Оглавление

Конспект лекция #1. Алгоритмический подход А.Н. Колмогорова к вычислению количества информации

Что нас ждет?

После прочтения этого материала Вы сможете:

1. дать определение понятий «колмогоровская сложность», «хаотическая последовательность»
2. объяснить иллюзорность разграничения царств порядка и хаоса
3. рассказать о исследованиях, выполненных Марковым А.А.
4. привести первый пример марковской цепи
5. сказать, есть ли разница между энтропией языка и энтропией речи
6. изложить комбинаторный подход Р. Хартли в терминах А.Н.Колмогорова
7. изложить вероятностный подход К. Шеннона в терминах А.Н.Колмогорова
8. изложить алгоритмический подход А.Н. Колмогорова

1. О колмогоровской сложности – «на пальцах»

В статье «Три подхода к определению количества информации», опубликованной в 1965 году в журнале «Проблемы теории информации», А.Н. Колмогоров предложил алгоритмический подход к вычислению количества информации, использующий теорию рекурсивных функций [1]. Он показал причины, по которым оценить энтропию конечных объектов (например, текстов в то время, веб-страниц - сейчас) не смогут ни комбинаторный, ни вероятностный подходы, и ввел понятие сложности конечных объектов (слов, состоящих из 1 и 0), впоследствии признанной всеми как «колмогоровская сложность».

Алгоритмический подход к вычислению количества информации Колмогорова А.Н.

Вот пример, с которого начинается доклад А.Н.Колмогорова и В.А. Успенского на Первом Всемирном конгрессе Общества математической статистики и теории вероятностей им. Бернулли (Ташкент, 1986 г.) [2]:

"Если кто-то скажет нам, что он подбросил "честную" монету 20 раз и, обозначив герб единицей, а решетку нулем, получил такой результат:

(I) 1000 101110 1111010000

или такой

(II) 0111101 1001101110001,

мы вряд ли будем удивлены. Однако, если нам скажут, что результат бросаний был таков:

(III) 00000000000000000000,

мы будем поражены или вообще не поверим или же усомнимся в корректности эксперимента. Возникает вопрос – почему?...

По-видимому, цепочки (I) и (II) воспринимаются, как случайные, а цепочки (III) – как неслучайные. Но что означают слова "воспринимается, как случайная?" Классическая теория вероятностей не дает ответа на этот важный вопрос. Не столь редко можно услышать следующее объяснение: вероятность цепочки (III) слишком мала, он равна 2^{-20}. Но ведь ровно такую же вероятность имеют и цепочки (I) и (II)".

Колмогоров предложил поразительную по красоте идею измерения порядка и хаоса: "…последовательность не хаотична, если существует ее простое описание, а если такового не существует, то есть, она достаточно сложна, то она несет в себе признаки случайной последовательности".

В рассмотренных примерах цепочка (III) не сложна, поскольку ее можно описать как ноль в периоде (0).

Как реализуется идея Колмогорова? Во множество конечных последовательностей вводится мера их сложности. Грубо говоря, сложность последовательности – это длина программы (или алгоритма, реализованного машиной Тюринга), описывающей данную последовательность. Есть определение еще проще: "сложность объекта есть не что иное как длина его наикратчайшего описания» [2]. Введенная Колмогоровым А.Н. мера сложности конечного объекта позволила определить понятия хаотической последовательности, конечной и бесконечной, и доказать, что хаотические (то есть, сложно устроенные) последовательности ведут себя как случайные, к ним "нельзя придраться" с точки зрения теории вероятностей. Кстати, эта теория А.Н. Колмогорова показала всю иллюзорность разграничения царств порядка и хаоса: на самом деле математический мир – един и существует огромное число детерминированных динамических систем, обладающих определенной неустойчивостью, начинают вести себя, как случайные, и наоборот, случайные явления подчиняются строгим законам.

А.Н. Колмогоров предложил так называемую алгоритмическую теорию информации, в которой *под энтропией понималась сложность объекта, равная сложности алгоритма, описывающего этот объект*. В качестве объектов, сложность которых надлежит установлению, могут рассматриваться, в частности, тексты. В работе [1] Колмогоров А.Н. пишет, что "… такие величины, как "сложность" текста романа "Война и мир", можно считать определенными с практической однозначностью". Колмогорова интересовал, в частности, вопрос о сложности литературных текстов, в том числе, какая доля сложности приходится на содержание текста, а какая – на те или иные литературные приемы, такие как рифма, метр, и т.п. Последние, кстати легче всего формализуются и вычленяются в поэзии.

Уместно заметить здесь, что теоретико-информационный подход Колмогорова оказался существенным вкладом в развитие современной теории информации, подробнее об этом будет в лекции 2.

2. Исследования русского текста Марковым А.А.

(по материалам статьи В.А. Успенского Предварение для читателей «Нового литературного обозрения» к семиотическим посланиям Андрея Николаевича

Алгоритмический подход к вычислению количества информации Колмогорова А.Н. Колмогорова//Новое литературное обозрение. № 24, 1997

http://kolmogorov.pms.ru/uspensky-predvarenie.html)

2.1. Роль вероятностных параметров букв для измерения содержащейся в тексте информации

Приведем следующую важную мысль Колмогорова: исследование *вероятностных закономерностей текстов* должно непременно предшествовать исследованию художественных приёмов, потому что при ином порядке исследования можно неизбежный статистический закон ошибочно принять за приём (ясно, например, что если та или иная ритмическая конструкция статистически характерна для русского ямба вообще, то её появление в каком-то месте текста вряд ли может считаться художественным приёмом). Вероятностно-статистическому анализу могут подлежать какие угодно детали текста - и лексемы, и падежи, и целые синтаксические обороты, и ритмические конструкции. Самый простой случай, когда статистика наводится на отдельные буквы и их сочетания.

