Р.Д.Аветисян, Д.О.Аветисян

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ИНФОРМАТИКИ

В книге рассматриваются теоретические основы наиболее часто встречающихся информационных процессов: аналого-цифрового преобразования (сканирование), сжатия (архивация), передачи но каналам связи, поиска и аналитико-синтетической обработки информации. Авторы старались найти то общее, что позволяет рассматривать эти проблемы с единой энтропийно-корреляционной точки зрения. Учитывая повышенный интерес к обеспечению секретности передаваемых но каналам связи сообщений с одновременной организацией электронной подписи, авторы рассматривают одну из наиболее перспективных криптосистем открытого шифрования — систему RSA.

ОГЛАВЛЕНИЕ

ОТ АВТОРОВ

УЖЕ НА ЭТАПЕ подготовки настоящего пособия нам пришлось серьезно задуматься над тем, насколько удачно выбрано нами его название. Ведь до сих пор различные специалисты по-разному интерпретируют содержание термина "информатика" (см. предисловие) и, использовав его, всегда надо быть готовым если не к жесткой оппозиции, то по крайней мере, к не всегда плодотворной дискуссии. Был момент, когда мы склонялись к тому, чтобы вовсе отказаться от этого термина и назвать книгу как-то иначе. Но, оценив всю курьезность ситуации, когда при преподавании дисциплины "Информатика" или даче наименований факультетам, институтам, академиям особо не заботятся о различном понимании этого термина, мы решили не отказываться от нашего понимания термина "информатика" и оставить название книги таким, каково оно есть.

В предисловии мы попытались обосновать наше понимание этого термина и показать роль и значимость компьютерных информационных технологий на современном этапе развития общества.

В шести главах рассматриваются исходные теоретические предпосылки при построении подсистем, реализующих аналого-цифровое преобразование, архивацию, защиту, передачу по каналам связи, поиск и аналитико-синтетическую обработку информации. Конкретнее, речь идет о следующих процедурах:

дискретизации непрерывных (аналоговых) сигналов путем их сканирования (развертки) и квантования;

сжатии (архивации) дискретных сообщений (текстов) с обеспечением максимально высокого уровня их сжатия, с одной стороны, и стойкости к помехам, с другой;

максимально быстрой и надежной передаче двоичной информации по каналам связи при наличии помех;

шифровании текстов с использованием открытого ключа и организации на этой основе электронной подписи;

поиске текстов;

функционировании систем, которые из больших баз данных извлекают информацию, рафинируют ее и в виде принципиально новой (доселе не существовавшей) аналитической информации выдают пользователям.

Естественно, что приведенный перечень далеко не полностью "покрывает" всю проблематику, связанную с проектированием и эксплуатацией информационных систем. Более того, стремление к доступности излагаемого материала при ограниченном объеме пособия предопре-

делило различную глубину и полноту охвата рассматриваемого круга вопросов.

Проблема дискретного представления аналоговых сигналов рассматривается лишь в классической ее постановке, а именно, определяется необходимый объем текста (числовых данных), достаточный для исчерпывающего описания (восстановления) аналоговых сигналов конечной продолжительности, с ограниченным частотным спектром и амплитудой (мощностью). Для сканирования же сигналов с неограниченным спектром частот предлагается использовать теорему Аветисяна Д.О. о полиномиальном сканировании, устанавливающую связь упомянутого выше объема текстов не с частотным спектром, а с характером поведения k-х производных сканируемых сигналов при $k \to \infty$.

Проблема архивации текстов рассматривается лишь в плане вероятностного ее анализа. Предметом особого рассмотрения служит понятие энтропии. Практически не затронут круг вопросов, связанный с алгебраической теорией кодирования.

Вопросы передачи двоичной информации рассматриваются в русле результатов, полученных К. Шенноном, а также в плане предложенной авторами схемы настройки самого канала связи к статистическим характеристикам передаваемого текста.

Проблема защиты информации при ее передаче по каналам связи рассматривается на примере анализа криптосистемы открытого шифрования RSA – одной из наиболее перспективных систем, позволяющей при открытом шифровании реализовать также электронную подпись.

В пятой и шестой главах пособия после фрагментарного изложения ряда общеизвестных фактов и определений из теории множеств и булевой алгебры приводятся результаты, полученные авторами.

При изложении текста мы старались оперировать математическим аппаратом, не выходящим за пределы разделов математики, которые читаются студентам технических и гуманитарных университетов.

Поскольку предметы рассмотрения различных глав носят достаточно самостоятельный характер, к каждой из них представлен свой список литературы.

Авторы приносят свою глубокую признательность руководству Университета, благодаря содействию которого стала возможной работа над данной книгой. Мы благодарны рецензенту – профессору Г.В. Россу за ценные замечания, а также редактору – О.Л. Глушковой.

ПРЕДИСЛОВИЕ

ВПЕРВЫЕ термин "информатика" был использован во Франции (Ф. Дрейфус) в июне 1962 г. Несколькими месяцами позже, в октябре того же года, в письме на имя директора ВИНИТИ (Всесоюзного института научной и технической информации) членом-корреспондентом АН СССР А.А. Харкевичем этот термин был рекомендован для наименования новой дисциплины – информатики.

В 1965 г. в рецензии на книгу А.И. Михайлова, А.И. Черного и Р.С. Гиляревского "Основы научной информации" профессор Я.Г. Дорфман писал: "Более логичным было бы назвать эту новую дисциплину, трактующую о принципах, методах и средствах сбора, переработки, хранения, поиска и распространения любого вида информации – информатикой" [6].

В третьем издании Большой Советской Энциклопедии (том 10, стр. 348) читаем [2]: "Информатика – дисциплина, изучающая структуру и общие свойства научной информации, а также закономерности ее создания, преобразования, передачи и использования в различных сферах человеческой деятельности".

Примерно то же самое определение термина "информатика" приведено в [7]: "Информатика – отрасль знаний, изучающая общие свойства и структуру научной информации, а также закономерности и принципы ее создания, преобразования, накопления, передачи и использования в различных областях человеческой деятельности".

Обратим внимание, что в обоих определениях предмет рассмотрения информатики ограничен лишь "научной" информацией, тогда как Я.Г. Дорфман предлагал расширить его до "любого вида" информации. В словаре же терминов по информатике на русском и английском языках на стр. 21 читаем [5]: "Информатика, теория научной информации. Informatics, information science. Отрасль знания, изучающая закономерности сбора, преобразования, хранения, поиска и распространения документальной информации и определяющая оптимальную организацию информационной работы на базе современных технических средств".

Здесь уже речь идет не об информации "любого вида" и даже не о "научной" информации, а именно о "документальной" информации. В этом определении информатики авторами специально сказано об "оптимальной организации информационной работы на базе современных технических средств". Такая оговорка представляется излишней, так как само собой разумеется, что эффективность тех или иных методов организации информационной работы в значительной степени зависит от конкретных технических средств реализации. Ведь нельзя даже

сопоставить методы организации информационной работы в библиотеках, оснащенных современными средствами вычислительной техники и множительной аппаратурой, и, скажем, в библиотеке ассирийского царя Ашшурбанипала в Ниневии (7 век до н.э.), где носителями информации служили лишь десятки тысяч глиняных плиток различных размеров. Естественно, что под "современными техническими средствами" прежде всего подразумевались современные средства вычислительной техники, что, кстати говоря, и привело к тому, что упомянутая выше оговорка со временем стала источником ряда заблуждений. Так, переставив акценты в определении информатики, ряд специалистов стал оперировать этим термином для обозначения основ вычислительной техники и программирования. Нечто близкое к этому и произошло со школьным курсом информатики, где под информатикой практически понимается совокупность вопросов по вычислительной технике и программированию, а что касается "закономерностей сбора, хранения, поиска...", то они лишь вскользь упоминаются либо вовсе упускаются из виду.

Перестановке акцентов от "закономерностей сбора..." на круг вопросов вычислительной техники и программирования в определенной мере способствовало то обстоятельство, что в отличие от традиции, сложившейся в Европе и России, в США термин "информатика" как бы не прижился, и зачастую, пытаясь трактовать содержание этого термина, специалисты рассматривают перевод не "information science", а "computer science".

Так, в [3] термин "информатика" интерпретируется как "computer science – общее название для группы дисциплин, занимающихся различными аспектами применения и разработки ЭВМ: программирование, прикладная математика, языки программирования и операционные системы, искусственный интеллект, архитектура ЭВМ".

Этот факт заслуживает сожаления, так как термины "вычислительная техника" и "программирование" (как и, впрочем, термины, им сопутствующие, такие как "архитектура ЭВМ", "языки программирования") прекрасно отражали и отражают смысл и содержание соответствующих дисциплин и использование здесь термина "информатика" скорее создает условия для дополнительной терминологической путаницы.

Действительно, если термин "информатика" включает программирование, то непонятным становится само название [3] – "Англо-русский словарь по программированию и информатике". Аналогично, если информатика включает "различные аспекты применения и разработки ЭВМ", в том числе "архитектуру ЭВМ", то становится непонятным название дисциплины "Основы информатики и вычислительной техники".

Естественно, возникает вопрос, а нужен ли термин "информатика" вообще и если да, то на каком этапе развития науки появилась объективная необходимость в нем? Ведь в течение многих лет вычислительная техника, а вместе с ней и программирование, развивались,

систематически совершенствовались, на ЭВМ решались сложнейшие задачи создания и совершенствования ядерного оружия, аэро- и гидродинамики, ряда задач математической физики и др., и вовсе не было необходимости оперировать термином "информатика". Это был период, когда ЭВМ монопольно использовались весьма узким кругом специалистов, а именно, программистами, математиками и представителями наук, где традиционно уже существовали более или менее четко сформулированные математические модели. ЭВМ отводилась довольно скромная роль – решение тех или иных конкретных математических задач. Со временем же выяснилось, что наличие соответствующих математических моделей вовсе не обязательно для применения ЭВМ и что ЭВМ можно поручить выполнять работы из самых различных сфер деятельности человека, причем это могут быть также сферы, где полностью отсутствуют и в ближайшем будущем вряд ли будут созданы какие-либо строгие математические модели. Если раньше человек общался с ЭВМ лишь через соответствующие математические модели, то со временем наличие таких моделей перестало быть необходимостью, ЭВМ стали ближе к человеку, чем сама математика. Ведь не секрет, что родители по-прежнему должны приложить немалые усилия, чтобы привить у детей вкус к математике, но значительно больше усилий от них требуется, чтобы "оторвать" детей от ЭВМ. Этому во многом способствовало появление и совершенствование операционных систем, различных сервисных программ, которые взяли на себя нетворческую, "черную" часть работы при программировании.

Важным фактором на пути "очеловечивания" ЭВМ стали и коммерческие соображения. Ведь фирмы – изготовители ЭВМ кровно заинтересованы в расширении круга пользователей, а значит, и покупателей. Отчасти, именно этим можно объяснить появление относительно недорогих персональных ЭВМ, что резко расширило круг покупателей. Возможность непосредственного общения с ЭВМ без помощи математических моделей создала реальные предпосылки для использования ЭВМ специалистами – представителями описательных, и прежде всего, гуманитарных наук, где основным средством анализа, вывода и принятия решений остается естественный язык. Ведь не секрет, что доля интеллектуальной деятельности человека, поддающаяся строгому математическому моделированию, мизерна. Если же учесть наличие порочной практики разработки математических моделей, обладающих сомнительной адекватностью исследуемым объектам, лишь с целью придания исследованию в целом большей авторитетности, "наукообразия", то станет очевидным, насколько важно непосредственное использование ЭВМ в обход этапа разработки соответствующих математических моделей.

На пути ЭВМ к "широким массам" возникла острая необходимость общаться с ЭВМ не на языке чисел и формул, а на языке естественном или хотя бы близком к естественному. Текстами естественных языков,

их кодированием, хранением и передачей по каналам связи занимались и ранее, причем на достаточно высоком научном уровне (вспомним хотя бы работы Фано, Хаффмэна, Шеннона). Но в этих исследованиях тексты естественного языка рассматривались лишь как соответствующие последовательности символов с теми или иными статистическими характеристиками. Особой необходимости рассмотрения этих текстов как носителей определенной семантики при этом не возникало.

Уже с появлением первых автоматизированных информационно-поисковых систем возникла острая необходимость в работе с текстами естественных языков с непременным учетом того, что они являются носителями определенной семантики. Стало неизбежным оперировать, наряду с понятием количества информации – центральным понятием статистической теории информации, понятием релевантности – центральным понятием информационного поиска. Глубокое же понимание релевантности было неразрывно связано с результатами, полученными в рамках ряда гуманитарных наук. В этом смысле (и не только в этом!) дальнейшее "очеловечивание" ЭВМ практически застопорилось бы, если бы не прибегли к помощи гуманитарных наук. Доминирующую значимость приобретают прикладная и структурная лингвистика, семантика, семиотика, психо- и этнолингвистика, различные алгоритмы обработки текстов на естественных языках. Приобретают характер первостепенной важности философские вопросы взаимоотношения языка и мышления. Ведь при общении с ЭВМ носителями смысловых категорий служат их языковые формулировки, путем идентификации которых, собственно, и приходится судить о расстоянии между различными семантическими категориями.

Именно вторжение ряда гуманитарных наук в область вычислительной техники и программирования привело к формированию новой научной дисциплины – информатики. Появилась объективная необходимость в новом термине – носителе факта взаимного диффундирования вычислительной техники и программирования, с одной стороны, и ряда традиционно гуманитарных наук – с другой. Таковым и оказался термин "информатика". Взаимопроникновение указанных дисциплин сопровождалось соответствующим терминологическим взаимопроникновением. Так, бывшие "технари" – специалисты в области вычислительной техники и программирования стали оперировать такими терминами, как "сценарий", "диалог", "дизайн" и др., которые ранее употреблялись лишь представителями гуманитарных наук.

Ряд специалистов справедливо обращает внимание на то, что появление этого термина совпадает по времени с появлением первых автоматизированных информационных систем [8]. Появление же последних стало возможным лишь благодаря тесному сотрудничеству представителей гуманитарных наук со специалистами по вычислительной технике и программированию.

При определении предмета исследования информатики специалисты справедливо ограничивают его закономерностями информационных про-

цессов именно в социальных коммуникациях, т.е. в искусственных информационных системах [8]. Иными словами, речь идет об информационных процессах, которые имеют место в информационых технологиях, созданных человеком.

Является ли термин "информатика" удачным или нет? Наверное, да, хотя бы потому, что в нем не содержится ничего "вычислительного" (лат. computare) или "счетного" (лат. calculatio). Тем самым подчеркивается, что на пути к информатике сняты все барьеры для людей самых различных профессий, в том числе людей, по природе своей испытывающих отвращение ко всякого рода счетам и вычислениям. Более того, наверное, настало время как-то по-иному называть и сами ЭВМ (компьютеры). Да, первые этапы их появления и эксплуатации неразрывно были связаны с процедурами сугубо вычислительного характера. В спектре выполняемых ими функций практически отсутствовали функции, непосредственно не связанные с вычислениями, и поэтому термины "ЭВМ" и "компьютеры" вполне адекватно отражали сущность обозначаемого. Со временем же этот спектр обогатился новыми функциями, и центр тяжести (доля выполняемых процедур) существенно сдвинулся в сторону процедур невычислительного характера. Так, уже вряд ли резонно называть "вычислительным" нечто, являющееся основой для построения мультимедиа-технологий.

Чтобы подчеркнуть, что информатика оперирует не числовой, а символьной информацией, занимается не собственно вычислениями, а обработкой текстовой информации, в определении термина "информатика, теория научной информации (informatics, information science)" специально оговорено, что речь идет о "документальной информации". Тем самым констатируется, что обработка цифровой информации и собственно вычисления касаются (но не являются предметом изучения) информатики лишь в той мере, в какой они нужны при обработке символьной информации. Само содержание термина "вычисление" в информатике расширяется, охватывая наряду с собственно вычислительными процедурами также и те, которые в той или иной мере связаны с кодированием и обработкой текстов, их поиском, шифровкой и передачей по различным каналам связи. О значимости обработки текстовой информации для мировой науки свидетельствуют данные, приведенные в [4] (см. табл. 1.1).

Таблица 1.1. Распределение банков данных США по типу хранимой информации

Тип данных	Относительная доля банков данных в %
библиографические	46
цифровые	34
комбинированные, текстуально-цифровые и др.	20

Проблема совершенствования средств хранения и передачи информации волнует человека с незапамятных времен. Но прошло много тысячелетий,прежде чем человек от наскальной живописи пришел к книгопечатанию. О значимости распространения информации вообще и о книгопечатании как о первой информационной революции, в частности, весьма образно высказывается отечественный специалист Г.Р. Громов [4]: «Книгопечатание выполняло для роста накапливаемых человечеством профессиональных знаний ту же роль, какую играет, например, для растений рассеяние семян. Массовое тиражирование для последующего "рассеяния" на больших пространствах зафиксированной на материальном носителе информации о новых знаниях значительно повышало вероятность событий, что хотя бы одно "семя знания попадет на благодатную почву", прозреет и в свою очередь даст "массовым тиражом" обогащенное новым знанием свое собственное "послание в будущее».

Потребовалось пять столетий со дня изобретения в 1445 г. печатного станка, чтобы человек пришел к безбумажным средствам хранения и распространения информации. Неслыханными темпами развиваются различные средства хранения и распространения информации на магнитных носителях. Появились и интенсивно совершенствуются лазерные диски, где отношение объема информации к геометрическим размерам ее носителя способно поражать самую смелую фантазию. Широкое внедрение безбумажных информационных технологий в хозяйственный механизм промышленно развитых стран создало реальные предпосылки для успешного решения уже давно назревшей проблемы оптимального территориального распределения отдельных звеньев крупных промышленных комплексов, все в большей степени приобретающих интернациональный характер. Этому во многом способствовало создание развитых сетей ЭВМ, объединенных через спутниковую связь в крупные информационные комплексы. Растет доля многоязычной информации в общем потоке. Первостепенное значение приобретают исследования в области криптографии и рациональной организации передачи информации через каналы связи. Среди этих исследований особое место занимают результаты, полученные выдающимся американским инженером-математиком К. Шенноном [9]. Им же получены чрезвычайно важные результаты, устанавливающие теоретические границы возможного при необходимости сжатия больших объемов информации для ее компактного хранения. Чтобы представить, насколько актуальна проблема сжатия текстов, вспомним, что в США "только за первые десятилетия после Второй мировой войны было накоплено 300 млн. страниц секретных документов" [4].

Еще недавно, до начала 80-х годов, деятельность крупных информационных центров характеризовалась параллельным наращиванием тематически ориентированных банков данных, с одной стороны, и

потоков информационных запросов внешних пользователей – с другой. Центральное место отводилось комплексу вопросов, в той или иной мере связанному с поиском в больших банках данных релевантной (соответствующей информационной потребности) информации и оперативным ее доведением до пользователей. Отдельные документы – элементы баз данных – рассматривались изолированно друг от друга. В последние же годы специалисты обнаружили, что при достижении объемами информации в базах данных некоторой критической отметки создается принципиальная возможность извлечения оттуда новой информации путем аналитико-синтетической обработки ранее накопленной информации.

При этом отдельные документы исходной базы данных рассматриваются не изолированно друг от друга, а как некие семантические компоненты принципиально новой информации, порождаемой в результате их совместного рассмотрения. Новая информация, полученная на основе исходного "сырья", порою оказывается настолько ценной и привлекательной, что ряд крупных информационных центров перешел к работе по замкнутому циклу, т.е. "на себя". После накопления больших объемов первичной информации осуществляется их аналитико-синтетическая обработка по различным "срезам" – заказам внутренних пользователей. Полученная при этом новая, синтетическая информация наряду с первоначально введенными документами выставляется на внешний рынок по значительно более высоким ценам по сравнению с исходной "сырой" информацией.

Г.Р. Громов образно формулирует сложившуюся вокруг деятельности крупных информационных центров экономическую ситуацию: «Разница между конечной экономической эффективностью в деятельности банков образца 70-х годов, торгующих в основном "сырой" информацией, и тем высокоавтоматизированным аналитическим комплексом, который начал складываться на их базе к концу 80-х годов, приблизительно такая же, как и между нефтяными компаниями развивающихся стран, занятыми добычей и продажей на мировом рынке в основном сырой нефти, и нефтяными гигантами, реализующими главным образом продукты многоуровневой нефтепереработки: бензин, пластмассы и т.д.».

Еще в начале 70-х годов нами был разработан специальный механизм динамического взаимодействия различных информационно-поисковых языков-стратегий, приводящего к порождению новой, синтетической информации в виде интеллектуальной подсказки по различным информационным запросам пользователей [1]. Позже удалось создать матричную модель такого взаимодействия, подробное описание которой приведено в главе 6 настоящего пособия.

ЛИТЕРАТУРА К ПРЕДИСЛОВИЮ

1. *Аветисян Д.О.* Проблемы информационного поиска. – М.: Финансы и статистика, 1981.
2. Большая Советская Энциклопедия. – 3-е изд. – М., 1972.
3. *Борковский А.Б.* Англо-русский словарь по программированию и информатике. – М.: Русский язык, 1989.
4. *Громов Г.Р.* Очерки информационной технологии. – М.: ИнфоАрт, 1993.
5. *Жданова Г.С., Колобродова Е.С., Полушкин В.А., Черный А.И.* Словарь терминов по информатике на русском и английском языках. – М.: Наука, 1971.
6. *Михайлов А.И., Черный А.И., Гиляревский Р.С.* Основы информатики. – М.: Наука, 1968.
7. Словарь иностранных слов / Под ред. А.Г. Спиркина, И.А. Акчурина, Р.С. Карпинской. – М.: Русский язык, 1987.
8. *Тараканов К.В.* Информатика. – М.: Книга, 1986.
9. *Шеннон К.* Работы по теории информации и кибернетике. – М.: Изд-во ин. лит., 1963.

ГЛАВА 1

ДИСКРЕТИЗАЦИЯ НЕПРЕРЫВНЫХ СООБЩЕНИЙ (АНАЛОГО-ЦИФРОВОЕ ПРЕОБРАЗОВАНИЕ)

АНАЛОГО-ЦИФРОВОЕ преобразование является одной из важнейших процедур современных информационных технологий. Нельзя даже представить себе, например, мультимедиа-технологии, которые смогли бы обойтись без аналого-цифрового преобразования. В результате такого преобразования аналоговые сигналы (например, речевые сигналы, кардио- или энцефалограммы, аэрофотоснимки, различные фильмы и др.) представляются в виде последовательностей цифр – своеобразных текстов, которые, в свою очередь, могут быть преобразованы в последовательности двоичных символов (см. главу 2) для последующей их обработки на ЭВМ, хранения, шифровки и передачи по каналам связи [4,13].

В общем случае каждый аналоговый сигнал можно представить некоторой непрерывной функцией от одного или более непрерывных аргументов. Мы будем рассматривать простейший случай одного аргумента, когда аналоговый сигнал представлен в виде зависимости (однозначной функции) $y = f(t)$, где в общем случае как t, так и y непрерывные величины.

1.1. СКАНИРОВАНИЕ (РАЗВЕРТКА) ФУНКЦИЙ НЕПРЕРЫВНОГО АРГУМЕНТА. ТЕОРЕМЫ ОТСЧЕТОВ И ПОЛИНОМИАЛЬНОГО СКАНИРОВАНИЯ

Первым этапом дискретизации функции $f(t)$ непрерывного аргумента t является ее замена функцией $F(n)$ дискретного аргумента $n = 0, \pm1, \pm2,...$, равной $F(n) = f(nT)$. Здесь значения $t = nT$ называются точками отсчетов, а величина T – шагом сканирования. С увеличением шага сканирования уменьшается число дискретных значений $f(nT)$,

представляющих заданную функцию $f(T)$. Но чрезмерное увеличение значения T может привести к потере возможности последующего точного восстановления функций $f(t)$ на основе значений $f(nT)$, и поэтому речь может идти лишь о той или иной точности восстановления. Естественно, возникает вопрос: существует ли для заданной функции $f(t)$ значение $T > 0$ такое, чтобы знание $f(nT)$ ($n = 0, \pm1, \pm2,...$) было достаточным для точного восстановления функции $f(t)$? Интуитивно ясно, что если для заданной функции удалось найти некоторое значение T, при котором имеет место ее восстановимость, то можно ожидать, что эта восстановимость будет иметь место также для всех значений T_* из интервала $0 < T_* \leq T$. Интуиция подсказывает также, что при прочих равных условиях, чем более "вялой" является зависимость $f(t)$ от аргумента t, тем большими могут оказаться допустимые значения T. И наоборот, при более "жестком", "резком" характере зависимости $f(t)$ мы будем вынуждены оперировать меньшими значениями T. В качестве формального, количественного показателя "жесткости" зависимости функции $f(t)$ от аргумента t могут служить различные параметры. Например, в качестве такого параметра может служить ширина спектра частот функции $f(t)$, т.е. частота самой высокой ее гармонической составляющей. В качестве другого параметра, характеризующего "жесткость" зависимости $f(t)$, может служить характер поведения ее p-й производной при $p \to \infty$. Естественно, что узкий спектр частот функции $f(t)$ свидетельствует о "вялом" характере зависимости $f(t)$. Об этом же свидетельствуют маленькие абсолютные значения p-х производных функции $f(t)$ при $p \to \infty$.

Приведенные здесь интуитивные соображения получили свою количественную формулировку в двух соответствующих теоремах – теореме отсчетов и теореме о полиномиальном сканировании. Теорема отсчетов была сформулирована и доказана в 1915 г. Уиттекером. Позже к доказательству этой теоремы и различным ее интерпретациям возвращались Неймарк (1924 г.), Котельников (1933 г.), Шеннон (1949 г.) и др. В литературе эту теорему называют также импульсной теоремой или теоремой Котельникова. Теорему о полиномиальном сканировании сформулировал и доказал Аветисян Д.О. в 1983 г. [1]. В отличие от теоремы отсчетов, которая оперирует частотными характеристиками функции $f(t)$, теорема о полиномиальном сканировании оперирует временными характеристиками этой функции, и потому эти две теоремы оказываются дополняющими друг друга. Предпочтительность и возможность применения одной или другой из этих теорем зависит от характера конкретных функций, подлежащих сканированию.

Прежде чем перейти к доказательству этих теорем, приведем без доказательства ряд фактов из теории рядов Фурье, интегральных преобразований Фурье и Лапласа, знание которых нам понадобится при доказательстве и физической интерпретации теоремы отсчетов [3, 12].

Читатели, знакомые с соответствующими разделами математики, эту часть текста могут пропустить.

Пусть $f(t)$ – некоторая периодическая функция от аргумента t с периодом 2π, имеющая на сегменте $[-\pi, \pi]$ не более конечного числа точек разрыва и абсолютно интегрируемая на этом сегменте. Тогда во всех точках дифференцируемости функции $f(t)$ имеет место

$$f(t) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} c_n e^{jnt}, \tag{1.1}$$

где

$$c_n = \int_{-\pi}^{\pi} f(t) e^{-jnt} dt \quad (n = 0, \pm 1, \pm 2, \ldots,). \tag{1.2}$$

Если $f(t)$ – периодическая функция с периодом $2l$, имеющая на сегменте $[-l, l]$ не более конечного числа точек разрыва и абсолютно интегрируемая на этом сегменте, то во всех точках дифференцируемости функции $f(t)$ имеет место

$$f(t) = \frac{1}{2l} \sum_{n=-\infty}^{\infty} c_n e^{jn\frac{\pi}{l}t}, \tag{1.3}$$

где

$$c_n = \int_{-l}^{l} f(t) e^{-jn\frac{\pi}{l}t} dt \quad (n = 0, \pm 1, \pm 2, \ldots,). \tag{1.4}$$

При рассмотрении случаев с $l \to \infty$, т.е. случаев, когда $f(t)$ не является периодической функцией, приходится иметь дело с континуальным аналогом формул (1.1) и (1.3), т.е. вместо суммирования по индексу n, пробегающему только целые значения (дискретный спектр частот), осуществлять интегрирование по непрерывно изменяющемуся параметру ω (непрерывный спектр частот).

Именно, пусть $f(t)$ – некоторая функция, определенная на всей числовой прямой, имеющая на каждом конечном сегменте не более конечного числа точек разрыва и абсолютно интегрируемая на $(-\infty, \infty)$, т.е. несобственный интеграл

$$\int_{-\infty}^{\infty} |f(t)| dt$$

есть конечная величина (интеграл сходится). Тогда во всех точках дифференцируемости этой функции имеет место

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} c(\omega) e^{j\omega t} d\omega, \tag{1.5}$$

где

$$c(\omega) = \int_{-\infty}^{\infty} f(t)e^{-j\omega t}dt. \qquad (1.6)$$

Правые части формул (1.1) и (1.3) называются комплексными формами ряда Фурье для функции $f(t)$ с периодами соответственно 2π и $2l$. Правая часть формулы (1.5) называется комплексной формой интеграла Фурье для функции $f(t)$. Формулы (1.6) и (1.5) называются также формулами прямого и обратного интегральных преобразований Фурье. Во всех этих формулах мы придерживались традиции отнесения множителей $1/2\pi$ и $1/2l$ к формулам (1.1), (1.3) и (1.5), а не к формулам (1.2), (1.4) и (1.6), как это можно встретить у ряда авторов (например, в [12]). Так мы поступили с целью более "плавного" перехода от интегральных преобразований Фурье (прямое и обратное преобразования Фурье) к рассматриваемым ниже прямому и обратному интегральным преобразованиям Лапласа, т.е. к операционному исчислению, где коэффициент $1/2\pi$ всегда относят к обратному преобразованию Лапласа – аналогу формул (1.1), (1.3) и (1.5). Заметим также, что в рамках настоящей главы вопросы, связанные с интегральными преобразованиями как Фурье, так и Лапласа, рассматриваются лишь фрагментарно, ровно в той мере, в какой это необходимо, чтобы уследить за доказательством теоремы отсчетов и понимать ее физическую сущность. Для более подробного ознакомления с кругом рассматриваемых вопросов (например, для более детального анализа необходимых и достаточных условий существования этих преобразований) можно рекомендовать специальную литературу, например, [3, 5, 9, 10].

Физическая интерпретация пары формул (1.5), (1.6) (как и, впрочем, пар формул (1.1), (1.2) и (1.3), (1.4)) заключается в представлении функции $f(t)$ как суммы (в общем случае бесконечной) ее гармонических (синусоидальных или косинусоидальных) составляющих с различными круговыми частотами ω. При этом значения амплитуды и фазы каждой слагаемой с заданным значением круговой частоты ω определяются формулой (1.6).

Замечательным свойством как рядов, так и интеграла Фурье является их физическая реальность, т.е. для каждого фиксированного значения ω значения амплитуды и фазы, полученные с помощью формул (1.2), (1.4) или (1.6), совпадают с их значениями, определенными экспериментальным путем, например, с помощью резонаторов, настроенных на данную частоту [2].

Своеобразным "свидетелем" соблюдения энергетического баланса при представлении функции $f(t)$ в виде суммы ее гармонических составляющих является теорема Парсеваля [2, 11]:

$$\int_{-\infty}^{\infty} [f(t)]^2 dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} |c(\omega)|^2 d\omega.$$

Если сравнить формулы (1.5) и (1.6) с формулами (1.3) и (1.4) (или (1.1) и (1.2)), то можно обнаружить, что если интеграл (1.4), определяющий спектральную последовательность, всегда существует (разумеется, если функция $f(t)$ интегрируема в конечном интервале $[-l, l]$), то существование интеграла (1.6), определяющего спектральную функцию $c(\omega)$, весьма проблематично. Действительно, в отличие от формулы (1.4) (или (1.2)) в формуле (1.6) мы имеем дело с несобственным интегралом, сходимость которого существенно зависит от поведения функции $f(t)$ в бесконечности. Оказывается, что для целого ряда достаточно простых и часто встречающихся функций интеграл (1.6) не сходится. В [7] в качестве примеров таких функций приводятся функции

$$f(t) = \text{const} \text{ и } f(t) = e^{j\omega t}.$$

Здесь же указываются пути преодоления сложностей, связанных со сходимостью интеграла (1.6). Проследим за ходом мысли Густафа Дёча при переходе от интегральных преобразований Фурье к интегральным преобразованиям Лапласа [7].

Выше мы молча предполагали, что аргумент функции $f(t)$ изменяется в интервале $-\infty < t < \infty$, тогда как в подавляющем большинстве практически важных задач вполне можно обойтись рассмотрением лишь интервала $0 \leqslant t < \infty$, т.е. принять, что в интервале $t < 0$ имеет место $f(t) = 0$. Тогда вместо (1.5) и (1.6) будем иметь

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} c(\omega)e^{j\omega t}d\omega = \begin{cases} f(t) & \text{при } t > 0 \\ 0 & \text{при } t < 0 \end{cases}, \qquad (1.5а)$$

где

$$c(\omega) = \int_{0}^{\infty} f(t)e^{-j\omega t}dt. \qquad (1.6а)$$

Поскольку в точках t_0 разрыва функции $f(t)$ значение интеграла (1.5а) равно среднему значению этой функции в точках левее $(t_0 - 0)$ и правее $(t_0 + 0)$ точки разрыва, то значение этого интеграла в точке $t = 0$ оказывается равным $f(0+)/2$. Далее наряду с функцией $f(t)$ будем рассматривать функцию

$$\Phi(t) = e^{-\alpha t} f(t) \quad (\alpha > 0).$$

Для этой функции из формул (1.5а) и (1.6а) будем иметь

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} c_\alpha(\omega)e^{j\omega t}d\omega = \begin{cases} e^{-\alpha t} f(t) & \text{при } t > 0 \\ 0 & \text{при } t < 0 \end{cases}, \qquad (1.5б)$$

где

$$c_\alpha(\omega) = \int_{0}^{\infty} f(t)e^{-(\alpha + j\omega)t}dt. \qquad (1.6б)$$

В силу наличия множителя $e^{-\alpha t}$ в подынтегральном выражении (1.6б) этот интеграл сходится для значительно более широкого класса функций $f(t)$, чем это имело место в интеграле (1.6а). Формулу (1.5б) можно переписать как

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} c_\alpha(\omega)e^{(\alpha+j\omega)t}d\omega = \begin{cases} f(t) & \text{при } t>0 \\ 0 & \text{при } t<0 \end{cases}. \qquad (1.5в)$$

Рассматривая далее пару формул (1.5в) и (1.6б), обнаруживаем, что имеем дело с комплексной переменной $p = \alpha + j\omega$, причем, поскольку в обеих этих формулах значение α остается неизменным, имеет место $dp = jd\omega$. Исходя из этого, формулы (1.5в) и (1.6б) можно переписать как

$$\frac{1}{2\pi j}\int_{\alpha-j\infty}^{\alpha+j\infty} L[f(t)]e^{pt}dp = \begin{cases} f(t) & \text{при } t>0 \\ 0 & \text{при } t<0 \end{cases}, \qquad (1.7)$$

где

$$L[f(t)] = c_\alpha(\omega) = \int_0^\infty f(t)e^{-pt}dt. \qquad (1.8)$$

Формула (1.8), которая заданной функции $f(t)$ (оригиналу) ставит в соответствие функцию $L[f(t)]$ от комплексного аргумента p, называется прямым преобразованием Лапласа. Функцию $L[f(t)]$ называют лапласовым изображением оригинала, т.е. функции $f(t)$. Формула (1.7) называется формулой обращения или обратным преобразованием Лапласа. Эта формула позволяет восстановить функцию $f(t)$ на основе ее изображения $L[f(t)]$.

