Курс лекций «Введение в ИИ.». Часть I. От психофизиологической проблемы до экспертных систем. Лекция 4. Информация, данные, знания

О.Г. Чанышев

Содержание

1 Информация

Информация - фундаментальная характеристика бытия, аксиоматическое понятие, такое, как материя, энергия, пространство-время и которое не может быть сведено к каким-либо более общим категориям.

Согласно Философскому энциклопедическому словарю (ФилЭС) [1], слово «информация» происходит от латинского слова informatio - ознакомление, разъяснение, представление, понятие и может обозначать:

1)сообщение, осведомление о положении дел, сведения о чем-либо, передаваемое людьми;

2)уменьшаемую, снимаемую неопределенность в результате получения сообщений;

3) сообщение, неразрывно связанное с управлением, сигналы в единстве синтаксических, семантических и прагматических характеристик;

4) передачу, отражение разнообразия в любых объектах и процессах (живой и неживой природы).

1.1 Определение количества информации по Хартли и Шеннону

Развитие средств связи (телефон, телеграф, радио) и в начале XX в. потребовало численных методов исследования характеристик трактов передачи сообщений, отвлекаясь от смысла сообщений.

1.1.1 Формула Хартли

Понятия неопределенности и вероятности взаимно обратимы. Выбор одного или нескольких вариантов из множества уменьшает неопределенность. Пусть некоторое событие имеет m равновероятных исходов. Таким событием может быть, например, появление любого символа из алфавита, содержащего m таких символов. Количество информации, которое может быть передано при помощи такого алфавита можно измерить, определив число N возможных сообщений, которые могут быть переданы при помощи этого алфавита. Если сообщение содержит n символов (n - длина сообщения), то N=mn. Для того, чтобы удовлетворить естественным требованиям равенства информации нулю при m=1 и чтобы количество информации, получаемое от двух независимых источников было равно сумме "информаций", Хартли предложил считать количество информации, приходящееся на одно сообщение, равным логарифму общего числа возможных сообщений:

$$I(N) = \log N$$

Если возможность появления любого символа алфавита равновероятна, то эта вероятность p = 1/m. Полагая, что N = m, получим:

$$I = \log N = \log m = \log \frac{1}{p} = -\log p$$

Количество информации на каждый равновероятный сигнал равно минус логарифму вероятности отдельного сигнала. Чем меньше вероятность получения сообщения, тем более оно информативно.

За единицу количество информации приняли ее количество, получаемое при выборе одного из двух взаимоисключающих вариантов. Для этого в формуле (2) следует взять логарифм по основанию 2. Тогда

$$I = -\log_2 p = -\log_2 \frac{1}{2} = log_2 2 = 1\ BIT(Binary\ uniT)$$

1.1.2 Определение информации Шенноном

На практике при определении количества информации необходимо учитывать как количество различных сообщений от источника, так и разную вероятность их получения.

Пусть имеем на достаточно длинном отрезке сообщения k элементарных различных сигналов в количестве $N_1, N_2.., N_k$, $\sum N_i = N$. $f_1, f_2, ..., f_k$ - частоты соответствующих сигналов ($f_i = \frac{Ni}{N}$). При возрастании длины отрезка сообщения каждая из частот стремится к фиксированному пределу, т.е.

$$\lim f_i = p_i, (i = 1, 2, ..., k),$$

и p_i можно считать вероятностью сигнала. Полное количество информации, доставаемое N сигналами, будет примерно равно (при достаточно большом N)

$$-N \sum_{i}^{k} p_i \log_2 p_i$$

Чтобы определить среднее количество информации, приходящееся на один сигнал, нужно это число разделить на N. В результате будет получено асимптотическое соотношение - формула Шеннона

В случае равной вероятности сигналов, формула формула Шеннона переходит в формулу Хартли.

Поскольку не всегда возможно установить перечень состояний системы и вычислить их вероятности, а также в силу ограниченности шенноновской теории только синтаксической стороной сообщения, были выдвинуты иные концепции и толкования понятия «информация».