Упомянем здесь исследования Маркова А.А. старшего, как наиболее ранний пример подхода к русскому (а возможно, и не только к русскому) художественному тексту, как к предмету статистики и теории вероятностей. Марков А.А. старший находит следующие вероятности для встречающихся в «Евгении Онегине» букв: для вероятности гласной буквы (т. е. для вероятности того, что выбранная наугад буква текста окажется гласной): $\mathbf{p}$ = 0,432; для вероятности согласной буквы: $\mathbf{q}$ = 0,568; для вероятности гласной буквы после гласной (т. е. для вероятности того, что выбранная наугад буква, следующая за гласной, окажется гласной): $\mathbf{p_1}$ = 0,128; для вероятности согласной буквы после гласной: $\mathbf{q_1}$ = 0,872; для вероятности гласной буквы после согласной: $\mathbf{p_0}$ = 0,663; для вероятности согласной буквы после согласной: $\mathbf{q_0}$ = 0,337; для вероятности гласной буквы после двух гласных: $\mathbf{p_{1,1}}$ = 0,104; для вероятности согласной буквы после двух согласных: $\mathbf{q_{0,0}}$ = 0,132; отсюда уже можно определить вероятность гласной буквы после двух согласных: $\mathbf{p_{0,0}}$ = 0,868.

Цепи Маркова, особенно их обобщение - марковские процессы, общая теория и классификация которых были даны Колмогоровым в 1930 г., находят свое место в широчайшем спектре естественно-научных и технических приложений.

Первый содержательный пример цепи Маркова, предложенный первооткрывателем этого понятия, был связан с изящной словесностью. И не исключено, что самоё понятие «цепь» родилось у Маркова из наблюдений над чередованиями букв в литературных текстах. Если эта гипотеза верна, то мы имеем впечатляющий пример того, как анализ текста приводит к рождению важного понятия математики. Посмотрим, что происходит, если учитывать частоты букв и их сочетаний.

2.1. Нулевое приближение

Вспомним инвентарь из 32 букв русской письменной речи (включая пробел!). Представим, что мы составили разрезную азбуку из этих 32 букв и поместили её в ящик (математики сказали бы “в урну”), тщательно перемешав. Будем теперь составлять из этой азбуки случайный текст, применяя следующую процедуру: мы вынимаем букву из ящика, записываем её, затем возвращаем в ящик, перемешиваем буквы, снова вынимаем букву, снова записываем (приписывая её к уже имеющемуся тексту), снова возвращаем, снова перемешиваем, снова вынимаем и т. д. Мы получим что-нибудь вроде:

СУХЕРРОБЬДЩ ЯЫХВЩИЮАЙЖТЛФВНЗАГФОЕНВШТЦР ПХГБКУЧТЖЮРЯПЧЬКЙХРЫС.

Про этот текст можно сказать лишь, что он составлен из русских букв. Но на русскую письменную речь он не похож: мы говорим, конечно, не об осмысленности (где уж!), а лишь о внешней похожести.

Дело в том, что в нашем эксперименте все буквы были *равновероятны* и потому в полученном тексте встречались с примерно одинаковыми частотами. В реальных же русских письменных текстах пробел и различные буквы встречаются с различными частотами* и потому ожидаются с различными вероятностями. Это, конечно, всем известный факт. Менее известен (хотя и очевиден) и потому будет сейчас воспроизведён следующий эффект: при учёте всё более и более глубоких статистических закономерностей, имеющихся в реальных текстах, экспериментальный искусственный текст делается всё более и более похожим на «настоящий». Тот искусственный текст, который мы получили, можно назвать приближением нулевого порядка к реальному тексту: здесь учитывается лишь состав алфавита и ноль статистических характеристик.

Таблица статистических характеристик (частот) приведена в справочнике по математике Ягломов, на с. 238. Сообщим сведения о 3-х чемпионах и 3-х аутсайдерах этой таблицы, указав для каждой из этих 6-ти букв, количество (в среднем!) их появлений на каждую тысячу букв текста. Первые три места занимают **пробел**, **о** и **е/ё**, имеющие, соответственно, 175, 90 и 72 появления на тысячу. Замыкают таблицу буквы **щ**, **э** и **ф**, имеющие, соответственно, 3, 3 и 2 появления на тысячу. Тут надо ясно отдавать себе отчёт в следующем. Говоря о частотах букв в русских текстах, имеют в виду частоты, вычисленные для большого корпуса разнообразных достаточно длинных текстов. Дело в том, что в отдельно взятом тексте могут наблюдаться значительные отклонения от средней нормы: такие отклонения может дать как текст специального содержания, ввиду обилия специальных терминов, так и поэтический текст, ввиду аллитераций и ассонансов.

2.2. Первое приближение

При приближении первого порядка учитываются частоты каждой из букв; иными словами, теперь требуется, чтобы в экспериментально построенном искусственном тексте буквы встречались с такими же (в идеале*) частотами, как и в реальных текстах. Тогда мы получаем, например, такой текст:

ЕЫНТ ЦИЯЬА ОЕРВ ОДНГ ЬУЕМЛОЛИЙК ЗБЯ ЕНВТША

Он уже более похож на настоящий: и длина слов нормальная, и нет того чудовищного преобладания согласных, как в тексте нулевого приближения.

Слова “в идеале” отражают философское различие между частотой и вероятностью. Частоты букв, встречающихся в реальных (и притом непременно длинных) текстах, приводят к возникающей в мозгу математика концепции о вероятностях этих букв; в качестве таковых берутся соответствующие частоты. Затем эти вероятности закладываются в схему эксперимента по созданию искусственного текста. После чего можно с высокой достоверностью надеяться, что частоты букв в экспериментально созданном искусственном тексте будут близки к указанным вероятностям.

2.3. Второе приближение

Приближение первого порядка не учитывает частот диграмм, то есть, сочетаний двух последовательно идущих букв. В приведённом тексте, например, встречаются диграммы **ЯЬ, ЬА** и **ЬУ**, частота которых в реальных текстах равна нулю. Учёт частот диграмм приводит к приближению второго порядка:

УМАРОНО КАЧ ВСВАННЫЙ РОСЯ НЫХ КОВКРОВ НЕДАРЕ.