На основе формул прямого и обратного преобразований Лапласа строится чрезвычайно важный раздел математики операционного исчисления, где устанавливается ряд соответствий между оригиналами и их лапласовыми изображениями, используемых при решении различных задач теоретического и прикладного характера.

Формулы интегральных преобразований Фурье и Лапласа, имеющие ключевое значение для современной теории связи и управления, нами будут использованы при доказательстве теоремы отсчетов.

Рассмотрим так называемую единичную функцию $1(t)$ (единичный скачок), равную нулю при $t < 0$ и единице при $t > 0$. Пользуясь формулой (1.8), можно показать, что

$$L[1(t)] = \frac{1}{p}. \qquad (1.9)$$

Нас будет интересовать также функция единичный импульс $\delta(t)$, которая формально определена как производная от единичной функции $1(t)$. Соответственно она равна нулю при $t \neq 0$ и стремится к бес-

конечности при $t = 0$. При этом считается, что

$$\int_{-\infty}^{\infty} \delta(t)dt = 1. \tag{1.10}$$

Пользуясь (1.8), можно показать, что

$$L[\delta(t)] = 1. \tag{1.11}$$

Мы будем пользоваться также двумя теоремами – теоремой умножения (свертки) в действительной области и теоремой запаздывания в действительной области. Обе эти теоремы доказываются на основе исходных формул (1.7) и (1.8) прямого и обратного преобразований Лапласа.

Теорема умножения в действительной области гласит, что если $L[f_1(t)] = L_1(p)$ и $L[f_2(t)] = L_2(p)$, то имеет место

$$L[f_1(t)f_2(t)] = \frac{1}{2\pi j}\int_{\alpha-j\infty}^{\alpha+j\infty} L_1(s)L_2(p-s)ds. \tag{1.12}$$

Теорема запаздывания в действительной области гласит, что если $L[f(t)] = L(p)$, то

$$L[f(t-\tau)] = e^{-p\tau}L(p). \tag{1.13}$$

Нам понадобится также формула Эйлера

$$e^{jt} = \cos t + j\sin t, \tag{1.14}$$

в справедливости которой легко убедиться путем разложения этих функций в соответствующие ряды Тейлора. Из (1.14) легко получить, в частности, формулу

$$\sin t = \frac{1}{2j}(e^{jt} - e^{-jt}), \tag{1.15}$$

которая также будет использована нами в ходе доказательства теоремы отсчетов [12].

Теорема отсчетов

Пусть функция $f(t)$ имеет ограниченный спектр частот, т.е. для этой функции можно найти такое конечное значение круговой частоты $\omega_0 = 2\pi/T_0$, чтобы для всех ω, удовлетворяющих условию $|\omega| > \omega_0$, имело бы место $c(\omega) = 0$ (см. формулу (1.6)). Тогда можно утверждать, что для произвольного $T > 0$, удовлетворяющего условию $T \leq T_0/2$, функцию $f(t)$ можно полностью восстановить [13] на основе совокупности ее значений в дискретных равноотстоящих точках $f(nT)$ ($n = 0, \pm1, \pm2,...,$).

Действительно, так как вне интервала $|\omega| \leq \omega_0$ функция $c(\omega)$ тождественно равна нулю, то мы можем осуществить ее периодическое продолжение по всей числовой оси $-\infty < \omega < \infty$, т.е. рассматривать $c(\omega)$ как периодическую функцию от ω с периодом $2\omega_0$. Разложив эту функцию в ряд Фурье, с помощью формулы (1.4) определим коэффициенты этого разложения:

$$c_n = \int_{-\omega_0}^{\omega_0} c(\omega) e^{-jn\frac{\pi}{\omega_0}\omega} d\omega \quad (n = 0, \pm 1, \pm 2, \ldots,). \tag{1.16}$$

Поскольку $c(\omega)$ является преобразованием Фурье функции $f(t)$, то согласно (1.5) с учетом ограниченности спектра этой функции имеем

$$f(t) = \frac{1}{2\pi} \int_{-\omega_0}^{\omega_0} c(\omega) e^{j\omega t} d\omega, \tag{1.17}$$

или, приняв $T = T_0/2 = \pi/\omega_0$,

$$f\left(n\frac{\pi}{\omega_0}\right) = \frac{1}{2\pi} \int_{-\omega_0}^{\omega_0} c(\omega) e^{jn\frac{\pi}{\omega_0}\omega} d\omega \quad (n = 0, \pm 1, \pm 2, \ldots,). \tag{1.17a}$$

Сопоставив формулы (1.16) и (1.17а), получим:

$$c_{-n} = 2\pi f\left(\frac{n\pi}{\omega_0}\right) \quad (n = 0, \pm 1, \pm 2, \ldots,). \tag{1.18}$$

Таким образом, значения $f(n\pi/\omega_0)$ при $n = 0, \pm 1, \pm 2, \ldots$ полностью определяют коэффициенты разложения в ряд Фурье функции $c(\omega)$, которые, в свою очередь, определяют саму эту функцию с помощью формулы (1.3):

$$c(\omega) = \begin{cases} \dfrac{\pi}{\omega_0} \displaystyle\sum_{n=-\infty}^{\infty} f\left(\frac{-n\pi}{\omega_0}\right) e^{jn\frac{\pi}{\omega_0}\omega} & \text{при } |\omega| \leq \omega_0 \\ 0 & \text{при } |\omega| > \omega_0 \end{cases}. \tag{1.19}$$

Подставляя значения $c(\omega)$ в (1.5), окончательно получим формулу для $f(t)$:

$$f(t) = \frac{1}{2\omega_0} \int_{-\omega_0}^{\omega_0} e^{j\omega t} \sum_{n=-\infty}^{\infty} f\left(\frac{-n\pi}{\omega_0}\right) e^{jn\frac{\pi}{\omega_0}\omega} d\omega. \tag{1.20}$$

Пользуясь формулой (1.15), формулу (1.20) можно переписать как

$$f(t) = \sum_{n=-\infty}^{\infty} f\left(\frac{n\pi}{\omega_0}\right) \frac{\sin(\omega_0 t - n\pi)}{\omega_0 t - n\pi}, \tag{1.20a}$$

или, что то же самое,

$$f(t)=\sum_{n=-\infty}^{\infty} f\left(n\frac{T_0}{2}\right)\frac{\sin\pi\left(\dfrac{2t}{T_0}-n\right)}{\pi\left(\dfrac{2t}{T_0}-n\right)}. \qquad (1.20б)$$

Таким образом, если для заданной функции $f(t)$ можно найти конечное значение ω_0 такое, чтобы для всех ω, удовлетворяющих условию $|\omega| > \omega_0$, имело бы место $c(\omega) = 0$, то функцию $f(t)$ можно восстановить на основе значений $f(nT)$ ($n = 0, \pm1, \pm2,...,$), если значение T принять равным $T = \pi/\omega_0 = T_0/2$. Покажем, что восстановимость функции $f(t)$ останется в силе, если в качестве шага сканирования взять произвольное другое значение T_* из интервала $0 < T_* < \pi/\omega_0$. Действительно, произвольному фиксированному значению T_* из этого интервала соответствует значение $\omega_* = \pi/T_* > \omega_0$. Поскольку $\omega_* > \omega_0$, то для всех ω, удовлетворяющих условию $|\omega| > \omega_*$, будет иметь место $c(\omega) = 0$, т.е. функцию $f(t)$ можно восстановить совокупностью значений $f(nT_*)$ ($n = 0, \pm1, \pm2,...,$), если, конечно, в формулах (1.20а) и (1.20б) вместо значений ω_0 и $T_0 = 2\pi/\omega_0$ использовать значения ω_* и $T_* = 2\pi/\omega_*$. В поисках же наибольшего допустимого шага сканирования для заданной функции $f(t)$ приходим к следующей формулировке.

Наибольшее допустимое значение шага сканирования для функций $f(t)$ равно

$$T_{max}=\frac{\pi}{\omega_{max}}, \qquad (1.21)$$

где ω_{max} – круговая частота самой высокой гармонической составляющей в разложении функции $f(t)$. Иными словами, ω_{max} – минимальная частота, для которой из условия $|\omega| > \omega_{max}$ следует $c(\omega) = 0$.

Фигурирующая в формулах (1.20а) и (1.20б) функция $\varphi(t) = \sin\omega_0 t/t$ (примерный характер этой функции приведен на рис. 1.1) замечательна тем, что ее прямое преобразование равно (см. выражение для разрывного множителя Дирихле [12]):

$$c(\omega)=\int_{-\infty}^{\infty}\frac{\sin\omega_0 t}{t}e^{-j\omega t}dt=\begin{cases}\pi & \text{при } |\omega|<\omega_0\\ \dfrac{\pi}{2} & \text{при } |\omega|=\omega_0,\\ 0 & \text{при } |\omega|>\omega_0\end{cases} \qquad (1.22)$$

т.е. частотный спектр строго ограничен.

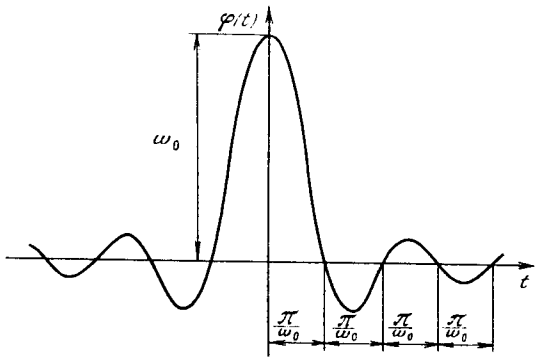

Рис. 1.1. Примерный характер функции $\varphi(t) = \frac{1}{t}\sin\omega_0 t$

Для более детального изучения того, что же происходит, когда нарушается условие $T \leq \pi/\omega_{max}$, определим частотный спектр функции

$$f_T(t) = f(t)\delta_T(t), \tag{1.23}$$

где

$$\delta_T(t) = \sum_{n=-\infty}^{\infty}\delta(t - nT). \tag{1.24}$$

Уравнение (1.23) можно рассматривать как выражение для амплитудной модуляции с несущей в виде последовательности единичных импульсов и модулирующей функции в виде функции $f(t)$.

Функция $\delta_T(t)$ представляет собой бесконечную последовательность единичных импульсов (площадь каждого из них равна единице), равноотстоящих во времени, начинающихся при минус бесконечности и продолжающихся до плюс бесконечности. Если же $f(t)$ тождественно равна нулю в области $t < 0$, то наличие или отсутствие в этой области единичных импульсов перестает иметь значение и поэтому в (1.23) функцию $\delta_T(t)$ можно принять равной

$$\delta_T(t) = \sum_{n=0}^{\infty}\delta(t - nT). \tag{1.24а}$$

Полагая, что функции $f(t)$ и $\delta_T(t)$ тождественно равны нулю в области $t < 0$, можно говорить об их лапласовых изображениях. Пользуясь теоремой запаздывания (1.13) и формулой для вычисления суммы бесконечно убывающей геометрической прогрессии, определим лапласово изображение для функции (1.24а) при $|e^{-pT}| < 1$:

$$L\left[\sum_{n=0}^{\infty}\delta(t - nT)\right] = \sum_{n=0}^{\infty}e^{-npT} = \frac{1}{1 - e^{-pT}}. \tag{1.24б}$$

Пользуясь теоремой умножения в действительной области (1.12) и

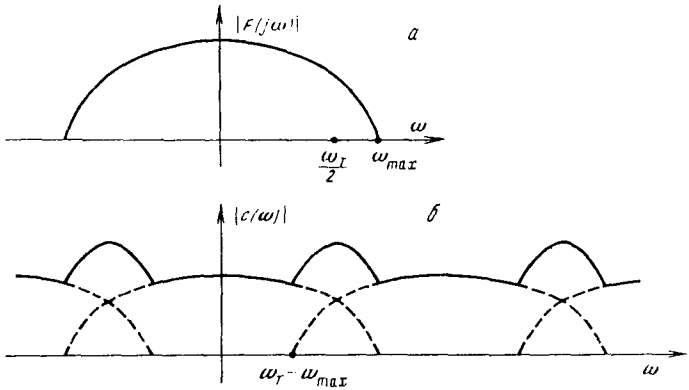

Рис. 1.2. Амплитудно-частотная характеристика функции f(t) до и после ее сканирования
а) Амплитудно-частотная характеристика функции f(t);
б) Амплитудно-частотная характеристика функции $f_T(t)$

приняв, что $L[f(t)] = F(p)$, с учетом (1.23) и (1.24б) получим

$$L\left[f_T(t)\right] = \frac{1}{2\pi j}\int_{\alpha-j\infty}^{\alpha+j\infty} F(s)\frac{1}{1-e^{-(p-s)T}}ds. \qquad (1.23б)$$

Пользуясь интегральной формулой Коши и подобрав соответствующий контур интегрирования, из (1.23б) можно получить (см., например, [8]) формулу

$$L\left[f_T(t)\right] = \frac{\omega_T}{2\pi}\sum_{k=-\infty}^{\infty} F(p + jk\omega_T), \qquad (1.23в)$$

где $\omega_T = 2\pi/T$ – круговая частота сканирования или частота следования единичных импульсов. Положив в (1.23в) $p = j\omega$, получим выражение для прямого преобразования Фурье от функции $f_T(t)$:

$$c(\omega) = \frac{\omega_T}{2\pi}\sum_{k=-\infty}^{\infty} E(j\omega + jk\omega_T). \qquad (1.25)$$

Пусть сканируемая функция $f(t)$ имеет ограниченный спектр частот, а модуль ее прямого преобразования Фурье имеет вид, представленный на рис. 1.2*а*. Выбрав в качестве шага сканирования некоторое конечное значение T, получим конкретное значение для круговой частоты сканирования, равное $\omega_T = 2\pi/T$, и, как это следует из (1.25), частотная характеристика функции $f_T(t)$ будет иметь вид, представленный на рис. 1.2*б*. Отсюда легко заметить, что в зоне $|\omega| < \omega_T - \omega_{max}$ частотные характеристики функции $f(t)$ и $f_T(t)$ совпадают (с точностью до постоянного множителя), а вне этой зоны в результате сканирования имеют место искажения. Чем больше значение $\omega_T - \omega_{max}$, тем больше неповрежденный участок частотной характеристики функции $f(t)$. Когда $\omega_T - \omega_{max} \geq \omega_{max}$, частотная характеристика функции $f(t)$ остается

невредимой во всем диапазоне своего существования $0 \leq |\omega| \leq \omega_{max}$. Именно это условие и есть условие восстановимости функции $f(t)$ на основе функции $f_T(t)$, т.е. для точного восстановления функции $f(t)$ во всем диапазоне существования ее частотной характеристики необходимо и достаточно, чтобы имело место

$$\omega_T \geq 2\omega_{max}. \tag{1.26}$$

Подставляя здесь $\omega_T = 2\pi/T$, окончательно получим условие, накладываемое на значение T для обеспечения восстановимости функции $f(t)$:

$$T \leq \frac{\pi}{\omega_{max}}. \tag{1.26а}$$

Если же это условие нарушено, то восстановленная на основе функции $f_T(t)$ функция будет иметь частотную характеристику, совпадающую с частотной характеристикой функции $f(t)$ лишь в диапазоне частот $0 \leq |\omega| < \omega_T - \omega_{max}$. Что же касается частот вне этого диапазона, то они будут искажены. Если, к примеру, шаг сканирования T выбран таким, что $\omega_T - \omega_{max} \leq 0$, то искажения будут иметь место во всем диапазоне частот.

Отсюда становится очевидным, что функции, для которых имеет место $\omega_{max} \to \infty$, т.е. функции с неограниченными спектрами частот, в результате сканирования будут искажены во всем диапазоне частот $0 \leq \omega < \infty$ при любом конечном значении шага сканирования T. А ведь на практике приходится иметь дело именно с сигналами, частотные спектры которых теоретически не ограничены. В рассматриваемом смысле применение теоремы отсчетов к реальным сигналам всегда сопровождается некоторой ошибкой. Вместе с тем, путем уменьшения шага сканирования нередко удается добиться приемлемого для практических целей уровня этих ошибок.

Дело в том, что при прохождении сигналов через однородные физические среды (например, через линейные динамические системы) они частично "теряют" свои высокие гармоники. Это обусловлено присущей таким средам инерционностью, что не позволяет им реагировать на "слишком быстрые", высокие гармоники, содержащиеся в этих сигналах. В результате эти среды становятся своеобразными фильтрами для высоких гармоник, амплитуды которых могут уменьшаться до пренебрежимо малых величин.

Пусть до того, как интересующий нас сигнал подвернется сканированию, в результате его прохождения через фильтрующие среды амплитуды всех гармоник этого сигнала в диапазоне частот $\omega_{max} < \omega < \infty$ уменьшились настолько, что ими можно пренебречь. Тогда мы практически имеем дело с сигналом, частотный спектр которого ограничен диапазоном $0 \leq |\omega| \leq \omega_{max}$.

При выборе подходящего шага сканирования руководствуемся следующими соображениями:

если планируется точное (во всем диапазоне частот) восстановление функции $f(t)$ на основе $f_T(t)$, то величина шага сканирования должна быть не более $T_{max} = \pi/\omega_{max}$;

если же при восстановлении функции $f(t)$ на основе $f_T(t)$ нас будут интересовать лишь гармоники из диапазона частот $0 \leq |\omega| \leq \omega_0$, где $\omega_0 < \omega_{max}$, то величина шага сканирования может быть увеличена до значения

$$T = \frac{2\pi}{\omega_0 + \omega_{max}}.$$

Выше уже говорилось о том, что все реальные сигналы имеют неограниченные спектры частот. Примером могут служит весьма часто встречающиеся сигналы, представленные функцией

$$f(t) = At^z e^{\sigma t} \sin(\omega t + \beta). \tag{1.27}$$

Эта функция представляет собой собственные колебания линейных динамических систем, обусловленные парой комплексно-сопряженных корней $p = \sigma \pm j\omega$ характеристических уравнений этих систем. Здесь $z + 1$ – кратность этой пары корней, а A и β – постоянные интегрирования. Частотный спектр этой функции при $\sigma \neq 0$ и/или $z \neq 0$ простирается до бесконечности, что свидетельствует о неправомочности применения к ней теоремы отсчетов. Вместе с тем, сканирование этой функции, как и целого ряда других функций, удовлетворяющих определенным условиям (см. ниже), становится вполне возможным, если пользоваться другой теоремой – теоремой о полиномиальном сканировании [1].

Теорема о полиномиальном сканировании

Пусть $f(t)$ – бесконечно дифференцируемая на всей числовой оси функция и существует такое $\lambda_0 = T_0/2 > 0$, что

$$\lim_{p\to\infty} \frac{1}{\sqrt{p}} \lambda_0^p \sup|f^{(p)}(\tau)| = 0, \tag{1.28}$$

где
$-p\lambda_0 \leq \tau \leq p\lambda_0$.

Тогда при любом $0 < T \leq T_0$ для любого фиксированного значения t справедливо

$$f(t) = \lim_{k\to\infty} \sum_{n=-k}^{k} Q\left(n,k,\frac{t}{T}\right) \; S\left(n,k,\frac{t}{T}\right) \; f(nT), \tag{1.29}$$

где

$$Q\left(n,k,\frac{t}{T}\right) = \frac{(k!)^2 \; \Gamma\left(k+1+\frac{t}{T}\right) \Gamma\left(k+1-\frac{t}{T}\right)}{(k-n)!\,(k+n)!\,\Gamma\left(k+1+\frac{t}{T}-n\right)\Gamma\left(k+1-\frac{t}{T}+n\right)}, \tag{1.29a}$$

$$S\left(n,k,\frac{t}{T}\right)=\prod_{j=1}^{k}\left[1-\left(\frac{t-nT}{jT}\right)^2\right], \tag{1.29б}$$

а $\Gamma(x)$ – гамма-функция от аргумента x, для которой, как известно, справедливы $\Gamma(1)=1$ и формула приведения (см., например, [12]):

$$\Gamma(x)=(x-1)(x-2)\ldots(x-d)\Gamma(x-d) \quad (d<x). \tag{1.29в}$$

Как сама формулировка этой теоремы, так и ее доказательство базируются на интерполяционной формуле Лагранжа, которая позволяет построить многочлен степени m, интерполирующий заданную функцию $f(t)$ в $m+1$ узлах интерполяции $t=t_i$ $(i=0, 1, \ldots, m)$

$$f(t)=\sum_{n=0}^{m} f(t_n)\prod_{\substack{j=0\\ j\neq n}}^{m}\left(\frac{t-t_j}{t_n-t_j}\right)+R_m(t), \tag{1.30}$$

где функция ошибки $R_m(t)$ равна нулю при всех $t=t_i$ $(i=0, 1, \ldots, m)$.

Выбрав в качестве узлов интерполяции $2k+1$ точки $t=0, \pm T, \pm 2T, \ldots$, где $T>0$ некоторая конечная величина, формулу (1.30) после несложных преобразований (с использованием формулы (1.29в)) можно представить как

$$f(t)=\sum_{n=-k}^{k} Q\left(n,k,\frac{t}{T}\right)\ S\left(n,k,\frac{t}{T}\right)\ f(nT)+R_{2k}(t). \tag{1.31}$$

Здесь интерполирующий многочлен Лагранжа мы специально выразили через функции $Q(n, k, t/T)$ и $S(n, k, t/T)$, чтобы подчеркнуть то обстоятельство, что при конечных значениях n и t/T имеют место

$$\lim_{k\to\infty} Q\left(n,k,\frac{t}{T}\right)=1, \tag{1.32}$$

$$\lim_{k\to\infty} S\left(n,k,\frac{t}{T}\right)=\frac{\sin\pi\left(\frac{t}{T}-n\right)}{\pi\left(\frac{t}{T}-n\right)} \tag{1.33}$$

(см., например, [2]), т.е. для любого фиксированного значения t при $k\to\infty$, когда формула (1.31) принимает вид

$$f(t)=\lim_{k\to\infty}\sum_{n=-k}^{k} Q\left(n,k,\frac{t}{T}\right)\ S\left(n,k,\frac{t}{T}\right)\ f(nT)+\lim_{k\to\infty} R_{2k}(t), \tag{1.31а}$$

слагаемые в сумме, соответствующие конечным значениям n, оказываются равными

$$\frac{\sin\pi\left(\frac{t}{T}-n\right)}{\pi\left(\frac{t}{T}-n\right)} f(nT). \tag{1.34}$$

Сопоставив это выражение с формулой (1.20б), легко обнаружить, что оно совпадает со слагаемыми, фигурирующими в этой формуле (если, конечно, положить $T = T_0/2$).

Проследим за доказательством теоремы о полиномиальном сканировании, приведенным в [1].

Выше уже говорилось о том, что во всех узлах интерполяции функция ошибки $R_m(t)$ равна нулю. Нас же будут интересовать значения этой функции для произвольных (не обязательно равных узлам интерполяции) значений t из некоторого интервала (a, b), включающего все узлы интерполяции. Известно (см., например, [6]), что при существовании у функции $f(t)$ всех производных до $(m + 1)$-й включительно имеет место

$$R_m(t) = \frac{1}{(m+1)!} f^{(m+1)}(\tau) \prod_{n=0}^{m} (t - t_n), \tag{1.35}$$

где

$$a < \tau = \tau(t) < b.$$

В случае, когда в качестве узлов интерполяции выбраны $2k + 1$ точек отсчета $t = 0, \pm T, \pm 2T, \ldots, \pm kT$, имеет место формула (1.31) с функцией ошибки $R_{2k}(t)$, равной

$$R_{2k}(t) = \frac{1}{(2k+1)!} f^{(2k+1)}(\tau) \prod_{n=-k}^{k} (t - nT), \tag{1.36}$$

где

$$\min(-kT, t) \leqslant \tau = \tau(t) \leqslant \max(kT, t).$$

Отсюда следует неравенство

$$\left|R_{2k}(t)\right| \leqslant \frac{1}{(2k+1)!} \left|\prod_{n=-k}^{k} (t - nT)\right| \sup\left|f^{(2k+1)}(\tau)\right|. \tag{1.37}$$

Для любого фиксированного значения t всегда можно найти достаточно большое значение k, такое, чтобы значение t оказалось в интервале $-kT < t < kT$. И тогда среди индексов $-k < n < k$ можно найти индекс $n = n(t)$ такой, чтобы имело место $t = T(n(t) - \alpha)$, где $0 \leqslant \alpha \leqslant 1$.

При этом будет иметь место

$$\prod_{n=-k}^{k}(t-iT)=T^{2k+1}\prod_{n=-k}^{k}(n(t)-\alpha-n)=$$
$$=T^{2k+1}(-1)^{k-n(t)+1}\cdot\frac{\Gamma(+)\,\Gamma(-)}{\Gamma(1-\alpha)\,\Gamma(\alpha)},\qquad(1.38)$$

где через Γ(+) и Γ(–) обозначены соответственно

$$\Gamma(+)=\Gamma(k+1+(n(t)-\alpha)),\qquad(1.38а)$$
$$\Gamma(-)=\Gamma(k+1-(n(t)-\alpha)).\qquad(1.38б)$$

Отсюда пользуясь свойствами гамма-функции (см., например, [12])

$$\Gamma(x)\,\Gamma(1-x)=\frac{\pi}{\sin\pi x},\qquad(1.39)$$

получим

$$\left|\prod_{n=-k}^{k}(t-nT)\right|\leqslant\frac{|\sin\pi\alpha|}{\pi}\,T^{2k+1}\,\Gamma(+)\,\Gamma(-).\qquad(1.40)$$

Пусть при некотором фиксированном значении t значение k стремится к бесконечности. Очевидно, из конечности t следует конечность также $n(t)$. И тогда, пользуясь формулами Стирлинга асимптотического разложения $\Gamma(x)$ и $x!$ (см., например, [11]), получим

$$\lim_{k\to\infty}|R_{2k}(t)|\leqslant\frac{\sqrt{2}\,|\sin\alpha\pi|}{\sqrt{\pi p}}\left(\frac{T}{2}\right)^{p}\sup\left|f^{(p)}(\tau)\right|(1+\varepsilon_p)$$
$$\left(-\frac{pT}{2}\leqslant\tau\leqslant\frac{pT}{2}\right),\qquad(1.41)$$

где $p=2k+1$, а $\varepsilon_p\to 0$ при $p\to\infty$.

Из (1.41) непосредственно следует справедливость теоремы о полиномиальном сканировании. Заметим, что в ряде случае вместо формулы (1.29) можно пользоваться эквивалентной стандартной формулой

$$f(t)=\lim_{k\to\infty}\frac{1}{T^{2k}}\sum_{n=-k}^{k}f(nT)\prod_{\substack{j=-k\\ j\neq n}}^{k}\left(\frac{t-jT}{n-j}\right).\qquad(1.42)$$

В [1] в качестве примера рассмотрена приведенная в (1.27) функция $f(t)=At^{z}e^{\sigma t}\sin(\omega t+\beta)$, которую, как уже говорилось выше, при $\sigma\neq 0$ и/или $z\neq 0$ нельзя сканировать с помощью теоремы отсчетов. Подставляя в формулу (1.28) выражение для p-й производной этой функции и переходя к пределу $p\to\infty$, можно найти уравнение для наибольших допустимых значений шага сканирования $T_0=2\lambda_0$ при заданных значе-

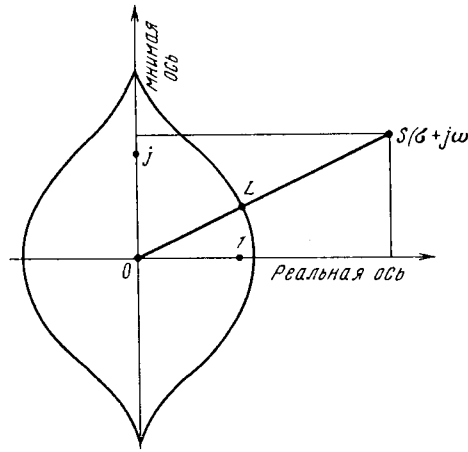

Рис. 1.3. Графический метод определения наибольшего допустимого шага сканирования

ниях σ и ω, независимо от значения $z = 0, 1, ...,$:

$$\lambda_0 R \exp\left(\lambda_0 R \left|\cos\varphi\right|\right) = 1, \tag{1.43}$$

где

$$R = \sqrt{\sigma^2 + \omega^2}, \quad \varphi = \operatorname{arctg}\frac{\omega}{\sigma}.$$

На рис. 1.3 приведена симметричная относительно мнимой и вещественной осей координат комплексной плоскости замкнутая кривая (лист)

$$0{,}5\rho \exp\left(0{,}5\rho\left|\cos\varphi\right|\right) = 1. \tag{1.44}$$

Из сопоставления (1.43) с (1.44) легко обнаружить весьма простой способ определения при заданных σ и ω наибольшего допустимого значения шага сканирования $T_0 = 2\lambda_0$, а именно, построить точку $s = \sigma + j\omega$, провести луч, проходящий через начало координат и точку s, определить точку L пересечения этого луча с листом. Значение T_0 при этом определяется как отношение длин двух отрезков:

$$T_0 = \frac{0L}{0S}$$

Следует особо отметить, что хотя в общем случае, когда $\sigma \neq 0$ и/или $z \neq 0$, применительно к функции (1.27) теорема отсчетов "не работает", в том единственном случае, когда $\sigma = z = 0$, т.е. когда речь идет о функции $f(t) = A\sin(\omega t + \beta)$ с ограниченным спектром частот и применение теоремы отсчетов становится возможным, именно ею и следует пользоваться. Дело в том, что теорема отсчетов при этом устанавливает лимит сверху на шаг сканирования, равный π/ω, тогда как значение лимита при полиномиальном сканировании оказывается равным $2/\omega$, т.е. в $\pi/2$ раза меньше.

1.2. КВАНТОВАНИЕ НЕПРЕРЫВНЫХ ЗНАЧЕНИЙ ФУНКЦИЙ

Для дискретного представления непрерывной функции $f(t)$ непрерывного аргумента после этапа ее сканирования (развертки) требуется квантование непрерывных значений $f(nT)$ ($n = 0, \pm 1, ...$), т.е. отображение вещественных чисел $f(nT)$ на некоторое счетное множество чисел, кратных шагу квантования [4]. Поскольку на практике мы всегда имеем дело со значениями $f(nT)$, которые ограничены по модулю, то такое отображение всегда приводит к конечному множеству, причем мощность этого множества тем больше, чем больше интервал возможных значений $f(nT)$ и чем меньше величина шага квантования, т.е. чем выше точность представления вещественных чисел.

Таким образом, при заданной величине шага квантования аналоговые сигналы, рассматриваемые в некотором конечном интервале значений аргумента t, несмотря на континуальный их характер, можно представить в виде некоторого текста (последовательности цифр) конечной длины. Причем, как мы убедимся во второй главе, этот текст можно подвергать дальнейшему сжатию до некоторой предельной длины l_{min}, которая, собственно, и является мерой количества информации (энтропии), содержащейся в данном сигнале. Разумеется, при прочих равных условиях значение l_{min} зависит от свойств самих сигналов. Например, из теоремы отсчетов непосредственно следует, что чем шире частотный спектр рассматриваемого сигнала, тем больше количество содержащейся в нем информации.

В заключение отметим, что на практике, при работе с конкретными сигналами, значительную помощь могут оказать сведения о специфических особенностях источников этих сигналов. Эти особенности, которые могут иметь самую различную природу (техническую, семантическую, физиологическую и др.), накладывают определенные ограничения на характер сигналов. Знание этих ограничений позволяет организовать рациональное (компактное) сканирование сигналов.

Еще большей компактности при сканировании непрерывных сигналов можно достичь, если исходя из конкретно поставленной практической задачи, т.е. из конкретных требований к процессу дискретизации, отказаться от варианта полной восстановимости функции $f(t)$ и ограничиться обеспечением восстановимости лишь тех параметров этой функции, которые действительно представляют практический интерес. Естественно, что при этом число дискретных значений, представляющих непрерывное сообщение, может оказаться значительно меньшим, чем это требовалось бы в русле рассмотренных выше теорем.

ЛИТЕРАТУРА К ГЛАВЕ 1

1. *Аветисян Д.О.* О представлении непрерывных функций одного класса дискретным множеством их значений // Проблемы передачи информации. – 1984. – Т. 20, вып. 3.
2. *Анго Андре.* Математика для электро- и радиоинженеров. – М.: Наука, 1967.
3. *Араманович И.Г., Лунц Г.Л., Эстольц Л.Э.* Функции комплексного переменного. Операционное исчисление. Теория устойчивости. – М.: Наука, 1965.
4. *Бауэр Ф., Гооз Г.* Информатика. – М.: Мир, 1976.
5. *Воробьев Н.Н.* Теория рядов. – М.: Наука, 1979.
6. *Гельфонд А.О.* Исчисление конечных разностей. – М.: Наука, 1967.
7. *Дёч Г.* Руководство к практическому применению преобразования Лапласа. – М.: Физматгиз, 1960.
8. *Джури Э.* Импульсные системы автоматического регулирования. – М.: Физматгиз, 1963.
9. *Диткин В.А., Прудников А.П.* Интегральные преобразования и операционное исчисление. – М.: Физматгиз, 1961.
10. *Диткин В.А., Прудников А.П.* Операционное исчисление. – М.: Высшая школа, 1966.
11. *Корн Г., Корн Т.* Справочник по математике для научных работников и инженеров. – М.: Наука, 1968.
12. *Романовский П.И.* Ряды Фурье. Теория поля. Аналитические и специальные функции. Преобразование Лапласа. – М.: Физматгиз, 1959.
13. *Шеннон К.* Работы по теории информации и кибернетике. – М.: Изд-во ин. лит., 1963.

ГЛАВА 2

СЖАТИЕ (АРХИВАЦИЯ) ТЕКСТОВ. ЭНТРОПИЯ КАК ПРЕДЕЛЬНАЯ МЕРА СЖАТИЯ ТЕКСТОВ. ИЗБЫТОЧНОСТЬ ТЕКСТОВ И СТЕПЕНЬ ИХ ЗАЩИЩЕННОСТИ. КОД Р. ХЭММИНГА

НАРЯДУ с привычным пониманием термина "текст" здесь под этим термином мы будем понимать любую последовательность символов независимо от их характера и назначения (цифры, буквы, знаки препинания и т.д.). Сюда войдут, например, тексты программ, цифровые представления различных изображений, аэрофотоснимков, музыки, мульт- или обычных фильмов, различных компьютерных игр и т.д. Для хранения текстов в памяти ЭВМ или на иных носителях информации часто возникает необходимость их двоичного кодирования, т.е. представления в виде тех или иных последовательностей нулей и единиц. При этом устанавливается взаимно однозначное соответствие между исходными символами (далее по тексту – буквами) и/или их различными комбинациями, с одной стороны, и отдельными двоичными символами и/или их различными комбинациями – с другой.

