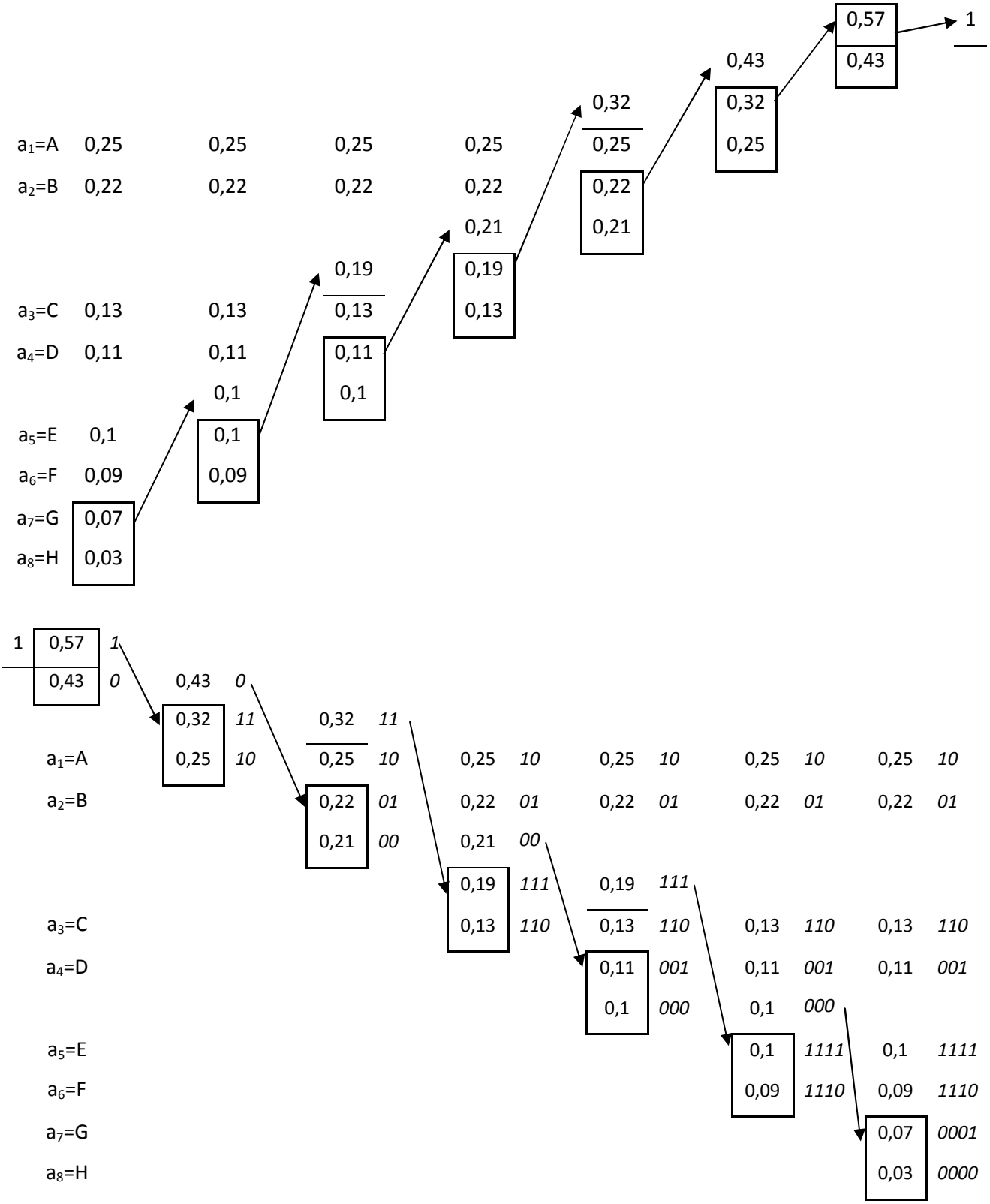
3) Последовательно объединить два символа с наименьшими вероятностями появления в новый составной символ b, вероятность появления которого равна сумме вероятностей составляющих его символов.