2.4. Третье приближение

Приближение третьего порядка, учитывает частоты триграмм, то есть, здесь трёхбуквенные сочетания должны встречаться примерно с теми же частотами, что и в реальных текстах:

ПОКАК ПОТ ДУРНОСКАКА НАКОНЕПНО ЗНЕ СТВОЛОВИЛ СЕ ТВОЙ ОБНИЛЬ

2.5. Четвёртое приближение

Приближение четвёртого порядка, учитывает частоты тетраграмм:

ВЕСЕЛ ВРАТЬСЯ НЕ СУХОМ И НЕПО И КОРКО

3. Роль вероятностных параметров слов для измерения содержащейся в тексте информации

Посмотрим, что происходит, если учитывать частоты слов и их сочетаний. Можно составлять экспериментальный текст не из букв, а сразу из слов (т. е. из словоформ). Естественно, при этом возникает ещё бОльшая похожесть на подлинный облик русского текста. Тут также возможны приближения разных порядков. Первое приближение на уровне слов, учитывающее частоты отдельных слов:

СВОБОДОЙ ДУШЕ ПРОТЯНУЛ КАК ГОВОРИТ ВСПОМНИТЬ МИЛОСТЬ КОМНАТАМ РАССКАЗА ЖЕНЩИНЫ МНЕ ТУДА ПОНЮХАВШЕГО КОНЦУ ИСКУСНО КАЖДОМУ РЯСАХ К ДРУГ ПЕРЕРЕЗАЛО ВИДНО ВСЕМ НАЧИНАЕТ НАД ДВУХ ЭТО СВЕТА ХОДУНОМ ЗЕЛЁНАЯ МУХА ЗВУК ОН БЫ ШЕЮ УТЁР БЕЗДАРНЫХ

Второе приближение, учитывающее частоты сочетаний из двух соседних слов:

ОБЩЕСТВО ИМЕЛО ВЫРАЖЕНИЕ МГНОВЕННОГО ОРУДИЯ К ДОСТИЖЕНИЮ ДОЛЖНОСТЕЙ ОДИН В РАСЧЁТЫ НА БЕЗНРАВСТВЕННОСТИ В ПОЭЗИИ РЕЗВИТЬСЯ ВСЕ ГРЫЗЁТ СВОИ БРАЗДЫ ПРАВЛЕНИЯ НАЧАЛА ЕГО ПОШЛОЙ

А вот приближение второго порядка на уровне слов, предложенное самим Шенноном - для английского языка:

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THE BEFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

Колмогоров учил, что статьи для энциклопедии надо писать так. В минимальном случае статья исчерпывается дефиницией (определением). Если же автору статьи дают ещё место, то сразу после дефиниции нужно написать несколько фраз, доступных человеку с начальным образованием. Если допустимый объём исчерпан, этим и следует ограничиться. Если же объём позволяет, надо написать абзац, требующий уже семиклассного образования, затем - десятиклассного. Если статья достаточно большая, можно перейти к сюжетам, предполагающим образование высшее, а в конце - даже требующим специальных знаний. Наконец, при очень большом объёме и в самом конце автор в качестве премии самому себе может поместить текст, который понимает он один.

Замечание: Как известно, энциклопедическая статья начинается с названия статьи, за которым идёт тире, затем определение и затем точка; определением как раз и называется текст, расположенный между сопровождающим заглавное слово тире и ближайшей точкой.

4. Свойства энтропии языка

Мы собираемся поступить с термином **энтропия** так, как, по слухам, поступил с термином “электрон” один из основателей квантовой механики английский физик Поль Дирак. Рассказывают, что, приехав в Москву и читая лекцию в Политехническом музее, он сказал: *“Поскольку никто не знает, что такое электрон, мы будем изучать, как он движется”*. Вот и мы не будем определять, что такое **энтропия языка** (хотя, в отличие от Дирака, не утверждаем, что этого никто не знает), а ограничимся выводами, вытекающими из сущности этого понятия.

О самой же сущности мы, по крайней мере, на первых порах, ограничимся следующим: энтропия - это численная мера гибкости языка, она отражает количество возможных вариантов текста с учётом вероятностей этих вариантов. Итак, энтропия языка - это, прежде всего, положительное число.

Как мы объявили заранее, мы не будем давать понятию энтропии определения, но укажем его наглядные свойства. Таких свойства мы укажем три (на самом деле, они вытекают друг из друга, но это уже чистая математика, вдаваться в которую мы не будем).

4.1 Первое свойство энтропии

Пусть энтропия языка равна **H**. Тогда существует примерно $\mathbf{2^{Hk}}$ текстов длины **k**, принадлежащих данному языку. Отсюда следует, что чем более узкий корпус текстов мы соотносим с представлением о языке, тем меньше будет энтропия языка; так, если взять энтропию языка русской художественной литературы или энтропию языка русского ямба, то каждая из них будет меньше энтропии русского языка в целом.

4.2 Второе свойство энтропии

Тексты можно закодировать, используя всего два каких-либо символа, наподобие того, как это делается в компьютерах с помощью нулей и единиц. Разумеется, кодировать надо так, чтобы исходный текст можно было восстановить по его коду. При этом, скорее всего, произойдёт удлинение текста (если не сделать оговорки о возможности декодирования, можно было бы, напротив, добиться укорочения текста, закодировав все тексты одним символом). Кодирование разумно производить так, чтобы указанное удлинение было бы - в среднем - как можно более маленьким (для этого следует частые тексты кодировать короткими цепочками, составленными из применяемых двух символов, а более длинные цепочки использовать для кодирования более редких текстов). Такой способ кодирования, при котором достигается минимальное, в среднем, удлинение, называется оптимальным. Так вот, если энтропия языка равна **H**, то при оптимальном способе кодирования каждый текст языка удлинится в среднем в **H** раз.

4.3. Третье свойство энтропии

Третье свойство энтропии имеет смысл в предположении, что каждому тексту языка приписана определённая вероятность - вероятность того события, что среди всех мыслимых текстов данной длины на свет появится именно рассматриваемый текст. Так вот, если энтропия языка равна **H**, то для подавляющего большинства текстов длины **k** эта вероятность равна $\mathbf{2^{-Hk}}$**.**

Замечание: три приведённые утверждения о свойствах энтропии выполняются тем точнее, чем больше длина тех текстов, к которым эти утверждения применяются.