В результате такого кодирования исходный текст преобразуется в последовательность двоичных символов, длина которой при заданном тексте в общем случае получается различной в зависимости от выбранного метода кодирования. Естественно, что при прочих равных условиях предпочтение следует отдавать тем методам кодирования, которые исходный текст преобразуют в последовательность двоичных символов меньшей длины. Тем самым достигается экономия средств памяти и времени передачи этих текстов по каналам связи.

Пусть объектом кодирования являются тексты, записанные на некотором (естественном или искусственном) языке, причем число букв в алфавите этого языка, включая (если есть такая необходимость) некоторые знаки препинания, знак пробела и т.п., равно n. Пусть далее, l – наименьшее натуральное число, удовлетворяющее условию $l \geqslant \log_2 n$. Тогда можно пользоваться простейшим из различных методов побуквенного кодирования, сводящимся к установлению взаимно однозначного соответствия между различными буквами исходного текста и различными кодовыми наборами двоичных символов фиксированной

длины, равной l. Например, если речь идет о текстах, записанных на русском языке, где число букв алфавита, включая знак пробела, $n = 34$, то, поскольку имеет место неравенство $5 < \log_2 34 < 6$, можно осуществить побуквенное кодирование, установив следующее соответствие:

Буква русского языка	Шестисимвольный кодовый набор	Десятичная запись
(пробел)	000000	0
а	000001	1
б	000010	2
.	. . . . .	.
л	001101	13
.	. . . . .	.
я	100001	33
.	. . . . .	.
.	111111	63

Декодирование при этом осуществляется очень просто: последовательность двоичных символов – закодированный текст – делится на блоки из шести символов и каждый блок заменяется соответствующей буквой алфавита исходного текста. Невооруженным глазом видно, что, будучи очень привлекательным по своей простоте, рассмотренный метод кодирования грешит определенной "расточительностью" (избыточностью). Об этом свидетельствует хотя бы то обстоятельство, что шестью двоичными символами мы смогли бы выразить не $n = 34$, а целых $n = 2^6 = 64$ букв алфавита. Чтобы улучшить положение, можно было, например, пойти на некоторую уступку, а именно, согласиться с тем, чтобы при кодировании и декодировании текстов пары букв "е"–"ё" и "ь"–"ъ" оказались "неразличимыми". Ведь люди, владеющие русским языком, все равно смогли бы восстановить это различие при работе с уже декодированным текстом. При наличии такого согласия число букв в алфавите русского языка (включая знак пробела) оказалось бы равным $n = 32$, и поэтому можно было бы обойтись кодовыми наборами постоянной длины, равной $l = \log_2 32 = 5$. Тем самым, из каждых шести двоичных символов один символ можно было сэкономить. Из этого примера легко сделать вывод, что при побуквенном кодировании букв исходного текста кодовыми наборами постоянной длины наиболее компактное (экономное) кодирование удается осуществить тогда, когда число букв в алфавите можно представить как целую степень двойки:

$$n = 2^l \ (l = 1, 2, \ldots,). \tag{2.1}$$

Нарушение этого условия при указанном методе кодирования непременно приводит к некоторой избыточности. Возникает вопрос, а имеются ли резервы для дальнейшего сокращения среднего числа двоичных символов, отводимых под одну букву? Оказывается, что такие резервы имеются, и даже тогда, когда n удовлетворяет условию (2.1), возможны варианты, когда кодирование можно осуществить таким

образом, чтобы среднее число двоичных символов, отводимых под одну букву, оказалось меньше $l = \log_2 n$.

Пусть алфавит исходного текста состоит из восьми букв A, B, C, D, E, F, G, H. Поскольку $n = 8 = 2^3$, т.е. $l = \log_2 n = 3$, то при рассмотренном только что методе кодирования каждой букве ставился бы в соответствие кодовый набор постоянной длины, равной трем.

Пусть нам известны значения вероятностей того, что наугад взятая буква из текстов этого языка окажется буквой A, B, C, D, E, F, G или H:

p (A) = 0,08	p (E) = 0,08
p (B) = 0,44	p (F) = 0,08
p (C) = 0,08	p (G) = 0,08
p (D) = 0,08	p (H) = 0,08

С учетом неравновероятности встречаемости различных букв алфавита представляется естественным отказаться от постоянства длины кодовых наборов и стараться осуществить такое кодирование, при котором наиболее часто встречающиеся буквы были бы закодированы возможно более короткими кодовыми наборами и, наоборот, наибольшую длину имели бы кодовые наборы, соответствующие наименее часто встречающимся буквам. В русле этих соображений специалистами были разработаны различные методы побуквенного кодирования.

В связи с переходом к переменной длине кодовых наборов возникает проблема установления границ между ними при декодировании. При этом крайне нежелательно, чтобы для установления границ были использованы какие-либо специальные разделительные символы, так как это привело бы к увеличению средней длины кодовых наборов. Коды (схемы, алгоритмы кодирования), где однозначность декодирования достигается без помощи каких-либо специальных разделительных символов, называются кодами без запятой. Среди них наиболее простыми и в то же время наиболее популярными являются так называемые префиксные коды, обладающие тем свойством, что кодовый набор ни одной буквы не является началом (префиксом) кодового набора другой буквы.

Пусть n – число букв в алфавите, n_k – число букв, кодовые наборы которых состоят из k двоичных символов, l_i – число двоичных символов в кодовом наборе i-й буквы алфавита, $L = \max(l_i)$.

Тогда, очевидно, $n = \sum_{k=1}^{L} n_k$, а для произвольного фиксированного значения k имеет место $n_k \leqslant 2^k$. Если же нам заданы значения $n_1, n_2, \ldots, n_{k-1}$, то, очевидно, будет иметь место неравенство

$$n_k \leqslant 2^k - 2n_{k-1} - 2^2 n_{k-2} - \ldots - 2^{k-1} n_1,$$

т.е.

$$\sum_{j=1}^{k} 2^{k-j} n_j \leqslant 2^k,$$

или, после деления обеих частей неравенства на 2^k,

$$\sum_{j=1}^{k} 2^{-j} n_j \leq 1.$$

Поскольку выбор значения k произвольный, то примем $k = L$, и тогда будем иметь:

$$\sum_{j=1}^{L} 2^{-j} n_j \leq 1.$$

Отсюда непосредственно следует

$$\sum_{i=1}^{n} 2^{-l_i} \leq 1. \qquad (2.2)$$

Неравенство (2.2) называется неравенством Крафта и имеет ключевое значение в теории кодирования. Хотя вывод этого неравенства мы осуществили применительно к двоичному префиксному коду, оно верно также для произвольного (не обязательно двоичного и не обязательно префиксного) кода без запятой.

Неравенство Крафта, собственно, и лимитирует наше желание оперировать как можно меньшими значениями l_i. Пусть, например, $n = 10$ и уже известны значения $l_1 = 2$, $l_2 = l_3 = \ldots = l_6 = 3$. Тогда, очевидно, значения $l_7 \div l_{10}$ должны удовлетворить неравенству

$$\sum_{i=7}^{10} 2^{-l_i} \leq 1 - 2^{-2} - 5 \cdot 2^{-3} = \frac{1}{8}.$$

Пусть, например, мы хотим, чтобы имело место

$$l_7 = l_8 = l_9 = l_{10} = l.$$

Тогда получим, что значение l должно удовлетворить неравенству $4 \cdot 2^{-l} \leq 1/8$, т.е. оно не может быть меньше пяти.

Префиксный код называется полным, если добавление к нему любого нового кодового набора нарушает свойство префиксности. Пусть, например, буквам A, B и C поставлены в соответствии кодовые наборы 00, 01 и 1. Тогда очевидно, что любая попытка закодировать еще хоть одну букву привела бы к нарушению свойства префиксности. Значит, код 00, 01, 1 является полным. Если же буквам A, B и C были поставлены в соответствие кодовые наборы 00, 01 и 10, то через ветвь 11... мы смогли бы, не нарушая свойство префиксности, закодировать сколько угодно новых букв. Мы также смогли бы без нарушения свойства префиксности через ветвь 01... закодировать сколько угодно новых букв, если бы буквам A, B и C были поставлены в соответствие кодовые наборы 000, 001 и 1. Значит, коды 00, 01, 10 и 000, 001, 1 являются неполными. Для полных префиксных кодов и только для них неравенство Крафта превращается в равенство. Естественно, что на

практике наибольший интерес представляют полные коды, так как при прочих равных условиях средняя длина кодовых наборов у полных кодов получается меньше, чем у неполных.

Перейдем к рассмотрению двух полных префиксных кодов, представляющих большой практический интерес.

2.1. СХЕМА ДВОИЧНОГО КОДИРОВАНИЯ ТЕКСТОВ ПО Р. ФАНО

Предложенная американским специалистом Р. Фано схема двоичного кодирования сводится к выполнению следующих операций.

1) Составить список букв алфавита (исходное множество букв) в порядке убывания значений соответствующих им вероятностей.

2) Разбить этот список на два подсписка (подмножества букв) таким образом, чтобы значения вероятностей того, что наугад взятая из рассматриваемого текста буква окажется в первом или во втором из этих подмножеств, были бы по возможности близки.

3) Приписать произвольному одному из этих подмножеств (подсписков) символ "0", а другому – "1".

4) Рассматривая каждое из этих подмножеств (подсписков) как исходное, применительно к каждому из них осуществить операции, указанные в пунктах (2) и (3).

5) Этот процесс продолжать до тех пор, пока в каждом из очередных подмножеств не окажется по одной букве.

6) Каждой букве приписать двоичный код, состоящий из последовательности нулей и единиц, встречающихся на пути из исходного множества букв ко множеству, состоящему из одной этой буквы.

Пользуясь схемой Р. Фано (см. рис. 2.1) применительно к приведенному выше примеру, легко установить наборы двоичных символов, соответствующие буквам исходного текста:

Буква	Двоичный код	Буква	Двоичный код
A	00	E	1011
B	01	F	110
C	100	G	1110
D	1010	H	1111

Если обозначить через $L_A = 2, L_B = 3, L_C = 3,\ldots$ числа двоичных символов в кодовых наборах, соответствующих буквам A, B, C, ..., то среднее число двоичных символов, отводимых под одну букву исходного алфавита, можно определить по формуле

$$l = p(\mathrm{A})l_A + p(\mathrm{B})l_B + \ldots + p(\mathrm{H})l_H = 2{,}8.$$

Таким образом, с переходом к переменной длине кодовых наборов, отводимых под каждую букву исходного текста, удается на 7% (2,80

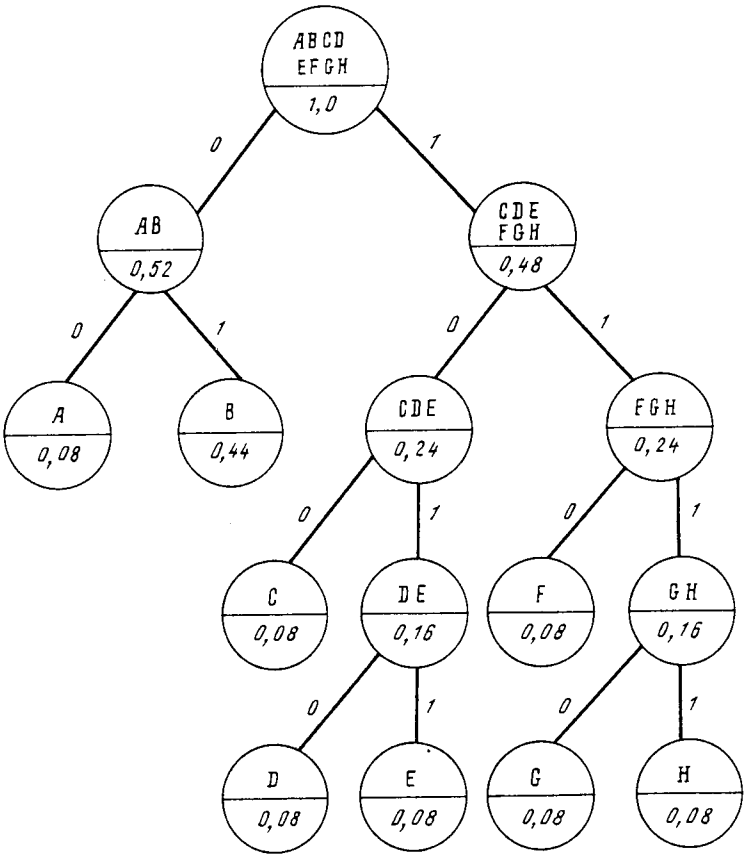

Рис. 2.1. Схема посимвольного кодирования по Р. Фано

вместо трех символов на одну букву) сократить число двоичных символов в закодированном тексте. Правда, это связано с некоторым усложнением процедур кодирования и декодирования. Будучи достаточно эффективной, схема кодирования. Р. Фано не всегда гарантирует, что при заданном наборе значений вероятностей средняя длина кодовых наборов l окажется наименьшей возможной. Такую гарантию дает другая схема кодирования, предложенная американским математиком Д. Хаффмэном. Исходные соображения здесь те же, что и при рассмотрении схемы Р. Фано, однако, оперируя более тонким механизмом кодирования, Д. Хаффмэну удалось достичь наименьшего возможного при побуквенном кодировании значения средней длины кодовых наборов.

2.2. СХЕМА ДВОИЧНОГО КОДИРОВАНИЯ ТЕКСТОВ ПО Д. ХАФФМЭНУ

Предложенная Д. Хаффмэном схема кодирования заключается в следующем (см. рис. 2.2).

Формируется первый (k = 1) столбик, где все буквы алфавита записываются в порядке убывания значений вероятностей их встречаемости в исходном, подлежащем кодированию тексте. Здесь же, напротив каждой буквы, пишется соответствующее ей значение вероятности. Две буквы, занявшие в столбике предпоследнюю и последнюю позиции, с левой стороны снизу отмечаются двоичными символами соответственно "0" и "1". На рис. 2.2 эти буквы отделены от других пунктирными линиями.

2) При известном k-м столбике строится (k + 1)-й столбик по тому же принципу, что и предыдущий, с той лишь разницей, что буквы, отмеченные в предыдущем столбике двоичными символами, в последующем столбике отсутствуют. В новом столбике их "представляет" одна составная буква со значением вероятности, равным сумме вероятностей слагаемых букв.

3) Этот процесс продолжается до тех пор, пока в очередном столбике под номером $k = n$ (число букв в алфавите) не окажется одна-единственная составная буква, "представляющая" весь алфавит исходного текста со значением вероятности, равным единице. Этот последний столбик выполняет лишь контрольную функцию.

4) Для определения кодового набора, соответствующего интересующей нас букве, поступаем следующим образом.

Начиная с последнего столбика, переходя поочередно к предыдущим столбикам, рассматриваем каждый раз только те буквы, которые занимают в очередных столбиках последние два места (участки ниже пунктирных линий). Если в указанном участке очередного столбика оказывается интересующая нас буква или какая-либо составная буква, включающая эту букву, то в очередном разряде ее кодового набора записываем двоичный символ, которым отмечена эта буква или включающая ее составная буква. Этот процесс продолжаем до тех пор, пока в очередном столбике, ниже пунктирной линии, не встретится интересующая нас буква в сольном варианте (не в составе какой-либо составной буквы).

Проследим, например, за получением кодового набора, соответствующего букве E. При движении от столбика с номером $k = 8$ к столбику с номером $k = 7$, в нем ниже пунктирной линии обнаруживаем составную букву $_0$(ACDEFGH), содержащую букву E. Это дает основание для записи в первом разряде кодового набора буквы E символа "0", которым отмечена составная буква $_0$(ACDEFGH). В следующий раз ниже пунктирной линии буква E встречается в составной букве $_0$(EFGH) в столбике под номером $k = 6$, поэтому во второй позиции кодового

$k = 1$	
B	0,44
A	0,08
C	0,08
D	0,08
E	0,08
F	0,08
- - - - - - - - - - - - - - -	
$_0$G	0,08
$_1$H	0,08

$k = 2$	
B	0,44
(GH)	0,16
A	0,08
C	0,08
D	0,08
- - - - - - - - - - - - - - -	
$_0$E	0,08
$_1$F	0,08

$k = 3$	
B	0,44
(GH)	0,16
(EF)	0,16
A	0,08
- - - - - - - - - - - - - - -	
$_0$C	0,08
$_1$D	0,08

$k = 4$	
B	0,44
(GH)	0,16
(EF)	0,16
- - - - - - - - - - - - - - -	
$_0$(CD)	0,16
$_1$A	0,08

$k = 5$	
B	0,44
(ACD)	0,24
- - - - - - - - - - - - - - -	
$_0$(GH)	0,16
$_1$(EF)	0,16

$k = 6$	
B	0,44
- - - - - - - - - - - - - - -	
$_0$(EFGH)	0,32
$_1$(ACD)	0,24

$k = 7$	
- - - - - - - - - - - - - - -	
$_0$(ACDEFGH)	0,56
$_1$B	0,44

$k = 8$	
(ABCDEFGH)	1,00

Рис. 2.2. Схема посимвольного кодирования по Д. Хаффмэну

набора буквы E записываем символ "0". Ниже пунктирной линии буква E встречается также в пятом столбике, в составной букве $_1$(EF), что дает основание для того, чтобы в третьей позиции буквы E записать символ "1". Далее буква $_0$E ниже пунктирной линии и уже в сольном варианте встречается во втором столбике. Исходя из этого, четвертую позицию кодового набора заполняем символом "0" и на этом останавливаемся, т.е. в итоге букве E приписываем кодовый набор 0010. Поступая аналогичным образом, получим кодовые наборы, соответствующие остальным буквам алфавита:

Буква	Двоичный код	Буква	Двоичный код
A	011	E	0010
B	1	F	0011
C	0100	G	0000
D	0101	H	0001

Среднее число двоичных символов, отводимых под одну букву исходного алфавита, здесь получается равным $l = 2{,}6 = l_{\min}$ (против $l = 2{,}8$ при схеме Р. Фано), что свидетельствует о том, что код Р. Фано хотя и экономный, но не оптимальный, так как, в отличие от схемы Д. Хаффмэна, схема кодирования по Р. Фано не всегда обеспечивает наименьшую возможную среднюю длину кодовых наборов.

Обе эти схемы ориентированы на то, чтобы ценою удлинения кодовых наборов менее вероятных букв достичь уменьшения длин кодовых наборов более вероятных букв. С этой задачей в общем-то справляются обе схемы кодирования, и поэтому с их помощью удается уменьшить среднее число двоичных символов, приходящихся на одну букву. Для того же, чтобы достичь наименьшую возможную средней длины кодового набора, требуется нечто большее, более тонкий механизм кодирования, нежели схема Р. Фано. Что же касается схемы Д. Хаффмэна, то в [1], например, приведено чрезвычайно простое и вместе с тем достаточно строгое доказательство того, что при любом наборе значений вероятностей эта схема обеспечивает наименьшую возможную при побуквенном кодировании среднюю длину кодовых наборов. В рассматриваемом смысле предложенная Д. Хаффмэном схема кодирования является оптимальной.

Следует особо подчеркнуть, что из оптимальности схемы Д. Хаффмэна вовсе не следует единственность варианта достижения этого оптимума. Рассмотрим, например, кодирование букв A, B, C, D и E при следующих значениях вероятностей их встречаемости:

P (A) = 0,49 P (D) = 0,16
P (B) = 0,17 P (E) = 0,01
P (C) = 0,17

С помощью схемы кодирования Д. Хаффмэна приходим к результату:

Буква	Двоичный код	Буква	Двоичный код
A	1	D	0010
B	01	E	0001
C	000		

т.е. к значению средней длины кодовых наборов, равному $l_{\min} = 2{,}02$ символа на букву.

Легко убедиться, что к такому же значению средней длины мы пришли бы при следующем варианте кодирования:

Буква	Двоичный код	Буква	Двоичный код
A	1	D	010
B	000	E	011
C	001		

Несмотря на свою неоптимальность, схема Р. Фано тем не менее обеспечивает значения l, достаточно близкие или в точности совпадающие с $l_{\min}$. В частности, легко убедиться в том, что в случае, когда

все $p(i)$ (i = A, B,...) можно представить как $p(i) = 2^{-m_i}$, где m_i – натуральные числа, в результате кодирования по этим схемам непременно приходим к одинаковым значениям l. Более того, в этих случаях, независимо от того, используется ли схема Р. Фано или Д. Хаффмэна, число двоичных символов в коде каждой i-й буквы алфавита оказывается равным $l_i = -\log_2 p(i) = m_i$, т.е. среднее число двоичных символов, приходящихся на одну букву, оказывается равным

$$l = l_{\min} = \sum_{i=1}^{n} p(i) l_i = -\sum_{i=1}^{n} p(i) \log_2 p(i).$$

Забегая вперед, отметим, что выражение, фигурирующее в правой части этого равенства, называется энтропией и имеет ключевое значение в теории информации. Это выражение остается в силе при произвольных значениях $p(i)$, вовсе не обязательно удовлетворяющих условию $p(i) = 2^{-m_i}$. Более подробно понятие энтропии мы рассмотрим чуть позже, в следующем параграфе этой главы.

Говоря об оптимальности кода Р. Хаффмэна, следует запомнить, что здесь пока речь идет лишь о побуквенном кодировании, и поэтому в общем случае схема кодирования по Д. Хаффмэну в том виде, в каком мы ее рассматривали, также не является пределом компактного представления исходных текстов последовательностью двоичных символов. Если, вслед за отказом от постоянства длин кодовых наборов, отказаться также от побуквенного кодирования, т.е. допустить кодирование сразу нескольких букв (комбинаций букв), то в общем случае можно добиться значительно большего эффекта сжатия (компрессии) исходных текстов. Заметим при этом, что при заданных статистических характеристиках исходного текста чем длиннее комбинации букв, подвергающиеся кодированию, тем, при прочих равных условиях, большим получается эффект сжатия. Естественно, что с увеличением длины этих последовательностей растет также время, расходуемое на их кодирование, и поэтому чрезмерно сильное сжатие исходных текстов не всегда оправдано, так как оно может быть связано с большими затратами машинного времени.

2.3. ПОНЯТИЕ ЭНТРОПИИ И ПРЕДЕЛЬНЫЕ ВОЗМОЖНОСТИ ПРИ СЖАТИИ ТЕКСТОВ

В основу оценки теоретических границ возможного при сжатии текстов с заданными статистическими характеристиками легли фундаментальные работы К. Шеннона, открывшие принципиально новую страницу в истории оптимального кодирования, шифровки и передачи текстов. Несколькими статьями К. Шеннона фактически было положено начало новой научной дисциплине – теории информации [3]. Ряд авторов так и называет эту дисциплину – "шенноновской теорией

информации" [2]. Центральным понятием этой теории является энтропия – количественная мера неопределенности, или, что то же самое, количество информации, необходимое для устранения имеющейся неопределенности.

С понятием энтропии мировой науке повезло дважды. В первый раз это случилось во второй половине прошлого столетия, когда Л. Больцман впервые ввел в рассмотрение это понятие для количественной оценки степени неопределенности, хаотичности состояния термодинамических систем. Во второй раз это произошло уже в конце первой половины нашего столетия, когда К. Шеннон использовал это понятие для оценки количества информации, необходимого для устранения имеющейся неопределенности. В обоих случаях понятие энтропии привело к коренному пересмотру существующих взглядов на соответствующие объекты исследования и формированию принципиально новых идей и взглядов.

В простейшем случае, когда значения вероятностей появления в тексте тех или иных букв алфавита не зависят от того, какие именно буквы им предшествовали, величину энтропии можно вычислить по формуле

$$H_1 = -\sum_{i=1}^{n} p(i) \cdot \log_2 p(i), \tag{2.3}$$

где под $p(i)$ подразумевается вероятность того, что наугад взятая из текста буква окажется i-й буквой алфавита. Если к тому же имеет место $p(i) = p(j) = 1/n$, то формула вычисления энтропии (2.3) упрощается и принимает вид

$$H_1 = \log_2 n. \tag{2.3а}$$

Значение энтропии H_1, вычисленное по формуле (2.3), устанавливает количественную меру "проблематичности" угадывания того, какой именно является наугад взятая буква, если известны все значения вероятностей $p(i)$ ($i = 1, 2,..., n$). Иными словами, это среднее количество информации, которое необходимо сообщить угадывающему, чтобы полностью ликвидировать имеющуюся у него неопределенность; количество информации, достаточное для того, чтобы он точно знал, какой именно является угадываемая буква. Формула (2.3) обладает рядом свойств, которые хорошо согласуются с интуитивными представлениями о степени "проблематичности" угадывания буквы. Так, H_1 является непрерывной функцией от $p(i)$ и при каждом фиксированном значении n достигает своего наибольшего возможного значения, когда все буквы равновероятны, т.е. когда

$$p(i) = p(j) = \frac{1}{n} \quad (i, j = 1, 2, \ldots, n). \tag{2.4}$$

При соблюдении условия (2.4) H_1 монотонно возрастает с увеличе-

нием n. И наконец, для функции H_1 справедливо

$$H_1 = -\sum_{i=1}^{k} p(i) \cdot \log_2 p(i) - \lambda_k \log_2 \lambda_k - \lambda_k \sum_{i=k+1}^{n} \frac{p(i)}{\lambda_k} \log_2 \frac{p(i)}{\lambda_k}, \qquad (2.5)$$

где

$$k < n, \quad \lambda_k = \sum_{i=k+1}^{n} p(i).$$

Это говорит о том, что исходную неопределенность можно представить как аддитивную сумму двух неопределенностей, а именно:

неопределенность угадывания того, является ли наугад взятая буква первой, второй,..., k-й или произвольной другой буквой из оставшихся $n - k$ букв алфавита;

при условии, что угадываемая буква является одной из $k + 1$, $k + 2$,..., n, букв алфавита, неопределенность, связанная с угадыванием того, какой же она окажется конкретно.

В приложениях важное место занимают случаи, когда алфавит рассматриваемого языка содержит две буквы. Примем, что ими являются двоичные символы "0" и "1". Для этого случая формула (2.3) примет вид

$$H_1 = -p(0)\log_2 p(0) - p(1)\log_2 p(1), \qquad (2.6)$$

или, так как $p(0) + p(1) = 1$,

$$H_1 = -p\log_2 p - (1-p)\log_2 (1-p), \qquad (2.7)$$

где через p обозначено какое-либо одно из значений $p(0)$ или $p(1)$, что, как легко заметить из формул (2.6) и (2.7), все равно. Функция (2.7) графически представлена на рис. 2.3. Наибольшее возможное значение $H_1 = 1$, которое достигается при $p = 0{,}5$, и принято за единицу измерения количества информации. Эту единицу по предложению Тьюки назвали бит (английское bit ⟨binary "двоичный" + digit "знак, цифра"⟩). Это название представляется не очень удачным, поскольку является источником весьма распространенного заблуждения, заключающегося в убеждении, что информационная нагрузка одного двоичного символа всегда равна одному биту.

Один бит – это количество неопределенности, которое имеет место при необходимости угадывания того, какое из двух равновероятных (и независимых от результатов предыдущих экспериментов) событий будет иметь место при очередном испытании. Иными словами, это количество информации, содержащееся в сообщении о том, какое именно из двух равновероятных событий имело место. Если же эти события не являются рав-

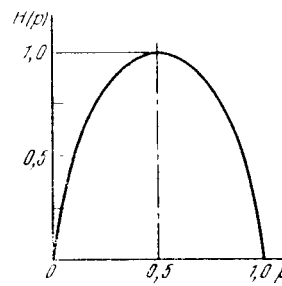

Рис. 2.3. Примерный характер функции К. Шеннона H(p) в области $0 \leqslant p \leqslant 1$

новероятными, то информационная нагрузка одного двоичного символа становится меньше одного бита и может уменьшаться до нуля.

В реальных текстах вероятности появления тех или иных букв алфавита в значительной степени зависят от того, какие именно буквы им предшествовали. Например, значение вероятности того, что очередная буква в произвольном русском тексте окажется мягким знаком, близка к нулю, если ей предшествовал набор букв "рассматрив?", и близка к единице, если предшествовал набор букв "рассматриват?". Интуитивно ясно, что знание предшествующих букв в общем случае облегчает угадывание очередной буквы. Ясно также, что чем больше число известных предшествующих букв, тем, при прочих равных условиях, в большей степени облегчается угадывание очередной буквы. Такая зависимость присутствует в любом связном тексте (будь это обычные тексты, цифровые представления изображений, музыки, фильмов или связный текст произвольного иного характера) и обусловлена синтаксисом рассматриваемого языка и (что важнее, хотя и менее очевидно) смысловой нагрузкой (семантикой) текста.

Чтобы осуществить количественную оценку указанного обстоятельства, наряду с (2.3) К. Шеннон рассматривал следующую формулу:

$$H_m = -\frac{1}{m}\sum_s p(B_{ms})\log_2 p(B_{ms}), \qquad (2.8)$$

где $p(B_{ms})$ – вероятность того, что наугад взятая из текста последовательность букв длины m окажется последовательностью B_{ms}. Суммирование в (2.8) ведется по всем n^m возможным последовательностям длины m. В качестве энтропии угадывания очередной буквы связного текста К. Шеннон предлагает пользоваться значением

$$H = \lim_{m\to\infty} H_m. \qquad (2.9)$$

Естественно, что H_m – монотонно убывающая функция от m. Лишь в том частном случае, когда вероятности появления в тексте тех или иных букв не зависят от того, какие именно буквы им предшествовали, значение H_m не зависит от m и равно $H_m = H_1$ (см. формулу (2.3)).

К. Шенноном была доказана теорема о том, что путем кодирования достаточно длинных последовательностей букв можно добиться того, чтобы среднее число двоичных символов (l), приходящееся на одну букву, было бы сколь угодно близким к значению H. Он же доказал, что величина H является нижним пределом значения l. Это и естественно, если энтропия угадывания одной буквы исходного текста равна H битам, то для того, чтобы ликвидировать это количество неопределенности, необходимо заполучить как минимум H бит информации. С другой стороны, известно (см. рис. 2.3), что каждый двоичный символ может содержать не более одного бита информации и поэтому необходимо "взглянуть" как минимум (в лучшем случае) на $l = H$ двоичных

символов, чтобы заполучить достаточное (т.е. равное H) количество информации. Несмотря на теоретическую достижимость условия $I = H$, на практике оно может оказаться труднореализуемым. Ведь, как это следует из рис. 2.3 и из формулы (2.7), для доведения информационной нагрузки одного двоичного символа до одного бита необходимо, чтобы соблюдались следующие два условия:

1) появления двоичных символов "0" и "1" в закодированном тексте должны быть равновероятными;

2) значения вероятностей появления этих символов не должны зависеть от того, какие именно символы им предшествовали.

Нарушение хотя бы одного их этих условий непременно приводит к тому, что значение I оказывается больше H, т.е. закодированный текст оказывается неоптимальным, избыточным.

Как уже говорилось выше, при кодировании последовательностей (B_{ms}) букв длины m значение H_m оказывается монотонно убывающей функцией от m. На практике, однако, приходится учесть, что с увеличением m растет также число всевозможных вариантов B_{ms}, что влечет большие затраты машинного времени. Практическое кодирование связного текста сводится к выбору некоторого компромисса, а именно, достижения приемлемого значения I за приемлемый промежуток времени.

Заметим, что указанный метод кодирования можно использовать для уменьшения I даже при независимости значений вероятностей $p(i)$ от того, какие буквы им предшествовали. Как уже говорилось выше, в этих случаях имеет место $H = H_m = H_1$ и для вычисления значения H вместо формулы (2.9) можно пользоваться формулой (2.3).

Рассмотрим пример.

Пусть требуется закодировать текст, записанный на языке, в алфавите которого имеются три буквы. Пусть далее известно, что независимо от того, какие буквы ей предшествовали, значения вероятностей того, что наугад взятая буква окажется буквой A, B или C, соответственно равны:

$p(\mathrm{A}) = 0{,}8$

$p(\mathrm{B}) = 0{,}1$

$p(\mathrm{C}) = 0{,}1$

Пользуясь формулой (2.3), определяем значение

$$H = -(p(\mathrm{A})\log_2 p(\mathrm{A}) + p(\mathrm{B})\log_2 p(\mathrm{B}) + p(\mathrm{C})\log_2 p(\mathrm{C})) = 0{,}92.$$

Применительно к этому примеру схемы побуквенного кодирования по Р. Фано и Д. Хаффмэну привели бы к одинаковым результатам:

A 0

B 10

C 11

т.е. к значению I, равному $I = 0{,}8 + 2 \cdot 0{,}2 = 1{,}2$. Поскольку при побук-

венном кодировании значение $I = 1,2$ оказалось заметно большим значения $H = 0,92$, то естественно попробовать уменьшить это значение путем кодирования двухбуквенных комбинаций букв. Так как значения появления в тексте тех или иных букв алфавита не зависят от того, какие буквы им предшествовали, то для различных двухбуквенных комбинаций получим следующие значения вероятностей:

$p(AA) = 0,64$	$p(BA) = 0,08$	$p(CA) = 0,08$
$p(AB) = 0,08$	$p(BB) = 0,01$	$p(CB) = 0,01$
$p(AC) = 0,08$	$p(BC) = 0,01$	$p(CC) = 0,01$

Рассматривая эти комбинации букв как самостоятельные буквы некоего гипотетического языка с девятью буквами в алфавите, мы сможем пользоваться схемами побуквенного кодирования Р. Фано и Д. Хаффмэна. Применительно к данному случаю эти схемы приводят к одинаковым результатам – среднее число двоичных символов, приходящихся на одну двухбуквенную комбинацию, оказывается равным $I = 1.92$.

Таким образом, среднее число двоичных символов, приходящихся на одну букву, оказывается равным $I = 0,96$. Поскольку даже по схеме Д. Хаффмэна значение I продолжает оставаться ощутимо большим значения H, то для дальнейшего его уменьшения следует осуществить кодирование трехбуквенных комбинаций $m = 3$ и так далее до достижения приемлемого значения $I \geqslant H$.

2.4. ИЗБЫТОЧНОЕ КОДИРОВАНИЕ. ИЗБЫТОЧНОСТЬ И УЯЗВИМОСТЬ ИНФОРМАЦИИ. ЗАЩИТА ИНФОРМАЦИИ ОТ СЛУЧАЙНЫХ ПОМЕХ. КОД Р. ХЭММИНГА

Говоря об оптимальном (в смысле максимального сжатия) кодировании текстов, мы имели в виду достижение условия $I = H$. Если же это условие не было достигнуто, то говорили, что имеет место неоптимальное, т.е. избыточное кодирование. Количественно избыточность можно оценить, например, разностью $\delta = I - H$ или, в процентах, $(\delta/I) \cdot 100\%$. Достижение условия $I = H$ обеспечивает максимально возможное сжатие исходных текстов (избыточность нулевая). При этом закодированный текст оказывается предельно сжатым и поэтому абсолютно беззащитным к случайным ошибкам. Если на уровне хоть одного двоичного символа оптимально закодированного текста произошла ошибка, то мы оказываемся теоретически лишенными возможности как-то обнаружить ее, а тем более исправить. Интуитивно ясно, что наличие некоторой избыточности создало бы принципиальную возможность обнаруживать (обнаруживающие коды), а в некоторых случаях и

исправлять (исправляющие коды) ошибки. Сказанное, однако, не означает, что сам факт наличия некоторой избыточности уже является достаточным для обнаружения или исправления ошибок. Наличие избыточности создает лишь теретическую, принципиальную возможность обнаружения или исправления ошибок. Для того же, чтобы она "работала на нас", всецело была направлена на обнаружение и исправление ошибок предполагаемого характера, эту избыточность следует специально "конструировать", что, собственно, и является предметом изучения чрезвычайно интересного и увлекательного раздела прикладной математики, занимающегося конструированием кодов, обнаруживающих и исправляющих ошибки. Там же устанавливаются количественные оценки того, на что именно мы вправе рассчитывать (обнаружение одной, двух... и т.д. ошибок, их исправление) при том или ином уровне избыточности.