1.2 Концепция разнообразия Эшби

Р. Эшби считал, что информация есть там, где есть неоднородность (разнообразие) и единицей измерения может быть различие между объектами в одном определенном свойстве. Чем больше различий, тем больше информации. Под разнообразием следует подразумевать характеристику степени несовпадения элементов некоторого множества. Если единицей обозначить факт одинаковости элементов, то логарифм единицы - ноль. Это будет соответствовать единичной вероятности выбора элемента множества, поскольку элементы неразличимы. По Эшби теория информации изучает «процессы передачи разнообразия» по каналам связи.

1.3 Алгоритмическое измерение количества информации - Колмогоров

Близка к «разнообразностной» идея алгоритмического измерения количества информации, выдвинутая в 1965 г. А.Н. Колмогоровым. Количество информации определяется как минимальная длина программы, позволяющей преобразовать один объект (множество) в другой (множество). Чем больше различаются два объекта между собой, тем сложнее (длиннее) программа перехода от одного объекта к другому. Длина программы при этом измеряется количеством команд. Этот подход, в отличие от подхода Шеннона, не базирующийся на понятии вероятности, позволяет, например, определить прирост количества информации, содержащейся в результатах расчета, по сравнению с исходными данными.

Вероятностная теория информации на этот вопрос не может дать удовлетворительного ответа. (См. также Алгоритмическая теория информации, Алгоритма сложность [2]).

1.4 Информация как мера неоднородности по Глушкову

Если понятие информации связывать с разнообразием, то причиной существующего в природе разнообразия, по мнению академика В.М. Глушкова, можно считать неоднородность в распределении энергии (или вещества, поскольку $E = m \times C^2$) в пространстве и во времени. Информация же есть мера этой неоднородности. Информация существует постольку, поскольку существуют сами материальные тела. С понятием информации в кибернетике не связано свойство ее осмысленности. Звезды существуют независимо от того, имеют люди информацию о них или нет. Объективное существование объекта создает неоднородность в распределении

вещества и поэтому является источником информации для когнитивной системы. Таким образом, по В.М. Глушкову, информация независима от нашего сознания. Но, в таком случае, следует ли вводить понятие информации для описания поведения объектов неживой природы?

1.5 Семантические и ценностные аспекты информации

Рассмотренные выше определения и толкования понятия «информации» в принципе не могут учесть ее содержательного и ценностного аспектов.

Попытки оценить не только количественную, но и содержательную сторону информации дали толчок к развитию семантической (смысловой) теории информации. Исследования в этой области теснее всего связаны с семиотикой - теорией знаковых систем. Одним из важнейших свойств информации, которое мы можем наблюдать, является ее неотделимость от носителя: во всех случаях, когда мы сталкиваемся с любыми сообщениями, эти сообщения выражены некоторыми знаками, словами, языками. Семиотика исследует знаки как особый вид носителей информации. Рассуждая о количестве, содержании и ценности информации, содержащейся в сообщении, можно исходить из возможностей соответствующего анализа знаковых систем, таких как естественные и искусственные языки, системы сигнализации, логические, математические и химические символы.

Знаковые системы рассматриваются с позиций синтактики, семантики и прагматики.

Синтактика изучает синтаксис знаковых структур - способы сочетаний знаков, правила образования сочетаний и преобразований безотносительно к их значениям.

Семантика изучает знаковые системы как средства выражения смысла, определенного содержания.

Прагматика концентрируется на изучении практической полезности сообщений для потребителя.

Основная идея семантической концепции информации заключается в возможности измерения содержания (предметного значения) суждений. Но содержание всегда связано с формой, хотя и не взаимно однозначно. Поэтому и исследования семантики базировались на понятии информации как уменьшении или устранении неопределенности.