4) Вставить этот символ в упорядоченный набор на подходящее место, чтобы упорядоченность значений вероятностей не нарушалась. В результате получится новый алфавит меньшей мощности $A = \{a_1, a_2, \ldots, a_{k-2}, b\}$, вероятности его символов – $p_1, p_2, \ldots, p_{k-2}, p_{k-1}+p_k$.

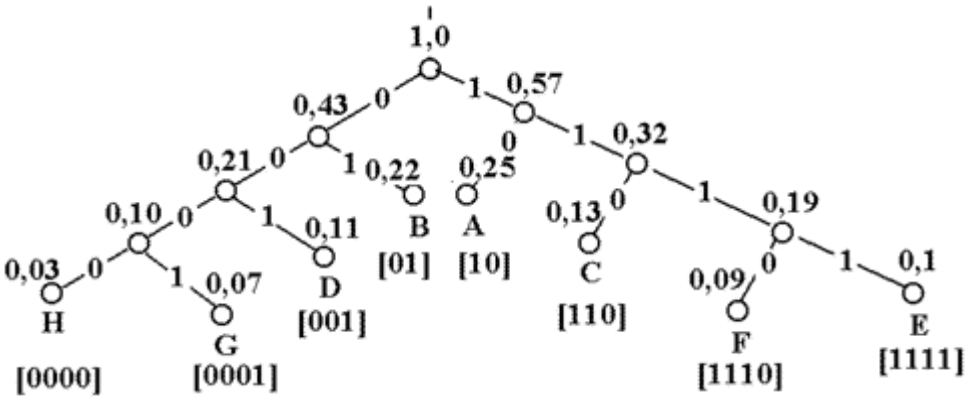
5) Пункт 4) повторять до тех пор, пока в алфавите не останется 1 символ.

6) Построить так называемое *кодовое дерево* и проследить в нем путь от корня к каждому листу дерева. Каждой ветви дерева, выходящей из данного узла, сопоставить символ двоичного алфавита (например, верхней – символ 0, а нижней – 1). Полученная последовательность символов вдоль ветвей от корня дерева до сообщения дает кодовое слово, отвечающее данному символу.

Пример.


a1=A 0,25 0,25 0,25 0,25 0,32 0,43 0,57 1
a2=B 0,22 0,22 0,22 0,22 0,25 0,32 0,43
0,21 0,22 0,25
a3=C 0,13 0,13 0,19 0,19 0,21
a4=D 0,11 0,11 0,13 0,13
0,1 0,1
a5=E 0,1 0,1
a6=F 0,09 0,09
a7=G 0,07
a8=H 0,03
1 0,57 1
0,43 0
0,43 0
0,32 11
0,32 11
a1=A 0,25 10 0,25 10 0,25 10 0,25 10 0,25 10 0,25 10
a2=B 0,22 01 0,22 01 0,22 01 0,22 01 0,22 01
0,21 00 0,21 00
0,19 111 0,19 111
a3=C 0,13 110 0,13 110 0,13 110 0,13 110
a4=D 0,11 001 0,11 001 0,11 001
0,1 000 0,1 000
a5=E 0,1 1111 0,1 1111
a6=F 0,09 1110 0,09 1110
a7=G 0,07 0001
a8=H 0,03 0000

На основании полученной таблицы можно построить кодовое дерево



Так как в процессе кодирования сообщениям сопоставляются только листья дерева, полученный код является префиксным и всегда однозначно декодируем.

Результат кодирования:

Символ	Вероятность	Код
a_1=A	0,25	*10*
a_2=B	0,22	*01*
a_3=C	0,13	*110*
a_4=D	0,11	*001*
a_5=E	0,1	*1111*
a_6=F	0,09	*1110*
a_7=G	0,07	*0001*
a_8=H	0,03	*0000*

В отличие от кода Шеннона–Фано, построение кодовых слов согласно алгоритму Хаффмена осуществляется в обратном порядке, т.е. от конца слова к его началу.

Цена кодирования:

$$C=0,25\cdot2+0,22\cdot2+0,13\cdot3+0,11\cdot3+0,1\cdot4+0,09\cdot4+0,07\cdot4+0,03\cdot4=$$
$$=2,82 \text{ двоичных символов/символ исходного сообщения}$$

4. Решение задач

Пример 1. Считая, что каждый символ кодируется одним байтом, определите, чему равен информационный объем следующего высказывания Жан-Жака Руссо: **Тысячи путей ведут к заблуждению, *к* истине - только один.**

Решение: Подсчитаем количество символов в данном высказывании, включая знаки препинания и пробелы. Получится 57. Для одного символа требуется I байт, или 8 бит.