4.4. Энтропия русского языка

Колмогоровым предложена оценка _____ (найти в первоисточнике!) для количества русских текстов длины **k**, она показывает (как дают несложные подсчёты), что он исходил из числа 1,33 в качестве значения для энтропии русского языка.

Шеннон указывает, что энтропия английского языка лежит приблизительно в пределах от 0,6 до 1,3

Колмогоров писал:

Вполне естественным является чисто комбинаторный подход к понятию «энтропии речи»*, если иметь в виду оценку «гибкости» речи - показателя разветвлённости возможностей продолжения речи при данном словаре и данных правилах построения фраз. Для двоичного логарифма числа **N** русских печатных текстов, составленных из слов, включённых в Словарь русского языка С. И. Ожегова и подчинённых лишь требованиям «грамматической правильности», длины **n**, выраженной в «числе знаков» (включая «пробелы»), М. Ратнер и Н. Светлова получили оценку **h** = 1,9 ± 0,1. Это значительно больше, чем оценки сверху для «энтропии литературных текстов», получаемые при помощи различных методов «угадывания продолжений». Такое расхождение вполне естественно, так как литературные тексты подчинены не только требованию «грамматической правильности».

* Под “чисто комбинаторным” разумеется подход, основанный лишь на статистистических характеристиках текста, без привлечения вероятностной модели.

4.5. Энтропия различных вариантов русского языка

Неудивительно, что величина энтропии для русского (да и для любого иного) языка зависит от того смысла, который вкладывается в само слово “язык”. При определении энтропии язык понимается как некоторая совокупность текстов. Именно такое понимание нужно для формулировки трёх свойств энтропии. Из первого свойства вытекает, что чем больше совокупность допустимых текстов заданной длины, тем больше и энтропия. При наиболее широком, пожалуй, понимании, русский язык понимается как корпус всех текстов, составленных из реальных русских словоформ по правилам русской грамматики. Мы получим ощутимое сужение корпуса допустимых текстов и - тем самым - уменьшение величины энтропии, если от произвольных русских текстов перейдём к литературным русским текстам, то есть от русского языка в полном объёме к литературному русскому языку. И дальнейшие сужение и уменьшение, если перейдём, скажем, к языку Гоголя или языку Чехова. Или к языку русского ямба (точнее, к литературному языку русского ямба, потому что возможны русские тексты, являющиеся ямбическими, но не являющиеся литературными). Все эти более частные (по сравнению с русским языком в его полном объёме) языки можно было называть подъязыками русского языка.

Одно важное обстоятельство требует высвечивания. Оно состоит в том, что тексты рассматриваются в своём потенциальном, а не актуальном качестве. Вот что это значит. Когда мы говорим о русских текстах, мы имеем в виду не только те реально существующие к настоящему времени тексты на русском языке, но также и те, которые могут быть составлены. Первые называются актуальными, вторые - потенциальными. Сказанное в полной мере справедливо и в отношении, скажем, литературных (т. е. написанных литературным языком) или ямбических текстов. Именно при таком взгляде на вещи будут иметь место три свойства энтропии, указанные выше.

На первый взгляд, кажется, что изложенную точку зрения невозможно применить к языку Гоголя или языку Чехова. Ведь язык того или иного писателя, понимаемый как совокупность текстов, исчерпывается академическим собранием его сочинений и потому состоит из одних только актуальных текстов. Но мыслим и более широкий подход, согласно которому Девятая повесть «Вечеров на хуторе близ Диканьки» написана языком Гоголя, хотя у Гоголя такой повести и нет. Текст - это пример потенциального гоголевского текста. Впрочем, этот пример не слишком показателен, поскольку состоит из готовых «актуально гоголевских» (т. е. созданных самим Гоголем) фраз. Под гоголевским текстом, в потенциальном смысле, понимается любой «гоголе-подобный» текст, т.е текст, составленный из характерных для Гоголя слов путём применения характерных для Гоголя синтаксических конструкций. При таком, широком понимании и становится возможным говорить об энтропии языка Пушкина, языка Гоголя, языка Толстого, языка Чехова.

На математическом уровне можно предложить следующее понимание термина “остаточная энтропия”. Пусть каждое предложение иностранного - скажем, английского - языка можно перевести на русский язык **n** способами. В предыдущей фразе содержится значительное огрубление ситуации, поскольку в действительности различные предложения можно перевести различным количеством способов. Но мы пойдём дальше и допустим ещё бОльшую идеализацию, приняв, что все русские переводы всех английских предложений состоят из одного и того же количества букв. Подобные огрубления, или идеализации, не слишком страшны, потому что они отражают вполне разумные представления о **среднем** числе переводов, приходящихся на одно предложение в достаточно длинном английском тексте и о **среднем** числе русских букв, приходящихся на один перевод. Итак, примем изложенную модель. Тогда текст, скажем, из ста предложений будет иметь $\mathbf{n}^{100}$ переводов. Остаточная энтропия - в том контексте, как она фигурирует в приведённой цитате из Колмогорова - есть двоичный логарифм числа переводов исходного текста, поделённый на длину любого из этих переводов (понимаемую, как число содержащихся в нём печатных знаков). При сделанных идеализирующих допущениях это число постоянно для всех исходных английских текстов. Иными словами, если остаточная энтропия равна **h**, то количество русских текстов длины **k**, являющихся переводами данного английского текста, равно примерно $\mathbf{2^{hk}}$.