Рассмотрим пример.

Пусть нам предстоит закодировать текст, записанный на некотором языке, таком, что число букв в алфавите этого языка $n = 2^m$ (m целое число), а появление в тексте тех или иных букв алфавита равновероятно и не зависит от того, какие буквы им предшествовали. Тогда имеем

$$p(i) = p(j) = \frac{1}{n}; \quad H = H_1 = \log_2 n = m.$$

Условия задачи таковы, что достичь оптимального кодирования можно самым незатейливым методом кодирования – побуквенным кодированием с постоянной длиной ($l = m$) кодовых наборов. При этом, однако, мы оказались бы лишенными какой-либо возможности обнаруживать, а тем более исправлять ошибки. Чтобы такая возможность появилась, необходимо отказаться от оптимальности кода, "раскошелиться" на несколько дополнительных двоичных символов на букву, т.е. умышленно ввести некоторую избыточность, которая смогла бы помочь нам обнаружить или исправить ошибки. Необходимое число дополнительно вводимых двоичных символов на одну букву обозначим через x, и тогда длина кодового набора станет равной $l = m + x$. Примем, что в результате помех (случайных или преднамеренных) лишь один или вовсе никакой из $m + x$ двоичных символов может превращаться из единицы в нуль или, наоборот, из нуля в единицу. Примем далее, что $1 + m + x$ событий, заключающиеся в том, что ошибка вообще не произойдет, произойдет на уровне первого, второго, ..., ($m + x$)-го символа кодового набора, равновероятны. Энтропию угадывания того, какое именно из этих $1 + m + x$ событий будет иметь место, в силу равновероятности этих событий можно определить по формуле (2.3а), т.е. она получается равной $H = \log_2 (1 + m + x)$ бит. Таким образом, для обнаружения самого факта наличия одиночной ошибки и установления ее позиции необходимо заполучить информацию в количестве не менее $H = \log_2 (1 + m + x)$ бит. Источником этой

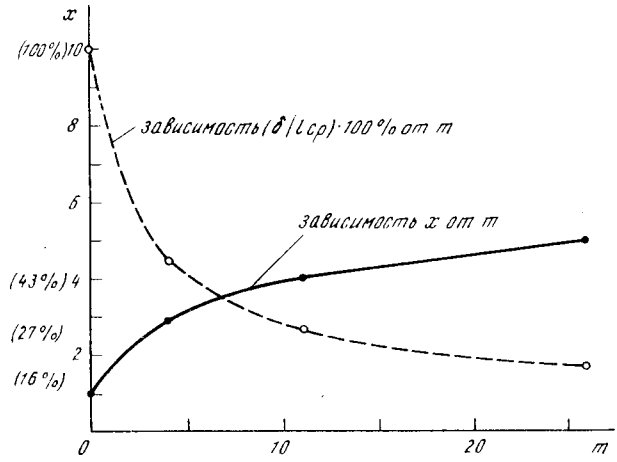

Рис. 2.4. Характер зависимости наименьшего допустимого значения параметра x от аргумента m (сплошная линия)
Характер зависимости параметра ($\delta/l_{ср}$). 100% от аргумента m (пунктирная линия)

информации служат лишь дополнительно введенные x двоичных символов, так как остальные m символов из-за оптимальности кодирования до предела заняты описанием самого текста. Выше уже говорилось о том, что x двоичных символов в лучшем случае могут содержать информацию в количестве x бит. Таким образом, при конструировании кода, обнаруживающего и исправляющего одиночную ошибку, следует учесть, что этого можно добиться лишь при значениях x, удовлетворяющих неравенству

$$x \geqslant \log_2(1 + m + x), \tag{2.10}$$

или

$$2^x - x - 1 \geqslant m. \tag{2.10а}$$

На рис. 2.4 приведена кривая, устанавливающая зависимость нижней границы допустимых значений x от m.

Р. Хэмминг разработал конкретную конструкцию кода, которая обеспечивает весьма элегантное обнаружение и исправление одиночных ошибок при минимально возможном числе дополнительно вводимых двоичных символов, т.е. при знаке равенства в (2.10) [3]. Проследим за построением этого кода, когда $m = 4$. Из рис. 2.4 следует, что при этом допустимое значение x равно трем, т.е. при числе основных (информационных) двоичных символов $m = 4$, число дополнительно введенных, т.е. контрольных символов должно быть не менее трех. Примем, что нам удалось "обойтись" именно тремя дополнительными символами, т.е. удалось сконструировать такой код, при котором каждый из дополнительно введенных трех символов дает нам максимально возможное количество информации, т.е. по одному биту. Тогда в расширенном ко-

довом наборе окажутся семь двоичных символов:

$\beta_1\beta_2\beta_3\beta_4$ $\beta_5\beta_6\beta_7$

(информационные символы) (контрольные символы)

Поскольку символы $\beta_1 \div \beta_4$ заняты кодированием собственно текста, то управлять их значениями нам не дано. Что же касается символов $\beta_5 \div \beta_7$, то они предназначены именно для обнаружения и исправления ошибок и поэтому их значения мы можем увязать со значениями информационных символов произвольными тремя функциями от аргументов $\beta_1 \div \beta_4$

$$\beta_5 = \beta_5(\beta_1 \div \beta_4), \tag{2.11}$$

$$\beta_6 = \beta_6(\beta_1 \div \beta_4), \tag{2.12}$$

$$\beta_7 = \beta_7(\beta_1 \div \beta_4) \tag{2.13}$$

такими, чтобы в последующем с помощью трех других функций от аргументов $\beta_1 \div \beta_7$

$$e_0 = e_0(\beta_1 \div \beta_7), \tag{2.14}$$

$$e_1 = e_1(\beta_1 \div \beta_7), \tag{2.15}$$

$$e_2 = e_2(\beta_1 \div \beta_7) \tag{2.16}$$

определить значения e_0, e_1, e_2, содержащие информацию о том, произошла ли ошибка вообще и если да, то на уровне какого именно из семи символов. Очевидно, имеется множество различных вариантов при выборе функций (2.11) ÷ (2.16). Р. Хэмминг поставил перед собой задачу выбора именно такой совокупности функций (2.11) ÷ (2.16), чтобы набор значений $e_2e_1e_0$ оказался двоичной записью позиции, где произошла ошибка. В случае же, когда ошибка не имела места, набор значений $e_2e_1e_0$ должен указать на "нулевую" позицию, т.е. на несуществующий символ β_0. Из двоичной эаписи этих позиций

0 0 0	(0)	1 0 0	(4)
0 0 1	(1)	1 0 1	(5)
0 1 0	(2)	1 1 0	(6)
0 1 1	(3)	1 1 1	(7)

легко уследить, что значение e_0 "несет ответственность" за позиции β_1, β_3, β_5 и β_7 и поэтому в качестве функции (2.14) берется зависимость

$$e_0 = \beta_1 + \beta_3 + \beta_5 + \beta_7 \qquad \text{mod } 2. \tag{2.14a}$$

Аналогично, обращая внимание на то, что значения $e1$ и $e2$ отвеча-

ют за позиции соответственно $\beta_2\beta_3\beta_6\beta_7$ и $\beta_4\beta_5\beta_6\beta_7$, получим

$$e_1 = \beta_2 + \beta_3 + \beta_6 + \beta_7 \qquad \text{mod } 2, \qquad (2.15а)$$

$$e_2 = \beta_4 + \beta_5 + \beta_6 + \beta_7 \qquad \text{mod } 2. \qquad (2.16а)$$

Обратим внимание, что систему (2.14а) ÷ (2.16а) можно рассматривать как развернутую запись матричного уравнения

$$\begin{vmatrix} e_0 \\ e_1 \\ e_2 \end{vmatrix} = \begin{vmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{vmatrix} \begin{vmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \end{vmatrix},$$

или

$$V_e = A \cdot V_a,$$

где V_e – вектор ошибки, указывающий на ее месторасположение; A – основная матрица, столбцы которой суть двоичные записи чисел от одного до семи.

Операция сложения во всех трех уравнениях (2.14а) ÷ (2.16а) осуществляется по модулю два. Подставляя в систему уравнений (2.14а) ÷ (2.16а) $e_0 = e_1 = e_2 = 0$, получим систему из трех уравнений

$$\beta_1 + \beta_3 + \beta_5 + \beta_7 = 0 \qquad \text{mod } 2, \qquad (2.14б)$$

$$\beta_2 + \beta_3 + \beta_6 + \beta_7 = 0 \qquad \text{mod } 2, \qquad (2.15б)$$

$$\beta_4 + \beta_5 + \beta_6 + \beta_7 = 0 \qquad \text{mod } 2, \qquad (2.16б)$$

Приняв в качестве неизвестных величины β_5, β_6 и β_7, получим систему из трех уравнений с тремя неизвестными:

$$\beta_5 + \beta_7 = \beta_1 + \beta_3 \qquad \text{mod } 2, \qquad (2.14в)$$

$$\beta_6 + \beta_7 = \beta_2 + \beta_3 \qquad \text{mod } 2, \qquad (2.15в)$$

$$\beta_5 + \beta_6 + \beta_7 = \beta_4 \qquad \text{mod } 2. \qquad (2.16в)$$

Эта система эквивалентна одному матричному уравнению

$$\begin{vmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{vmatrix} \begin{vmatrix} \beta_5 \\ \beta_6 \\ \beta_7 \end{vmatrix} = \begin{vmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix} \begin{vmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{vmatrix}, \qquad (2.17)$$

или

$$CV_c = IV_i, \qquad (2.17а)$$

где V_c и V_i векторы-столбцы, координаты которых представлены соответственно контрольными и информационными разрядами; C и I – так называемые контрольная и информационная матрицы. Столбцы этих матриц суть двоичные записи номеров соответственно контрольных и информационных разрядов.

Решение системы (2.14в) ÷ (2.16в), или, что то же самое, матричного уравнения (2.17) относительно β_5, β_6 и β_7 приводит к конкретным выражениям для функций (2.11) ÷ (2.13):

$$\beta_5 = \beta_2 + \beta_3 + \beta_4 \qquad \text{mod 2}, \qquad (2.11а)$$

$$\beta_6 = \beta_1 + \beta_3 + \beta_4 \qquad \text{mod 2}, \qquad (2.12а)$$

$$\beta_7 = \beta_1 + \beta_2 + \beta_4 \qquad \text{mod 2}. \qquad (2.13а)$$

Заметим, что сам Р. Хэмминг в качестве контрольного берет не набор символов β_{m+1}, β_{m+2}, ..., β_{m+x}, а набор символов, индексы которых представляют целые степени двойки. В случае, когда число контрольных символов равно трем, эти индексы равны $2^0 = 1$, $2^1 = 2$ и $2^2 = 4$, т.е. речь идет о наборе символов $\beta_1\beta_2\beta_4$, относительно которых решение системы (2.14б) ÷ (2.16б) чрезвычайно упрощается:

$$\beta_1 = \beta_3 + \beta_5 + \beta_7 \qquad \text{mod 2},$$

$$\beta_2 = \beta_3 + \beta_6 + \beta_7 \qquad \text{mod 2},$$

$$\beta_4 = \beta_5 + \beta_6 + \beta_7 \qquad \text{mod 2}.$$

Это и естественно, поскольку в данном случае вместо (2.17) имеем дело с матричным уравнением

$$\begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} \begin{vmatrix} \beta_1 \\ \beta_2 \\ \beta_4 \end{vmatrix} = \begin{vmatrix} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{vmatrix} \begin{vmatrix} \beta_3 \\ \beta_5 \\ \beta_6 \\ \beta_7 \end{vmatrix},$$

где контрольная матрица C всегда равна единичной матрице.

Отметив, что при указанной рекомендации Р. Хэмминга контрольная матрица всегда (независимо от m и x) оказывается равной единице, подробное обсуждение этой рекомендации оставим на потом, продолжая рассматривать в качестве контрольных $\beta_5\beta_6\beta_7$, а качестве информационных – $\beta_1\beta_2\beta_3\beta_4$.

Рассмотрим, к примеру, набор информационных символов $\beta_1\beta_2\beta_3\beta_4$ = = 1 0 1 1. С помощью зависимостей (2.11а) ÷ (2.13а) определим набор контрольных (дополнительно введенных, избыточных) символов $\beta_5\beta_6\beta_7$ = 0 1 0. Пусть ошибка произошла на уровне символа β_5, т.е. вместо истинного расширенного кодового набора 1 0 1 1 (0) 1 0 получен код 1 0 1 1 (1) 1 0. Тогда с помощью зависимостей (2.14а) ÷ (2.16а) найдем

$$e_0 = 1 + 1 + 1 + 0 = 1 \qquad \text{mod } 2$$

$$e_1 = 0 + 1 + 1 + 0 = 0 \qquad \text{mod } 2$$

$$e_2 = 1 + 1 + 1 + 0 = 1 \qquad \text{mod } 2$$

Набор значений $e_2e_1e_0$ = 1 0 1 является двоичной записью числа "пять", т.е. указывает именно на пятую позицию (на символ β_5), где, собственно, и произошла ошибка.

Приведенная схема Р. Хэмминга по конструированию кода, обнаруживающего и исправляющего одиночную ошибку, универсальна, и аналогичный код может быть построен для произвольной пары значений m и x, удовлетворяющих уравнению

$$2^x - x - 1 = m. \qquad (2.10б)$$

Заметим также, что вовсе не обязательно, чтобы набор из m информационных символов представлял собой код какой-то определенной буквы, как это имело место в только что рассмотренном примере. На практике сначала можно осуществить оптимальное (или близкое к оптимальному) кодирование текста. Затем уже закодированный текст можно делить на блоки по m двоичных символов в каждом, причем из возможных значений $m = 2^x - x - 1$ (x = 3, 4, ...) его конкретное значение следует выбирать исходя из эксплуатационной необходимости. При прочих равных условиях значение m должно быть тем меньшим, чем больше значимость информации и чем больше уровень помех. После выбора конкретного значения m каждый блок из m информационных символов следует наращивать $x = x(m)$ контрольными символами, предназначенными для обнаружения и исправления одиночных ошибок в рамках данного блока.

А теперь вернемся к рассмотрению вопроса о том, почему Р. Хэмминг в качестве контрольных берет именно символы, индексы которых равны целым степеням двойки, т.е. 1, 2, 4, 8, 16,.... Во-первых, как уже об этом говорилось выше, при таком выборе контрольная матрица всегда оказывается равной единице, т.е. фактически снимается вопрос

решения системы (2.14б) ÷ (2.16б) относительно контрольных символов, так как ее "решение" сводится к простому переписыванию соответствующих уравнений. Но это не главное, так как систему (2.14б) ÷ (2.16б) приходится решать только один раз и далее при каждом акте кодирования мы пользуемся лишь системой (2.11а) ÷ (2.13а) – решением системы (2.14б) ÷ (2.16б) относительно контрольных символов. При реализации процедур кодирования и декодирования на ЭВМ сам факт, что контрольные символы разобщены (не следуют подряд друг за другом), создает определенные неудобства при каждом акте кодирования и декодирования. Естественно поэтому желание выбрать контрольные символы таковыми, чтобы они следовали подряд друг за другом, пусть даже ценою того, чтобы один раз решить систему (2.14б) ÷ (2.16б). Именно так поступали мы, когда вопреки рекомендации Р. Хэмминга взять в качестве контрольных символы β_1, β_2 и β_4 взяли в качестве таковых символы β_5, β_6 и β_7. Хотя это и вынудило нас решить систему (2.14в) ÷ (2.16в) относительно переменных β_5, β_6 и β_7, но зато при каждом акте кодирования и декодирования мы смогли оперировать "пачками" контрольных символов, а не "выковыривать" их среди информационных символов.

Возникает вопрос: а всегда ли, при любом числе информационных символов мы смогли бы поступать аналогичным образом? Нет, не смогли бы, если по-прежнему хотим, чтобы двоичный набор символов $e_{x-1}, e_{x-2},..., e_0$ указывал на адрес ошибки. Потому что уже когда число контрольных символов больше трех, мы не имеем права взять в качестве контрольных последние x символов. Легко убедиться, что при этом контрольная матрица непременно оказалась бы вырожденной, т.е. значение ее детерминанта оказалась бы равным нулю. Более того, даже в рассмотренном нами случае, когда число контрольных символов равно трем, мы не смогли бы в качестве контрольных взять, например, первые три символа. Во всех этих случаях определители контрольных матриц (вспомним, что столбцы этой матрицы суть двоичные записи номеров выбранных нами контрольных символов) оказываются равными нулю. Пусть, например, мы выбрали в качестве контрольных не пачку символов β_5, β_6, β_7, а символы β_1, β_2, β_3. Тогда нам пришлось бы иметь дело с квадратной матрицей третьего порядка, столбцы которой являются двоичными формами записи чисел 1, 2 и 3:

$$C = \begin{vmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{vmatrix}.$$

Равенство нулю детерминанта этой матрицы свидетельствует о том, что систему (2.14б) ÷ (2.16б) нельзя решить относительно переменных β_1, β_2 и β_3.

Таким образом, при выборе среди $m + x$ символов x контрольных следует заботиться о том, чтобы определитель контрольной матрицы порядка x, столбцы которой представляют собой двоичные записи номеров выбранных символов, не оказался равным нулю. Именно чтобы избавиться от этих забот, Р. Хэмминг рекомендует в качестве контрольных взять символы с индексами 1, 2, 4, 8 и т.д. Легко обнаружить, что при таком выборе контрольных символов мы всегда (независимо от их числа) будем иметь дело с единичной матрицей.

Кроме зависимости (2.10а), на рис. 2.4 приведена также зависимость относительной избыточности $(\delta/l_{ср})\cdot 100\%$ от m. Легко заметить, что с увеличением m требуемый процент избыточности для обнаружения и исправления одиночной ошибки резко уменьшается. Столь неестественный результат является следствием искусственного, далекого от реальности допущения, что в рамках каждого кодового набора независимо от его длины $m + x$ может произойти не более одной ошибки. Если же допустить возможность двух и более ошибок, то задача их обнаружения, и тем более исправления усложняется. Построить для этих случаев коды столь же элегантные, как код Р. Хэмминга для одиночной ошибки, пока не удалось.

В заключение подытожим те основные результаты, которые при двоичном кодировании текстов представляются наиболее важными и будут использованы в последующих главах. В общем случае, когда значения вероятностей появления различных букв алфавита в исходном тексте зависят от того, какие буквы им предшествовали, значение энтропии H вычисляется по формуле (2.9). Если же такая зависимость отсутствует, то значение энтропии H определяется по более простой формуле (2.3). Формула вычисления энтропии еще более упрощается и принимает вид (2.3а), если к тому же имеет место $P(i) = P(j) = 1/n$. При заданном значении H среднее число двоичных символов, приходящихся на одну букву исходного текста, всегда (при любом методе кодирования) больше или равно H. Знак равенства достигается лишь при оптимальном кодировании текста, т.е. при максимальном его сжатии, когда избыточность доводится до нуля и каждый двоичный символ закодированного текста предельно загружен – содержит один бит информации. При этом появления в закодированном тексте символов "0" и "1" равновероятны и не зависят от того, какие символы им предшествовали.

При прочих равных условиях, чем больше избыточность текста, тем легче осуществить его несанкционированное декодирование, точнее дешифровку. В этом смысле оптимально закодированные тексты характеризуются большей защищенностью. В то же время эти тексты абсолютно беззащитны к случайным и/или умышленно введенным ошибкам – достаточно хоть одной ошибки на уровне какого-либо двоичного символа оптимально закодированного текста, и уже не только "противник", но и "свой" адресат лишится возможности декодировать –

восстановить исходный текст. Чтобы предоставить адресату хоть какую-то возможность обнаружить, а тем более исправить имеющие место ошибки, приходится отказаться от предельного сжатия текста и ввести некоторую избыточность. Но эта избыточность должна быть специально сконструирована, т.е. она должна быть нацелена на обнаружение, а если это возможно, то и исправление ошибок.

ЛИТЕРАТУРА К ГЛАВЕ 2

1. *Аршинов М.Н., Садовский Л.Е.* Коды и математика. – М.: Наука, 1983.
2. *Бауэр Ф., Гоос Г.* Информатика. – М.: Мир, 1976.
3. *Шеннон К.* Работы по теории информации и кибернетике. – М.: Изд-во ин. лит. 1963.

ГЛАВА 3

ПЕРЕДАЧА ТЕКСТОВ ПО КАНАЛАМ СВЯЗИ. ПРОПУСКНАЯ СПОСОБНОСТЬ КАНАЛОВ СВЯЗИ

В УСЛОВИЯХ параллельного функционирования целого ряда информационно-вычислительных центров (ИВЦ) страны, находящихся в различных стадиях внедрения и эксплуатации, вопросы, связанные с оптимальной организацией обмена информацией между ними, приобретают характер первостепенной важности. Сложилась ситуация, когда информационная изоляция, пусть даже подкрепленная какими-то внутрисистемными преимуществами данного ИВЦ, чревата опасностью его замораживания. Наряду с традиционными методами в настоящее время широко практикуется обмен информацией на машиночитаемых носителях.

Возросшие требования к оперативности оповещения, необходимость обеспечения концептуальной, информационной и технологической общности информационных систем различного назначения настоятельно требуют организации "перекачки" информации по соответствующим каналам связи путем создания сети взаимосвязанных центров информации. Именно отсутствие развитой сети взаимосвязанных ИВЦ, осуществляющей промышленную эксплуатацию информационных систем в масштабе страны, привело к тому, что, замкнувшись в сфере собственных лабораторных исследований, ряд специалистов направил свои усилия на решение тех или иных частных задач, потеряв много ценного времени на противопоставление в общем-то не сильно различающихся между собой подходов к их решению.

Вместе с тем очевидные успехи в области проектирования и эксплуатации сетей в ряде развитых стран, интенсивная эксплуатация информационных систем, объединенных в международные сети, позволяют надеяться, что уже начатое с этими странами многостороннее научно-техническое и экономическое сотрудничество приведет в ближайшее время к объединению в единую информационную среду дислоцированные по всему миру центры научно-технической информации. Такое объединение связано с преодолением ряда проблем экономического,

технического, технологического, программного, математического и иного характера, среди которых все отчетливее выделяется комплексная проблема защиты информации (криптография, помехоустойчивое кодирование и др.). Жесткие требования предъявляются к надежности каналов связи, при анализе которых наиболее плодотворным остается разработанный К. Шенноном аппарат статистической теории информации [2].

Каналы связи, через которые осуществляется передача информации от источника к приемнику, в общем случае могут иметь самую различную физическую природу. Абстрагируясь от природы, назначения и характера реализации конкретных каналов связи, мы будем рассматривать канал как некий "черный ящик", который в ответ на каждый поданный на его вход двоичный символ выдает на выходе соответствующий двоичный символ. Если на все поданные на вход двоичные символы канал отвечает такими же символами, то его будем называть идеальным (не шумящим, абсолютно надежным и т.д.). Если на все поданные на вход нули канал отвечает единицами и, наоборот, на все поданные на вход единицы – нулями, то его будем называть идеальным инвертором. При анализе работы реальных каналов связи приходится иметь дело с вероятностным описанием процесса передачи двоичных символов.

3.1. ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

Рассмотрим матрицу сопряженности "вход-выход" канала связи (рис. 3.1). Пусть на вход канала было подано достаточно большое число $n = a + b + c + d$ двоичных символов, из коих $a + c$ символов оказались единицами, а $b + d$ символов – нулями. В a случаях из $a + c$ случаев подачи на вход канала единиц на выходе появились единицы, в остальных c случаях – нули. Аналогично, в d случаях из $b + d$ случаев

		Вход канала	
		$x = 1$	$x = 0$
Выход канала	$y = 1$	a	b
	$y = 0$	c	d

Рис. 3.1. Матрица сопряженности "вход–выход" двоичного канала связи

подачи на вход канала нулей на выходе появились нули, в остальных b случаях – единицы. Иными словами, мы имеем дело с ансамблем двух случайных величин x и y, первая из которых характеризует вход канала, а вторая – выход. При достаточно больших $a + c$ и $b + d$ мы можем говорить о следующем наборе статистических оценок соответ-

ствующих значений вероятностей:

вероятность того, что наугад взятый из входной последовательности символ окажется единицей, –

$$p(x=1)=\omega=\frac{a+c}{n}, \tag{3.1}$$

вероятность того, что наугад взятый из входной последовательности символ окажется нулем, –

$$p(x=0)=1-\omega=\frac{b+d}{n}. \tag{3.2}$$

вероятность того, что наугад взятый из выходной последовательности символ окажется единицей, –

$$p(y=1)=\lambda=\frac{a+b}{n}, \tag{3.3}$$

вероятность того, что наугад взятый из выходной последовательности символ окажется нулем, –

$$p(y=0)=1-\lambda=\frac{c+d}{n}, \tag{3.4}$$

вероятность того, что на поданную ко входу единицу канал ответит единицей, –

$$p(y=1/x=1)=\lambda_1=\frac{a}{a+c}, \tag{3.5}$$

вероятность того, что на поданную ко входу единицу канал ответит нулем, –

$$p(y=0/x=1)=1-\lambda_1=\frac{c}{a+c}, \tag{3.6}$$

вероятность того, что на поданный ко входу нуль канал ответит нулем, –

$$p(y=0/x=0)=\lambda_2=\frac{d}{b+d}, \tag{3.7}$$

вероятность того, что на поданный ко входу нуль канал ответит единицей, –

$$p(y=1/x=0)=1-\lambda_2=\frac{b}{b+d}, \tag{3.8}$$

вероятность того, что наугад взятая выходная единица окажется ответом на входную единицу, –

$$p(x=1/y=1)=\omega_1=\frac{a}{a+b}, \tag{3.9}$$

вероятность того, что наугад взятая выходная единица окажется ответом на входной нуль, –

$$p(x=0/y=1)=1-\omega_1=\frac{b}{a+b}, \tag{3.10}$$

вероятность того, что наугад взятый выходной нуль окажется ответом на входной нуль, –

$$p(x=0/y=0)=\omega_2=\frac{d}{c+d}, \tag{3.11}$$

вероятность того, что наугад взятый выходной нуль окажется ответом на входную единицу, –

$$p(x=1/y=0)=1-\omega_2=\frac{c}{c+d}, \tag{3.12}$$

вероятность того, что наугад взятая пара "входной-выходной символы" окажется типа "единица – единица", –

$$p(1,1)=\frac{a}{n}, \tag{3.13}$$

вероятность того, что наугад взятая пара "входной-выходной символы" окажется типа "нуль – единица", –

$$p(0,1)=\frac{b}{n}, \tag{3.14}$$

вероятность того, что наугад взятая пара "входной-выходной символы" окажется типа "единица – нуль", –

$$p(1,0)=\frac{c}{n}, \tag{3.15}$$

вероятность того, что наугад взятая пара "входной-выходной символы" окажется типа "нуль – нуль", –

$$p(0,0)=\frac{d}{n}. \tag{3.16}$$

В дальнейшем изложении приведенные соотношения будем принимать за значения соответствующих вероятностей, не оговаривая каждый раз, что речь идет лишь о статистических их оценках. Примем также, что появление каждого символа как входной, так и выходной последовательностей не зависит от того, какие символы ему предшествовали.

С целью упрощения последующего изложения в некоторых из формул (3.1)÷(3.16) наряду с обозначениями типа $p(\quad)$ будем использовать также обозначения ω, ω_1, ω_2, λ, λ_1 и λ_2.

При вычислении значений вероятностей (3.1)÷(3.16) исходными являются четыре независимые переменные: a, b, c и d. Если ввести в рассмотрение набор новых переменных $\bar{a} = a/n$ $\bar{b} = b/n$, $\bar{c} = c/n$ и $\bar{d} = d/n$, то, очевидно, будет иметь место

$$\bar{a} + \bar{b} + \bar{c} + \bar{d} = 1, \tag{3.17}$$

т.е. из четырех вновь введенных переменных $\bar{a}$, $\bar{b}$, $\bar{c}$ и $\bar{d}$ только три окажутся независимыми. С другой стороны, поскольку все значения вероятностей (3.1)÷(3.16) можно выразить через эти вновь введенные переменные, то можно утверждать, что лишь три из всех значений вероятностей (3.1)÷(3.16) являются независимыми. Например, при заданных значениях ω, λ_1 и λ_2 можно определить все остальные значения вероятностей. Так,

$$\lambda = 1 - \lambda_2 + \omega(\lambda_1 + \lambda_2 - 1), \tag{3.18}$$

$$\omega_1 = \frac{\omega\lambda_1}{1 - \lambda_2 + \omega(\lambda_1 + \lambda_2 - 1)}, \tag{3.19}$$

$$\omega_2 = \frac{(1-\omega)\lambda_2}{\lambda_2 - \omega(\lambda_1 + \lambda_2 - 1)}. \tag{3.20}$$

Легко заметить, что пара значений λ_1 и λ_2 является вероятностной характеристикой собственно "черного ящика" – канала связи, тогда как значение ω характеризует входную последовательность двоичных символов. Естественно, что задание тройки значений λ_1, λ_2 и ω полностью предопределяет дальнейший ход событий, т.е. значения всех остальных вероятностей.

3.2. ЭНТРОПИЙНАЯ ТЕОРИЯ ПЕРЕДАЧИ ИНФОРМАЦИИ. ПРОПУСКНАЯ СПОСОБНОСТЬ КАНАЛА СВЯЗИ

На основе значений вероятностей (3.1)÷(3.16) можно определить значения следующих энтропий:

энтропия ансамбля (x, y) случайных величин x и y –

$$H[x,y] = -(p(1,1)\log_2 p(1,1) + p(0,1)\log_2 p(0,1) + p(1,0)\log_2 p(1,0) + p(0,0)\log_2 p(0,0)), \tag{3.21}$$

энтропия случайной величины x –

$$H[x] = -(\omega\log_2\omega + (1-\omega)\log_2(1-\omega)), \tag{3.22}$$

энтропия случайной величины y –

$$H[y] = -(\lambda\log_2\lambda + (1-\lambda)\log_2(1-\lambda)), \tag{3.23}$$

условная (остаточная) энтропия случайной величины x при известном значении y –

$$H[x/y] = -\lambda(\omega_1 \log_2 \omega_1 + (1-\omega_1)\log_2(1-\omega_1)) - \\ -(1-\lambda)(\omega_2 \log_2 \omega_2 + (1-\omega_2)\log_2(1-\omega_2)), \quad (3.24)$$

условная (остаточная) энтропия случайной величины y при известном значении x –

$$H[y/x] = -\omega(\lambda_1 \log_2 \lambda_1 + (1-\lambda_1)\log_2(1-\lambda_1)) - \\ -(1-\omega)(\lambda_2 \log_2 \lambda_2 + (1-\lambda_2)\log_2(1-\lambda_2)). \quad (3.25)$$

Из приведенных в (3.21)÷(3.25) определений легко убедиться, что

$$H[x, y] = H[x] + H[y/x] = H[y] + H[x/y]. \quad (3.26)$$

Величина

$$I[x, y] = H[x] - H[x/y] \quad (3.27)$$

равна среднему количеству информации о случайной величине x, содержащемуся в одном сообщении о том, какое именно конкретное значение получила случайная величина y. Аналогично, величина

$$I[y, x] = H[y] - H[y/x] \quad (3.28)$$

равна среднему количеству информации о случайной величине y, содержащемуся в одном сообщении о том, какое именно конкретное значение получила случайная величина x. Из (3.26) и (3.27) с учетом (3.21) легко заметить, что имеет место

$$I[x, y] = I[y, x] = H[x] + H[y] - H[x, y]. \quad (3.29)$$

Информационная сущность канала передачи заключается в том, чтобы, заполучив конкретные значения случайной величины y, можно было бы судить о том, какое конкретное значение при этом имела случайная величина x. Степень "проблематичности" угадывания входного символа, когда известен выходной символ, характеризуется остаточной энтропией $H[x/y]$.

С помощью (3.19)÷(3.20) легко установить, что когда канал связи является идеальным, т.е. при условии

$$\lambda_1 = \lambda_2 = 1, \quad (3.30)$$

имеет место

$$\omega_1 = \omega_2 = 1,$$

т.е. $H[x/y] = 0$.

К такому же результату мы приходим в случае идеального инвертора, когда имеют место

$$\lambda_1 = \lambda_2 = 0, \quad (3.31)$$

С помощью тех же (3.19)÷(3.20) отсюда получим:

$$\omega_1 = \omega_2 = 0,$$

т.е. $H[x/y] = 0$. В обоих этих случаях значение выходного символа оказывается достаточным, чтобы точно знать, каким является входной символ. Это и естественно, так как при этом количество информации о входном символе, содержащееся в каждом сообщении о том, какое именно конкретное значение получил выходной символ, равно

$$I[x, y] = H[x] - H[x/y] = H[x], \tag{3.32}$$

т.е. ровно столько, чтобы полностью ликвидировать равную $H[x]$ исходную неопределенность по угадыванию входного символа.

При работе с реальными каналами связи значение $H[x/y]$ оказывается большим нуля, вследствие чего количество информации

$$I[x, y] = H[x] - H[x/y]$$

оказывается меньше $H[x]$, т.е. значение выходного символа оказывается недостаточным для внесения полной ясности в вопрос о том, каким именно является входной символ. Это означает, что с просмотром каждого очередного выходного символа адресат получает о входном символе информацию, количественно равную не $H[x]$, как это имело место при идеальном канале (или идеальном инверторе), а $I[x, y] < H[x]$.

Значение $H[x]$ зависит только от параметра ω – вероятностной характеристики входной последовательности. При заданном значении ω величина $I[x, y]$ зависит также от пары значений λ_1 и λ_2 – вероятностных характеристик собственно канала.

Так, из (3.29) следует, что

$$I[x, y] = I[y, x],$$

т.е. с учетом (3.28) и (3.23) имеем:

$$I[x, y] = -\lambda \log_2 \lambda - (1-\lambda)\log_2(1-\lambda) - H[y/x]. \tag{3.33}$$

Из (3.25) следует, что $H[y/x]$ зависит только от параметров λ_1, λ_2 и ω, т.е. достаточно подставить в (3.33) значение $\lambda = 1 - \lambda_2 + \omega(\lambda_1 + \lambda_2 - 1)$ (см. (3.18)), чтобы получить зависимость $I[x, y]$ от параметров λ_1, λ_2 и ω:

$$I[x, y] = f_I(\lambda_1, \lambda_2, \omega). \tag{3.34}$$

Кроме двух идеальных случаев, когда имеют место (3.30) или (3.31) и поэтому $I[x, y] = H[x]$, во всех остальных случаях значение $I[x, y]$ оказывается меньше значения $H[x]$.

