**Измерение информации**

**Основной теоретический материал**.

***Основные понятия темы:***

* единицы измерения информации;
* содержательный (вероятностный) и алфавитный подходы к измерению информации;
* алфавит, мощность алфавита, информационный вес символа алфавита;
* формула Шеннона, формула Хартли;
* информационный объем сообщения, информационного объекта.

***В результате изучения данной темы вы должны будете:***

***знать:***

* названия единиц измерения информации;
* формулы расчета количества информации;

***уметь:***

* переводить количество информации из одних единиц в другие;
* определять информационный объем информационного объекта.

*Для учащихся основной школы (8-9 классы) для сдачи ОГЭ достаточно знать единицы измерения информации и уметь рассчитывать количество информации с точки зрения алфавитного подхода. В ОГЭ содержится одно задание по этой теме. В ЕГЭ два задания: одно базового уровня и одно повышенного уровня.*

**Вероятностный и алфавитный подходы к измерению информации**

Измерение информации является одной из важнейших задач теоретической информатики. В курсе информатике рассматриваются два подхода к измерению информации: содержательный (вероятностный) и алфавитный.

**Содержательный (Вероятностный) подход**

Вероятностный подход используется для определения количества информации, содержащейся в сообщении о каком-либо объекте или событии и связан с содержанием информационного сообщения.

Он основан на следующих соображениях:

* те или иные события имеют некоторую вероятность (возможность произойти или не произойти);
* событие, которое совершается всегда, имеет вероятность равную единице (например, восход Солнца); событие, которое не совершается никогда, имеет вероятность равную 0 (например, восход Солнца на западе); в остальных случаях вероятность совершения события есть дробное число от 0 до 1;
* получая сообщение о совершении (или не совершении) некоторого события, мы получаем некоторое количество информации, которое определяется снятой с ее помощью неопределенностью наших знаний об указанном событии:
	+ если вероятность совершения события точно равна 1 или 0 (т.е. мы точно знаем, что событие произойдет (или не произойдет), то никакой неопределенности в наших знаниях нет, и сообщение о таком событии несет нулевое количество информации);
	+ для равновероятных событий чем больше их количество (т.е шире возможный выбор вариантов и потому меньше вероятность каждого из них), тем большее количество информации несет сообщение о совершившемся конкретном событии или можно сформулировать так: чем меньше вероятность некоторого события, тем больше информации содержит сообщение об этом событии;
	+ количество информации в сообщении о совершении (не совершении) *нескольких* независимых событий равно *сумме* количеств информации, содержащейся в сообщениях о каждом отдельном таком событии.

**Вычисление количества информации для**

**равновероятных событий**

Сообщение о том, что произошло одно событие из двух равновероятных, несёт 1 бит информации. Сообщение о том, что произошло одно событие из четырёх равновероятных, несёт 2 бита информации. Откуда взялись именно такие величины?

Р.Хартли в 1928 г. сформулировал законы, которым должно подчиняться количество информации:

1) Если сообщение несет заранее известную информацию, количество информации

равно нулю.

2) Чем больше количество возможных вариантов событий, тем больше информации

содержится в сообщении о наступлении конкретного события.

3) Количество информации в сообщении о нескольких независимых событиях

должно быть равно сумме количеств информации, содержащейся в сообщениях о каждом из этих событий.



Количество информации ***i***, содержащееся в сообщении о том, что произошло одно из **N** равновероятных событий, определяется из решения показательного уравнения (**формулы Хартли**) **2*i*=N**, где ***i***=0, 1, 2 ...