1.5.1 Семантическая информация (Карнап и Бар-Хиллел)

Первую попытку построения теории семантической информации предприняли Р. Карнап и И. Бар-Хиллел. Они предложили определять величину семантической информации посредством так называемой логической вероятности, представляющей собой степень подтверждения той или иной гипотезы. При этом количество семантической информации, содержащейся в сообщении, возрастает по мере уменьшения степени подтверждения априорной гипотезы. Если вся гипотеза построена на эмпирических данных, полностью подтверждаемых сообщением, то такое сообщение не приносит получателю никаких новых сведений. Логическая вероятность гипотезы при этом равна единице, а семантическая информация оказывается равной нулю. Наоборот, по мере уменьшения степени подтверждения гипотезы, количество семантической информации, доставляемой сообщением, возрастает. Концепция Карнапа - Бар-Хиллела является только началом исследований в области измерения содержания передаваемой информации. Она позволяет, например, выявить связь гипотезы с начальным достоверным значением, в частности, сделать заключение о степени подтверждения гипотезы. Однако, она приводит к парадоксам типа того, что высказывание «Снежный человек существует» информативно, а «Эйнштейн - известный ученый» - не информативно, поскольку является достоверным.

1.5.2 Прагматика

Для всех прагматических подходов характерно стремление связать понятие информации с целенаправленным поведением и выдвинуть те или иные количественные меры ценности информации.

А.А. Харкевич предложил связать меру ценности информации с изменением вероятности достижения цели при получении этой информации:

$$I = \log \frac{p_1}{p_0} = \log p_1 - \log p_0,$$

где p_0 и p_1 - вероятность достижения цели соответственно до и после получения информации.

А.А. Харкевич первым подчеркнул фундаментальный характер связи прагматических свойств информации с категорией цели, понимаемой как опережающее отражение, модель будущего результата деятельности.

1.6 Информация и Энтропия

Информация в термодинамике появилась как величина, обратная по знаку энтропии. Тела могут «самонагреваться» и «самоостывать», однако, с подавляющей вероятностью они будут все же остывать. Энтропия - это мера необратимого рассеяния энергии, мера деградация системы на пути от порядка к хаосу. Информация, следовательно, отражает обратное движение, и их сумма строго равна нулю.

Но все наблюдаемые процессы развития от простого к сложному сопровождаются ростом информации и ростом энтропии как меры рассеваемой энергии.

Этот парадокс нашел свое разрешение в рамках нелинейной термодинамики, созданной И. Пригожиным. При избытке энергии, подводимой к системе извне («свободной энергии»), система выходит из равновесия, сохраняя вероятность к нему вернуться, но в ином качестве - повысив сложность (уровень организации), и, следовательно, количество информации. Энтропия также увеличивается, но во внешней среде, к которой только и применимо это понятие.

1.7 Информация как свойство живой материи

Но вот что пишет [3] по этому поводу академик Н.Н. Моисеев:

«Во-первых, я полагаю, что строгого и достаточно определенного понятия информации не только нет, но оно и вряд ли возможно. Во вторых, это понятие представляется мне в некотором смысле «историческим». Необходимость его введения возникает лишь при описании довольно поздних этапов развития материального мира, лишь тогда, когда в нем зарождается жизнь.

Если описывать последовательное развитие материального мира, опираясь на принцип «лезвия Оккама», то информация появится в нем лишь тогда, когда мы начнем изучать объекты с целеполаганием, то есть объекты, способные к целенаправленным действиям. Именно только такие системы предполагают необходимость использования термина «информация», без которого нельзя описать процедуры принятия решений, то есть целенаправленного поведения, и изучать зависимость характера принимаемых решений от изменений внешних условий.»

Определение информации как меры «упорядоченности» является одним из вариантов «разнообразностного» ее понимания. Упорядоченность всегда связана с ограничением разнообразия

как следствием управления. Управление невозможно без получения информации. Но имеет ли смысл говорить об о процессах управления в неживой природе, если ее объекты не обладают целеполаганием? Управление и, следовательно, информационные процессы имеют место только в кибернетических и биологических системах.

1.8 Что же такое «информация»

ТЕЗИСЫ ИЗ СТАТЬИ СТОЛЯРОВА «Онтологический и метонимический смыслы понятия информация»: Согласно Столярову, существуют шесть основных философских концепций информации, как научного понятия.

Первое понятие относится с отрицанием к существованию информации. Информация воспринимается как призрак, ошибочное представление науки, как то, чего никто никогда не видел, ощущал или фиксировал с помощью какой-либо аппаратуры.