Образно говоря, каждый двоичный символ, загруженный на входе канала $H[x]$ бит информации, доводит до конца канала лишь часть этой информации, а именно, $I[x, y]$ бит. Остальная часть информации, т.е. $H[x/y]$ бит, теряется в пути. Лишь в случае идеального канала (или

идеального инвертора) в пути ничего не теряется, $H[x/y] = 0$, и каждый символ доводит до конца канала $I[x, y] = H[x]$ бит информации. Пусть теперь вероятностные характеристики λ_1 и λ_2 канала связи таковы, что имеет место

$$\lambda_1 + \lambda_2 = 1. \tag{3.35}$$

Тогда, как это следует из (3.19) и (3.20), имеют место

$$\omega_1 = \omega, \tag{3.36}$$

$$\omega_2 = 1 - \omega. \tag{3.37}$$

Подставляя эти значения в (3.24), с учетом (3.22) получим

$$H[x/y] = H[x], \tag{3.38}$$

или, с учетом (3.27):

$$I[x, y] = 0. \tag{3.39}$$

Таким образом, при соблюдении условия (3.35) каждый двоичный символ, независимо от того, каким количеством информации он был загружен на входе канала, весь свой "информационный груз" теряет в пути и к выходу канала приходит "с пустыми руками".

Пусть известно, что энтропия угадывания одной буквы передаваемого текста равна H бит. Тогда на входе канала, где информационная нагрузка каждого двоичного символа равна $H[x]$ бит, для распознавания каждой буквы в среднем необходимо $H/H[x]$ символов. Столько же символов потребовалось бы на выходе канала, если бы канал был идеальным. Если же канал не идеальный, то среднее число символов на выходе, необходимое для распознавания одной буквы, оказывается равным $H/I[x, y] > H/H[x]$. В частности, когда имеет место условие (3.35), т.е. когда $I[x, y] = 0$, число символов на выходе, необходимых для распознавания одной буквы, стремится к бесконечности. В рассматриваемом случае, независимо от того, сколько двоичных символов можно передавать через канал связи за единицу времени, количество информации, передаваемое через него за любой конечный отрезок времени, равно нулю. Далее этот случай исключим из рассмотрения, т.е. примем, что

$$\lambda_1 + \lambda_2 \neq 1.$$

Рассмотрим, например, случай, когда подлежит передаче оптимально закодированный текст, причем, прежде чем подать этот текст непосредственно на вход канала, его делят на блоки по m двоичных символов в каждом и каждый блок "снабжают" s дополнительно вводимыми (контрольными) символами, нацеленными на обнаружение и исправление возможных (одиночных) ошибок в канале связи. При этом значения m и s выбраны такими, что имеет место $2^s - s - 1 = m$. Иными словами, речь идет о передаче оптимально закодированного текста,

снабженного кодом Р. Хэмминга, подробно рассмотренным во второй главе. Из схемы построения кода Р. Хэмминга и характера предполагаемых ошибок легко сделать вывод, что, как и для каждого информационного символа, для каждого контрольного символа при этом справедливо следующее: значения вероятностей того, что данный контрольный символ окажется нулем или единицей, равны между собой и не зависят от того, какие символы предшествовали данному символу.

Из вышеизложенного следует, что каждый двоичный символ входной последовательности несет информационную нагрузку, равную одному биту.

Оценим количество информации, которое доводится до конца канала связи очередным блоком из $m + s$ двоичных символов (m информационных и s контрольных).

Из (3.27) имеем

$$I[x, y] = H[x] - H[x/y], \tag{3.27а}$$

где $H[x]$ – количество неопределенности, соответствующее входному блоку из $m + s$ двоичных символов; $H[x/y]$ – количество остаточной неопределенности, соответствующее входному блоку из $m + s$ двоичных символов после того, как известен результат его приема на выходе канала.

Из оптимальности входной последовательности непосредственно следует, что $H[x] = (m + s)\cdot 1$ бит. Что же касается значения остаточной неопределенности $H[x/y]$, то для его вычисления мы руководствуемся следующими соображениями.

Поскольку речь идет лишь об одиночной ошибке, то после приема очередного блока из $m + s$ двоичных символов, для внесения полной ясности в то, каким был этот блок на входе канала, необходимо ответить на два вопроса:

имела ли место ошибка вообще;

если ошибка имела место, то на уровне какого именно из $m + s$ символов она произошла.

Пользуясь формулой (2.5) второй главы, общее количество неопределенности $H[x/y]$ определяем как сумму двух неопределенностей, соответствующих этим двум вопросам:

$$H[x/y] = \left(-\frac{1}{1+m+s}\log_2\frac{1}{1+m+s} - \frac{m+s}{1+m+s}\log_2\frac{m+s}{1+m+s}\right) +$$

$$+\frac{m+s}{1+m+s}\log_2(m+s) = \log_2(1+m+s) = s. \tag{3.27б}$$

Здесь мы учли, что $m + s + 1 = 2^s$, а множитель $(m + s)/(1 + m + s)$ при втором слагаемом выполняет ту же роль, что множитель λ_k в формуле (2.5).

Подставляя значения $H[x] = m + s$ и $H[x/y] = s$ в формулу (3.27а), получим значение $I[x, y] = m$ количества информации, которое доводится до конца канала очередным блоком из $m + s$ символов. Таким образом, каждый блок из $m + s$ символов, будучи загруженным на входе канала $m + s$ бит информации, из-за наличия помех в канале связи часть этой информации, а именно, s бит теряет в пути и до конца канала доводит лишь m бит информации. Каждый двоичный символ доводит до конца канала $m/(m + s)$ бит информации, и поэтому, если передаваемый текст характеризуется энтропией, равной H бит, то для передачи через канал связи одной буквы исходного текста потребуются $(m + s)\,H/m$ двоичных символов. Здесь мы еще раз убеждаемся в справедливости тех результатов, которые были получены во второй главе при рассмотрении кода Р. Хэмминга.

Введем в рассмотрение коэффициент проводимости канала связи, определив его как

$$\varkappa[x, y] = I[x, y] / H[x] = f_{\varkappa}(\lambda_1, \lambda_2, \omega). \tag{3.40}$$

Значение этого коэффициента численно равно той доле входной информационной нагрузки, которое двоичному символу удается провести через канал связи до его выхода. Значения коэффициента проводимости $\varkappa[x, y]$ нормированы в интервале

$$0 \leqslant \varkappa[x, y] \leqslant 1, \tag{3.41}$$

причем нулевое значение коэффициента проводимости соответствует каналу с $\lambda_1 + \lambda_2 = 1$, а единичное значение – идеальному каналу (или идеальному инвертору). В [1] нами было доказано существование и единственность при заданных значениях λ_1 и λ_2 ($\lambda_1 + \lambda_2 \neq 1$) значения $\omega = \omega_{\varkappa}$, обеспечивающего наибольшее возможное значение коэффициента проводимости $\varkappa[x, y] = \varkappa_{max}$. Там же приведено доказательство неравенства

$$\varkappa_{max} \leqslant (\sqrt{\lambda_1\lambda_2} - \sqrt{(1-\lambda_1)(1-\lambda_2)})^2 \tag{3.42}$$

В общем случае, при произвольных λ_1 и λ_2 ($\lambda_1 + \lambda_2 \neq 1$), значение $\omega_{\varkappa}$ может отличаться от $\omega = 0{,}5$. Тогда при работе в режиме $\omega = \omega_{\varkappa}$, т.е. при желании свести к минимуму долю потерь информации в канале, приходится "отойти" от значения $\omega = 0{,}5$, т.е. от режима оптимального кодирования, обеспечивающего возможно наибольшую (1 бит) информационную нагрузку каждого символа на входе канала. В результате информационная нагрузка каждого двоичного символа на входе канала может оказаться настолько малой, что даже при минимальных потерях в пути все равно до выхода канала дойдет крайне малое количество информации. Придерживаясь же значения $\omega = 0{,}5$, мы, хотя и обеспечим максимально возможную информационную нагрузку на входе канала, большая доля этой информации будет потеряна в пути и в итоге

до выхода канала дойдет опять-таки небольшое количество информации. Интуитивно ясно, что значение $\omega = \omega_I$, при котором достигается возможно наибольшее значение $I[x, y] = I_{max}$, должно находиться между двумя значениями: $\omega = 0{,}5$, обеспечивающее максимально возможную информационную нагрузку (1 бит) на входе канала, и $\omega = \omega_x$, обеспечивающее наилучшие условия передачи, т.е. условия, когда доля потерь информации в пути достигает минимума.

В [1] приведено доказательство неравенства

$$|\omega_I - 0{,}5| \leq |\omega_x - 0{,}5|. \tag{3.43}$$

При заданных значениях λ_1 и λ_2 значение ω_I легко определить из условия (см. (3.34))

$$\partial f_I(\lambda_1, \lambda_2, \omega) / \partial\omega = 0, \tag{3.44}$$

откуда следует, что

$$\omega_I = \frac{1}{(\lambda_1 + \lambda_2 - 1)}\left(\lambda_2 - \frac{\alpha}{\alpha + \beta}\right), \tag{3.45}$$

где

$$\alpha = \exp\left(\frac{H(\lambda_1)\ln 2}{\lambda_1 + \lambda_2 - 1}\right) \tag{3.45а}$$

$$\beta = \exp\left(\frac{H(\lambda_2)\ln 2}{\lambda_1 + \lambda_2 - 1}\right) \tag{3.45б}$$

Подставляя значение ω_I в (3.34), определим максимально возможное значение $I[x, y] = I_{max}$ при заданных значениях λ_1 и λ_2 ($\lambda_1 + \lambda_2 \neq 1$):

$$I_{max} = \log_2\left(\frac{\alpha + \beta}{\alpha^{\lambda_1} \cdot \beta^{\lambda_1}}\right). \tag{3.46}$$

Из (3.45) легко установить, что

$$\omega_I(\lambda_1, \lambda_2) = 1 - \omega_I(\lambda_2, \lambda_1) = \omega_I[(1 - \lambda_1), (1 - \lambda_2)]. \tag{3.47}$$

Зависимость (3.45) графически представлена на рис. 3.2. Здесь хорошо прослеживаются свойства (3.47) симметричности $\omega_I(\lambda_1, \lambda_2)$ относительно прямой $\lambda_1 = \lambda_2$ и точки $\lambda_1 = 0{,}5$, $\lambda_2 = 0{,}5$. В [1] нами было доказано, что область существования величины ω_I ограничена интервалом значений

$$\frac{1}{e} \leq \omega_I \leq 1 - \frac{1}{e}. \tag{3.48}$$

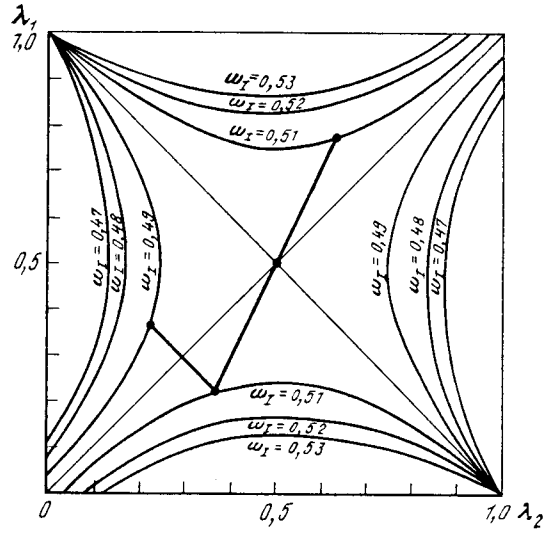

Рис. 3.2. Кривые, иллюстрирующие характер зависимости ω_I от аргументов λ_1 и λ_2.

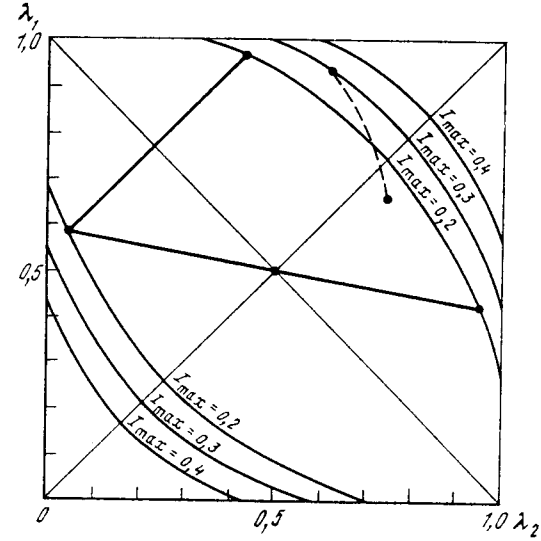

Рис. 3.3. Кривые, иллюстрирующие характер зависимости I_{max} от аргументов λ_1 и λ_2.

Из (3.46) легко установить, что

$$I_{\max}(\lambda_1, \lambda_2) = I_{\max}(\lambda_2, \lambda_1) = I_{\max}[(1-\lambda_1),(1-\lambda_2)]. \qquad (3.49)$$

Эти свойства симметричности хорошо прослеживаются на рис. 3.3, где приведено графическое представление зависимости $I_{\max}$ от λ_1 и λ_2.

Рассмотрим случай, когда вероятностная характеристика канала связи задана не в виде фиксированной пары значений λ_1 и λ_2, а в виде параметрической зависимости

$$\lambda_1 = \lambda_1(t), \qquad (3.50)$$

$$\lambda_2 = \lambda_2(t), \qquad (3.51)$$

где t – некий параметр, значения которого можно изменять в определенном диапазоне. Каждое фиксированное значение t из этого диапазона задает соответствующую точку на кривой зависимости $\lambda_1 = \lambda_1(\lambda_2)$. Представляет интерес задача подбора из всего диапазона возможных значений t такого значения (такой точки на кривой зависимости $\lambda_1 = \lambda_1(\lambda_2)$), при котором можно достичь максимального возможного значения $I_{\max 0}$. Пример графического решения этой задачи приведен на рис 3.3, где пунктиром представлена зависимость $\lambda_1 = \lambda_1(\lambda_2)$. Решение задачи осуществляется в два этапа:

1) из всех кривых постоянного уровня $I_{\max} = \text{const}$, имеющих общую точку с кривой $\lambda_1 = \lambda_1(\lambda_2)$, выбирается та, которой соответствует наибольшее значение $I_{\max} = I_{\max 0}$. Общая точка этой кривой и кривой $\lambda_1 = \lambda_1(\lambda_2)$ принимается за рабочую точку $t = t0$ ($\lambda_1 = \lambda_{1\,0}$, $\lambda_2 = \lambda_{2\,0}$).

2) пара значений $\lambda_1 = \lambda_{1\,0}$ и $\lambda_2 = \lambda_{2\,0}$, соответствующих рабочей точке $t = t0$, подставляется в формулу (3.45), и вычисляется значение ω_I, при котором достигается значение $I_{\max} = I_{\max 0}$. Заметим, что при известных $\lambda_1 = \lambda_{1\,0}$ и $\lambda_2 = \lambda_{2\,0}$ значение $\omega = \omega_I$ можно определить также графически, с помощью кривых, приведенных на рис. 3.2.

Таким образом, после выбора рабочей точки $t = t0$ решение рассматриваемой задачи сводится к решению ранее рассмотренной задачи, а именно, при заданной паре значений λ_1 и λ_2 найти соответствующее значение $\omega = \omega_I$, обеспечивающее максимальное возможное значение $I_{\max}$. В [1] задачи с аналогичной постановкой называются задачами согласования входа.

Определенный научно-практический интерес представляют случаи, когда при заданном фиксированном значении $\omega = \omega_0$ требуется достичь наибольшего возможного значения $I[x, y]$ путем подбора соответствующей рабочей точки на заданной кривой зависимости $\lambda_1 = \lambda_1(\lambda_2)$. Группу задач, связанных с подбором на кривой $\lambda_1 = \lambda_1(\lambda_2)$ рабочих точек, которым при заданном значении $\omega = \omega_0$ соответствуют максимальные возможные значения тех или иных параметров (в нашем случае таким параметром является количество информации $I[x, y]$), будем называть

задачами настройки системы. Более подробное рассмотрение задач настройки можно найти в [1]. Их решение сводится к следующему:

путем подстановки в выражение $I = f_I(\lambda_1, \lambda_2, \omega)$ значения $\omega = \omega_0$ получаем выражение $I = \Phi(\lambda_1, \lambda_2)$. Далее каким-либо из стандартных методов максимизации функций от двух аргументов находим значения $\lambda_1 = \lambda_0$ и $\lambda_2 = \lambda_{2\,0}$, при которых достигается наибольшее возможное значение $I[x, y]$. В частности, в [1] рекомендуется графический метод нахождения значений $\lambda_1 = \lambda_{1\,0}$ и $\lambda_2 = \lambda_{2\,0}$, обеспечивающих наибольшее возможное значение $I[x, y] = I_{\max}$. Для этого в выражении $I = \Phi(\lambda_1, \lambda_2)$ задаемся различными значениями $I = \text{const}$ из интервала [0, 1] и для каждого из этих значений в координатах (λ_1, λ_2) строим кривые постоянного уровня $I = \text{const}$. Из всех кривых постоянного уровня, имеющих общую точку с кривой $\lambda_1 = \lambda_1(\lambda_2)$, выбирается та, которой соответствует наибольшее значение $I_{\max}$. Общая точка этой кривой и кривой зависимости $\lambda_1 = \lambda_1(\lambda_2)$ и принимается за искомую рабочую точку с координатами $\lambda_1 = \lambda_{1\,0}$ и $\lambda_2 = \lambda_{2\,0}$.

Если максимальное возможное число двоичных символов, которое можно передавать через данный канал связи за единицу времени, обозначить через F, то произведение $V = F \cdot I_{\max}$ окажется численно равным количеству информации, больше которого через данный канал связи нельзя передать за единицу времени. Эту величину К. Шеннон называет пропускной способностью канала связи. Если, например, энтропия угадывания одной буквы в передаваемом тексте равна H, а пропускная способность канала связи – V, то потребуется не менее H/V единиц времени на каждую передаваемую букву.

Таким образом, кроме скорости передачи самих двоичных символов, скорость передачи информации по каналу связи зависит еще и от того, какое количество информации может довести до выхода канала каждый двоичный символ.

ЛИТЕРАТУРА К ГЛАВЕ 3

1. *Аветисян Д.О.* Проблемы информационного поиска. – М.: Финансы и статистика, 1981.
2. *Шеннон К.* Работы по теории информации и кибернетике. – М.: Изд-во ин. лит., 1963.

ГЛАВА 4

ПЕРЕДАЧА КОНФИДЕНЦИАЛЬНЫХ СООБЩЕНИЙ ПО ОТКРЫТЫМ КАНАЛАМ СВЯЗИ. ОТКРЫТОЕ ШИФРОВАНИЕ И ОРГАНИЗАЦИЯ ЭЛЕКТРОННОЙ ПОДПИСИ

СОВРЕМЕННЫЕ средства связи и вычислительной техники создали благоприятные условия для дальнейшего совершенствования и широкого распространения сетевых информационных технологий. Интенсивно растет число абонентов, вовлеченных в глобальные информационные сети. Примером может служить сеть INTERNET, охватывающая сотни тысяч абонентов из различных стран, независимо от их местонахождения, идеологической или религиозной приверженности. Глобальные сети ЭВМ создали уникальную возможность общения между пользователями сети, преследующими самые различные цели (культурные, коммерческие, узкопрофессиональные и др.). В соответствии с этим, по одним и тем же информационным сетям циркулирует информация самого различного характера и назначения, начиная от любительских фильмов и детских рисунков и кончая информацией о результатах контроля за вооружением. Все чаще встречаются ситуации, когда абоненты, единственными средствами общения между которыми являются открытые, доступные всем глобальные информационные сети, нуждаются в регулярном обмене конфиденциальной информацией.

Классические схемы организации обмена конфиденциальной информацией по открытым каналам связи предполагают непременное наличие у обменивающихся сторон некоторого секретного ключа шифрования – дешифрования, на основе которого отправитель информации осуществляет шифрование конфиденциальных сообщений и уже полученную шифрограмму по открытым каналам связи посылает получателю информации. Последний с помощью секретного ключа расшифровывает полученную шифрограмму, востановив тем самым исходный текст конфиденциального сообщения. При этом, естественно, секретный ключ должен быть таким, чтобы от третьих сторон (их называют также злоумышленниками), пытающихся осуществить несанкционированный доступ к конфиденциальным сообщениям, для этого потребовалось бы достаточно много усилий. Сейчас трудно устновить точную

дату появления первых шифрограмм. Можно лишь с уверенностью утверждать, что они появились значительно раньше пятого века до нашей эры, когда патриарх истории древнего мира Геродот упоминал о текстах, понятных лишь для одного адресата [3]. Попытки разгадывания секретных ключей начались практически одновременно с появлением первых шифрограмм. За все время своего существования различные методы шифрования систематически совершенствовались. Но столь же последовательно совершенствовались методы разгадывания секретных ключей. Одно лишь осталось неизменным – необходимость наличия секретного ключа, известного лишь двум сторонам – отправителю конфиденциальных сообщений и их получателю. А для обмена секретными ключами всегда требовалось наличие некоторого закрытого, недоступного для третьих сторон канала связи.

Так было до 1976 года, когда американские специалисты У. Диффи и М. Хеллман предложили новый принцип шифрования конфиденциальных сообщений – принцип открытого шифрования [12].

Чтобы оценить значимость, место и роль принципа открытого шифрования, вкратце рассмотрим некоторые простейшие алгоритмы шифрования, ориентированные на использование секретных ключей шифрования.

4.1. О КРИПТОСИСТЕМАХ, ИСПОЛЬЗУЮЩИХ СЕКРЕТНЫЕ КЛЮЧИ ШИФРОВАНИЯ

Одним из наиболее ранних методов шифрования является метод простой замены символов исходного текста конфиденциальных сообщений другими символами, согласно некоторой подстановке, выполняющей функции секретного ключа. В простейшем случае такая подстановка сводится к сдвигу всех букв алфавита влево или вправо на некоторую постоянную величину. Следующий пример иллюстрирует работу алгоритма постоянного сдвига применительно к русским текстам при допущении, что в передаваемых сообщениях отсутствуют заглавные буквы и знаки препинания, а буква ё отождествлена с буквой е. Иными словами, предполагается, что в сообщениях, подлежащих шифрованию, могут встречаться 33 различных символа, а именно: 32 буквы русского языка и символ пробела. Приняв, что сдвиг букв осуществляется вправо, а постоянная сдвига равна восьми буквам, получим следующую подстановку:

а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п	р	с
щ	ъ	ы	ь	э	ю	я	–	а	б	в	г	д	е	ж	з	и	й
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я	–
к	л	м	н	о	п	р	с	т	у	ф	х	ц	ч	ш
18	19	20	21	22	23	24	25	26	27	28	29	30	31	32

Обратим внимание, что символы, вышедшие в результате сдвига за

рамки алфавита (в приведенной выше подстановке эти символы подчеркнуты), занимают освободившиеся 8 позиций в начальной части алфавита. Естественно, что число таких символов равно постоянной сдвига (в нашем случае оно равно восьми). Если буквы алфавита пронумеровать в порядке возрастания занимаемых ими позиций в алфавите ('а' = 0, 'б' = 1, 'в' = 2, ..., 'я' = 31, '–' = 32), то очередная буква исходного текста, которая в алфавите занимает i-ю позицию, будет заменена буквой, занимающей j-ю позицию алфавита, где

$$j = (i - k) \bmod(N), \tag{4.1}$$

k – постоянная сдвига, равная в нашем примере восьми, а N – число различных символов, равное в нашем примере тридцати трем. Например, в результате такого шифрования исходный текст "встреча у букиниста" примет форму шифрограммы "ы й к и ю п щ ш л ш ъ л в а е а й к щ". Чтобы расшифровать такую шифрограмму, необходимо очередную букву шифрограммы, занимающую i-ю позицию алфавита, заменить буквой, занимающей в алфавите j-ю позицию, где

$$j = (i + k) \bmod(N). \tag{4.2}$$

Поступая таким образом, из полученной только что шифрограммы получим исходный текст "встреча у букиниста".

В общем случае в качестве секретного ключа шифрования могут быть использованы произвольные подстановки, определенные на множестве из N символов. Примером может служить, например, подстановка

$$X = \begin{array}{cccccccccccccccc} а & б & в & г & д & е & ж & з & и & й & к & л & м & н & о & п & р \\ й & – & а & б & э & ю & я & ъ & ы & ь & т & у & ф & х & ц & ч & ш \end{array}$$

$$\begin{array}{cccccccccccccccc} с & т & у & ф & х & ц & ч & ш & щ & ъ & ы & ь & & э & ю & я & – \\ с & в & г & д & е & ж & з & и & щ & к & л & м & & н & о & п & р \end{array}$$

Для расшифрования текстов, зашифрованных согласно некоторой подстановке X, используется обратная ей подстановка $Y = X^{-1}$, т.е. такая, чтобы имело место $X \cdot Y = \overline{E}$, где $\overline{E}$ – единичный элемент, представляющий тождественную подстановку:

$$\overline{E} = \begin{array}{cccccccccccccccc} а & б & в & г & д & е & ж & з & и & й & к & л & м & н & о & п & р \\ а & б & в & г & д & е & ж & з & и & й & к & л & м & н & о & п & р \end{array}$$

$$\begin{array}{cccccccccccccccc} с & т & у & ф & х & ц & ч & ш & щ & ъ & ы & ь & э & ю & я & – \\ с & т & у & ф & х & ц & ч & ш & щ & ъ & ы & ь & э & ю & я & – \end{array}$$

Подстановочные алгоритмы шифрования в классическом варианте их использования оказались легко взламываемыми. Основным "виновником" этого является присущая естественным языкам избыточность, в

данном случае выражающаяся в том, что различные буквы алфавита имеют различную вероятность встречаемости в текстах естественных языков. Путем компьютерной обработки достаточно представительных объемов текстов на английском, французском, венгерском, армянском, арабском, русском, украинском и др. языках нами получены таблицы вероятностей встречаемости различных букв алфавитов этих языков в соответствующих текстах. Некоторые из этих таблиц приведены, например, в [1, 2]. Применительно к русским текстам, например, на основе результатов статистической обработки свыше 0,5 млн. символов можно утверждать, что если какая-то буква имеет наименьшую встречаемость в шифрограмме, то скорее всего эта буква – результат преобразования какой-либо одной из букв 'ъ', 'ф', 'э' или 'щ'. И, наоборот, если какая-то буква имеет наивысшую встречаемость в шифрограммах, то скорее всего эта буква – результат преобразования какой-либо одной из букв 'о', 'е', 'и' или 'п'. Руководствуясь аналогичными соображениями, методом проб и ошибок удается сравнительно легко взламывать подстановочные алгоритмы. Если же учесть, что для анализа могут быть использованы огромные возможности современных компьютеров, то, по крайней мере, в наши дни криптостойкость этих алгоритмов следует признать чрезвычайно низкой.

Простейшим представителем другого направления шифрования – перестановочных алгоритмов может служить перестановочный алгоритм шифрования с шагом в k букв. В рамках этого алгоритма буквами исходного текста поочередно заполняются клетки прямоугольника из k строк в порядке, указанном ниже на примере шифрования исходного текста "встреча у букиниста" (значение k принято здесь равным трем).

в р а – к и а
с е – б и с –
т ч у у н т –

и в качестве шифрограммы принимается последовательность букв "в р а – к и а с е – б и с – т ч у у н т –". Иными словами, используется правило "запись по столбцам – шифрограмма по строкам". Расшифровка шифрограммы осуществляется в обратном порядке "запись по строкам – исходный текст по столбцам", а именно, буквами шифрограммы поочередно заполняются клетки прямоугольника из $m = M/k$ столбцов (M – число букв в шифрограмме), а исходный текст формируется поочередным чтением букв по столбцам.

Описанный только что алгоритм шифрования не обладает сколько-нибудь серьезным уровнем криптостойкости. Но здесь налицо основной принцип работы перестановочных алгоритмов – перестановка позиций, которые занимают буквы в исходных текстах. Путем усложнения правил перестановки в ряде случаев удается добиться достаточно высокого уровня крипстойкости. И тогда для взламывания криптосистемы приходится прибегать к более мощным средствам, например, с исполь-

зованием значений вероятностей встречаемости различных пар букв, триад букв и т.д., с учетом даже позиций этих пар, триад в словах естественных языков.

Не вдаваясь в подробности анализа различных хитроумных решений в каждой конкретной криптосистеме, где могут сочетаться подстановочный и перестановочный принципы шифрования, отметим лишь, что основным инструментом их взлома является использование тех или иных проявлений избыточности текстов естественных языков.

В рассматриваемом смысле наиболее продуктивными оказались алгоритмы шифрования, которые в наибольшей степени защищены от средств взлома, базирующихся на использовании избыточности естественных языков. В этом плане большой научно-практический интерес представляют алгоритмы, являющиеся различными вариантами реализации криптосистемы, предложенной аббатом из Вюрцбурга Иоганном Тритемиусом [3]. Сущность системы заключается в использовании в качестве секретного ключа некоторой конечной последовательности букв. При работе с этой криптосистемой все символы, которые могут встречаться в передаваемых сообщениях, предварительно произвольным образом нумеруются. В частности, порядок нумерации может совпадать с алфавитным порядком. И тогда, если условиться, что, как и выше, в передаваемых сообщениях могут встречаться лишь 33 различных символа, то нумерация букв будет такой же, как и в рассмотренных выше примерах ('а' = 0, 'б' = 1, ...). Далее выбирается произвольная конечная последовательность букв в качестве секретного ключа шифрования. Обычно, чтобы легко было запоминать, в качестве таковой берется некоторый осмысленный текст или слово. Примем, например, что в качестве секретного ключа выбрано слово "аура", а передаче подлежит тот же текст "встреча у букиниста". Для шифрования этого текста под ним записывается секретный ключ, как это показано ниже:

в с т р е ч а – у – б у к и н и с т а
а у р а а у р а а у р а а у р а а у р

и далее осуществляется побуквенное сложение этих строк по модулю 33. Например, результат сложения буквы 'в' с буквой 'а' определится как

('в' = 2) + ('а' = 0) = 2mod(33) = 2 = 'в'.

Аналогично определяется результат сложения

('ч' = 23) + ('у' = 19) = 42mod(33) = 9 = 'й'.

Поступая таким образом, в качестве шифрограммы получим следующую последовательность букв:

в г б р е й р – у т с у к ы э и с д р

Чтобы отсюда восстановить исходный текст, получатель информации

записывает под шифрограммой секретный ключ, как это показано ниже:

в г б р е й р – у т с у к ы э и с д р
а у р а а у р а а у р а а у р а а у р

и затем осуществляет побуквенное вычитание этих строк по модулю 33. Например, результат вычитания буквы 'р' от буквы 'б' определится как

('б' = 1) – ('р' = 16) = –15mod(33) = 18 = 'т'.

Поступая таким образом, в качестве разницы двух строк получим исходный текст "встреча у букиниста".

Легко заметить, что взламывание алгоритма Тритемиуса путем использования избыточности языка значительно сложнее по сравнению с подстановочными и перестановочными алгоритмами. Вместе с тем, сам факт многократного повторения секретного ключа вкупе с тем, что секретный ключ имеет собственную избыточность, в ряде случаев приводит к тому, что алгоритм Тритемиуса в рассмотренном только что варианте его реализации все же удается взламывать. Чтобы полностью (или почти полностью) избавить криптограммы от избыточности, присущей естественным языкам, К. Шеннон рекомендует конфиденциальные тексты предварительно архивировать с помощью какого-либо из эффективных алгоритмов сжатия текстов (Фано, Хаффмэн, Шеннон) и уже архивированный текст подвергать шифрованию по схеме Тритемиуса, используя в качестве секретного ключа случайную последовательность символов алфавита данного естественного языка. При этом практически исключаются случаи взламывания, но такой алгоритм вряд ли можно признать приемлемым с практической точки зрения, поскольку при его использовании требуется посылать адресату не только шифрограмму, но и секретный ключ – случайную последовательность букв, длина которой в данном случае оказывается равной длине архивированного текста [3, 11].

В заключение настоящего раздела отметим наиболее характерные черты криптосистем, ориентированных на использование секретных ключей шифрования. Все они, независимо от конкретной их реализации, непременно требуют наличия закрытого канала связи для обмена секретными ключами. На основе этих ключей осуществляется шифрование конфиденциальных сообщений и полученная в результате этого шифрограмма, уже непонятная для третьих сторон, передается адресату по открытым каналам связи. Адресат же восстанавливает исходный текст путем расшифрования криптограммы с помощью того же (или почти того же) ключа шифрования.

Принято считать, что во всех рассмотренных выше криптосистемах при шифровании и расшифровании конфиденциальных текстов используется один и тот же секретный ключ. В принципиальном же плане такое утверждение не совсем верно. Вернемся, например, к рассмот-

рению простейшего подстановочного алгоритма, реализованного путем постоянного сдвига всех букв алфавита на 8 позиций вправо (см. выше). В рамках этой системы при шифровании исходных текстов буква'а' заменяется буквой 'щ', тогда как при расшифровании криптограмм та же буква 'а' заменяется буквой 'и'. Иными словами, ключи шифрования и расшифрования в общем-то различны, но переход от ключа шифрования к ключу расшифрования настолько прост, что эти ключи практически отождествляют. Для человека, владеющего ключом шифрования, никакого труда не представляет расшифрование шифрограмм. Иными словами, если $y = f(x)$ – функция шифрования, то нахождение функции $x = f^{-1}(y)$ настолько просто, что функции $f(x)$ и $f^{-1}(y)$ практически отождествляют и говорят, что имеют дело с одним и тем же секретным ключом.

Существенно по-иному обстоит дело в криптосистемах открытого шифрования, где определение значения функции $f^{-1}(y)$ на основе значения функции $f(x)$ чрезвычайно затруднено. В следующем разделе мы проанализируем принцип построения алгоритмов открытого шифрования, где ключи шифрования и расшифрования различны настолько, что знание ключа шифрования вовсе не является достаточным для расшифрования криптограмм.

4.2. ОБ ОДНОСТОРОННИХ ФУНКЦИЯХ И О КРИПТОСИСТЕМАХ ОТКРЫТОГО ШИФРОВАНИЯ

Еще в начале шестидесятых годов для предотвращения несанкционированного доступа к различным объектам (ЭВМ, базы данных, файлы и т.д.),конкретнее, для организации парольного доступа к объектам, применялся метод так называемых односторонних функций. Под ним подразумевают функции $y = f(x)$, такие, что вычисление значения y при заданном x не представляет особого труда, тогда как нахождение x, соответствующего заданному значению y, чрезвычайно трудно, точнее, связано с чрезмерно большим объемом вычислений, реализация которых за обозримый промежуток времени не удается. Пусть, к примеру, рассматривается функция

$$y = f(x) = A^x \bmod(N), \tag{4.3}$$

где x и N – чрезмерно большие числа, а A – произвольное число из интервала $[2, N - 2]$. Здесь при заданном x значение y вычисляется относительно просто, тогда как для вычисления значения $x = f^{-1}(y)$ связано с реализацией чрезмерно большого объема вычислений.

