ЧАСТНОЕ ПРОФЕССИОНАЛЬНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ «СЕВЕРО-КАВКАЗСКИЙ КОЛЛЕДЖ ИННОВАЦИОННЫХ ТЕХНОЛОГИЙ»

ЛЕКЦИЯ 4. ТЕМА 2.1. ПОДХОДЫ К ПОНЯТИЮ И ИЗМЕРЕНИЮ ИНФОРМАЦИИ

1. Понятие информации: многоаспектность термина

Термин «информация» (от лат. *informatio* — разъяснение, изложение) принадлежит к числу фундаментальных понятий современной науки, однако строгое универсальное определение ему не дано. В различных областях знания содержание термина уточняется по-разному, что порождает ряд взаимодополняющих подходов.

Содержательный (вероятностный) подход. В 1948 году Клод Элвуд Шеннон в работе «Математическая теория связи» предложил рассматривать информацию как снятую неопределённость. Количественная мера информации связывается с вероятностью события: чем менее вероятно сообщение, тем больше информации оно несёт. Основой теории служит формула для энтропии дискретного источника:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i,$$

где p_i — вероятность i-го сообщения. При равновероятных событиях ($p = 1/n$) энтропия максимальна и равна $\log_2 n$. В частном случае равновероятного выбора из двух исходов единицей информации является **бит** (binary digit) — количество информации, снимающее неопределённость выбора одного из двух равновозможных вариантов.

Алфавитный (объёмный, технический) подход. В вычислительной технике количество информации оценивается не через смысл сообщения, а через его объём в знаках некоторого алфавита. Если алфавит содержит N символов, то информационный вес одного символа определяется соотношением:

$$i = \log_2 N \quad (\text{бит}),$$

а объём сообщения из K символов составляет $I = K \cdot i$. Данный подход операционален и потому лежит в основе инженерных расчётов ёмкости носителей и каналов связи.

Комбинаторный подход. Ральф Хартли (1928) предложил измерять информацию числом возможных равновероятных состояний: если некоторая система может принимать одно из N*N* состояний, количество информации равно log2*N*. Комбинаторный подход исторически предшествовал вероятностному и близок к алфавитному.

Алгоритмический (колмогоровский) подход. А. Н. Колмогоров (1965) ввёл меру информации как минимальную длину программы (в битах), позволяющей однозначно восстановить объект (сообщение). Данная мера, известная как колмогоровская сложность, инвариантна относительно выбора универсальной машины с точностью до аддитивной константы и позволяет оценивать сложность индивидуальных объектов, а не только множеств сообщений.

Семантический и прагматический подходы. Ю. И. Шнейдер (Шнейдер—Клини? — уточнение не требуется; далее следуют работы А. А. Харкевича и других исследователей) — точнее, семантическая теория информации развивается с 1960-х годов в работах Р. Карнапа, Й. Бар-Хиллела, Ю. И. Шнейдера и других. Семантический подход оценивает содержание сообщения, прагматический — его полезность для достижения цели получателя. Оба направления сохраняют преимущественно теоретическое значение, тогда как в технике применяются объёмная и вероятностная меры.

2. Информационные объекты различных видов

Информационный объект — совокупность данных, оформленная как единое целое и пригодная для хранения, обработки и передачи средствами вычислительной техники. Информационные объекты классифицируются по форме представления данных.

Числовая информация. Числа представляются в позиционных системах счисления; внутри ЭВМ используется двоичная система с представлением целых чисел (беззнаковых и со знаком — в прямом, обратном и дополнительном кодах) и вещественных чисел в формате с плавающей точкой (стандарт IEEE 754).

Текстовая информация. Каждый символ кодируется целым числом. Исторически применялись однобайтовые кодировки (ASCII, KOI-8, Windows-1251), позднее вытесненные универсальной кодировкой Unicode, в которой распространены форматы UTF-8 (переменная длина, 1–4 байта на символ) и UTF-16.

Графическая информация. Растровое изображение представляет собой матрицу пикселей, каждому из которых сопоставлен код цвета; при глубине цвета i*i* бит на пиксель число различимых цветов равно 2i. Векторное изображение описывается математическими объектами (отрезками, кривыми Безье) и их атрибутами.

Звуковая информация. При аналого-цифровом преобразовании сигнал дискретизируется по времени и квантуется по уровню; качество оцифровки определяется частотой дискретизации и разрядностью квантования. Согласно теореме Котельникова (Найквиста—Шеннона), частота дискретизации должна не менее чем вдвое превышать верхнюю частоту спектра сигнала.

Видеоинформация. Представляет собой последовательность кадров растровой графики со звуковой дорожкой; объём данных сокращается методами сжатия (кодеки H.264, H.265 и др.).