Вторая концепция основана на тезисах, что информация существует, но не в нашем физическом мире. Эта доктрина объясняет природу телепатии, вспышек, привидений и т.д., которая не признается ортодоксальной наукой.

Третья точка зрения касается существования чистой информации без какой-либо формы разновидности.

Четвертым является утверждение, что информация имеет материальную природу, которая сама по себе очень информативна.

Пятая гипотеза может быть охарактеризована как панинформистская теория. Согласно ей, информация является первичной, а материя - вторичной.

Шестая теория представляет информацию как субъективную реальность. В объективном мире существуют разнообразные свойства и отношения между субстанцией и энергией. Часть их воспринимается нашими органами чувств, распознается, и субъективно воспринимается как информация.

Истина, как представляется, была обнаружена еще Норбертом Винером: информация - «это обозначение содержания (сигналов), полученного из внешнего мира в процессе нашего приспосабливания к нему и приспосабливания к нему наших чувств».

В объективном мире существуют бесчисленные свойства и отношения внутри вещества и энергии, а также между веществом и энергией. Часть их воспринимается органами чувств, распознается, т.е. превращается в образ (семантически преобразуясь в модель объективной реальности) и субъективно осознается как информация. Информация в сущности есть достояние только субъективного сознания, за его пределами информации не существует. Онтологически информация есть субъективная реальность.

Информация, ушедшая в глубины сознания, называется памятью. Информация может быть объективирована - если она представлена в знаковой форме и перенесена на внешний носитель.

Информация, переработанная субъектом, упорядоченная, наложенная на прежние представления и сохраненная, называется знанием. Знание в онтологическом смысле - это тоже исключительно субъективная реальность.

Понятие «информация» неотделимо от воспринимающего субъекта. Один и тот же объект может быть рассмотрен различными субъектами (или одним и тем же, в зависимости от объема знаний и потребностей) с различных точек зрения. Понятия «информация» рассматривается с двух основных точек зрения: коммуникативной и физической (термодинамической). Пункты 1-3 толкования информации в ФилЭС как раз отражают коммуникативный взгляд. Пункт 4 «прокладывает дорожку» к физическому толкованию.

Но Мир, отраженный в коллективном сознании Человечества, един. (Или связен, если Знание о мире в целом рассматривать как граф, состоящий из подграфов различных предметных областей.) Предложение академика Моисеева относительно использования термина «информация» только применительно к объектам живой природы можно совместить с принципом связности знания на основе философской теории отражения. Обратимся вновь к ФилЭС.

«Отражение, всеобщее свойство материи, заключающееся в воспроизведении признаков, свойств и отношений отображаемого объекта... Способность к отражению, а также характер ее проявления зависит от уровня организации материи. ... Взаимодействие различных материальных систем имеет своим результатом взаимоотражение, которое выступает в виде простой механической деформации, сокращения или расширения в зависимости от колебаний окружающей температуры ...».

Эволюционируя, природа проходит некоторый порог сложности, за которым появляются субъекты, различно интерпретирующие одни и те же внешние явления в зависимости от свойств и состояния нервной системы. (Следует иметь в виду, что интерпретируется воспринятый образ).

2 Тезаурусная модель коммуникации

Для учета эффектов, связанных с «субъективными» различиями приемника и передатчика в коммуникационных процессах, являющимися следствиями различных объемов знаний в предметной области сообщения, была предложена тезаурусная модель. Она интересна тем, что сохраняя идеологическую связь с психологическим подходом к процессам мышления и феномену понимания, позволяет очень наглядно и просто проиллюстрировать эти процессы.

Слово THESAURUS означает сокровище, богатство, запас. Термин был применен впервые в 13-м веке учителем Данте флорентийцем Бруннет Латини (1220-1294) как название энциклопедии. [4] В наше время этот термин был введен Кэмбриджской группой по изучению языка (Великобритания) в 1956 г.[5].

Понятие «тезаурус» вначале использовалось для обозначения словарей, предназначенных для поиска слов по их смыслу. Толковые словари в электронном виде, используемые для описания терминологии какой-либо отрасли знаний в автоматизированных системах поиска информации, получили название информационно-поисковых тезаурусов.