Одним из возможных приложений этой или любой другой односторонней функции может служить упомянутый выше пример организации парольного доступа к ЭВМ или иным объектам ограниченного доступа. В традиционных схемах его организации таблица паролей хра-

нится в памяти ЭВМ и для доступа к ней от каждого i-го пользователя требуется назвать свой пароль $x(i)$. Наличие названного $x(i)$ в таблице паролей является достаточным для того, чтобы допустить данного пользователя к ЭВМ. Если, к примеру, противнику удалось завладеть таблицей паролей, то, называя те или иные пароли, он может беспрепятственно получить доступ к ЭВМ, имитируя любого пользователя. Если же в таблице доступа хранить не сами значения паролей $x(i)$, а только значения соответствующих им $y(x(i))$, где $y = f(x)$ – некоторая односторонняя функция, то доступ к ЭВМ можно разрешить лишь после того, как в таблице окажется вычисленное на основе предъявленного данным пользователем пароля $x(i)$ значение $y(x(i))$. При такой постановке интерес противника к этой таблице сразу же отпадет, поскольку на основе приведенных там значений y значения самих паролей, т.е. значения $x(i)$ из-за односторонности функции $y = f(x)$, он не может вычислить.

Из приведенного примера легко заметить, насколько важными для практического применения являются односторонние (мы бы их назвали храповыми) функции. Но то, что ввели в рассмотрение сначала в теоретическом плане, а потом и в плане практического применения У. Диффи и М. Хеллман, повлекло за собой настоящую революцию в современной криптографии. В 1976 г. они опубликовали статью "Новые направления в криптографии", где впервые ввели в рассмотрение понятие односторонних функций с ловушкой (лазейкой). Как и все остальные односторонние функции $y = f(x)$, это функции, где вычисление $y = f(x)$ легко осуществимо, тогда как вычисление $x = f^{-1}(y)$ связано с практически непреодолимыми трудностями. Но в отличие от других односторонних функций, односторонние функции с ловушкой обладают тем специфическим свойством, что при знании определенной информации (и только при этом!) вычисление $x = f^{-1}(y)$ становится легко реализуемым. Иными словами, для лиц, владеющих этой информацией, функция $y = f(x)$ становится легко обратимой, тогда как для всех остальных лиц, не владеющих этой информацией, она остается практически необратимой. Именно эта информация и выполняет роль той ловушки (лазейки), с помощью которой удается обращать функции такого типа.

В своей статье "Новые направления в криптографии" У. Диффи и М. Хеллман заявили: "Сегодня мы находимся накануне революции в криптографии". И они были правы. Своей публикацией они положили начало новому научному направлению. Начался интенсивный поиск односторонних функций с ловушкой с одновременным "прощупыванием" почвы для возможных их приложений. Как и следовало ожидать, на этом пути были взлеты и падения, но за прошедшие со дня публикации упомянутой статьи У. Диффи и М. Хеллмана двадцать лет односторонние функции с ловушкой заняли свое прочное и, пожалуй, наиболее перспективное место в современной криптографии. За это время были

найдены различные процедуры-функции, обладающие свойствами односторонних функций с ловушкой. Особенно продуктивными оказались попытки создания таких процедур в русле возведения чисел в большие степени по большому модулю, а также в русле создания кодов, обнаруживающих и исправляющих ошибки. Примерами алгоритмов, относящихся к первому направлению исследований, могут служить алгоритмы, предложенные Р. Ривестом, А. Шамиром, Л. Адлеманом, Т. Эль-Гамалем и другими [8, 13, 14]. Примерами же алгоритмов, относящихся ко второму направлению, могут служить алгоритмы Мак-Элиса, Нидеррайтера, Габидулина, Крука и других [4].

4.3. КРИПТОСИСТЕМА ОТКРЫТОГО ШИФРОВАНИЯ RSA

Из известных нам криптосистем, базирующихся на односторонних функциях с ловушкой, наибольшую популярность получила криптосистема RSA, относящаяся к первому направлению исследований – направлению возведения чисел в большие степени по модулю, также являющемуся большим числом. Свое название этот алгоритм получил по первым буквам фамилий его создателей (Rivest, Shamir, Adleman). Популярность алгоритма RSA, по-видимому, можно объяснить возможностью довольно элегантной реализации в рамках этого алгоритма как передачи конфиденциальных сообщений, так и организации электронной подписи. Механизм функционирования криптосистемы RSA заключается в следующем [4, 5, 6, 8, 9, 10, 14].

Каждый i-й абонент сети независимо от других абонентов генерирует два больших простых числа $q(i)$ и $p(i)$ и вычисляет число $N(i) = q(i)p(i)$. Порядок величин $q(i)$ и $p(i)$ определяется двумя соображениями:

– с увеличением этих чисел скорость шифрования, передачи по каналам связи и расшифрования конфиденциальных сообщений уменьшается;

– при прочих равных условиях с увеличением простых чисел $q(i)$ и $p(i)$ криптостойкость системы RSA растет.

Обычно рекомендуется в качестве $q(i)$ и $p(i)$ выбрать простые числа, состоящие из 150÷200 десятичных знаков каждое. Естественно, что эти рекомендации не следует принимать за догму, и в зависимости от эксплуатационной необходимости эти числа могут быть выбраны значительно меньшими, или, наоборот, большими. Более того, как мы убедимся ниже, криптостойкость системы RSA зависит не только от самих величин этих чисел, но и от их характера, и в этом смысле представляется актуальной задача оптимального выбора этих чисел, с тем, чтобы при приемлемых скоростных показателях обеспечить максимально возможный уровень криптостойкости. При выборе и проверке на простоту больших чисел обычно пользуются малой теоремой Ферма,

а именно, число S считают простым, если для произвольно выбранного числа $M < S$ имеет место

$$M^{S-1} = 1 \bmod(S). \tag{4.4}$$

Хотя условие (4.4) является лишь необходимым, но не достаточным условием, чтобы число S признать простым, тем не менее, после соответствующих допроверок теорема Ферма способствует выбору простых чисел $q(i)$ и $p(i)$. После определения числа $N(i)$, i-й абонент сети вычисляет число Эйлера от аргумента $N(i)$, которое при простых $q(i)$ и $p(i)$ определяется по формуле

$$F(N(i)) = (q(i) - 1)\,(p(i) - 1). \tag{4.5}$$

Далее i-м абонентом выбирается произвольное достаточно большое и взаимно простое с $F(N(i))$ число $e(i)$, после чего выбирается произвольное число $d(i)$ такое, чтобы имело место

$$e(i) \cdot d(i) = 1 \bmod(F(N(i))) \tag{4.6}$$

(к вопросу о том, насколько правомочна произвольность выбора чисел $q(i)$ и $p(i)$, мы вернемся ниже). После того, как i-м абонентом определены числа $q(i)$, $p(i)$, $N(i)$, $F(N(i))$, $e(i)$ и $d(i)$, он уже готов к приему конфиденциальных сообщений. Для этого он помещает в общедоступный справочник числа $N(i)$ и $e(i)$ в качестве открытого ключа шифрования, а число $d(i)$ хранит у себя в качестве секретного ключа расшифрования. Поскольку при известных числах $e(i)$ и $N(i)$ знания любого из чисел $q(i)$, $p(i)$ или $F(N(i))$ достаточно для того, чтобы вычислить число $d(i)$ – секретный ключ расшифрования, то числа $q(i)$, $p(i)$ и $F(N(i))$ следует хранить в тайне, либо же вообще "уничтожить", поскольку далее они этому абоненту не нужны.

Для передачи конфиденциальных сообщений в адрес i-го абонента пользователи сети предварительно архивируют передаваемые сообщения с помощью какого-либо общедоступного архиватора, затем полученный архивированный текст делят на фрагменты (если в этом есть необходимость) так, чтобы численное представление каждого из этих фрагментов оказалось меньше числа $N(i)$. Численные представления X каждого из этих фрагментов и есть образы конфиденциальных сообщений, подлежащих передаче в адрес i-го абонента. Заметим, что процедура предварительной архивации текстов не является обязательной, хотя она и создает дополнительные сложности для злоумышленников, пытающихся рассекретить систему RSA. Предварительная архивация текстов полезна еще и потому, что в результате архивации исходные тексты уменьшаются, в результате чего уменьшается также число фрагментов, на которые делятся исходные тексты, с тем, чтобы численные представления каждого из этих фрагментов оказались меньше

числа $N(i)$. Число $N(i)$, лимитирующее сверху допустимый объем каждого передаваемого фрагмента текста (независимо от того, является ли этот текст архивированным или нет), зависит от того, насколько большими выбраны числа $q(i)$ и $p(i)$. Например, если числа $q(i)$ и $p(i)$ содержат по 100 десятичных знаков каждое, то число $N(i)$ будет состоять из 200 десятичных знаков, что эквивалентно 660 битам, т.е. при таком выборе чисел $q(i)$ и $p(i)$ объем каждого фрагмента шифруемого текста сверху лимитирован 660 битами.

В качестве односторонней функции с ловушкой в системе RSA служит функция

$$y = f(X) = X^{e(i)} \bmod(N(i)). \qquad (4.7)$$

Эта функция признается односторонней в силу того, что пока не известны результаты, позволяющие при достаточно больших числах $e(i)$ и $N(i)$ на основе числа y (т.е. на основе криптограммы) определить число X (т.е. исходное сообщение). Иными словами, при заданном аргументе X вычисление $y = f(X)$ не представляет особого труда, тогда как обращение функции $f(X)$, т.е. вычисление значения

$$X = f^{-1}(y) \qquad (4.8)$$

при известном y связано с большим объмом вычислительных работ. Заметим, что утверждение об отсутствии эффективных методов обращения функции (4.7), равно как и утверждение об отсутствии эффективных методов разложения больших чисел $N(i)$ на простые множители $q(i)$ и $p(i)$,скорее являются предположениями, нежели утверждениями в строгом математическом смысле. По крайней мере, нам не известны публикации, где бы приводилось доказательство этих предположений, которые скорее строятся не на строгих доказательствах, а лишь на отсутствии работ, где бы приводились эффективные методы обращения функции (4.7) или разложения числа $N(i)$ на простые множители. Это обстоятельство является одной из слабых сторон, присущих всем криптосистемам открытого шифрования, базирующихся на возведении чисел в большие степени по большому модулю.

Несмотря на вышесказанное, в дальнейшем изложении мы все же будем придерживаться предположения о том, что обращение функции (4.7), равно как и разложение чисел $N(i)$ на простые множители, представляются достаточно сложными задачами и поэтому перехват злоумышленником числа y не позволит ему восстановить конфиденциальное сообщение X. В этом, собственно, и заключается односторонность функции (4.7). Что же касается ловушки для этой функции, то ее роль в данном случае выполняет секретный ключ $d(i)$, поскольку с его помощью обращение функции (4.7) существенно упрощается. Так, легко

убедиться, что имеет место

$$y^{d(i)}\text{mod}(N(i)) = X^{e(i)d(i)}\text{mod}(N(i)) = \tag{4.9}$$

$$= X^{1+k\text{F}(N(i))}\text{mod}(N(i)) = X \cdot X^{k\text{F}(N(i))}\text{mod}(N(i)).$$

Если X является взаимно простым с $N(i)$ числом, то согласно теореме Эйлера

$$X^{\text{F}(N(i))} = 1\text{mod}(N(i)) \tag{4.10}$$

и поэтому с учетом (4.9) имеем

$$X = y^{d(i)}\text{mod}(N(i)). \tag{4.11}$$

Можно, однако, доказать, что соотношение (4.11) имеет место для любого $X < N(i)$, независимо от того, является ли X взаимно простым с числом $N(i)$ или нет, т.е. независимо от того, имеет ли место формула (4.10) или нет. Во всех случаях число

$$X^{\text{F}(N(i))}\text{mod}(N(i))$$

является единичным элементом некоторой группы относительно операции умножения – группы, которой принадлежит элемент X. А это означает, что при любых $X < N(i)$ формулу (4.9) можно переписать как

$$y^{d(i)}\text{mod}(N(i)) = X \cdot \overline{E}\ \text{mod}(N(i)) = X, \tag{4.12}$$

где $\overline{E}$ единый элемент группы, которой принадлежит элемент X. Иными словами, абонент, владеющий секретным ключом $d(i)$, с помощью формулы

$$X = f^{-1}(y) = y^{d(i)}\text{mod}(N(i)) \tag{4.13}$$

относительно легко восстановит исходное сообщение X, расшифровывая тем самым криптограмму y.

В этом, собственно, и заключается сущность криптосистемы RSA, где шифрование исходных сообщений X осуществляется с использованием открытого ключа – пары чисел $e(i)$ и $N(i)$ и сводится к вычислению криптограммы y с помощью формулы (4.7). Расшифрование криптограммы y осуществляется с использованием секретного ключа – числа $d(i)$ и сводится к вычислению исходного сообщения X с помощью формулы (4.11).

Пример 4.1.

Пусть в качестве простых чисел $q(i)$ и $p(i)$ выбраны числа $q(i) = 17$ и $p(i) = 23$. Тогда $N(i) = 17 \cdot 23 = 391$, а число $F(N(i))$ определится по формуле (4.5), т.е.

$$F(N(i)) = 16 \cdot 22 = 352 = 2^5 \cdot 11.$$

В качестве $e(i)$ при этом можно брать произвольное взаимно простое с

$F(N(i))$ число, например, число $e(i) = 85$. Тогда в качестве числа $d(i)$ можно выбрать произвольное число, удовлетворяющее условию (4.6), например, число $d(i) = 29$. Легко проверить, что пара чисел $e(i) = 85$ и $d(i) = 29$ удовлетворяет условию (4.6). Очередным конфиденциальным сообщением может служить произвольное число X, удовлетворяющее условию

$$2 \leqslant X \leqslant N(i) - 2$$

(обратим внимание, что из интервала возможных значений X мы исключили числа $X = 1$ и $X = N(i) - 1$). Пусть $X = 35$. Тогда шифрограммой будет служить число

$$y = 35^{85} \mathrm{mod}(391) = 307.$$

Именно число $y = 307$ и посылается по открытому каналу связи в адрес i-го абонемента – получателя информации. Чтобы восстановить исходное сообщение, т.е. число X, i-й абонент возводит число $y = 307$ в степень $d(i) = 29$ по тому же модулю $N(i) = 391$:

$$X = 307^{29} \mathrm{mod}(391) = 35.$$

Аналогично, если $X = 51$ (обратим внимание, что число $X = 51$ кратно числу $q(i) = 17$), то

$$y = 51^{85} \mathrm{mod}(391) = 306,$$

$$X = 306^{29} \mathrm{mod}(391) = 51.$$

Важно отметить, что абонент – отправитель конфиденциальных чисел владеет лишь открытым ключом шифрования – парой чисел $N(i)$ и $e(i)$, знание которых не является достаточным для расшифрования криптограмм. Если допустить, например, что после шифрования очередного сообщения X его отправитель потерял это сообщение, то на основе им же вычисленной криптограммы y с помощью открытого ключа шифрования он уже не может восстановить исходное сообщение X. В этом и заключается специфика криптосистем открытого шифрования. И поскольку знание ключа шифрования вовсе не является достаточным для того, чтобы восстановить исходное сообщение, то отпадает необходимость держать этот ключ в секрете. А коль скоро снимается необходимость держать его в секрете, то отпадает необходимость и в его индивидуализации с каждым потенциальным отправителем. Тем самым становится возможным не только "открывать" ключ шифрования, но и сделать его единым для всех отправителей. В соответствии с этим становится единым и секретный ключ расшифрования, т.е. одним и тем же числом $d(i)$ расшифровываются все засекреченные тексты, независимо от того, от какого именно отправителя они получены.

4.4. ОРГАНИЗАЦИЯ ЭЛЕКТРОННОЙ ПОДПИСИ В КРИПТОСИСТЕМЕ RSA

Важным преимуществом криптосистем открытого шифрования вообще и криптосистемы RSA в частности является возможность довольно простой организации в ее рамках электронной подписи. Раньше, когда сторонами, обменивающимися секретными сообщениями, были дипломаты, военные и др., можно было говорить о надежности партнеров по связи, об их взаимном доверии друг к другу. Основной заботой обменивающихся сторон служило лишь то, чтобы в конфиденциальную связь не смогли вклиниться третьи стороны. Практически были исключены случаи, когда после получения очередного конфиденциального сообщения адресат вел бы себя недобросовестно и по каким-либо соображениям объявлял о неполучении им этого сообщения. Или же, наоборот, когда абонент объявлял бы о получении им некоторой информации, хотя в действительности такую информацию он не получал. Иными словами, речь шла об обмене конфиденциальными сообщениями между "своими", которые пользовались безграничным взаимным доверием. При такой постановке вполне приемлемыми оказались криптосистемы, базирующиеся на использовании секретных ключей шифрования.

Принципиально иная картина складывается сейчас, когда обмен документами (сообщениями) осуществляется между абонентами, которые заведомо не доверяют друг другу. Например, когда речь идет об обмене информацией (пусть даже конфиденциальной) между коммерческими фирмами, банками или иными подобными организациями, надеяться на добросовестность, порядочность партнера по связи было бы наивным и, главное, опасным. Здесь должны быть предусмотрены дополнительные меры, доказывающие факт посылки или получения соответствующих сообщений. Именно здесь проявляется одно из важных преимуществ односторонних функций и реализованных на них криптосистем с открытым ключом шифрования. С помощью односторонних функций удается организовать электронную подпись, которая по своей надежности вполне может конкурировать с обычными подписями на бумажных носителях.

Проследим, например, за механизмом организации электронной подписи в рамках системы RSA.

Пусть имеется необходимость в том, чтобы j-м абонентом в адрес i-го абонента было послано некоторое сообщение (некоторый текст) X и чтобы к тому же j-й абонент подписался под этим текстом, с тем, чтобы в последующем у i-го абонента было неопровержимое (или почти неопровержимое) доказательство того, что данный текст был послан не кем иным, как именно j-м абонентом.

Следуя за [7], будем рассматривать вариант реализации электронной подписи с использованием хеш-функции от аргумента X, т.е. функции $h(X)$, обладающей следующими свойствами:

– хеш-функция $h(X)$ должна быть чувствительна ко всевозможным модификациям (изменениям) аргумента X, таким, как вставка, выбросы, перестановки и т.п.;

– функция $h(X)$ должна обладать свойством необратимости, т.е. задача подбора текста X, который обладал бы данной $h(X)$, должна быть чрезвычайно сложной (вычислительно неразрешимой);

– вероятность того, что значения $h(X)$ двух различных текстов совпадут, должна быть ничтожно мала.

Электронную подпись с использованием хеш-функции $h(X)$ в рамках системы RSA можно реализовать следующим образом.

1. Отправитель информации (j-й абонент) вычисляет хеш-функцию $h(X)$ от аргумента X – передаваемого сообщения.

2. В зависимости от того, является сообщение X конфиденциальным или нет:

а) шифрует сообщение X, т.е. вычисляет число

$$y(X) = X^{e(i)} \bmod(N(i)) \tag{4.14}$$

и по открытому каналу посылает его в адрес i-го абонента,

б) в адрес i-го абонента посылает число X.

3. Ставит свою подпись под $h(X)$, т.е. вычисляет число

$$S(h(X)) = (h(X))^{d(j)} \bmod(N(j)) \tag{4.15}$$

и посылает его в адрес i-го абонента.

Получатель подписанного текста (i-й абонент) при необходимости, т.е. когда имеет место случай (а), расшифровывает текст с помощью формулы

$$X = (y(X))^{d(i)} \bmod(N(i)), \tag{4.16}$$

вычисляет хеш-функцию $h^*(X)$ от аргумента X и сверяет ее значение с результатом расшифрования криптограммы $S(h(X))$. Иными словами, проверяется условие

$$h^*(X) = (S(h(X))^{e(j)} \bmod(N(j)), \tag{4.17}$$

соблюдение которого и есть доказательство того, что сообщение X с его хеш-функцией $h(X)$ в адрес i-го абонента было послано именно j-м абонентом. Ведь никто другой, кроме абонента, владеющего секретным ключом $d(j)$, не может вычислить число $S(h(X))$ такое, чтобы оно удовлетворило равенству (4.17). Заметим, что в результате "перехвата" числа $S(h(X))$ злоумышленник сможет восстановить хеш-функцию $h(X)$, поскольку число $e(j)$, т.е. открытый ключ шифрования j-го абонента, общеизвестно. Но это не поможет ему в деле подделки подписи. Для этого ему необходимо владеть закрытым ключом шифрования, т.е. числом $d(j)$. Только тогда он сможет имитировать посылку в адрес любого абонента произвольного текста от имени (за подписью)

j-го абонента. Исходя из этого, можно заключить, что предъявление арбитру со стороны i-го абонента текста X, его хеш-функции $h(X)$ и числа $S(h(X))$ является достаточно убедительным доказательством того, что текст X он получил именно от j-го абонента.

4.5. ВОЗМОЖНЫЕ АТАКИ НА СИСТЕМУ RSA И НЕКОТОРЫЕ ВОПРОСЫ ЕЕ КРИПТОСТОЙКОСТИ

Как при передаче конфиденциальных сообщений, так и при организации электронной подписи в системе RSA объектами атаки со стороны злоумышленников в одинаковой степени могут быть числа q, p, $F(N)$ и d (ниже индексы, указывающие на абонентов, при этих числах будем ставить лишь при необходимости), поскольку знание любого из этих чисел при известных e и N вполне достаточно для рассекречивания системы RSA.

В ряде случаев, особенно тогда, когда программное обеспечение системы RSA приобретено на стороне, имеется определенная вероятность того, что противнику будет известен интервал значений, откуда могут быть выбраны числа q и p. Покажем, что при определенных условиях это может представлять серьезную угрозу с точки зрения разложения противником числа N на множители q и p.

Пусть известно (в том числе и противнику), что простые числа q и p выбираются из интервала $[a, b]$, причем $q < p$. Естественно полагать, что противнику известно также число $N = q \cdot p$, которое может оказаться меньшим, равным или большим произведения $a \cdot b$.

Если $N = a \cdot b$, то значения q и p удовлетворяют неравенствам

$$a \leqslant q < \sqrt{a \cdot b}, \tag{4.18}$$

$$\sqrt{a \cdot b} < p \leqslant b, \tag{4.18а}$$

т.е. диапазоны изменения чисел q и p равны соответственно

$$\Delta q = \sqrt{a \cdot b} - a, \tag{4.19}$$

$$\Delta p = b - \sqrt{a \cdot b}. \tag{4.19а}$$

Отсюда следует, что

$$\Delta p - \Delta q = (\sqrt{b} - \sqrt{a})^2 > 0. \tag{4.20}$$

Если $N < a \cdot b$, то значения q и p удовлетворяют неравенствам

$$a \leqslant q < \sqrt{N}, \tag{4.21}$$

$$\sqrt{N} < p \leqslant N/a, \tag{4.21а}$$

т.е. диапазоны изменения чисел q и p равны соответственно

$$\Delta q = \sqrt{N} - a, \tag{4.22}$$

$$\Delta p = N/a - \sqrt{N}. \tag{4.22а}$$

Отсюд следует, что

$$\Delta p - \Delta q = (\sqrt{N} - a)^2/a) > 0. \tag{4.23}$$

Если же $N > a \cdot b$, то значения q и p удовлетворяют неравенствам

$$N/b \leqslant q < \sqrt{N}, \tag{4.24}$$

$$\sqrt{N} < p \leqslant b, \tag{4.24а}$$

т.е. диапазоны изменения чисел q и p равны соответственно

$$\Delta q = \sqrt{N} - N/b, \tag{4.25}$$

$$\Delta p = b - \sqrt{N}. \tag{4.25а}$$

Отсюда следует, что

$$\Delta p - \Delta q = (b - \sqrt{N})^2/b > 0. \tag{4.26}$$

Таким образом, как это следует из (4.20), (4.23) и (4.26), во всех трех случаях имеет место $\Delta q < \Delta p$, т.е. диапазоны возможных значений q оказываются меньше диапазонов возможных значений p.

В случаях, когда $N \leqslant a \cdot b$, область возможных значений q определяется формулой (4.21), а в случае, когда $N > a \cdot b$, эта область определяется формулой (4.24). Если у противника не имеется иных, более эффективных методов разложения числа N на множители q и p, чем метод прямого перебора, то при $N \leqslant a \cdot b$ ему предстоит осуществить $\Delta q = \sqrt{N} - a$ (см. формулу (4.22)) проверок, а при $N > a \cdot b$ необходимое число проверок определяется по формуле (4.25). Если рассматривать зависимость Δq от аргумента $\sqrt{N}$, то в области

$$a \leqslant \sqrt{N} \leqslant \sqrt{a \cdot b} \tag{4.27}$$

эта зависимость линейная, а в области

$$\sqrt{a \cdot b} \leqslant \sqrt{N} < b \tag{4.28}$$

имеет место параболическая зависимость. На рис. 4.1 приведен примерный характер зависимости необходимого числа проверок от числа $\sqrt{N}$ при известных значениях a и b. Здесь случаи (а), (б) и (в) характеризуют эту зависимость в случаях, когда соответственно $a < b/4$, $a = b/4$ и $a > b/4$.

Из рис. 4.1 легко заметить, что с приближением числа $\sqrt{N}$ к значе-

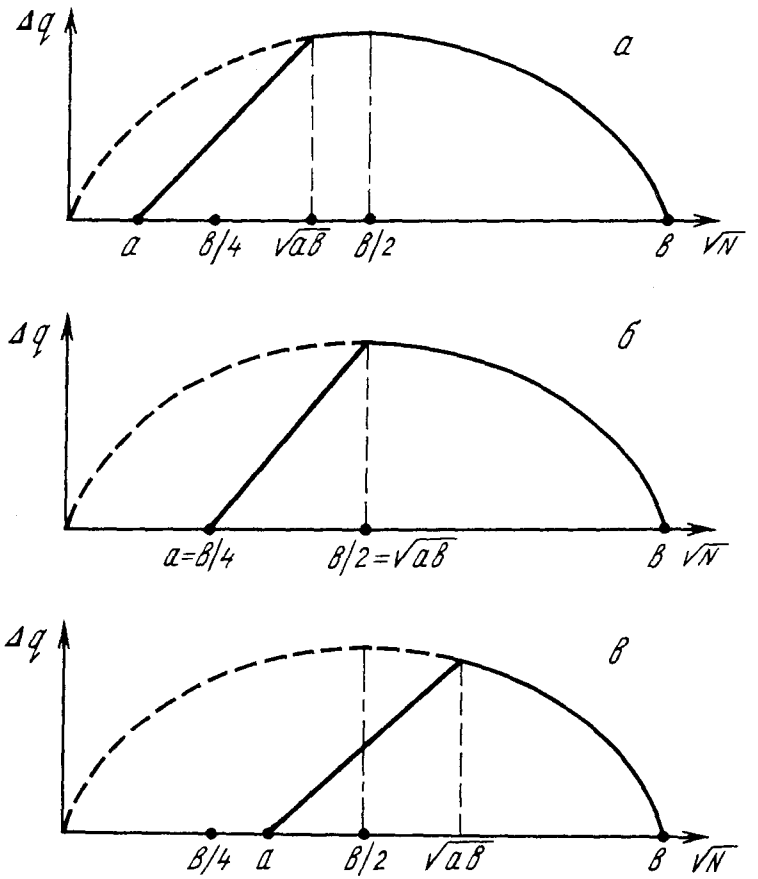

Рис. 4.1. Характер зависимости необходимого числа проверок (Δq) от числа $\sqrt{N}$
а) при a<b/4
б) при a=b/4
в) при a>b/4

ниям a или b необходимое число проверок, т.е. величина Δq уменьшается. А это таит в себе большую опасность, поскольку пользователи, придерживаясь принципа "выбрать числа q и p как можно большими", могут значительно облегчить задачу злоумышленника. Если к примеру, в качестве чисел q и p пользователь возьмет числа, существенно близкие к верхней границе, т.е. к числу b, то тем самым значительно уменьшит необходимое число проверок для злоумышленника.

Таким образом, при заданных числах a и b, чтобы максимально усложнить задачу злоумышленника, числа q и p необходимо выбрать такими, чтобы число $\sqrt{N}$ оказалось близким к числу $b/2$, если $a \leqslant b/4$ и числу $\sqrt{a \cdot b}$, если $a > b/4$ (см. рис. 4.1).

Пример 4.2. Пусть известно (в том числе и злоумышленнику), что числа q и p выбираются из итервала чисел, содержащих по 3 десятичных знака. Пользуясь таблицей простых чисел, можно заметить, что числа a и b при этом окажутся равными, соответственно, $a = 101$ и $b = 997$. Очевидно, отношение $b/a = 997/101$ больше четырех, т.е. имеет место случай (a) (см. рис. 4.1). Пусть число p принято равным $p = 983$ и необходимо выбрать число q такое, чтобы максимально усложнить задачу злоумышленника по вычислению чисел q и p. В табл. 4.1 для различных значений q приведены значения $\sqrt{N}$ и Δq. Численные значения Δq определены с помощью формул (4.22) или (4.25), в зависимости от того, имеет ли место условие $N \leqslant a \cdot b$ или нет.

Таблица 4.1

q	101	103	211	251	257	313	491	911	967	977
$\sqrt{N}$	315	318	455	498	503	554	694	946	975	979
Δq	214	217	247	250	249	246	211	49	22	18
	$N < ab$	$N > a \cdot b$								

В рассматриваемом примере к линейной зоне (формула (4.22)) относится лишь значение $q = 101$ и далее, начиная со значения $q = 103$, начинается параболическая зона (формула (4.25)). В этой зоне с ростом q число Δq сначала растет до достижения значения $\Delta q = b/4 = 250$ (при этом $\sqrt{N}$ оказывается равным $b/2 = 498$), после чего дальнейший рост числа q влечет уменьшение числа Δq, облегчив тем самым задачу потенциальных злоумышленников. Если, к примеру, пользователь стремится выбрать число q как можно большим, надеясь максимально усложнить этим задачу злоумышленника, то это приводит к диаметрально противоположному эффекту: необходимое число испытаний для

злоумышленника может уменьшаться из числа $\Delta q = 250$ (при $q = 251$) до числа $\Delta q = 18$ (при $q = 977$).

Другим объектом прямой атаки (не через числа q и p) со стороны злоумышленника может служить секретный ключ расшифрования – число d. Для анализа возможных угроз его рассекречивания будем рассматривать кольцо сообщений $C(N)$, подразумевая под ним множество классов вычетов по модулю N с определенными на его элементах операциями сложения и умножения по этому модулю. Под периодом кольца $T(C(N))$ будем понимать минимальное значение числа $z > 0$, при котором для всех сообщений (всех элементов кольца $C(N)$) $X \in C(N)$ имеет место

$$X^{z+1} = X \bmod(N). \tag{4.29}$$

Можно показать, что говорить о периоде кольца $C(N)$ можно лишь применительно к случаям, когда в каноническом разложении числа N

$$N = p_1^{\alpha_1} p_2^{\alpha_2} \dots p_n^{\alpha_n} \tag{4.30}$$

все α_i ($i = 1, 2, 3, \dots, n$) равны единице. Во всех остальных случаях среди элементов $X \in C(N)$ всегда окажутся такие, для которых формула (4.29) не имеет места ни при каких $z > 0$. Поскольку в системе RSA число N определяется как произведение двух простых чисел, то можно говорить о периоде кольца $C(N)$, значение которого равно $T(C(N))$. Можно показать, что в рассматриваемом нами случае число T равно наименьшему общему кратному чисел $F(p) = p - 1$ и $F(q) = q - 1$, т.е.

$$T(C(N)) = [p - 1, q - 1]. \tag{4.31}$$

Введем в рассмотрение понятие кольца степеней $C(T)$, подразумевая под ним множество классов вычетов по модулю $T(C(N))$ с определенными на нем операциями сложения и умножения по этому модулю. Можно показать, что множество элементов кольца степеней, которые взаимно просты с числом $T(C(N))$, образует группу относительно операции умножения по этому модулю. Эту группу будем называть группой допустимых степеней, имея в виду то обстоятельство, что числа e и d всегда выбираются такими, чтобы они были взаимно простыми с числом

$$F(N) = (q - 1) \cdot (p - 1). \tag{4.32}$$

Отсюда с учетом (4.31) легко сделать вывод, что числа e и d являются взаимно простыми также с числом $T(C(N))$. Иными словами, элементы группы допустимых степеней характеризуются тем, что числа e и d всегда можно представить как

$$e = E + k_1 \cdot T, \tag{4.33}$$

$$d = D + k_2 \cdot T, \tag{4.34}$$

где элементы E и D принадлежат этой группе, которую далее

обозначим через $G_1(T)$. Из (4.33) и (4.34) следует, что в формулах (4.7) и (4.11), согласно которым в системе RSA осуществляется шифрование и расшифрование конфиденциальных сообщений, числа e и d всегда можно заменить функционально эквивалентными им числами E и D, которые принадлежат группе $G_1(T)$ и в общем случае меньше чисел соответственно e и d. В рассматриваемом смысле для злоумышленника вовсе не обязательно найти именно число d, а вполне достаточно найти соответствующее ему число D, которое в общем случае может оказаться существенно меньшим числа d. Заметим также, что элементы E и D группы $G_1(T)$ являются обратными друг другу, т.е. имеет место

$$E \cdot D = 1 \bmod(T). \tag{4.35}$$

Порядки элементов E и D равны между собой, т.е. $g(E) = g(D)$. Легко заметить, что именно этим порядком и определяется необходимое число операций для рассекречивания числа d – секретного ключа расшифрования. Это обстоятельство необходимо учесть как на этапе выбора чисел q и p, так и особенно на этпе выбора чисел e и d. Выбор конкретной пары чисел q и p уже предопределяет число m – период группы $G_1(T)$. Легко заметить, что именно число m является верхней границей числа $g(D)$. Можно показать, например, что, если каноническое разложение числа T имеет вид

$$T = t_1^{\beta_1} t_2^{\beta_2} \ldots t_k^{\beta_k}, \tag{4.36}$$

то число m определяется по формуле

$$m = [\mathrm{F}^*(t_1^{\beta_1}),\ \mathrm{F}(t_2^{\beta_2}),\ldots,\ \mathrm{F}(t_k^{\beta_k})], \tag{4.37}$$

где

$$\mathrm{F}^*(t_1^{\beta_1}) = 2^{\beta_1 - 2}, \tag{4.37а}$$

если $t_1 = 2$ и $\beta_1 \geq 3$, и

$$\mathrm{F}^*(t_1^{\beta_1}) = \mathrm{F}(t_1^{\beta_1}) = (t_1 - 1) \cdot t_1^{\beta_1 - 1} \tag{4.37б}$$

во всех остальных случаях.