Общим для всех перечисленных форм является их **двоичное кодирование**: любой информационный объект в памяти компьютера существует как последовательность нулей и единиц.

3. Универсальность дискретного (цифрового) представления информации

Компьютер — универсальное устройство обработки информации, что достигается за счёт **дискретного (цифрового) представления** любых данных. Дискретизация — преобразование непрерывного сигнала в последовательность дискретных значений. Сигнал называется дискретным, если его параметр принимает конечное число значений.

Универсальность двоичного представления обусловлена следующими свойствами:

1. **Технологическая простота реализации.** Двум значениям (0 и 1) соответствуют два устойчивых физических состояния элемента: наличие или отсутствие тока (напряжения), намагниченность или её отсутствие, наличие или отсутствие отражения света на оптическом носителе. Реализация элементов с двумя устойчивыми состояниями технически надёжна и дёшева.
2. **Помехоустойчивость.** Различение двух состояний требует меньшего запаса по помехам, чем различение множества состояний, что повышает достоверность передачи и хранения данных.
3. **Алгебраическая полнота.** Над двоичными переменными определён аппарат булевой алгебры (И, ИЛИ, НЕ), позволяющий строить любые логические устройства.
4. **Универсальность кодирования.** Любая форма информации — числа, текст, изображение, звук, видео — представима последовательностью бит. Благодаря этому один и тот же процессор обрабатывает данные различной природы, а единый канал связи передаёт разнородные сообщения.

Переход к многозначным (например, троичным) системам исследуется, однако двоичное представление сохраняет статус промышленного стандарта благодаря экономичности и отработанности элементной базы.

4. Представление информации в двоичной системе счисления

Система счисления — совокупность приёмов и правил записи чисел с помощью определённого набора символов (цифр). Системы счисления делятся на непозиционные (римская) и позиционные, в которых значение цифры зависит от её позиции в числе. В позиционной системе с основанием q число A представляется в виде:

$$A = \sum_{i=-m}^{n} a_i q^i,$$

где a_i — цифры алфавита системы, $0 \le a_i < q$.

Двоичная система ($q = 2$) использует алфавит {0, 1}; разряд называется битом. Достоинствами системы являются минимальный алфавит, простота аппаратной реализации и прямое соответствие булевым функциям.

Перевод числа из десятичной системы в двоичную осуществляется последовательным делением на 2 с фиксацией остатков; остатки, записанные в обратном порядке, образуют двоичную запись числа. Например:

$$19_{10} = 10011_2,$$

поскольку $19 = 16 + 2 + 1 = 1 \cdot 2^4 + 0 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0$.

Перевод из двоичной системы в десятичную выполняется суммированием весов разрядов:

$$10011_2 = 1 \cdot 2^4 + 0 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0 = 19_{10}.$$

Системы с основанием 2^n. Поскольку $8 = 2^3$ и $16 = 2^4$, между двоичной и восьмеричной (шестнадцатеричной) системами существует простое соответствие: каждая триада (тетрада) двоичных разрядов заменяется одной цифрой. Так, $1101\,1011_2 = \mathrm{DB}_{16}$. Шестнадцатеричная система (алфавит 0–9, A–F) широко применяется как компактная форма записи двоичных кодов, например адресов памяти и машинных команд.

Единицы измерения. Минимальная единица — бит; группа из 8 бит образует **байт**. Для кратных единиц используются десятичные приставки СИ (1 Кбайт = 1000 байт по стандарту МЭК 80000-13) и двоичные приставки (1 КиБ = 1024 байта); в образовательной практике традиционно применяется соотношение 1 Кбайт = 210210 = 1024 байта.

Представление чисел в памяти ЭВМ. Целые числа без знака хранятся в прямом двоичном коде; целые со знаком — в дополнительном коде, что позволяет сводить вычитание к сложению. Вещественные числа хранятся в формате с плавающей точкой: знак, мантисса, порядок (стандарт IEEE 754: одинарная точность — 32 бита, двойная — 64 бита). Дискретность представления порождает принципиальное ограничение: большинство вещественных чисел непредставимо точно, что проявляется в погрешностях машинных вычислений.

Многообразие подходов к понятию информации — вероятностного, объёмного, комбинаторного, алгоритмического, семантического — отражает многоаспектность понятия и обеспечивает его применимость в различных областях: от теории связи до теории вычислимости. Техническая реализация обработки информации основана на универсальности дискретного двоичного представления: любые данные — числовые, текстовые, графические, звуковые — сводятся к последовательностям бит, что и определяет единство архитектуры современных вычислительных систем.