Относительно недавно «тезаурус» стал обозначать структурированное знание. Графически тезаурус - это семантическая сеть. Поэтому тезаурусную модель коммуникационных процессов можно было бы с полным основанием назвать «знаниевой», однако - это уже дело вкуса.

Человек получает информацию только в том случае, когда в его знаниях, т.е. в его тезаурусе после получения сообщения произошли какие-либо изменения. И чем больше изменений в тезаурусе приемника, тем большее количество информации он получил из этого сообщения.

Используя введенные ранее обозначения, полагаем, что тезаурус $T = (I, C, \Gamma)$.

Пусть передатчик, обладающий тезаурусом T_{out}, передает какое-то фрагмент своего тезауруса $T_f \subseteq T_{out}$ сообщение приемнику, обладающему тезаурусом T_{inp}.

T_f при приеме сообщения сравнивается с T_{inp}. Рассмотрим возможные варианты этого процесса.

Если $T_f \subset T_{inp}$, то никаких изменений в T_{inp} не происходит, следовательно, приемник не извлекает из сообщения никакой информации.

Если $R = T_f \bigcap T_{inp} \neq \emptyset,\ T_f \neq T_{inp}$, то $D = T_{out} \ominus T_{inp}$ может быть воспринята приемником. Подчеркнем «может быть». Это будет зависеть как от «желания» приемника получить эту

информацию, так и от возможностей реконструкции тезауруса.

Если $R = \emptyset$, то приемник не извлечет из сообщения никакой информации и изменения его тезауруса не произойдет.

Чем больше T_{inp}, тем больше вероятность того, что $T_f \subset T_{inp}$, и количество информации, получаемое из сообщения, будет зависеть в итоге от величины T_{inp}. Таким образом, минимальному значению величины тезауруса $T_{inp,min}$ соответствует нулевое количество полученной из сообщения информации. Такое же (нулевое) количество информации соответствует и максимальному тезаурусу $T_{inp,max}$. Наибольшее же количество информации I_{max} извлекается приемником из сообщения при величине его тезауруса, близкой к средней.

При значительной разнице тезаурусов источника и приемника информации количество информации, извлекаемое из сообщения приемником, невелико. Например, если тезаурус ученого, работающего в какой-либо области науки, значительно шире среднего тезауруса специалиста в этой области, то знакомящиеся с его работами коллеги вероятнее всего не смогут извлечь из них сколь-нибудь значительного количества информации, т.е. не поймут их. (Примеры: Лобачевский, Эварист Галуа, Циолковский)

Очевидно, что можно говорить о тезаурусе человечества как о сумме накопленных им знаний.

Можно исследовать как тезаурусы отдельных специалистов, так и тезаурусы областей знания.

Объекты внешнего мира и отношения между ними отражаясь мозгом человека, образуют его тезаурус *плана содержания*. Вербализованная часть плана содержания (слова,поставленные в соответствие элементам плана содержания - информационным единицам - узлам и дугам - отношениям) составляет тезаурус *плана выражения*. Планы содержания и планы выражения не обязательно идентичны, поскольку «слово не покрывает понятия». Однако исследовать тезаурус плана выражения существенно легче, чем структуру нейронных ансамблей - физических (физиологических) носителей информации плана содержания.

Список литературы

[1] Философский энциклопедический словарь. М.: Советская энциклопедия, 1983.

[2] Математическая энциклопедия, т. 1. М: Советская энциклопедия, 1977.

[3] Н. Моисеев. человек и ноосфера. - М.:Молодая гвардия, 1990.

[4] В.П.Леонов. Применение статистики в статьях и диссертациях по медицине и биологии. часть III. проблемы взаимодействия "автор - редакция - читатель". Международный журнал медицинской практики, 1999, вып 12, стр.7-13.

[5] Михайлов А.И., Черный А.И., Гиляревский Р.С. Научные коммуникации и информатика. М.: Наука, 1976

[6] Терминологический словарь по вычислительной технике. М.: Машиностроение, 1989.

Следующая лекция

Лекция 5. ЯЛП ПРОЛОГ. Часть I