Но следует иметь в виду, что m – это лишь верхняя граница для чисел $g(D)$, т.е. это число, для которого при произвольном выборе чисел e и d имеет место

$$g(D) \leq m. \tag{4.38}$$

При случайном же выборе чисел e и d, какими бы большими они ни были, вполне может оказаться, что число $g(E) = g(D)$, характеризующее в данном случае фактическую криптостойкость системы RSA, окажется недопустимо маленьким. В рассматриваемом смысле при выборе чисел e и d необходимо заботиться, чтобы достаточно большими оказа-

лись не сами эти числа, а соответствующие им числа $g(E)$, которыми, собственно, и определяется необходимое число операций, необходимых для определения секретного ключа расшифрования. Опасность же того, что эти числа могут оказаться недопустимо малыми даже при чрезвычайно больших значениях e и d, вполне реальна, поскольку при случайном выборе чисел q и p недопустимо малым может оказаться даже число m – верхний предел значений $g(E)$. Бытующую же практику случайного выбора простых чисел q и p, с одной стороны, и чисел e и d – с другой, никак нельзя признать приемлемой с точки зрения обеспечения приемлемого уровня криптостойкости. Одно лишь очевидно, что, если числа e и d оказываются бóльшими соответствующих им элементов E и D, т.е., когда числа k_1 и k_2 в формулах (4.33) и (4.34) оказываются бóльшими нуля, то это приводит лишь к ухудшению скоростных показателей системы, ни на йоту не увеличивая фактическую криптостойкость системы.

Пример 4.3. Пусть числа q и p выбраны равными соответственно $q = 1601$ и $p = 4801$, т.е. $N = 7686401$. При этом $F(N) = 7680000$, а число $T(C(N))$ определим по формуле (4.31), т.е.

$$T = [1600, 4800] = 4800 = 2^6 \cdot 3 \cdot 5^2.$$

Число различных вариантов выбора числа E при этом равно $F(T) = 1280 = 2^8 \cdot 5$. Пользуясь формулой (4.37), можно определить верхнюю границу числа $g(E)$, т.е. число m

$$m = [2^4, 2, 2^2 \cdot 5] = 80.$$

Это означает, что при любом выборе чисел e и d для рассекречивания данной системы потребуется не более 80 операций. При случайном же выборе числа e необходимое число операций для рассекречивания системы может оказаться значительно меньшим. В табл. 4.2 для рассматриваемого случая приведены некоторые конкретные значения числа E и соответствующие им значения $g(E)$.

Таблица 4.2

E	799	1343	1921	929	497
$g(E)$	2	4	5	10	20

Из приведенных в табл. 4.2 численных данных легко убедиться, что при заданной паре чисел q и p случайный выбор числа e чреват опасностью существенного уменьшения криптостойкости системы RSA. Обратим внимание, что фактическая криптостойкость системы определяется числом $g(E)$ и она не изменится, если вместо числа E выбрать в качестве e произвольное другое число, определяемое по формуле (4.33).

4.6. О НАДЕЖНОСТИ СИСТЕМЫ RSA. ШИФРУЕМЫЕ И НЕШИФРУЕМЫЕ СООБЩЕНИЯ

В ряде случаев, говоря о надежности тех или иных криптосистем, имеют ввиду их криптостойкость, т.е. число операций, необходимых для рассекречивания этих систем. Применительно к системе RSA, говоря о ее надежности, мы будем иметь в виду круг вопросов, связанных с анализом того, может ли случиться так, что шифрограмма очередного сообщения совпадет с исходным текстом. Конкретнее, речь идет о том, существуют ли сообщения, для которых имеет место

$$X^e \bmod(N) = X, \tag{4.39}$$

и, если да, то какова вероятность того, что наугад взятый элемент кольца сообщений окажется именно таким.

Важность рассмотрения этого круга вопросов очевидным образом следует из того факта, что сообщения, удовлетворяющие условию (4.39), после их шифрования остаются такими же, какими они были до шифрования. В результате может случиться так, что конфиденциальные сообщения будут передаваться по открытому каналу связи без какой-либо их модификации.

Это обстоятельство накладывает дополнительные требования при выборе числа e – оно должно быть таким, чтобы по возможности меньшее число элементов кольца сообщений удовлетворяли условию (4.39). Сообщения, удовлетворяющие условию (4.39), будем называть нешифруемыми. Забегая вперед, отметим, что невозможно подобрать числа q, p и e такими, чтобы все элементы кольца сообщений оказались шифруемыми. И, наоборот, всегда можно подобрать такую тройку чисел q, p и e, чтобы все элементы кольца сообщений оказались нешифруемыми. Например, легко убедиться, что при любом выборе простых чисел q и p, если число e принять равным

$$e = 1 + k \cdot [F(q), F(p)], \tag{4.40}$$

где k – произвольное натуральное число, то все элементы кольца сообщений окажутся нешифруемыми.

Для элементов кольца $C(N)$, которые взаимно просты с числом N, условие (4.39) эквивалентно условию

$$X^{e-1} \bmod(N) = 1, \tag{4.41}$$

поскольку относительно операции умножения по модулю N эти элементы образуют группу $G_1(N)$ с единичным элементом, равным $\overline{E} = 1$. Очевидно, число элементов этой группы равно $F(N)$. Среди остальных элементов кольца $C(N)$ не могут оказаться элементы, удовлетворяющие условию (4.41), так как они распределены по двум другим группам, единичные элементы которых отличны от $\overline{E} = 1$. При заданной паре

чисел q и p значение периода $T(C(N))$ можно определить по формуле (4.31). Можно показать, что при заданном значении e число нешифруемых сообщений группы $G_1(N)$, т.е. число элементов этой группы, удовлетворяющих условию (4.41), можно определить по формуле

$$\upsilon_1(e-1) = \gamma_q(e-1) \cdot \gamma_p(e-1), \tag{4.42}$$

где числа $\gamma_q(e-1)$ и $\gamma_p(e-1)$ определены как

$$\gamma_q(e-1) = (q-1, e-1), \tag{4.43}$$

$$\gamma_p(e-1) = (p-1, e-1). \tag{4.43а}$$

Когда число N является произведением двух простых чисел q и p, все $N-1$ ненулевых элемента кольца $C(N)$ распределяются по трем группам. Это группа $G_1(N)$, включающая $F(N)$ взаимно простых с N элементов кольца, группа $G_q(N)$, включающая $p-1$ элементов кольца $C(N)$, кратных числу q, и группа $G_p(N)$, включающая $q-1$ элементов кольца $C(N)$, кратных числу p. Естественно, что среди элементов групп $G_q(N)$ и $G_p(N)$ не могут оказаться элементы, которые являются решением уравнения (4.41), так как единичные элементы этих групп отличны от элемента $\overline{E} = 1$. В то же время условие (4.41) является достаточным, но не необходимым для того, чтобы очередной элемент оказался нешифруемым. Действительно, ведь для того, чтобы очередной элемент X оказался нешифруемым, необходимо, чтобы имело место

$$X^e = X \bmod(N). \tag{4.44}$$

Чтобы это условие имело место, необходимо и достаточно, чтобы имело место

$$X^{e-1} = \overline{E} \bmod(N), \tag{4.45}$$

где под $\overline{E}$ подразумевается единичный элемент той группы, которой принадлежит элемент X. В случае, когда речь шла об элементах группы $G_1(N)$, единичный элемент оказался равным единице и вместо уравнения (4.45) мы имели дело с уравнением (4.41). В общем же случае следует учесть конкретные значения единичных элементов. Так, можно показать, что число элементов группы $G_q(N)$, удовлетворяющих уравнению (4.45), равно $\gamma_q(e-1)$. Аналогично, число элементов группы $G_p(N)$, удовлетворяющих условию (4.45), равно $\gamma_p(e-1)$.

Таким образом, при заданных значениях e и N, из $N-1$ ненулевых элементов кольца $C(N)$

$$V(e-1) = \upsilon_1(e-1) + \gamma_q(e-1) + \gamma_p(e-1) \tag{4.46}$$

элементы окажутся нешифруемыми.

Пример 4.4. Пусть в качестве простых чисел q и p выбраны числа, соответственно, $q = 151$ и $p = 251$. Тогда число N получится равным $N = 37901$, а числа элементов в группах, соответственно, $G_1(N)$, $G_q(N)$

Таблица 4.3

e	7	11	13	17	19	23	29	31
$\upsilon_1(e-1)$	12	100	12	4	12	4	4	300
$V(e-1)$	20	120	20	8	20	8	8	340

e	37	41	43	49	77	121	151	173
$\upsilon_1(e-1)$	12	100	12	12	4	300	7500	4
$V(e-1)$	20	120	20	20	8	340	7700	8

e	193	211	251	751
$\upsilon_1(e-1)$	12	300	12500	37500
$V(e-1)$	20	340	12800	37900

и $G_p(N)$ окажутся равными $F(N) = (q-1)\cdot(p-1) = 37500 = 2^2\cdot 3\cdot 5^5$, $p-1 = 250 = 2\cdot 5^3$, $q-1 = 150 = 2\cdot 3\cdot 5^2$. Значение периода кольца сообщений $C(N)$ при этом определяется как

$$T = [q-1, p-1] = [150, 250] = 750 = 2\cdot 3\cdot 5^3.$$

Число функционально различных вариантов выбора числа e, т.е. число различных E равно

$$F(T) = 200 = 2^3\cdot 5^2.$$

Порядок каждого из этих чисел E, т.е. каждое число типа $g(E)$ является делителем числа m, определяемого по формуле

$$m = [F(2), F(3), F(5^3)] = 100 = 2^2\cdot 5^2.$$

Это означает, что при любом выборе числа e для рассекречивания системы потребуется не более 100 операций. При случайном же выборе числа e число $g(E)$ – характеристика фактической криптостойкости системы – может оказаться значительно меньшим числа $m = 100$.

Пусть в качестве e выбрано число $e = 1301$, взаимно простое с числом $F(N) = 37500$. Число $e-1$ при этом окажется равным $e-1 = 1300 = 2^2\cdot 5^2\cdot 13$, т.е.

$$\gamma_q(e-1) = (2\cdot 5^3, 2^2\cdot 5^2\cdot 13) = 50 = 2\cdot 5^2,$$

$$\gamma_p(e-1) = (2\cdot 3\cdot 5^2, 2^2\cdot 5^2\cdot 13) = 50 = 2\cdot 5^2.$$

Подставляя эти значения в формулы (4.42) и (4.46), получим числа

$$\upsilon_1(e-1) = 2500,$$

$$V(e-1) = 2500 + 50 + 50 = 2600.$$

Таким образом, при принятых конкретных значениях $q = 151$, $p = 251$

и $e = 1301$ из общего числа $N - 1 = 37900$ ненулевых элементов кольца сообщений $V(e - 1) = 2600$ элементов окажутся нешифруемыми, т.е., в среднем каждое из 15-ти сообщений окажется нешифруемым. Естественно, что такое положение дел не может устраивать пользователей.

В заключение приведем значения $\upsilon_1(e - 1)$ и $V(e - 1)$ при некоторых других вариантах выбора числа e (см. табл. 4.3):

ЛИТЕРАТУРА К ГЛАВЕ 4

1. *Аветисян Д.О.* Проблемы информационного поиска: (Эффективность, автоматическое кодирование, поисковые стратегии). – М.: Финансы и статистика, 1981. – 208 с., ил.
2. *Аветисян Д.О.* Статистические методы при решении прикладных задач документального поиска. Автореф. дис. д-ра техн. наук. – Новосибирск, 1975.
3. *Аршинов М.Н., Садовский Л.Е.* Коды и математика – М.: Наука, 1983. – (Б-чка "Квант"; Вып. 30).
4. *Аснис И.Л., Федоренко С.В., Шабунов К.Б.* Краткий обзор криптосистем с открытым ключом // Защита информации "Конфидент". – 1994. – № 2.
5. *Герасименко В.А.* Защита информации в автоматизированных системах обработки данных. В 2-х кн.: Кн. 1 и 2. – Энергоатомиздат, 1994.
6. *Лебедев А.Н.* Криптография "с открытым ключом" и возможности ее практического применения // Защита информации. – 1992. – Вып. 2.
7. *Мельников Ю.Н.* Электронная цифровая подпись. Возможность защиты // Защита информации "Конфидеит". – 1995. – № 6/4.
8. *Мошонкин А.Г.* Что такое криптография с открытым ключом? // Защита информации "Конфидент". – 1994. – № 1.
9. *Расторгуев С.* Программные методы защиты информации в компьютерах и сетях. – М.: Издательство Агентства "Яхтсмен". – 1993. – 188 с.
10. *Спесивцев А.В., Вегнер В.А., Крутяков А.Ю., Серегин В.В., Сидоров В.А.* Защита информации в персональных ЭВМ. – М.: Радио и связь, МП "ВЕСТА", 1978.
11. *Шеннон К.* Работы по теории информации и кибернетике. – М.: Иностранная литература, 1963.
12. *Diffie W., Hellman M.* New Direction in Criptography // IEEE Trans. Inform. Theory. – 1976. – Vol. 22, Nov. – P. 644–654.
13. *ElGamal T.* A public key cryptosystem and signature scheme based on discrete logarithms // IEEE Trans., Inform., Th. 1985. – Vol. IT-31. – № 4, July. – P. 469–472.
14. *Rivest R., Shamir A., Adleman L.* A method for obtaining digital signatures and public-key cryptosystems // Comm. of the ACM. – 1978. – Vol. 21. – P. 120–126.

ГЛАВА 5

ПОИСК ТЕКСТОВ. МАТЕМАТИЧЕСКИЕ МОДЕЛИ ДОКУМЕНТАЛЬНОГО ПОИСКА

ПРОБЛЕМА кодирования текстов, их максимального сжатия и возможно быстрой, надежной и экономной передачи через различные каналы связи с давних времен была и остается предметом интенсивных исследований специалистов различных стран. Эти исследования естественным образом привели к созданию К. Шенноном теории информации, в рамках которой большинство задач, в той или иной мере касающихся этой проблемы, удается довести до достаточно строгих формальных постановок [15].

Что же касается проблемы документального поиска, то она стала предметом активных научных исследований лишь с шестидесятых годов нашего столетия, после появления первых автоматизированных систем документального поиска (АСДП). Поскольку документальный поиск сводится к установлению семантического расстояния между различными идеями, понятиями путем сопоставления их языковых формулировок, то с самого начала в центре внимания оказались гносеологические вопросы взаимоотношения языка и мышления. Ряд специалистов категорически опровергал возможность формального математического описания отдельных процедур, в совокупности реализующих предварительную обработку, хранение и поиск документальной информации. Со временем эти страсти улеглись и большинство специалистов пришло к заключению, что разработка формальных математических средств для описания и решения если не проблемы в целом, то хотя бы отдельных ее задач не только возможна, но и необходима. В частности, была доказана эффективность применения для решения различных задач документального поиска теории вероятностей и математической статистики, теории нечетких подмножеств, корреляционного анализа, теории матриц и, что представляется наиболее важным, теории информации.

В разделе 5.1 настоящей главы вкратце рассматривается гносеологическая проблематика, сложившаяся вокруг основного понятия документального поиска – понятия релевантности. В следующих разде-

лах рассматриваются множественные, энтропийная, корреляционная и матричные модели документального поиска. Глава завершается рассмотрением круга вопросов технико-экономической и функциональной эффективности автоматизированных систем документального поиска.

5.1. РЕЛЕВАНТНОСТЬ КАК ЦЕНТРАЛЬНОЕ ПОНЯТИЕ ТЕОРИИ ДОКУМЕНТАЛЬНОГО ПОИСКА

Релевантность характеризует степень соответствия содержания документа, найденного в результате информационного поиска, содержанию информационного запроса. Степень же соответствия содержания найденного документа информационной потребности пользователя, сформулированной в виде информационного запроса, характеризуется понятием пертинентности [7].

В литературе по информатике имеются также другие определения этих понятий. Ряд авторов, например, рассматривает пертинентность не как самостоятельное понятие, а как нечто, уточняющее понятие релевантности. Сложность количественной оценки релевантности обусловлена тем, что при ее оценке приходится судить о семантическом расстоянии между содержанием документа и информационной потребностью лишь на основе сопоставления (в большинстве случаев третьей стороной) их языковых формулировок, а именно текста документа и информационного запроса. Чрезвычайно сложное психологическое явление информационной потребности не всегда удается точно, однозначно и исчерпывающе сформулировать в виде информационного запроса. Под одним и тем же информационным запросом различные специалисты в принципе могут подразумевать различные информационные потребности. В общем случае нельзя также рассчитывать на однозначность при интерпретации содержания текстов документов.

Элементы субъективизма, неминуемо сопровождающие процедуры языковой формулировки различных семантических категорий, стали причиной многочисленных научных споров вокруг понятия релевантности (пертинентности). Указывая на неизбежную терминологическую путаницу "релевантность" – "пертинентность" при обобщенном определении этого многоаспектного понятия, ряд авторов справедливо находит целесообразным дифференцированное его рассмотрение, позволяющее разграничить понятия релевантности и пертинентности:

1) формальная релевантность (наличие в документе контекстных ситуаций, затребованных пользовательским запросом),

2) содержательная или действительная релевантность (соответствие содержания документа информационной потребности пользователя),

3) индивидуально-прагматическая релевантность, или пертинентность.

Несмотря на существенное различие в понимании понятия релевантности в смысле ее различных определений, в гносеологическом плане последние отражают диалектическое единство чувственного образа восприятия объективной реальности и его языковой формы выражения. Здесь уместно вспомнить, что еще в глубокой древности стоики Хрисипп и Кратес Малосский различали "обозначаемое", "обозначающее" и "объект". Под "объектом" они подразумевали внешний субстрат, понимание которого, или отражение которого в нашем рассудке, они отождествляли с "обозначаемым". Под "обозначающим" они понимали звук, т.е. материализованную оболочку "обозначаемого" – идей, понятий. Таким образом, "объект" и "обозначающий" телесны, тогда как "обозначаемое" бестелесно и может быть истинным или ложным. В рассматриваемом нами случае в качестве "обозначающего" выступают текст документа и информационный запрос, тогда как "обозначаемыми" являются содержание документа и информационная потребность [4].

Второе и отчасти третье определения релевантности неразрывно связаны с гносеологическими основами взаимоотношения языка и мышления. Образную формулировку этого взаимоотношения приводит крупнейший немецкий языковед XVIII столетия В. Гумбольдт: "...понятие не может отрешиться от слова, как человек не может скинуть с себя своей физиономии. Слово есть индивидуальная физиономия понятия, которое, захотев сбросить ее с себя, только переменило бы одно слово на другое и, стало быть, все же явилось бы в слове" [4].

Как бы в продолжение этой мысли у отечественного специалиста В.З. Панфилова читаем: "Язык есть средство осуществления человеческого мышления, это последнее не может протекать вне и помимо естественного языка или других знаковых систем. Язык и мышление неотделимы друг от друга как в своем возникновении, так и в своем существовании".

И далее: "Язык и мышление образуют такое диалектически противоречивое единство, в котором язык, при определяющей роли мышления, представляет собой относительно самостоятельное явление, в свою очередь оказывающее определенное обратное воздействие на мышление" [12].

В результате абсолютизации относительной самостоятельности языка В. Гумбольдт делает вывод о том, что наши отношения к предметам и явлениям объективного мира обусловлены не свойствами этих предметов и явлений, а языком. Позже представители различных направлений неогумбольдтианства и отдельные представители структурализма рассматривали язык как некую имманентную сущность, выступающую в качестве первичной не только по отношению к мышлению, но и по отношению к самой объективной действительности. "Реальный мир, – пишет американский специалист по этнолингвистике Э. Сепир, – в значительной степени бессознательно строится на основе языковых норм данной группы... Мы видим, слышим и воспринимаем

так или иначе те или другие явления главным образом благодаря тому, что языковые нормы нашего общества предполагают данную форму выражения" [5].

Другой крайностью является концепция специалистов, допускающих, что мышление может протекать в чистом виде, без помощи языка. "Иногда думают, – пишет представитель логического позитивизма Б. Рассел, – что не может быть мысли без языка, но я не могу с этим согласиться: я считаю, что может быть мысль и даже истинное и ложное верование без языка" [13].

Сторонники "чистого мышления" утверждают, что язык отягощает и искажает мысль, возникшую в процессе чистого мышления. Примечательно, что нечто подобное встречается и у В. Гумбольдта: "...слова стесняют внутреннее чувство, которое всегда полнее их содержания, и часто угрожают подавить его особенные оттенки своей природой, которая слишком материальна по звуку и слишком абстрактно обща по значению" [4].

Приверженцы такой концепции фактически ставят под сомнение возможность сопоставления продуктов мышления (идеи, понятия, информационная потребность и т.д.) на уровне их языковых формулировок (статьи, рефераты, запросы и т.д.). Менее категоричным представляется высказывание И. Бар-Хиллела: "Не будет неправильным сказать, что эти символы – будь то отдельные слова, целые фразы, предметные заголовки, дескрипторы, унитермы и т.д. – ... как-то обозначают информационное содержание документа. Однако скорей в заблуждение вводят часто встречающиеся утверждения, что такие символы сами содержат – например, в конденсированном виде – часть информации, приведенной в этом документе" [16].

При рассмотрении понятия релевантности важное место отводится также третьей ее компоненте. У американского математика М. Таубе читаем: "Релевантность – это психологический предикат, описывающий признание или непризнание им (потребителем) определенного сходства между смыслом или содержанием документа и смыслом или содержанием запроса" [19].

С определенными оговорками приведенное здесь определение релевантности можно отнести к ее третьей компоненте, т.е. пертинентности. Очевидно, что пертинентность носит явно субъективный характер. Частная ее оценка отражает индивидуальную реакцию каждого пользователя на данную информационную ситуацию. Указывая на субъективный характер понятия пертинентности, И. Бар-Хиллел считает "непостижимой" или "дьявольски трудной" задачей разработку функций, которые дали бы полезную меру для оценки смыслового расстояния между темами: "Понятие степени релевантности следует отличать от вероятности быть релевантным, хотя интуиция подсказывает, что между ними могут существовать какие-то плохо определенные связи" [17].

Идя дальше, М. Таубе ставит под сомнение правомочность применения метода экспертных оценок, решительно возражая против того, чтобы на основе субъективного понятия пертинентности строились какие-либо математические модели для оценки эффективности информационного поиска.

Аналогичные рассуждения приводят в логический тупик, единственным выходом из которого является признание наличия кантианских трансцендентных "вещей в себе". Классики диалектической теории познания подвергли резкой критике агностицизм, отрицающий возможность познания объективной реальности, возможность научных методов познания окружающего нас мира. "Если из опасения заблуждаться, – пишет в известном труде "Феноменология духа" Гегель, – проникаются недоверием к науке, которая, не впадая в подобного рода мнительность, прямо берется за работу и действительно познает, то неясно, почему бы не проникнуться, наоборот, недоверием к самому недоверию и почему бы не испытать опасения, что сама боязнь заблуждаться есть уже заблуждение" [6].

Действительно, давая себе отчет о присутствии субъективного компонента в понятии релевантности, целый ряд авторов, однако, полагает, что при умелом использовании из него можно извлечь определенную пользу. Л. Дойл, например, считает, что релевантность – это умственный костыль, с которым мы, хотя и не строго, можем мыслить о проблеме информационного поиска, а иначе мы совсем бы не смогли мыслить об этом [18]. Свидетельством тому служат многочисленные теоретико-экспериментальные исследования отечественных и зарубежных специалистов, широко применяющих статистические методы обработки данных при анализе психологических аспектов понятия релевантности-пертинентности.

Метод экспертных оценок в настоящее время широко применяется для вычисления параметров, на основе которых можно прогнозировать полезность привлечения индивидуума к процессам обработки информации. Удается выделять типологические группы индексаторов с присущими им особенностями склада мышления, стиля работы и т.д. Экспериментальные исследования показали принципиальную возможность изучения психологической природы индексирования и личности индексатора (внутренние факторы) путем анализа результатов индексирования, полученных при варьировании внешних факторов.

Для более точного и целенаправленного информационного обслуживания пользователей важное значение приобретают вопросы их подразделения на категории, типологии пользователей.

Разработаны множественно-статистические алгоритмы вычисления "меры близости" экспертов. Предлагаются способы ее увеличения путем взаимно согласованного (дискуссионного) принятия решения о степени релевантности (пертинентности) найденной информации поставленному вопросу.

Исходя из вышеизложенного мы еще раз убеждаемся в целесообразности раздельного рассмотрения понятия релевантности в соответствии с тремя ее различными определениями. Будучи предназначенными для описания единого понятия релевантности, эти определения существенно различны в гносеологическом плане их рассмотрения.

Формулировка каждым индивидуумом своей информационной потребности в виде запроса и оценка меры соответствия найденной информации своей информационной потребности носят субъективный характер и относятся к третьему определению релевантности. Субъективный фактор присутствует также при оценке каждым экспертом смыслового расстояния заданной пары документ – запрос. В определенной мере такая оценка носит случайный характер. Однако эта случайность является лишь формой проявления действительной, истинной меры смыслового соответствия заданной пары документ – запрос, относящейся ко второму определению релевантности. При заданной паре документ – запрос мера действительной релевантности вычисляется как среднее значение экспертных оценок. В процессе усреднения случайные составляющие, носящие субъективный характер, взаимно компенсируются, уравновешиваются. В результате этого среднее значение высвобождается от субъективного компонента, как бы объективируясь, и отражает истинную, объективную меру смыслового расстояния заданной пары документ – запрос. Характер сходимости среднего значения к истинной мере релевантности зависит от ряда факторов: от числа экспертов, от их компетентности в рассматриваемой теме, от их психологической совместимости и др.

Как индивидуальная, так и экспертная оценка смыслового расстояния заданной пары документ – запрос осуществляются на основе сопоставления их текстовых компонентов, т.е. на основе анализа формальной релевантности, или релевантности в соответствии с ее первым определением. Трансформация формальной релевантности через интеллектуальный потенциал экспертных групп или отдельных индивидуумов порождает соответственно меры релевантности, относящиеся ко второму и третьему ее определениям.

Задачей проектировщиков автоматизированных ИПС является имитация с помощью ЭВМ тех интеллектуальных возможностей человека, наличие которых позволяет ему осуществить переход от формальной релевантности к истинной, действительной релевантности. В дальнейшем объектами нашего рассмотрения будут истинная и автоматная (машинная) меры релевантности, где под "автоматной" будем понимать меру релевантности, генерированную ЭВМ на основе анализа формальной релевантности, т.е. на основе сопоставления тех или иных компонентов текстов документов и запросов. При этом ЭВМ оперирует арсеналом логических и лингвистических средств идентификации, совокупность которых принято называть критериями оценки смысловой (семантической) близости документов и запросов.

5.2. МНОЖЕСТВЕННЫЕ МОДЕЛИ ДОКУМЕНТАЛЬНОГО ПОИСКА. ОБЫЧНЫЕ И НЕЧЕТКИЕ ПОДМНОЖЕСТВА РЕЛЕВАНТНОСТИ И ВЫДАЧИ, ИХ ВЕКТОРНЫЕ ПРЕДСТАВЛЕНИЯ

Представим базу данных как множество документов N, состоящее из n элементов $s \in N$ – документов базы данных. Элементы этого множества образуют 2^n различных подмножеств $\beta_i \subset N$ этого множества, в том числе пустое подмножество ϕ, не содержащее ни одного документа, и полное подмножество, равное N. Остальные $2^n - 2$ подмножества, непустые и не равные N, называются собственными подмножествами множества N.

Множество 2^n подмножеств $\beta_i \subset N$ ($i = 1, 2, ..., 2^n$) множества N обозначим через M и в нем определим бинарные операции логического сложения (объединения) и логического умножения:

объединение $\beta_i \cup \beta_j$ есть подмножество всех документов, содержащихся либо в подмножестве β_i, либо в подмножестве β_j, либо и в β_i, и в β_j.

пересечение $\beta_i \cap \beta_j$ есть подмножество всех документов, содержащихся и в β_i, и в β_j.

Множество (класс) M объектов β_i ($i = 1, 2, ..., 2^n$), в котором определены операции логического сложения и логического умножения, называется булевой алгеброй и обладает следующими свойствами [9] для всех β_i:

1) M содержит $\beta_i \cup \beta_j$ и $\beta_i \cap \beta_j$ – замкнутость;

2) $\beta_i \cup \beta_j = \beta_j \cup \beta_i$, $\beta_i \cap \beta_j = \beta_j \cap \beta_i$ – коммутативность;

3) $\beta_i \cup (\beta_j \cup \beta_q) = (\beta_i \cup \beta_j) \cup \beta_q$

$\beta_i \cap (\beta_j \cap \beta_q) = (\beta_i \cap \beta_j) \cap \beta_q$ – ассоциативность;

4) $\beta_i \cap (\beta_j \cup \beta_q) = (\beta_i \cap \beta_j) \cup (\beta_i \cap \beta_q)$,

$\beta_i \cup (\beta_j \cap \beta_q) = (\beta_i \cup \beta_j) \cap (\beta_i \cup \beta_q)$ – дистрибутивность;

5) $\beta_i \cup \beta_i = \beta_i$, $\beta_i \cap \beta_i = \beta_i$ – идемпотентность;

6) $\beta_i \cup \beta_j = \beta_j$ в том и только в том случае, когда

$\beta_i \cap \beta_j = \beta_i$ – совместимость;

7) класс M содержит элементы 1 и ϕ такие, что для всякого элемента из M

$$\beta_i \cup \phi = \beta_i, \quad \beta_i \cap 1 = \beta_i, \quad \beta_i \cap \phi = \phi, \quad \beta_i \cup 1 = 1;$$

8) для каждого элемента β_i класс M содержит элемент $\overline{\beta}_i$ (допол-

нение элемента β_i) такой, что

$$\beta_i \cup \overline{\beta}_i = 1, \quad \beta_i \cap \overline{\beta}_i = \phi.$$

В общем случае каждой конкретной информационной потребности на множестве N соответствует одно из его 2^n возможных подмножеств (в том числе это могут быть пустое или полное подмножества). Подмножество $X \subset N$, соответствующее данной информационной потребности, включает документы, содержание которых соответствует данной информационной потребности. Умение пользователей грамотно сформулировать свою информационную потребность в виде запросов вкупе с арсеналом логико-лингвистических средств АСДП должны обеспечить по возможности большую близость с этим подмножеством подмножества $Y \subset N$ автоматно релевантных документов, т.е. документов, которые АСДП признает соответствующими информационному запросу. Ниже подмножества X и Y будем называть подмножествами соответственно релевантных и выданных (выдача) документов. Совместное рассмотрение пары подмножеств X и Y позволяет выделить на множестве N следующие подмножества:

подмножество релевантных документов, оказавшихся в выдаче,

$$a = X \cap Y; \tag{5.1}$$

подмножество нерелевантных документов, оказавшихся в выдаче,

$$b = \overline{X} \cap Y; \tag{5.2}$$

подмножество релевантных документов, не оказавшихся в выдаче,

$$c = X \cap \overline{Y}; \tag{5.3}$$

подмножество нерелевантных документов, не оказавшихся в выдаче,

$$c = \overline{X} \cap \overline{Y}. \tag{5.4}$$

На рисунке 5.1 приведена матрица сопряженности "релевантность – выдача", где число элементов в подмножествах a, b, c и d обозначено теми же буквами, что и сами эти подмножества.

	релевантные документы X	нерелевантные документы $\overline{X}$
выданные документы Y	a	b
невыданные документы $\overline{Y}$	c	d

Рис. 5.1. Матрица сопряженности "релевантность–выдача"

Представляется естественным подмножества X и Y признать тем более близкими, чем, при прочих равных условиях, большее число доку-

ментов содержится в подмножествах a и d и меньшее – в подмножествах b и c. В пределе, когда при конечных a и d имеет место $b = c = 0$, подмножества X и Y совпадают и поэтому говорят, что имеет место идеальное качество поиска. Вопросы количественной оценки степени близости подмножеств X и Y будут рассмотрены в следующих разделах, а сейчас перейдем к рассмотрению нечетких (размытых) подмножеств множества документов N.

Пусть при оценке степени истинной и/или автоматной релевантности каждого документа информационному запросу шкала возможных значений релевантности не ограничивается двумя значениями (1 – релевантен, 0 – не релевантен), как это имело место при рассмотрении обычных подмножеств X и Y множества N. Тогда мы будем иметь дело с нечеткими (размытыми) подмножествами множества N [8]:

нечетким подмножеством истинной релевантности будем называть нечеткое подмножество $X(\nu)$ множества N с функцией принадлежности $\nu = \nu(i)$, где $\nu(i)$ – мера истинной релевантности i-го документа информационному запросу;

нечетким подмножеством автоматной релевантности будем называть нечеткое подмножество $Y(\lambda)$ множества N с функцией принадлежности $\lambda = \lambda(i)$, где $\lambda(i)$ – автоматная мера соответствия i-го документа информационному запросу.

Иногда нечеткие подмножества $X(\nu)$ и $Y(\nu)$ будем называть проще – нечеткими подмножествами соответственно релевантности и выдачи.

Множество T, состоящее из всех терминов, которые хоть раз встретились в каком-либо документе базы данных, будем называть множеством терминов базы данных. Пусть множество T состоит из m элементов $t \in T$. Тогда, как и в случае множества документов N, 2^m подмножеств множества T образуют булеву алгебру. Так же, как и в случае со множеством N, на множестве T могут быть определены различные нечеткие подмножества с различными функциями принадлежности.

Каждому элементу $s \in N$ множества документов можно поставить в соответствие одно из подмножеств множества T такое, которое состоит из элементов множества T – терминов, хоть раз встретившихся в рассматриваемом документе.

На множестве терминов T можно определить также нечеткие подмножества, соответствующие различным элементам множества документов. Например, каждому документу можно поставить в соответствие нечеткое подмножество множества T с функцией принадлежности $\alpha = \alpha(i)$, равной числу встречаемости i-го термина в рассматриваемом документе.

Аналогично, каждому элементу $t \in T$ множества терминов можно поставить в соответствие одно из 2^n подмножеств множества N такое,

которое состоит из элементов этого множества – документов, содержащих рассматриваемый термин.

На множестве документов N можно определить также нечеткие подмножества, соответствующие различным элементам множества терминов T. Например, каждому термину можно поставить в соответствие нечеткое подмножество множества N с функцией принадлежности $\gamma = \gamma(i)$, равной числу встречаемости рассматриваемого термина в i-м документе.

На этапах как проектирования, так и эксплуатации АСДП часто возникает необходимость в оценке степени близости двух подмножеств (обычных и/или нечетких) одного и того же множества. Для такой оценки в ряде случаев оказывается удобным оперировать векторными представлениями этих подмножеств, т.е. векторами, находящимися во взаимно однозначном соответствии с этими подмножествами [1].