57 байт = 57 · 8 бит = 456 бит.

Ответ: 456 бит.

Пример 2. Для записи текста используются только строчные буквы русского алфавита и пробел для разделения слов. Какой информационный объем имеет текст, состоящий из 2000 символов?

Решение. В русском алфавите 33 буквы. Сократив его на две буквы (например, «ё» и «й») и введя символ пробела, получаем очень удобное число символов – 32. Используя приближение равной вероятности символов, запишем формулу Хартли:

$$I_1 = \log_2 32 = 5.$$

Таким образом, 5 бит – информационный вес каждого символа русского алфавита. Тогда информационный объем всего текста равен:

$$I = 2000 \cdot 5 = 10\,000 \text{ бит}.$$

Пример 3. Вычислить в килобайтах информационный объем текста размером в 2000 символов, в записи которого использован алфавит компьютерного представления текстов мощностью 256.

Решение. В данном алфавите информационный вес каждого символа равен 1 байту. Следовательно, информационный объем текста равен 2000 байт.

Переведем полученное значение в килобайты:

$$I = \frac{2000}{1024} \approx 1{,}9531 \text{ Кб}.$$

Ответ: ≈ 1,9531 Кб.

Пример 4. Объем сообщения, содержащего 2048 символов, составил $\frac{1}{512}$ часть мегабайта. Какова мощность алфавита, с помощью которого записано сообщение?

Решение. Переведем информационный объем сообщения из мегабайтов в биты:

$$I = \frac{1}{512} \cdot 1024 \cdot 1024 \cdot 8 = 16384 бит.$$

Поскольку такой объем информации несут 2048 символов (K), то на один символ приходится:

$$I_1 = \frac{I}{K} = \frac{16384}{2048} = 8 бит.$$

Отсюда следует, что мощность использованного алфавита равна $2^8 = 256$ символов.

Ответ: 256 символов.

Пример 5. В алфавите племени Мумба-Юмба всего 4 буквы (А, У, М, Б), один знак препинания (точка) и для разделения слов используется пробел. Подсчитали, что в популярном романе «Мумбаюм» содержится всего 10 000 знаков, из них: букв А – 4000, букв У – 1000, букв М – 2000, букв Б – 1500, точек – 500, пробелов – 1000. Какой объем информации содержит книга?

Решение. Поскольку объем книги достаточно большой, то можно допустить, что вычисленная по ней частота встречаемости в тексте каждого из символов алфавита характерна для любого текста на языке племени Мумба-Юмба. Подсчитаем вероятность появления каждого символа алфавита во всем тексте книги и их информационные веса:

$$P_A = \frac{4000}{10000} = 0{,}4; \qquad P_B = \frac{1000}{10000} = 0{,}1;$$

$$P_M = \frac{2000}{10000} = 0{,}2; \qquad P_Б = \frac{1500}{10000} = 0{,}15;$$

$$P_{точка} = \frac{500}{10000} = 0{,}05; \qquad P_{пробел} = \frac{1000}{10000} = 0{,}1.$$

Объем информации, содержащийся в одном символе, вычислим по формуле К. Шеннона:

$$I_1 = -\sum_{i=1}^{4} P_i \cdot \log_2 P_i = -0{,}4 \cdot \log_2 0{,}4 - 0{,}1 \cdot \log_2 0{,}1 - 0{,}2 \cdot \log_2 0{,}2 - 0{,}15 \cdot \log_2 0{,}15 -$$
$$- 0{,}05 \cdot \log_2 0{,}05 - 0{,}1 \cdot \log_2 0{,}1 = 890 \textit{ бит}.$$

Общий информационный объем:

$$I = I_1 \cdot 10000 = 890 \cdot 10000 = 8900000 \text{ бит} = 1{,}061 \text{ Мб}$$

Ответ: 1,061 Мб.

Пример 6. На автобусной остановке останавливаются два маршрута автобусов: № 5 и № 7. Ученику дано задание: определить, какое количество информации содержится в сообщении о том, что к остановке подошел автобус № 5, и какое количество информации содержится в сообщении о том, что подошел автобус № 7.