Попробуем пояснить смысл величины 0,4. Мы рассматриваем параллельно два языка: русский язык в его полном объёме и ограниченный русский язык, являющийся подъязыком предыдущего и состоящий из русских текстов, удовлетворяющих некоторым жёстко фиксированным требованиям к форме. Пусть энтропия первого языка есть **A**, энтропия второго языка есть **B**, так что примерное количество текстов длины **k** в первом языке есть 2^{Ak}, а во втором языке есть 2^{Bk}. Очевидно, **A** > **B**, так что **A** = **B** + α, или **B** = **A** – α, где $\alpha > 0$. Эта величина α характеризует ту долю, которую среди всех русских текстов данной длины занимают тексты допустимые: для текстов длины **k** вторых в $2^{\alpha k}$ раз меньше, чем первых. Делается допущение, что эта доля не зависит от содержания текстов. А тогда она должна быть такой же для ситуации, в которой рассмотрение ограничивается текстами, являющимися переводами заданного иноязычного текста. Иными словами, если для данного исходного текста имеется **N** русских переводов длины **k**, то допустимых переводов (т. е. переводов, являющихся допустимыми текстами) должно быть в $2^{\alpha k}$ раз меньше. Поэтому в случае, если $2^{\alpha k} > N$, допустимых переводов не будет вовсе. Значит, чтобы допустимые переводы существовали, должно быть $N \geq 2^{\alpha k}$. А это последнее неравенство равносильно такому: $h \geq \alpha$, где **h** - остаточная энтропия (ведь, как указано в последней фразе предыдущего примечания, $N = 2^{\alpha k}$. Итак, выполнение условия $h \geq \alpha$ необходимо для того, чтобы перевод, удовлетворяющий желаемым требованиям формы, был возможен. Для обсуждаемых требований (четырёхстопный рифмованный ямб с некоторыми ограничениями) значение параметра α есть, по мнению Колмогорова, 0,4. Так что количество текстов длины **k**, подчинённых этим требованиям, в $2^{0,4k}$ раз меньше, чем количество всех текстов длины **k**. Разумеется, все приводимые здесь (как и в аналогичных местах нашего изложения) числа (**N,** $2^{\alpha k}$ и прочие) не следует понимать слишком буквально. Точные количества и доли вообще не могут быть точно определены. Числа эти, здесь и в аналогичных местах, надо воспринимать как не более чем способ математического моделирования реальной ситуации.

5. Колмогоровская сложность

В качестве объектов, сложность которых подлежит установлению, могут рассматриваться, в частности, тексты. *“...Такие величины, как* *«сложность» текста романа «Война и мир», можно считать определёнными с практической однозначностью”*

Ясно, что длина описания может существенно зависеть от того объёма знаний, которые разрешается использовать при составлении описания. Поэтому важное место в теории Колмогорова занимает представление об *условной сложности* при тех или иных исходных данных - это есть сложность, вычисленная при условии, что указанные данные уже известны и могут быть использованы при составлении описаний. Очевидно, что условная сложность чего бы то ни было не может быть больше абсолютной (т. е. не условной) сложности того же самого. Можно, для примера, сравнить бОльшую абсолютную сложность текста «Войны и мира» как текста русского языка и его меньшую условную сложность при условии, что заранее известно, что текст написан Львом Толстым, т. е. принадлежит языку Толстого.

В применении к текстам приобретает смысл понятие удельной сложности. Удельная сложность текста есть его сложность как целого, поделённая на длину текста (это, так сказать, сложность, приходящаяся в среднем на один знак). Оказывается, что для длинных текстов их удельная сложность не превосходит энтропии того языка, на котором эти тексты написаны. Точнее, удельная сложность не превосходит суммы энтропии с некоторой добавкой, которая стремится к нулю при увеличении длины текста. Оценивание, о котором идёт речь в приводимой ниже цитате, как раз и представляет собою оценивание сверху удельной условной сложности с помощью энтропии, вычисляемой, в свою очередь, методом угадывания продолжений.

Эксперименты по угадыванию продолжений литературных текстов позволяют оценить сверху [удельную] условную сложность при заданном запасе «априорной информации» (о языке, стиле, содержании текста), которой располагает угадывающий. В опытах, проводившихся на кафедре теории вероятностей Московского гос. ун-та, такие оценки сверху колебались между 0,9 и 1,4. Оценки порядка 0,9–1,1, получившиеся у Н. Г. Рычковой, вызывали у менее удачливых угадчиков разговоры о её телепатической связи с авторами текстов.

5.1. Статья «Три подхода к определению понятия "количество информации"»

В 1965 году в журнале "Проблемы передачи информации" появилась работа "Три подхода к определению понятия "количество информации". Именно в этой работе А.Н. Колмогоров ввел и математически обосновал понятие, впоследствии получившее название "колмогоровской сложности". Повторим за А.Н. Колмогоровым его рассуждения о измерении количества информации в одном объекте (X) о другом объекте (Y), пример взаимоотношений между объектами показан в таблице.

5.1.1. Комбинаторный подход Р. Хартли с точки зрения А.Н. Колмогорова

Пусть переменное x способно принимать значения, принадлежащие конечному множеству X, которое состоит из N элементов. Говорят, что «энтропия» переменного x равна

$$H(x) = \log 2\ N.$$

Указывая определенное значение $x = a$ переменного x, мы «снимаем» эту неопределенность (энтропию), сообщая «информацию» :

$$I = \log_2 N$$

Если переменные $x_1, x_2, \ldots, x_к$ способны независимо пробегать множества, которые состоят соответственно из N_l , $N_2, \ldots, N_k$ элементов, то

$$H(x_1, x_2, \ldots, x_k) = H(x_1) + H(x_2) + ,\ldots, + H(x_k) \qquad (1)$$

Для передачи количества информации I приходится употреблять следующее количество двоичных знаков:

$$I' = \begin{cases} I\_при\_I\_целом \\ [I]\_при\_I\_дробном \end{cases}$$

Например, число различных «слов», состоящих из K нулей и единиц и одной двойки, равно

$$2^k(k+1)$$

Поэтому, количество информации в такого рода сообщении равно

$$I = k + \log_2(k+1),$$

то есть для «кодирования» такого рода слов в чистой двоичной системе требуется количество нулей и единиц, равное приблизительно:

$$I' \approx k + \log_2 k$$

Далее в статье автор подчеркнул логическую независимость от каких бы то ни было вероятностных допущений, приведя в пример задачу кодирования сообщений, записанных в алфавите, состоящем из *s* букв, при условии, что известны частоты появления отдельных букв в сообщении длины *n:*

$$p_r = \frac{s_r}{s} \qquad (2)$$

При этом эти частоты удовлетворяют неравенству:

$$\chi = -\sum_{r=1}^{8} p_r \log_2 p_r \le h \qquad (3)$$

Легко подсчитать, что при больших *n* двоичный логарифм числа сообщений, подчиненных требованию (2), имеет асимптотическую оценку:

$$H = \log N \approx nh$$

Поэтому при передаче такого рода сообщений достаточно употребить примерно nh двоичных знаков.