Решение этого уравнения выглядит как ***i*=log2N,**

где log – *функция логарифма по основанию 2*, обратная возведению значения основания логарифма в степень, равную I, т.е. из формулы Хартли следует зависимость

**N = 2i**

Для облегчения вычислений для значений N, представляющих собой степени числа 2, можно составить таблицу

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| N | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |
| I (бит) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

Для значений N, не равных степени двойки, при определении количества информации в битах из вышеприведенной таблицы берется ближайшее большее значение N, равное степени 2. Например, для 48 равновозможных событий, количество информации, которое содержится в сообщении о совершении конкретного события, принимается равным 6 бит (так как ближайшее большее значение, равное степени числа 2 равно 64).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Сводная таблица исчисления количества информации для равновероятных событий | Объект с двумя гранями (монетка) | Объект с четырьмя гранями (тетраэдр) | Объект с шестью гранями (куб) | Объект с восемью равновероятными состояниями (шляпа с разноцветными шариками внутри) |
| Неопределённость знаний о результате N | 2 | 4 | 6 | 8 |
| Вероятность события 1/N | 1/2 | 1/4 | 1/6 | 1/8 |
| Количество информации в сообщении о событии i | 1 бит | 2 бита | 2 бита < i < 3 бита | 3 бита |
| Формула Хартли | 2¹=2log22=1 | 2²=4log24=2 | 22,584962501...=6log26≈2,584962501 | 2³=8log28=3 |

Из таблицы видно, что когда число возможных исходов событий равно целым степеням двойки, показательное уравнение решается в уме. Если же число возможных исходов событий не равно целым степеням двойки, то приходится пользоваться таблицей логарифмов.

**Рассмотрим примеры заданий на вычисление количества информации равновероятных событиях**

**Задание 1**

«Вы выходите на следующей остановке?» – спросили человека в автобусе. «Нет», – ответил он. Какое количество информации содержит его ответ?

***Решение:***

Человек мог ответить только «Да» и «Нет», т.е выбрать один ответ из двух возможных, поэтому N=2, значит 2=2I, (можно записать 21=2I), откуда I=1

Ответ: 1 бит.

**Задание 2**

Группа школьников пришла в бассейн, в котором 4 дорожки для плавания. Тренер сообщил, что группа будет плавать на дорожке номер 3. Сколько информации получили школьники из этого сообщения?

***Решение:***

Из 4 дорожек необходимо выбрать одну, т.е. N = 4. Значит I = 2, т.к. 4 = 22.

Пояснение: номер дорожки (3) не влияет на количество информации, так как вероятности событий в этих задачах мы приняли считать одинаковыми.

Ответ: 2 бита.

**Вычисление количества информации для неравновероятных событий.**

 **Формула Шеннона**

При измерении количества информации Р.Хартли не учитывал вероятность наступления события. К.Шеннон в своих работах предложил учитывать вероятность наступления события при измерении информации. Основная идея заключалась в том, что сообщение о наступлении маловероятного события несет большее количество информации, чем сообщение о наступлении более вероятного.

**Если N – общее число возможных событий, из них интересующее нас событие может произойти K раз, то вероятность этого события можно оценить как K/N.**

Вероятность выражается в долях единицы. Вероятность достоверного события равна 1. Достоверным называют событие, которое обязательно произойдет. Например, в корзине 30 шаров, все они красного цвета. Событие «из корзины вынут красный шар» достоверное. Вероятность невозможного события равна нулю (из корзины с 30-ю красными шарами вынут белый шар).

Допустим, контрольную работу писал отличник. Пусть вероятность того, что он получит оценку 5, равна 1/2, оценку 4 – 1/4, оценки 3 и 2 – по 1/81. Для каждого возможного сообщения в отдельности формула Хартли справедлива и должна применяться. Если вероятность некоторого j-го события равна pj, то количество информации о наступлении этого события равно –log2 pj. Таким образом, сообщение о том, что ученик получил оценку 5, несет –log2(1/2) = 1 бит, оценку 4 –log2(1/4) = 2 бита, оценку 2 или 3 –log2(1/8) = 3 бита информации.

Средневзвешенное количество информации, которая может быть получена, определяется по формуле I = 1/2 \* 1 бит + 1/4 \* 2 бита + 1/8 \* 3 бита + 1/8 \* 3 бита = 1,75 бита.