Решение. Ученик провел исследование. В течение всего рабочего дня он подсчитал, что к остановке автобусы подходили 100 раз. Из них 25 раз подходил автобус № 5 и 75 раз подходил автобус № 7. Сделав предположение, что с такой же частотой автобусы ходят и в другие дни, ученик вычислил вероятность появления на остановке автобуса № 5:

$$p_5 = 25/100 = 1/4,$$

и вероятность появления автобуса № 7:

$$p_7 = 75/100 = 3/4.$$

Отсюда, количество информации в сообщении об автобусе № 5 равно:

$$i_5 = \log_2 4 = 2 \text{ бита}.$$

Количество информации в сообщении об автобусе № 7 равно:

$$i_7 = \log_2(4/3) = \log_2 4 - \log_2 3 = 2 - 1{,}58496 = 0{,}41504 \text{ бит}.$$

Ответ: 0,41504 бит.

Пример 7. На остановке останавливаются автобусы № 5 и № 7. Сообщение о том, что к остановке подошел автобус № 5, несет 4 бита

информации. Вероятность появления на остановке автобуса № 7 в 2 раза меньше, чем вероятность появления автобуса № 5. Сколько бит информации несет сообщение о появлении на остановке автобуса № 7?

Решение. Запишем условие задачи в следующем виде:

$$i_5 = 4 \text{ бита}, \; p_5 = 2 \cdot p_7.$$

Вспомним связь между вероятностью и количеством информации:

$$2^i = \frac{1}{p}.$$

Отсюда: $p = 2^{-i}$.

Подставляя в равенство из условия задачи, получим:

$$2^{i_5} = 2 \cdot 2^{-i_7},$$
$$2^{-4} = 2 * 2^{-i_7} = 2^{1-i_7},$$
$$i_7 - 1 = 4,$$
$$i_7 = 5.$$

Ответ: 5 бит.

Вывод: уменьшение вероятности события в 2 раза увеличивает информативность сообщения о нем на 1 бит. Очевидно, и обратное правило: увеличение вероятности события в 2 раза уменьшает информативность сообщения о нем на 1 бит.

5. Вопросы для контроля и задачи для самостоятельного решения

1. Дайте несколько определений понятию информация. Чем отличаются данные от информации?
2. Охарактеризуйте виды информации и их использование при формировании информационного ресурса.
3. Обоснуйте логарифмическую меру информации.
4. Что такое равновероятные сообщения, и как на их основе получают количественную оценку информации?
5. Что дает свойство аддитивности меры информации при определении количества информации в неравновероятных сообщениях?
6. Докажите, что формула Р. Хартли является частным случаем формулы К. Шеннона.
7. При угадывании целого числа в некотором диапазоне было получено 8 бит информации. Сколько чисел содержит этот диапазон?
8. Происходит выбор одной карты из колоды в 32 карты. Какое количество информации мы получаем в зрительном сообщении о выборе определённой карты?
9. Какое количество информации получит второй игрок при игре в крестики-нолики на поле 8×8, после первого хода первого игрока, играющего крестиками?
10. Какой объем информации содержит сообщение, уменьшающее неопределенность знаний в 4 раза?
11. В корзине лежат 8 шаров. Все шары разного цвета. Сколько информации несет сообщение о том, что из корзины достали красный шар?
12. В классе 30 человек. За контрольную работу по математике получено 6 пятерок, 15 четверок, 8 троек и 1 двойка. Какое количество информации в сообщении о том, что Иванов получил четверку?
13. Известно, что в ящике лежат 20 шаров. Из них 10 – черных, 5 – белых, 4 – желтых и 1 – красный. Какое количество информации несут сообщения о том, что из ящика случайным образом достали черный шар, белый шар, желтый шар, красный шар?
14. За четверть ученик получил 100 оценок. Сообщение о том, что он получил четверку, несет 2 бита информации. Сколько четверок ученик получил за четверть?

15. Информационное сообщение объемом 1,5 Кбайта содержит 3072 символа. Сколько символов содержит алфавит, при помощи которого было записано это сообщение?