Универсальный метод кодирования, который позволит передавать любое достаточно длинное сообщение в алфавите из *s* букв, употребляя не многим более, чем nh двоичных знаков, не обязан быть чрезмерно сложным, в частности, не обязан начинаться с определения частот p_r для всего сообщения. Чтобы понять это, достаточно заметить: разбивая сообщение *S* на *m* отрезков $S_1, S_2, \ldots, S_m$, получим неравенство:

$$\chi \geq \frac{1}{n}\left[n_1\chi_1 + n_2\chi_2 + \ldots + n_m\chi_m\right] \qquad (4)$$

Вполне естественным является чисто комбинаторный подход к понятию
«энтропии речи», если иметь в виду оценку «гибкости» речи — показателя разветвленности возможностей продолжения речи при данном словаре и данных правилах построения фраз. Для двоичного логарифма числа N русских печатных текстов, составленных из слов, включенных в Словарь русского языка С. И. Ожегова и подчиненных лишь требованию «грамматической правильности» длины *n,* выраженной в «числе знаков» (включая «пробелы»), М. Ратнер и Н. Светлова получили оценку:

$$h = \frac{\log N}{n} = 1{,}9 \pm 0{,}1$$

Это значительно больше, чем оценки сверху для «энтропии литературных текстов», получаемые при помощи различных методов «угадывания продолжений». Такое расхождение вполне естественно, так как литературные тексты подчинены не только требованию «грамматической правильности».

Труднее оценить комбинаторную энтропию текстов, подчинённых определённым содержательным ограничениям. Представляло бы, например, интерес оценить энтропию русских текстов, могущих рассматриваться как достаточно точные по содержанию переводы заданного иноязычного текста. Только наличие такой «остаточной энтропии» (как числовой характеристики той гибкости языка, которая всё ещё остаётся после того, как на текст наложены жёсткие ограничения, связанные с его содержанием) делает возможным стихотворные переводы, где «затраты энтропии» на следование избранному метру и характеру рифмовки могут быть довольно точно подсчитаны. Можно показать, что классический русский рифмованный ямб с некоторыми естественными ограничениями на частоту «переносов» и тому подобным требует допущения свободы обращения со словесным материалом, характеризуемой «остаточной энтропией» порядка 0,4 (при указанном выше условном способе измерения длины текста по «числу знаков, включая пробелы»). Если учесть, с другой стороны, что стилистические ограничения жанра, вероятно, снижают приведённую выше оценку «полной» энтропии с 1,9 до не более чем 1,1–1,2, то ситуация становится примечательной, как в случае перевода, так и в случае оригинального поэтического творчества.

А.Н. Колмогоров отметил при этом, что более широкая проблема оценки количеств информации, с которыми имеет дело творческая человеческая деятельность, имеет очень большое значение.

Далее он предложил посмотреть, в какой мере чисто комбинаторный подход позволяет оценивать «количество информации», содержащееся в переменном *x* относительно связанного с ним переменного *y*. Связь между переменными *x* и у, пробегающими соответственно множества *X* и У, заключается в том, что не все пары *x, y,* принадлежащие прямому произведению *X x Y*, являются «возможными». По множеству возможных пар *U* определяются при любом $a \in x$ множества Y_a тех *y,* для которых справедливо выражение:

$$(a, y) \in U$$

Естественно определить условную энтропию равенством:

$$H(y/a) = \log(Y_a) \qquad (5),$$

(где *N* (Y_x) — число элементов в множестве Y_x), а информацию в *x* относительно у формулой:

$$I(x:y) = H(y) - H(y/x)$$

Например, для случая, изображенного в таблице 3, имеем:

$$I(x=1:y)=0$$

$$I(x=2:y)=1$$

$$I(x=3:y)=2$$

Таблица 3 - Пример соотношений между объектами X и Y

x	**y**			
	1	**2**	**3**	**4**
1	+	+	+	+
2	+	-	+	-
3	-	+	-	-

Понятно, что $H(y/a)$ и $I(x:y)$ являются функциями от **x** (в то время, как **y** входит в их обозначение в виде «связанного переменного»). Таким образом, понятно, что в чисто комбинаторной концепции без труда вводится представление о «количестве информации, необходимом для указания объекта x при заданных требованиях к точности указания».

Очевидно, что

$$H(x/x)=0 \text{ и } I(x:y)=H(x)$$

Очевидно также, что

$$H(x/x)=0 \text{ и } I(x:x)=H(x) \qquad (7)$$

5.1.2. Вероятностный подход К. Шеннона с точки зрения А.Н. Колмогорова

Колмогоров отмечает, что придание переменным x и y характера «случайных переменных», обладающих определенным совместным распределением вероятностей, позволяет получить значительно более богатую систему понятий и соотношений.

Вводятся соотношения (оставим их такими, как их ввел А.Н.Колмогоров):

$$H_W(x)=-\sum_x p(x)\log p(x) \qquad (8)$$

$$H_W(y/x) = -\sum_y p(y/x)\log p(y/x) \quad (9)$$

$$I_W(x:y) = H_W(y) - H_W(y/x) \quad (10)$$

По-прежнему $H_W(y/x)$ и $I_W(x:y)$ являются функциями от x. Имеют место также неравенства:

$$H_W(x) \le H(x),$$

$$H_W(y/x) \le H(y/x) \quad (11),$$

переходящие в равенства при равномерности соответствующих распределений (на X и на Y_x). Величины $I_W(x:y)$ и $I(x:y)$ не связаны неравенством определенного знака. Как и ранее,

$$H_W(x/x) = 0 \text{ и } I_W(x:x) = H_W(x) \quad (12)$$

Но отличие заключается в том, что можно образовать математические ожидания

$$MH_W(y/x) \text{ и } MI_W(x:x),$$

а величина

$$I_W(x,y) = MI_W(x:y) = MI_W(y:x) \quad (13)$$

характеризует «тесноту связи» между x и y симметричным образом.