В общем случае среднее количество информации, получаемой при неравновероятных событиях, определяется по формуле Шеннона:



Эту формулу можно расписать в таком виде:



Если все эти события равновероятны, т.е. р1= р2=……=рN = р, то очевидно, что формула Шеннона преобразуется в формулу Хартли, которая является частным случаем формулы Шеннона.

**Связь между количеством информации и вероятностью события.**

Для N равновероятных событий **вероятность одного отдельного события р=1/N**.

С учетом этого формула Хартли может быть преобразована в соотношение:

**I = log2 (1/р)**

В этом случае вычисление количества информации можно производить по данным таблицы, представленной выше, предварительно вычислив значение N как величину, обратную значению р.

Например, для события, вероятность которого (р) составляет 0,018, получается N=1/0,018 = 55,56, тогда берется ближайшее большее значение N, кратное 2 (N=64) и по таблице определяем, что I = 6 бит.

**Рассмотрим примеры заданий на вычисление количества информации при неравновероятных событиях**

**Задание 3**

Синдбад-мореход в числе 10 других купцов плывёт в дальние страны на корабле с командой из 40 человек. Корабль захватили пираты, разграбили груз, разделили одежды пленников. Их привели на рынок невольников для продажи на галеры. Торг только начался. Какова вероятность пиратам продать на рынке невольников бывшего моряка? Велика ли вероятность пиратам продать на галеры бывшего купца? С какой вероятностью на галеры продадут именно Синдбада?

***Решение***

Обозначим вероятность (по-английски *probability*) продажи бывшего купца **pm**, а вероятность продажи бывшего моряка **ps**. Тогда **pm**=10/50=0,2, а **ps**=40/50=0,8. **pm**/**ps**=1/4.Значит, вероятность продажи бывшего купца в 4 раза меньше вероятности продажи бывшего моряка.

Все возможные продажи 50 пленников составят тогда 0,2+0,8=1. Значит, поскольку Синдбад – один из 50 пленников, вероятность, что на галеры продадут именно его, равна 1/50.

**Задание 4**

В озере обитает 12500 окуней, 25000 пескарей, а карасей и щук по 6250. Сколько информации мы получим, когда поймаем какую-нибудь рыбу.

***Решение:***

1. Найдем общее количество рыб в озере: S = 12500 + 25000 + 2\*6250 = 50000.
2. Найдем вероятность попадания на удочку каждого вида рыб:

Ро= 12500/50000 = 0,25,

Pк = 25000 /50000 = 0,5,

Pп =6250/50000 = 0,125,

Pщ = 6250/50000 = 0,125.

3. Найдем количество информации:

**i** = - (0,25\*log20,25 + 0,5\*log20,5 + 0,125\*log20,125 + 0,125\*log20,125) = - (0,25\*(-2) + 0,5\*(-1) + 0,125\*(-3) + 0,125\*(-3)) = -(-0,5-0,5-0,375-0,375)= -(-1,75)= 1,75 бит.

 *Ответ:* 1,75 бит информации.

**Единицы измерения информации**

Принято считать, что минимально возможное количество информации соответствует такому сообщению, получение которого уменьшает неопределенность в 2 раза (пример, орел-решка). Это минимальное количество информации получило название «**бит»**.

В вычислительной технике бит соответствует одному *двоичному разряду*, который может принимать одно из двух возможных значений: 0 или 1. в качестве более крупной величины принят **байт**, соответствующий двоичному числу из 8 разрядов (битов). В оперативной памяти компьютера минимальный объем ячейки памяти, выделяемой для хранения какой-либо величины, как правило, равен 1 байту.