16. Для записи текста использовался 256-символьный алфавит. Каждая страница содержит 30 строк по 70 символов в строке. Какой объем информации содержат 5 страниц текста?

17. Сообщение занимает 2 страницы и содержит 1/16 Кбайта информации. На каждой станице записано 256 символов. Какова мощность использованного алфавита?

18. Два сообщения содержат одинаковое количество символов. Количество информации в первом тексте в 1,5 раза больше, чем во втором. Сколько символов содержат алфавиты, с помощью которых записаны сообщения, если известно, что число символов в каждом алфавите не превышает 10, и на каждый символ приходится целое число битов?

19. ДНК человека (генетический код) можно представить себе как некоторое слово в четырехбуквенном алфавите, где каждой буквой помечается звено цепи ДНК, или нуклеотид. Сколько информации (в битах) содержит ДНК человека, содержащий $1{,}5 \cdot 10^{23}$ нуклеотидов?

20. Выяснить, сколько бит информации несет каждое двухзначное число (отвлекаясь от его конкретного числового значения).

21. Имеется 25 монет одного достоинства; 24 из них имеют одинаковый вес, а одна – фальшивая – несколько легче остальных. Спрашивается, сколькими взвешиваниями на чашечных весах без гирь можно обнаружить эту фальшивую монету.

22. Пусть рассматривается алфавит из двух символов русского языка – «К» и «А». Относительная частота встречаемости этих букв равна, соответственно, p_1=0,028, p_2=0,062. Возьмем произвольное слово длины N из k букв «К» и m букв «А». Какое количество информации содержится в таком слове?

23. Два исполнителя – Шалтай и Болтай проставляют 0 или 1 в каждую из имеющихся в их расположении клеточек и таким образом кодируют символы. Шалтай может закодировать 512 символов, и у него на 2 клеточки больше, чем у Болтая. Сколько клеток было в распоряжении Болтая?

24. На остановке останавливаются автобусы с разными номерами. Сообщение о том, что подошел автобус № 1 несет 4 бита информации. Вероятность появления автобуса № 2 в два раза меньше, чем вероятность появления автобуса № 1. сколько информации несет сообщение о появлении автобуса № 2?

25. Алфавит племени состоит из 4 букв – a, b, c, d. В тексте из 64 встречаются: «a» – 32 раза, «b» – 16, «c» – 8, «d» – 8. Какое количество информации несет в себе любое слово из этого текста, составленное из 5 букв?

26. Цветное растровое графическое изображение, палитра которого включает в себя 65536 цветов, имеет размер 100×100 точек. Какой объем памяти в килобайтах занимает это изображение?

27. В корзине 32 шара: 2 белых, 16 красных, 4 синих, 8 черных, 2 зеленых. Какое количество информации несет сообщение о том:

a. какого цвета достали шар;
b. что достали красный шар.

28. В школьной библиотеке 16 стеллажей с книгами. На каждом стеллаже 8 полок. Библиотекарь сообщил Пете, что нужная ему книга находится на пятом стеллаже на третьей сверху полке. Какое количество информации библиотекарь передал Пете?

29. Сообщение записано в виде десятичного числа из 5 цифр, причем предполагается, что все цифры равновероятны и независимы. Какое количество информации несет это сообщение? Во сколько раз меньшее количество информации содержало бы сообщение, состоящее из 5 двоичных цифр?

30. Считая, что каждый символ кодируется одним байтом, определите, чему равен информационный объем следующего высказывания Алексея Толстого: «*Не ошибается тот, кто ничего не делает, хотя это и есть его основная ошибка*».

Литература

1. Большой энциклопедический словарь: в 2 т. / гл. ред. А. М. Прохоров. – М. : Сов. энцикл., 1991. – 2 т.

2. Гуров В. В., Чуканов В. О. Основы теории и организации ЭВМ: учебное пособие. – М: БИНОМ. Лаборатория знаний. Интернет–Университет Информационных технологий, 2006.

3. Могилев А. В., Пак Н. И., Хеннер Е. К. Информатика: учебное пособие для вузов. – М.: Академия, 2004.

4. Острейковский В. А. Информатика: учебник. – М.: Высшая школа, 2000.

5. Стариченко Б. Е. Теоретические основы информатики. – М.: Горячая линия – Телеком, 2004.