В вероятностном подходе Колмогоров отмечает один «…парадокс: величина $I(x:y)$ при комбинаторном подходе всегда неотрицательна, как это и естественно при наивном представлении о «количестве информации», величина же $I_W(x:y)$ может быть и отрицательной. Подлинной мерой «количества информации» теперь становится лишь осредненная величина $I_W(x,y)$.

Колмогоров подтверждает полезность использования вероятностного подхода в теории передачи информации по каналу связи «массовой» информации, состоящей из большого числа не связанных или слабо связанных между собой сообщений, подчиненным определенным вероятностным закономерностям.

Он говорит, что «…практически безвредно и укоренившееся в прикладных работах смешение вероятностей и частот в пределах одного достаточно длинного временного ряда (получающее строгое оправдание при гипотезе достаточно быстрого «перемешивания»)». Практически можно считать, например, вопрос об «энтропии» потока поздравительных телеграмм и «пропускной способности» канала связи, требующегося для своевременной и неискаженной передачи, корректно поставленным в его вероятностной трактовке и при обычной замене вероятностей эмпирическими частотами. Если здесь и остается некоторая неудовлетворенность, то она связана с известной расплывчатостью наших концепций, относящихся к связям между математической теорией вероятностей и реальными «случайными явлениями» вообще.

Но какой реальный смысл имеет, например, говорить о «количестве информации», содержащемся в тексте «Войны и мира»? Можно ли включить разумным образом этот роман в совокупность «возможных романов» да еще постулировать наличие в этой совокупности некоторого распределения вероятностей? Или следует считать отдельные сцены «Войны и мира» образующими случайную последовательность с достаточно быстро затухающими на расстоянии нескольких страниц «стохастическими связями»?

5.1.3. Алгоритмический подход А.Н. Колмогорова

А.Н. Колмогоров предлагает, как наиболее содержательное, представление о количестве информации «в чем-либо» (**x**) «о чем-либо» (**y**). Не случайно именно оно в вероятностной концепции получило обобщение на случай непрерывных переменных, для которых энтропия бесконечно, но в широком круге случаев может быть вычислено и конечно:

$$I_W(x, y) = \iint P_{xy}(dxdy) \log \frac{P_{xy}(dxdy)}{P_x(dx) P_y(dy)}$$

Реальные объекты, подлежащие изучению с точки зрения определения количества информации (например, поисковые системы в Интернете), очень, а по выражению Колмогорова, неограниченно сложны. Поэтому при рассмотрении связей между двумя реально существующими объектами используется, как правило, схематизированное их описание. Если географическая карта дает нам значительную информацию об участке земной поверхности, то все же микроструктура бумаги и краски, нанесенной на бумагу, никакого отношения не имеет к микроструктуре изображенного участка земной поверхности.

Чаще всего нас практически интересует количество информации в индивидуальном объекте **x** относительно индивидуального объекта **y**. Правда, уже заранее ясно, что такая индивидуальная оценка количества информации может иметь разумное содержание лишь в случаях достаточно больших количеств информации. Не имеет, например, смысла спрашивать о количестве информации в последовательности цифр

0110

относительно

1100.

Но если мы возьмем вполне конкретную таблицу случайных чисел обычного в статистической практике объема и выпишем для каждой ее цифры цифру единиц ее квадратов по схеме:

0 1 2 3 4 5 6 7 8 9

0 **1 4 9 6 5 6 9 4 1**,

то новая таблица будет содержать примерно

$$(\log_2 10 - 8/10)*n$$

информации о первоначальной, где **n** – число цифр в столбцах.

Далее А.Н. Колмогоров предложил определение величины

$$I_A(x:y),$$

заведомо оговорив некоторую ее приблизительность и неопределенность. Разные равноценные варианты этого определения будут приводить к значениям, эквивалентным лишь в смысле $I_{\Lambda A_1} \approx I_{\Lambda_2}$, то есть:

$$I_{A_1} - I_{A_2} \leq C_{A_1,A_2},$$

где константа C_{A_1,A_2} зависит от положенных в основу двух вариантов определения универсальных методов программирования A_1 и A_2.

Будем рассматривать «нумерованную область объектов», т.е. счетное множество

X = {x},

каждому элементу которого поставлена в соответствие в качестве «номера» **n(x)** конечная последовательность нулей и единиц, начинающаяся с единицы. Обозначим через **l(x)** длину последовательности **n(x).**

Будем предполагать, что

1) соответствие между **X** и множеством **D** двоичных последовательностей описанного вида взаимно однозначно;
2) **D** $\subset$ **X,** функция **n(x)** на **D** общерекурсивна, причем, для **x** $\in$ **D** имеет место неравенство: $l(n(x)) \leq l(x) + C$, где **C** – некоторая константа.
3) вместе с x и y в X входит упорядоченная пара (x,y), номер этой пары есть общерекурсивная функция номеров x и y и

$$l(x,y) \leq C_x + l(y),$$

где C_x - зависит только от x.

Не все эти требования существенны, но они облегчают изложение. Конечный результат построения инвариантен по отношению к переходу к новой нумерации **n'(x),** обладающей теми же свойствами и выражающейся общерекурсивно через старую, и по отношению к включению системы **X** в более обширную систему X’ (в предположении, что номера **n**’ в расширенной системе для элементов первоначальной системы общерекурсивно выражаются через первоначальные номера **n**. При всех этих преобразованиях новые «сложности» и количества информации остаются эквивалентными первоначальным в смысле **≈** (приблизительности).

«Относительной сложностью» объекта **y** при заданном **x** будем считать минимальную длину **l(p)** программы **p** получения **y** из **x**. Сформулированное так определение зависит от «метода программирования». Метод программирования есть не что иное, как функция

$$\varphi(p,x)=y\,,$$

ставящая в соответствие программе **p** и объекту **x** объект **y** , причем функцию $\varphi(p,x)$ считаем частично рекурсивной.