Для обозначения количества информации, больших чем байт, приняты следующие производные единицы:

1 Килобайт (Кбайт) = 210 байт = 1024 байт,

1 Мегабайт (Мбайт) = 210 Кбайт = 1024 Кбайт,

1 Гигабайт (Гбайт) = 210 Мбайт = 1024 Мбайт,

1 Терабайт (Тбайт) = 210 Гбайт = 1024 Гбайт,

1 Петабайт (Пбайт) = 210 Тбайт = 1024 Тбайт,

1 Экзабайт (Эбайт) = 210 Пбайт = 1024 Пбайт.

**! Внимание**

*В отличие от одноименных приставок в кратных величинах в математике, изменение величин в вычислительной технике происходит на каждом шаге вышеуказанной шкалы на* ***210=1024****, а не на 103=1000.*

**Алфавитный подход**

При ***алфавитном подходе*** к определению количества информации отвлекаются от содержания (смысла) информации и рассматривают информационное сообщение как последовательность знаков определенной знаковой системы.

Проще всего разобраться в этом на примере текста, написанного на каком-нибудь языке. Для нас удобнее, чтобы это был русский язык.

Все множество используемых в языке символов будем традиционно называть алфавитом. Обычно под алфавитом понимают только буквы, но поскольку в тексте могут встречаться знаки препинания, цифры, скобки, то мы их тоже включим в алфавит. В алфавит также следует включить и пробел, т.е. пропуск между словами.

**Алфавит – это множество символов, используемых при записи информации.**

**Мощность (размер) алфавита – это полное количество символов в алфавите.**

Мощность алфавита обозначается буквой **N.**

*Например:*

* мощность алфавита из русских букв равна 33;
* мощность алфавита из латинских букв – 26;
* мощность алфавита текста набранного с клавиатуры равна 256 (строчные и прописные латинские и русские буквы, цифры, знаки    арифметических операций, скобки, знаки препинания и т.д.);
* мощность двоичного алфавита равна 2.

При алфавитном подходе считается, что каждый символ текста имеет информационный вес. Информационный вес символа зависит от мощности алфавита.

Алфавит, с помощью которого записано сообщение состоит из **N** знаков. В простейшем случае, когда длина кода сообщения составляет один знак, отправитель может послать одно из N возможных сообщений, которое будет нести количество информации **I**.

Тогда в формуле N = 2i ,где

**N** – количество знаков в алфавите знаковой системы, **I** – количество информации, которое несет каждый знак*.*

*Например*, из формулы можно определить количество информации, которое несет знак в двоичной знаковой системе



Информационная емкость знака двоичной знаковой системы составляет 1 бит.

А теперь давайте попробуем определить, какое количество информации несет буква русского алфавита (без буквы ё).

Решение:



Буква русского алфавита несет 5 битов информации.

Формула связывает между собой количество возможных событий и количество информации, которое несёт полученное сообщение. В рассматриваемой ситуации *N* – это количество знаков в алфавите, знаковой системы, а *I* – количество информации, которое несёт один знак.

Сообщение состоит из последовательности знаков, каждый из которых несет определенное количество информации. Количество информации в сообщении можно посчитать, умножив количество информации, которое несет один знак на количество знаков в сообщении.

,

где Ic – количество информации в сообщении

 – количество информации, которое несет один знак

 – количество знаков в сообщении

**! Внимание**

*Мощность алфавита определяется не набором знаков, используемых в конкретном сообщении, а количеством знаков, которые вообще могут быть использованы в сообщениях, кодируемых в соответствии с данным алфавитом.*

**Алгоритм определения количества информации в сообщении**

1. Определяется мощность используемого алфавита N
2. Определяется количество информации, приходящееся в алфавите на один его знак
	1. если использование всех знаков равновероятно, то используется формула Хартли
	2. если известны вероятности использования тех или иных знаков (на основе составленной таблицы частоты встречаемости этих знаков), то используется формула Шеннона.

3. Вычисленное количество информации (I), приходящееся на один знак умножается на количество знаков в сообщении.

Ic = I\*К

**Для решения различных задач на определение количества информации вы должны помнить следующее:**

* **отношение количества вариантов (или чисел) N к количеству информации I, которую несет в себе один из вариантов: N=2I**
* **полный информационный объем сообщения V равен количество символов в сообщении K умноженное на количество информации на каждый символ I: V=K\*I**
* **Формула Шеннона для равновероятных событий: I=log2N**
* **если алфавит имеет мощность ( количество символов в этом алфавите) М, то количество всех возможных «слов» (символьных цепочек) длиной *N*(без учета смысла) равно K=MN; для двоичного кодирования (мощность алфавита *M*–2 символа) получаем известную формулу:K=2N**

**Таблица степеней двойки, покажет сколько вариантов можно закодировать с помощью N бит:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **N бит** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **K вариантов** | **2** | **4** | **8** | **16** | **32** | **64** | **128** | **256** | **512** | **1024** |

**Рассмотрим примеры заданий на вычисление количества информации с точки зрения алфавитного подхода**

**Задание 5.**

Какое количество информации содержит слово «ПРИВЕТ», если считать, что алфавит состоит из 32 букв?

***Решение.***

Что нам требуется найти в данной задаче? Нам нужно найти какое количество информации содержит слово «ПРИВЕТ».

Что нам для этого дано?

Дано: количество знаков в сообщение и мощность алфавита.

Количество знаков в сообщении равно 6, а мощность данного алфавита равна 32.

Что нам нужно найти? Нам нужно найти какое количество информации содержит слово «ПРИВЕТ».

Посмотрим на наше сообщение, оно содержит несколько знаков, значит для того чтобы найти количество информации нашего сообщения, нам нужно умножив количество информации, которое несет один знак, на количество знаков в сообщении, т.е. воспользоваться формулой Ic = I\*К.

Но мы еще не можем воспользоваться формулой, т.к. не знаем, какое количество информации несет один знак. Для этого воспользуемся формулой Хартли. Сообщение записано с помощью алфавита, мощность которого равна 32 (без буквы ё), т.е. N равно 32. Мы получили уравнение. 32=2I. 25=2I. Решив это уравнение, мы получили, что количество информации, которое несет один знак нашего алфавита, равно 5 бит. Зная количество информации, которое несет один знак нашего алфавита, и количество знаков в сообщении, мы можем найти какое количество информации содержит наше сообщение.

Итак, наше сообщение содержит 30 бит.

Ответ 30 бит.

**Задание 6**

Мощность алфавита равна 256. Сколько Кбайт памяти потребуется для сохранения 160 страниц текста, содержащего в среднем 192 символа на каждой странице?

Решение:

1. 256 различных символов закодированы 256 кодами, по формуле 2i=N определим кол-во бит соответствующее одному символу 2i=256, по таблице степеней числа 2 определяем, что i=8 бит
2. I=i\*(160\*192), где (160\*192) – это кол-во символов во всем тексте
I=8 бит\*160\*192=8\*30720=245760
3. Чтобы перейти к Кбайтам эту величину надо поделить на 8, (т.к. 8 бит = 1 байт) и поделить на 1024 (т.к. 1 Кбайт = 1024 байта)
Ответ 30 Кбайт

**Вопросы и задания**

*1. Что такое алфавит?
2. Что такое мощность алфавита?
3. Как определяется информационный объем текста при использовании алфавитного подхода?
4. Текст составлен с использованием алфавита мощностью 64 символа и содержит 100 символов. Каков информационный объем текста?
5. Что такое байт, килобайт, мегабайт.
6. Информационный объем текста, подготовленного с помощью компьютера, равен 3,5 Кб. Сколько символов содержит этот текст?
7. Два текста содержат одинаковое количество символов. Первый текст составлен в алфавите мощностью 32 символа, второй — мощностью 64 символа, Во сколько раз отличаются информационные объемы этих текстов?*