Для любой такой функции полагаем:

$$K_\varphi(y/x)=\begin{cases}\min\limits_{\varphi(p,x)=y} l(p),\\ \infty,\_\_\_\_\_\_\_\_ если\_нет\_такого\_p\\ \_\_\_\_\_\_\_\_\_\_ что\_\varphi(p,x)=y\end{cases}$$

При этом функция

$$v=\varphi(u)$$

от **u∈X** со значениями **v∈X** называется частично рекурсивной, если она порождается частично рекурсивной функцией преобразования номеров:

$$n(v)=\Psi[n(u)]$$

Для понимания определения важно заметить, что частично рекурсивные функции, вообще говоря, не являются всюду определенными. Не существует регулярного процесса для выяснения того, приведет применение программы **p** к объекту **x** к какому-либо результату, или нет. Поэтому функция

$K_\varphi \langle y \mid x \rangle$ не обязана быть эффективно вычислимой (общерекурсивной) даже в случае, когда она заведомо конечна при любых **x** и **y.**

5.2 Основная теорема Колмогорова

Существует такая частично рекурсивная функция $A(p, x)$, что для любой другой частично рекурсивной функции $\varphi(p, x)$ выполнено неравенство:

$$K_\Lambda \langle y \mid x \rangle \le K_\varphi \langle y \mid x \rangle + C_\varphi ,$$

где константа C_φ не зависит от **x** и **y**.

Доказательство опирается на существование **универсальной** частично рекурсивной функции

$$\Phi(n, u) ,$$

обладающей тем свойством, что, фиксируя надлежащий номер **n,** можно получить по формуле

$$\varphi(u) = \Phi(n, u)$$

любую другую частично рекурсивную функцию. Нужная нам функция $A(p, x)$ определяется формулой:

$$A((n, q), x) = \Phi(n, (q, x)) .$$

В самом деле, если

$$y = \varphi(p, x) = \Phi(n, (p, x)) ,$$

то
$$A((n, q), x) = y ,$$

$$l(n, p) \le l(p) + C_n .$$

Функции $A(p, x)$, удовлетворяющие требованиям основной теоремы, названы (как и определяемые ими методы программирования) *асимптотически оптимальными*. Очевидно, что для них при любых **x** и **y** «сложность» $K_\Lambda \langle y \mid x \rangle$ конечна. Для любых таких функций A_1 и A_2

$$\left| K_{\Lambda_1} \langle y \mid x \rangle - K_{\Lambda_2} \langle y \mid x \rangle \right| \le C_{\Lambda_1 \Lambda_2} ,$$

где $C_{\Lambda_1\Lambda_2}$ не зависит от **x** и **y,** то есть, $K_{\Lambda_1}\langle y|x\rangle \approx K_{\Lambda_2}\langle y|x\rangle$.

Наконец,

$$K_{\Lambda}(y) = K_{\Lambda}\langle y|1\rangle$$

можно считать просто «сложностью объекта **y**» и определить «количество информации в **x** относительно **y**» формулой:

$$\mathrm{I}_{\Lambda}(x:y) = K_{\Lambda}(y) - K_{\Lambda}\langle y|x\rangle$$

Легко доказать, что величина эта всегда в существенном положительна:

$$\mathrm{I}_{\Lambda}(x:y) \geq\approx 0,$$

что понимается в том смысле, что $\mathrm{I}_{\Lambda}(x:y)$ не меньше некоторой отрицательной константы C, зависящей лишь от условностей избранного метода программирования. Как уже говорилось, вся теория рассчитана на применение к большим количествам информации, по сравнению с которым $|\mathrm{C}|$ будет пренебрежимо мал.

Наконец,

$$K_{\Lambda}\langle x|x\rangle \approx 0, \quad I_{\Lambda}(x:x) \approx K_{\Lambda}(x)$$

Конечно, можно избегнуть неопределенностей, связанных с константами C_{φ} и т.д., остановившись на определенных областях объектов **X**, их нумерации и функции **A**, но сомнительно, чтобы это можно было сделать без явного произвола. Следует, однако, думать, что различные представляющиеся здесь «разумные» варианты будут приводить к оценкам «сложностей», расходящимся на сотни, а не на десятки тысяч бит. Поэтому, такие величины, как «сложность» текста романа «Война и мир», можно считать определенными с практической однозначностью.

В статье [1] Колмогоров самокритично указывает на «…один существенный недостаток: подход не учитывает «трудности» переработки программы **p** и объекта **x** в объект **y**».

Однако, недостаток этот не умаляет главного вывода статьи: «Если конечное множество **M** из очень большого числа элементов **N** допускает определение при помощи программы длины, пренебрежимо малой по сравнению с $\log_2 N$, то почти все элементы множества **M** имеют сложность **K(x)**, близкую к $\log_2 N$. Элементы $x \in M$ этой сложности и рассматриваются как «случайные» элементы множества **M.**

Контрольные вопросы к лекции #1

1. Дайте определение понятий «колмогоровская сложность», «хаотическая последовательность»
2. Объясните иллюзорность разграничения царств порядка и хаоса
3. Расскажите о исследованиях, выполненных Марковым А.А., которые привели к созданию теории марковских цепей
4. Приведите пример марковской цепи, отличный от приведенного в лекции
5. Объясните разницу между энтропией языка и энтропией речи

Литература

1. В.А. Успенский Предварение для читателей «Нового литературного обозрения» к семиотическим посланиям Андрея Николаевича Колмогорова // Новое литературное обозрение. № 24, 1997 г. http://www.kolmogorov.info/uspensky-predvarenie.html

2. А.*Н. Колмогоров.* Три подхода к определению понятия «количество информации» // Проблемы передачи информации. - 1965. - Т. 1. - № 1. - С. 3–11.

3. А. Н. Колмогоров, В. А. Успенский Алгоритмы и случайность *Теория вероятности и ее применения*, 1987, том 32:3, с. 425–455.