Рассмотрим множество A с q элементами $a \in A$. Пусть $B \subset A$ – некоторое обычное подмножество этого множества и в q-мерном пространстве векторов определен вектор b такой, что значение i-й его координаты равно единице, если i-й элемент множества A принадлежит подмножеству B, и нулю – в противном случае. Тогда можно утверждать, что установлено взаимно однозначное соответствие (отображение) между 2^q подмножествами множества A и 2^q вершинами q-мерного единичного куба. В частности, пустому подмножеству множества A соответствует начало координат, а подмножеству $B = A$ – вершина куба с координатами (1, 1, ..., 1).

Пусть теперь $B(\beta)$ – некоторое нечеткое подмножество множества A с функцией принадлежности $\beta = \beta(i)$, где $\beta(i)$ – мера принадлежности i-го элемента множества A подмножеству $B(\beta)$. В q-мерном пространстве определим вектор b такой, что значение его i-й координаты равно $\beta(i)$. Тогда можно утверждать, что установлено взаимно однозначное соответствие (отображение) между нечеткими подмножествами множества A и точками q-мерного пространства.

Заметим, что обычные подмножества $B \subset A$ являются частными случаями нечетких подмножеств с функцией принадлежности $\beta = \beta(i)$, где $\beta(i)$ равно единице, если i-й элемент множества A принадлежит данному обычному подмножеству, и нулю – в противном случае.

Приведенные здесь отображения позволяют в дальнейшем вместо обычных и размытых подмножеств множества A с q элементами рассматривать соответствующие им q-мерные векторы. Так, при рассмотрении подмножеств множества документов N можно говорить о векторах релевантности и выдачи. Аналогично, при рассмотрении подмножеств множеств терминов T можно говорить о векторах, представляющих различные документы.

В последующем изложении векторы, представляющие обычные подмножества, иногда будем называть бинарными.

5.3. ЭНТРОПИЙНАЯ МОДЕЛЬ ДОКУМЕНТАЛЬНОГО ПОИСКА

В третьей главе мы убедились в эффективности использования аппарата статистической теории информации и, в частности, понятия энтропии при анализе работы каналов связи. Последние, будучи несемантическими компонентами сети, призваны объединить звенья сети – документальные системы, которые имеют отчетливую семантическую природу. В рассматриваемом смысле представляется заманчивым разработать энтропийную модель документального поиска, с тем, чтобы математический анализ сети в целом (включая ее звенья семантической природы) осуществить в рамках единой статистической теории информации.

В ходе разработки энтропийной модели документального поиска мы попытаемся установить границы возможных аналогий между процесом передачи информации по шумящим каналам связи и документальным поиском, максимально придерживаясь обозначений, которыми оперировали в третьей главе.

Несмотря на то, что проблема передачи информации по каналам связи тесно переплетается с проблемой ее оптимального кодирования, представляется возможным раздельное рассмотрение задач, связанных с оптимальным кодированием и передачей уже закодированной информации.

Из совместного рассмотрения матриц сопряженности "вход-выход" и "релевантность-выдача", приведенных на рис. 3.1 и 5.1, легко обнаружить практически полное их совпадение. Это наводит на мысль, что для описания работы АСДП можно использовать разработанный К. Шенноном математический аппарат статистической теории информации. Ниже мы убедимся в адекватности математического аппарата статистической теории информации для описания документального поиска хотя бы в поведенческом плане его рассмотрения. Если отвлечься от конкретных технических, логико-лингвистических и программных средств реализации АСДП, то, как и при рассмотрении каналов связи, анализ АСДП можно свести к анализу работы некоего "черного ящика", который на каждый двоичный символ, поданный к его входу, "отвечает" соответствующим выходным двоичным символом. Входные и выходные двоичные символы мы будем интерпретировать следующим образом [2]:

входная единица будет интерпретирована как релевантный документ, поданный на вход АСДП для анализа его релевантности;

входной нуль будет интерпретирован как нерелевантный документ, поданный на вход АСДП для анализа его релевантности;

выходная единица будет интерпретирована как документ, признанный АСДП релевантным, независимо от того, является ли на самом деле этот документ релевантным или нет;

выходной нуль будет интепретирован как документ, признанный АСДП нерелевантным, независимо от того, является ли на самом деле этот документ релевантным или нет.

Если рассматриваемый нами "черный ящик" таков, что в ответ на каждый поданный к его входу двоичный символ отвечает тем же символом, то он описывает идеальную систему документального поиска. Если же рассматриваемый "черный ящик" является идеальным инвертором, то он описывает систему, которая все релевантные документы признает нерелевантными, и, наоборот, все нерелевантные – релевантными.

В общем же случае о характере работы АСДП можно судить путем анализа матрицы сопряженности "релевантность-выдача", рассматривая следующие значения вероятностей, характеризующие ансамбль случайных величин X и Y (см. рис. 5.1):

вероятность того, что наугад взятый из входного потока документ окажется релевантным (в литературе по информатике называется коэффициентом релевантности базы данных поставленному запросу), –

$$p(x = 1) = \omega = (a + c)/n; \tag{5.5}$$

вероятность того, что наугад взятый документ, поданный на вход АСДП, будет признан системой релевантным (в литературе по информатике иногда называется коэффициентом выдачи), –

$$p(y = 1) = \lambda = (a + b)/n; \tag{5.6}$$

вероятность того, что поданный на вход АСДП релевантный документ будет признан системой релевантным (в литературе по информатике называется коэффициентом полноты поиска), –

$$p(y = 1/x = 1) = \lambda_1 = a/(a + c); \tag{5.7}$$

вероятность того, что поданный на вход АСДП нерелевантный документ будет признан системой как нерелевантный (в литературе по информатике называется коэффициентом специфичности), –

$$p(y = 0/x = 0) = \lambda_2 = d/(d + b); \tag{5.8}$$

вероятность того, что документ, признанный системой релевантным, на самом деле окажется релевантным (в литературе по информатике называется коэффициентом точности), –

$$p(x = 1/y = 1) = \omega_1 = a/(a + b); \tag{5.9}$$

вероятность того, что документ, признанный системой нерелевантным, на самом деле окажется нерелевантным, –

$$p(x = 0/y = 0) = \omega_2 = d/(c + d). \tag{5.10}$$

Естественно, что при конечных значениях a, b, c и d речь может идти не о самих значениях соответствующих вероятностей, а о тех или иных их статистических оценках. Поэтому далее будем считать, что

величины $a + b$, $a + c$, $b + d$ и $d + c$ столь велики, что значения (5.5)–(5.10) можно отождествлять со значениями соответствующих вероятностей.

Далее, пользуясь теми же формулами, что и в третьей главе, а именно, формулами (3.22), (3.24), (3.27), на основе элементов a, b, c и d матрицы сопряженности "релевантность – выдача" определим значения различных энтропий, которые будем интерпретировать в русле документального поиска, например:

$H[x]$ интерпретируется как "проблематичность" угадывания того, является ли наугад взятый из исходного потока документ релевантным или нет;

$H[x/y]$ – условная (остаточная) энтропия угадывания того, каким (релевантным или нерелевантным) является очередной документ входного потока, если известен результат оценки его релевантности системой;

$I[x, y]$ – количество информации о том, является ли очередной документ входного потока релевантным или нет, содержащееся в среднем в одном сообщении о том, каким (релевантным или нерелевантным) признан данный документ системой.

В случае, когда

$$\lambda_1 = \lambda_2 = 1, \tag{5.11}$$

имеют место

$$H[x/y] = 0, \quad I[x, y] = H[x] - H[x/y] = H[x], \tag{5.11a}$$

т.е. количества $I[x, y]$ информации, содержащегося в среднем в одном сообщении о том, каким признан системой очередной документ, полностью хватает для "снятия" исходной, допоисковой неопределенности $H[x]$ о том, является ли на самом деле этот документ релевантным или нет поставленному запросу. Именно, если документ признан системой релевантным, то он и на самом деле является релевантным, и наоборот, если система данный документ признала нерелевантным, то и на самом деле этот документ нерелевантный.

Аналогичную картину мы наблюдаем при

$$\lambda_1 = \lambda_2 = 0, \tag{5.12}$$

когда также имеют место (5.11а), с той лишь разницей, что в этом случае признание системой очередного документа релевантным или нерелевантным дает полную гарантию того, что на самом деле этот документ окажется соответственно нерелевантным или релевантным.

Таким образом, если в случае (5.11) АСДП является идеальной поисковой системой, то в случае (5.12) мы имеем дело лишь с идеальным инвертором. В этом случае мы легко можем достичь идеального результата поиска, если после работы АСДП подмножества выдачи и невыдачи поменять местами.

В случае, когда

$$\lambda_1 + \lambda_2 = 1, \tag{5.13}$$

имеют место формулы

$$H[x/y] = H[x], \quad I[x, y] = H[x] - H[x/y] = 0, \tag{5.14}$$

т.е. в сообщениях о признании или непризнании системой данного документа релевантным не содержится никакой информации о том, каким является данный документ на самом деле. Работа АСДП эквивалентна случайной выборке. Естественно, что такая АСДП не может быть пригодна для документального поиска.

Наряду с $I[x, y]$ работу АСДП можно характеризовать ее коэффициентом относительно уменьшения исходной неопределенности или проще – коэффициентом проводимости (см. главу 3):

$$\varkappa[x, y] = I[x, y]/H[x]. \tag{5.15}$$

Как и при рассмотрении каналов связи, при заданных λ_1 и λ_2 значения $I[x, y]$ и $\varkappa[x, y]$ достигают своих наибольших возможных значений при значениях ω, равных ω_I и $\omega_\varkappa$ соответственно. Естественно, что при рассмотрении АСДП остаются в силе все формулы, полученные в главе 3.

Заметим, что при анализе энтропийных моделей документального поиска задачи согласования входа, рассмотренные в главе 3, представляют лишь теоретический интерес, так как варьировать значения ω в данном случае мы не можем – они зависят от конкретных запросов пользователей.

С другой стороны, при анализе энтропийных моделей документального поиска чрезвычайно актуальной становится задача настройки информационно-поисковых систем, т.е. задача подбора на кривой $\lambda_1 = \lambda_1(\lambda_2)$ рабочей точки, обеспечивающей наибольшее возможное значение $I[x, y]$. В качестве параметра настройки t (см. формулы (3.50)–(3.51)) здесь могут быть использованы глубина индексирования запросов (при работе, например, с классификационными информационно-поисковыми языками), глубина терминологического наращивания запросов с помощью дескрипторных словарей, тезаурусов и др.

5.4. КОРРЕЛЯЦИОННАЯ МОДЕЛЬ ДОКУМЕНТАЛЬНОГО ПОИСКА

В разделе 5.2 мы уже говорили о возможности векторного представления как обычных, так и нечетких подмножеств релевантности и выдачи. Как и при рассмотрении энтропийной модели, в этом разделе мы, не вдаваясь в подробности логико-лингвистических и программных средств реализации информационного поиска, будем рассматривать АСДП как некий "черный ящик", который в ответ на каждый

поданный к его входу вектор X "отвечает" соответствующим выходным вектором Y. Если рассматриваемый нами "черный ящик" таков, что в ответ на каждый поданный к его входу вектор X выдает вектор Y, допускающий представление в виде

$$Y = \alpha X + \beta E, \tag{5.16}$$

где $\alpha > 0$ и β произвольные (действительные) скаляры, а через E обозначен вектор (1, 1,..., 1), то будем говорить, что имеем дело с идеальной АСДП. При рассмотрении реальных АСДП условие (5.16) обычно нарушается и нашей задачей будет оценить степень расхождения работы реальной и идеальной АСДП.

Рассмотрим *операции центрирования и нормирования векторов* [1].

Под центрированием вектора $A(A_1, A_2,..., A_n)$ будем понимать замену вектора A вектором $A_0(A_{01}, A_{02},..., A_{0n})$, где

$$A_{0i} = A_i - m_A, \tag{5.17}$$

$$m_A = \frac{1}{n}\sum_{i=1}^{n} A_i. \tag{5.18}$$

Очевидно, имеет место

$$A_0 = A - m_A E.$$

Под нормированием вектора $A(A_1, A_2,..., A_n)$ будем понимать замену этого вектора вектором γA, где

$$\gamma = 1 \Big/ \sqrt{\sum_{i=1}^{n} A_i^2}\,. \tag{5.19}$$

Очевидно, нулевой вектор операции нормирования не подлежит.

Из (5.17) ÷ (5.19) легко заметить, что вектор A_0 не что иное, как векторная проекция вектора A на гиперплоскость, перпендикулярную вектору E. Скалярное произведение A_0E произвольного центрированного вектора A_0 на вектор E равно нулю. Модуль произвольного нормированного вектора равен единице.

Обозначим через a вектор, полученный из вектора A путем последовательного применения к нему операций центрирования и нормирования. Геометрически последовательное применение к вектору A операций центрирования и нормирования сводится к отображению n-мерного пространства векторов на поверхность сферы с центром в начале координат и радиусом, равным единице. Радиус-вектор произвольной точки этой сферы перпендикулярен вектору E. Очевидно, значение i-й координаты вектора a окажется равным

$$a_i = \frac{A_i - m_A}{\sqrt{\sum_{i=1}^{n} (A_i - m_A)^2}}, \tag{5.20}$$

где значение m_A определяется по формуле (5.18). Также очевидно, что

скалярное произведение aE всегда равно нулю, а модуль вектора a всегда равен единице.

Из (5.20) легко заметить, что если последовательное применение к вектору A операций центрирования и нормирования отображает его в точку x, то произвольный вектор B, который можно представить как

$$B = \alpha A + \beta E \qquad (\alpha > 0), \tag{5.16а}$$

также отобразится в точку x.

Последовательное применение операций центрирования и нормирования к паре векторов X и Y, где X – вектор релевантности, а Y – вектор выдачи, приводит к паре векторов соответственно x и y с координатами, равными

$$x_i = \frac{X_i - m_x}{\sqrt{\sum_{i=1}^{n}(X_i - m_x)}}; \qquad y_i = \frac{Y_i - m_y}{\sqrt{\sum_{i=1}^{n}(Y_i - m_y)}} \tag{5.21}$$

Скалярное произведение векторов x и y, очевидно, равно

$$xy = \frac{\sum_{i=1}^{n}(X_i - m_x)(Y_i - m_y)}{\sqrt{\sum_{i=1}^{n}(X_i - m_x)^2}\sqrt{\sum_{i=1}^{n}(Y_i - m_y)^2}}. \tag{5.22}$$

Рассмотрим систему случайных величин x и y и обозначим через X_i и Y_i конкретные значения, которые принимают случайные величины x и y в i-м эксперименте из n независимых экспериментов, проведенных в одинаковых условиях. Для оценки степени связанности случайных величин x и y обычно оперируют коэффициентом линейной корреляции между этими случайными величинами,

$$r_{xy} = \frac{R_{xy}}{\sqrt{D_x}\sqrt{D_y}} = \frac{R_{xy}}{\sigma_x \sigma_y}, \tag{5.23}$$

где

$m_x = M[X_i]$ – математическое ожидание случайной величины x;

$R_{xy} = M[(X_i - m_x)(Y_i - m_y)]$ – корреляционный момент или момент связи между случайными величинами x и y;

$D_x = R_{xx}$ – дисперсия случайной величины x;

$\sigma_x = \sqrt{D_x}$ – среднее квадратичное отклонение случайной величины x.

Для статистической оценки значения коэффициента линейной корреляции r_{xy} между случайными величинами x и y на основе данных n независимых экспериментов, проведенных в одинаковых условиях, поль-

зуются формулой

$$r_{xy} = \frac{\sum_{i=1}^{n} (X_i - m_x)(Y_i - m_y)}{\sqrt{\sum_{i=1}^{n} (X_i - m_x)^2} \sqrt{\sum_{i=1}^{n} (Y_i - m_y)^2}}, \qquad (5.22a)$$

где m_x – статистическая оценка математического ожидания случайной величины x. Ее значение определяется по формуле (5.18).

Естественно, что формула (5.22а) имеет смысл лишь при отличных от нуля значениях D_x и D_y, т.е. при соблюдении условий

$$\sum_{i=1}^{n} (X_i - m_x)^2 \neq 0 \quad \text{и} \quad \sum_{i=1}^{n} (Y_i - m_y)^2 \neq 0.$$

Формулы (5.22) и (5.22а) совпадают, т.е. если отождествлять меры истинной и автоматной релевантности i–го документа с конкретными значениями случайных величин x и y (далее их будем называть случайными величинами релевантности и выдачи) в i-м эксперименте, общее число экспериментов отождествлять с числом n документов в базе данных, а под "одинаковыми условиями" понимать "проведение экспериментов в рамках одной и той же АСДП применительно к фиксированному запросу", то в качестве степени связанности (коэффициента линейной корреляции) случайных величин релевантности и выдачи x и y можно использовать формулу (5.22) скалярного произведения векторов x и y.

Можно показать (см., например, [10]), что значения r_{xy} ограничены интервалом

$$-1 \leqslant r_{xy} \leqslant 1, \qquad (5.24)$$

причем $|r_{xy}| = 1$ тогда и только тогда, когда векторы X и Y связаны формулой

$$Y = \alpha X + \beta E, \qquad (5.25)$$

где α и β произвольные скаляры. Заметим, что при этом имеет место

$$r_{xy} = \operatorname{sgn}(\alpha). \qquad (5.26)$$

Легко заметить также, что в общем случае $r_{xy} = r_{yx}$, а при наличии связи (5.25) имеет место

$$r_{zx} = r_{zy} \operatorname{sgn}(\alpha). \qquad (5.27)$$

Подавляющее большинство промышленных АСДП работают с обычными подмножествами множества N, т.е. имеют дело с бинарными векторами X и Y, координаты которых могут принимать одно из двух значений, а именно, нуль или единица. При этом, как мы в этом неоднократно убедились выше, исчерпывающей характеристикой работы

АСДП может служить матрица сопряженности "релевантность – выдача", представленная на рис. 5.1. Действительно, пользуясь формулой (5.22), путем несложных преобразований можно показать, что в частном случае, когда речь идет о простых подмножествах релевантности и выдачи (X и Y), т.е. когда речь идет о бинарных векторах, значение r_{xy} зависит от элементов матрицы сопряженности "релевантность – выдача" по формуле [2]:

$$r_{xy} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}. \qquad (5.28)$$

Очевидно, эта формула имеет смысл лишь при соблюдении условий $a + b \neq 0$, $a + c \neq 0$, $b + d \neq 0$ и $c + d \neq 0$.

Обратим внимание на то, что знак r_{xy} совпадает со знаком определителя (детерминанта) матрицы сопряженности, причем $r_{xy} = 0$ тогда и только тогда, когда значение определителя равно нулю, т.е. когда

$$ad = bc, \qquad (5.29)$$

или, что то же самое,

$$a/(a + b) = (a + c)/n. \qquad (5.29а)$$

Равенство нулю определителя матрицы свидетельствует о том, что работа АСДП эквивалентна случайной выборке, когда концентрация, доля релевантных документов в выдаче оказывается равной концентрации релевантных документов в исходном множестве N документов. При этом эффект присутствия АСДП равен нулю.

Отрицательные значения определителя свидетельствуют о том, что концентрация релевантных документов в выдаче меньше, чем их концентрация в исходном множестве документов. Эффект присутствия АСДП отрицательный, АСДП оказывает нам "медвежью услугу".

Лишь положительные значения определителя свидетельствуют о положительном эффекте присутствия АСДП, так как лишь при этом концентрация релевантных документов в выдаче оказывается большей, чем их концентрация в исходном множестве.

Пользуясь (5.28), легко убедиться в справедливости следующих равенств

$$\begin{aligned} \partial r_{xy}/\partial a \geq 0, \qquad & \partial r_{xy}/\partial c \leq 0, \\ \partial r_{xy}/\partial b \leq 0, \qquad & \partial r_{xy}/\partial d \geq 0. \end{aligned} \qquad (5.30)$$

Легко обнаружить также, что r_{xy} достигает своего наибольшего возможного значения $r_{xy} = 1$ (идеальный поиск), когда подмножества X и Y совпадают и минимально возможного $r_{xy} = -1$ (идеальный инвертор), когда подмножества X и Y дополняют друг друга до исходного (универсального) множества N. Во всех остальных случаях имеет место $|r_{xy}| < 1$.

Из (5.28) легко установить, что имеют место

$$r_{xy} = r_{\bar{x}\bar{y}} = -r_{x\bar{y}}. \qquad (5.31)$$

Переходя в (5.28) к переменным $\bar{a} = a / n$, $\bar{b} = b / n$, $\bar{c} = c / n$ и $\bar{d} = d / n$ и вспомнив (см. главу 3), что из этих переменных только три являются независимыми, нетрудно убедиться в принципиальной возможности выразить значение r_{xy} через произвольные три независимые переменные из ω, ω_1, ω_2, λ, λ_1, λ_2 (связь этих переменных с элементами матрицы сопряженности приведена, например, в разделе 5.3).

Из (5.28), например, нетрудно вывести формулу

$$r_{xy} = (\lambda_1,\ \lambda_2,\ \omega) = \frac{(\lambda_1 + \lambda_2 - 1)\sqrt{\omega(1-\omega)}}{\sqrt{[\lambda_2 - \omega(\lambda_1 + \lambda_2 - 1)][1 - \lambda_2 + \omega(\lambda_1 + \lambda_2 - 1)]}}, \qquad (5.32)$$

где знак выражения $\lambda_1 + \lambda_2 - 1$ совпадает со знаком определителя матрицы сопряженности. В реально функционирующих промышленных АСДП имеет место $\lambda_1 + \lambda_2 - 1 > 0$,и поэтому в дальнейшем именно эти случаи и будут предметом нашего рассмотрения. Из (5.32) следует, что при фиксированных характеристиках $\lambda_1 = a/(a + c)$ и $\lambda_2 = d/(d + b)$ собственно АСДП значение r_{xy} зависит еще и от характеристики среды, а именно от $\omega = (a + c)/n$. Можно показать, например, что при заданных характеристиках системы λ_1 и λ_2 ($\lambda_1 + \lambda_2 \neq 1$) существует единственное решение уравнения $\partial r_{xy}/\partial\omega = 0$, а именно [2]:

$$\omega = \omega_r = \frac{\sqrt{\lambda_2(1-\lambda_2)}}{\sqrt{\lambda_2(1-\lambda_2)} + \sqrt{\lambda_1(1-\lambda_1)}}, \qquad (5.33)$$

при котором значение $r\ (\lambda_1, \lambda_2, \omega)$ достигает своего экстремального значения, равного

$$r(\lambda_1,\ \lambda_2,\ \omega_r) = \sqrt{\lambda_1\lambda_2} - \sqrt{(1-\lambda_1)(1-\lambda_2)}. \qquad (5.34)$$

Эта величина больше нуля и равна максимуму $r\ (\lambda_1, \lambda_2, \omega)$, если $\lambda_1 + \lambda_2 > 1$. Если же $\lambda_1 + \lambda_2 < 1$, то эта же величина меньше нуля и равна минимуму $r\ (\lambda_1, \lambda_2, \omega)$. Иными словами, независимо от того, больше или меньше нуля выражение $\lambda_1 + \lambda_2 - 1$, абсолютное значение r_{xy} достигает своего максимума при $\omega = \omega_r$. При заданных характеристиках АСДП λ_1 и λ_2 ($\lambda_1 + \lambda_2 \neq 1$) поисковую среду с $\omega = \omega_r$ будем называть согласованной с данной АСДП средой. Очевидно, при работе АСДП с согласованной средой поисковый эффект от ее работы получается наибольшим (разумеется, если $\lambda_1 + \lambda_2 > 1$).

Особый интерес представляет зависимость значения r_{xy} от параметров λ_1, λ_2, и ω_1. Напомним, что в литературе по информатике эти параметры называются коэффициентами полноты, специфичности и точности, и ряд специалистов для оценки качества работы АСДП

пользуется именно этими параметрами [11]. Из (5.28) нетрудно вывести формулу:

$$r(\lambda_1,\ \lambda_2,\ \omega_1) = \frac{(\lambda_1+\lambda_2-1)\sqrt{\omega_1(1-\omega_1)}}{\sqrt{\lambda_1\lambda_2-\omega_1(\lambda_1+\lambda_2-1)}}. \qquad (5.35)$$

Как и в предыдущем случае, примем, что характеристики собственно АСДП, а именно λ_1 и λ_2 ($\lambda_1 + \lambda_2 \neq 1$) нам заданы, и проанализируем зависимость $r(\lambda_1, \lambda_2, \omega_1)$ от параметра ω_1. Используя (5.35), легко показать, что при этом существует единственное решение уравнения $\partial r_{xy}/\partial\omega_1 = 0$, а именно:

$$\omega_1 = \omega_{1r} = \frac{\sqrt{\lambda_1\lambda_2}}{\sqrt{\lambda_1\lambda_2}+\sqrt{(1-\lambda_1)(1-\lambda_2)}}, \qquad (5.36)$$

при котором значение $r(\lambda_1, \lambda_2, \omega_1)$ достигает своего экстремального значения, равного

$$r(\lambda_1,\ \lambda_2,\ \omega_{1r}) = \sqrt{\lambda_1\lambda_2} - \sqrt{(1-\lambda_1)(1-\lambda_2)}. \qquad (5.34а)$$

Как и при рассмотрении $r(\lambda_1, \lambda_2, \omega)$, эта величина больше нуля и равна максимуму $r(\lambda_1, \lambda_2, \omega_1)$, если $\lambda_1 + \lambda_2 > 1$, и меньше нуля и равна минимуму $r(\lambda_1, \lambda_2, \omega_1)$, если $\lambda_1 + \lambda_2 < 1$.

Рассмотрим численный пример.

Пусть сопоставляются результаты трех экспериментов, при проведении которых были определены значения коэффициентов полноты (λ_1), специфичности (λ_2), и точности (ω_1). Определим на их основе значения r_{xy}, ω_{1r} и r_{max}.

Таблица 5.1. Результаты экспериментов по определению характеристик работы АСДП

№ п/п	Параметр		Эксперимент		
	обозн.	название	I	II	III
1	λ_1	коэф. полноты	0,60	0,65	0,65
2	λ_2	коэф. специфичности	0,80	0,85	0,85
3	ω_1	коэф. точности	0,71	0,97	0,76
4	r_{xy}	коэф. корреляции	0,41	0,34	0,57
5	ω_{1r}	значение ω_1, при котором достигается максимально возможное значение коэф. корреляции	0,71	0,76	0,76
6	r_{max}	максимально возможное значение коэф. корреляции	0,41	0,57	0,57

Сначала сравним результаты первого и третьего экспериментов. В обоих этих экспериментах характеристики ω поисковых сред оказались согласованными с характеристиками собственно АСДП, т.е. в обоих

экспериментах имели место $\omega_1 = \omega_{1r}$. Это привело к тому, что реальные значения $r(\lambda_1, \lambda_2, \omega_1)$ оказались равными их максимальным возможным значениям $r(\lambda_1, \lambda_2, \omega_1)$, т.е. в обоих случаях имели место $r(\lambda_1, \lambda_2, \omega_1) = r(\lambda_1, \lambda_2, \omega_{1r})$. Представляется вполне естественным, что значение коэффициента корреляции в третьем эксперименте оказалось выше значения этого коэффициента в первом эксперименте. Ведь коэффициенты полноты, специфичности и точности в третьем эксперименте оказались большими по сравнению с этими коэффициентами в первом эксперименте.

А теперь сравним результаты второго и третьего экспериментов. В этих экспериментах характеристики собственно АСДП (λ_1 и λ_2) оказались одинаковыми, но значение коэффициента точности во втором эксперименте ($\omega_1 = 0{,}97$) оказалось больше согласованного значения ($\omega_{1r} = 0{,}76$), которое имело место в третьем эксперименте. Казалось бы, значение итоговой оценки эффективности для второго эксперимента должно было получиться выше этого значения в третьем эксперименте. На самом же деле значение коэффициента корреляции $r(\lambda_1, \lambda_2, \omega_1)$ во втором эксперименте (0,34) оказалось меньше, чем в третьем (0,57).

Это обстоятельство еще раз свидетельствует о неправомочности использования значения коэффициента точности (ω_1) в качестве параметра, характеризующего АСДП. При фиксированных значениях λ_1 и λ_2 значение этого параметра является лишь своего рода опосредованной оценкой поисковой среды. Что же касается коэффициента корреляции, то он характеризует качество работы АСДП и при фиксированных значениях λ_1 и λ_2 зависит от степени согласованности поисковой среды (параметр ω) с характеристиками λ_1 и λ_2 собственно АСДП.

Из (5.28) легко установить, что если при конечных a, b и c параметр d стремится к бесконечности, то значение r_{xy} приближенно можно определить по формуле

$$r_{xy} \approx \frac{a}{\sqrt{(a+b)(a+c)}} = \sqrt{\omega_1 \lambda_1}. \qquad (5.37)$$

Именно к такому результату мы пришли бы, если бы вычислили значение косинуса угла между векторами X и Y:

$$\cos(X \hat{} Y) = \frac{X \cdot Y}{|X| \cdot |Y|}. \qquad (5.38)$$

Из сопоставления формул (5.37) и (5.38) легко заметить, что в общем случае, когда значение d соизмеримо со значениями a, b и c, использование критерия (5.38), весьма популярного среди проектировщиков АСДП, нельзя признать корректным, хотя бы потому, что здесь не учитываются реальные значения величины d.

Нельзя признать корректным также использование для оценки эффективности работы АСДП в качестве критерия скалярного произведения векторов X и Y, так как при рассмотрении, например, бинарных случаев величина этого критерия равна a, что, естественно, не может считаться приемлемым. Это обстоятельство следует особо подчеркнуть, так как при построении различных матричных моделей АСДП ряд авторов оперирует обычной операцией умножения матриц, в основе которой лежит операция скалярного произведения векторов. Тем самым, хотя и неявно, но допускается грубая ошибка, что естественно ставит под сомнение адекватность этих моделей исследуемым объектам [14].

В заключение настоящего раздела отметим, что как и при рассмотрении энтропийных моделей, при рассмотрении корреляционных моделей АСДП остаются актуальными задачами их настройки, подробный анализ которых можно найти в [2].

5.5. СВЯЗЬ МЕЖДУ ПАРАМЕТРАМИ, ХАРАКТЕРИЗУЮЩИМИ ЭНТРОПИЙНУЮ И КОРРЕЛЯЦИОННУЮ МОДЕЛИ (БИНАРНЫЙ СЛУЧАЙ)

В [2] энтропийная и корреляционная модели документального поиска рассматриваются в самых общих постановках. В рамках же настоящего пособия мы в основном ограничились рассмотрением частного, бинарного случая, который характеризуется простотой реализации и получил наибольшее распространение. Бинарные случаи характеризуются еще и тем, что при их рассмотрении удается установить определенные связи между параметрами энтропийной и корреляционной моделей.

При рассмотрении энтропийной модели документального поиска применительно к бинарному случаю центральную роль играет представленная на рис. 5.2 зависимость энтропии от вероятности появления одного из двух возможных событий

$$H(p) = -p \log_2 p - (1-p) \log_2 (1-p). \quad (5.39)$$

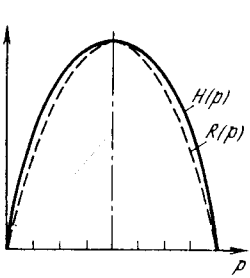

Рис. 5.2. Замена функции К. Шеннона H(p) (сплошная линия) параболической зависимостью R(p) (пунктирная линия)

На рис. 5.2 пунктиром представлена также зависимость

$$R(p) = 4p(1-p), \quad (5.40)$$

которая в области $0 \leq p \leq 1$ отличается от зависимости (5.39) лишь незначительно и поэтому может рассматриваться как некое (параболическое) приближение к функции $H(p)$.

При рассмотрении энтропийной моде-

ли (бинарный случай) основными параметрами являются (см. формулы (3.22), (3.24), (3.27) и (3.40):

$$H\ [x] = H\ (p), \tag{5.41}$$

$$H[x\,/\,y] = \lambda H(\omega_1) + (1-\lambda)H(\omega_2), \tag{5.42}$$

$$I[x,\ y] = H[x] - H[x\,/\,y], \tag{5.43}$$

$$\varkappa[x,\ y] = I[x,\ y]\,/\,H[x]. \tag{5.44}$$

Заменив в формулах (5.41) ÷ (5.44) значения функции H значениями функции R с теми же аргументами и введя соответствующие обозначения, получим

$$R\ [x] = R\ (p), \tag{5.45}$$

$$R[x\,/\,y] = \lambda R(\omega_1) + (1-\lambda)R(\omega_2), \tag{5.46}$$

$$I_R[x,\ y] = R[x] - R[x\,/\,y], \tag{5.47}$$

$$\varkappa_R\ [x,\ y] = I_R[x,\ y]\,/\,R[x]. \tag{5.48}$$

Выразив в (5.45) ÷ (5.48) значения ω, ω_1, ω_2 и λ через соответствующие элементы матрицы сопряженности, получим:

$$R[x] = 4(a+c)(b+d)\,/\,n^2, \tag{5.49}$$

$$R[x\,/\,y] = 4(ab\,/\,(a+b) + cd\,/\,(c+d))\,/\,n, \tag{5.50}$$

$$I_R[x,\ y] = 4((a+c)(b+d) - abn\,/\,(a+b) - cdn\,/\,(c+d)), \tag{5.51}$$

$$\varkappa_R[x,\ y] = \frac{(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}. \tag{5.52}$$

Сравнивая (5.52) с (5.28), легко обнаружить, что величина $\varkappa_R[x, y]$ не что иное, как квадрат коэффициента линейной корреляции. Иными словами, величина $r_{xy}^2 = \varkappa_R[x,\ y]$, нормированая, как и $\varkappa[x, y]$, в интервале $0 \leqslant r_{xy}^2 \leqslant 1$, может рассматриваться как некое приближение к этому коэффициенту, названному в разделе 5.3 коэффициентом проводимости АСДП.

Аналогично, величина $I_R[x, y]$ может рассматриваться как некое приближение к значению $I[x, y]$ – количеству информации, полученной в результате информационного поиска.

Таким образом, замена функции H функцией R позволяет установить соответствующие аналогии между параметрами, характеризующими энтропийную и корреляционную модели АСДП. Совместный анализ

этих параметров оказывается особенно плодотворным при рассмотрении ряда оптимизационных задач, когда прямое вычисление некоторых величин не удается и приходится ограничиться лишь их приближенными оценками. При этом удается с единой точки зрения интерпретировать параметры, характеризующие энтропийную и корреляционную модели, а иногда и установить между ними прямые аналитические (не приближенные) связи.

Рассмотрим пример.

В третьей главе мы показали, что величина $I[x, y]$ может быть представлена как функция трех аргументов

$$I[x,\ y] = f_I(\lambda_1,\ \lambda_2,\ \omega).$$

Пусть теперь нас интересует характер зависимости величины $I[x, y]$ от аргумента ω при фиксированных значениях λ_1 и λ_2. При исследовании этой зависимости удается установить, что справедлива формула

$$\partial^2 I[x,\ y] / \partial\omega^2 = r_{xy}^2 d^2 H / d\omega^2, \tag{5.53}$$

но поскольку

$$d^2 H / d\omega^2 = -\log_2 e / \omega(1-\omega) < 0, \tag{5.54}$$

то из (5.53) непосредственно следует выпуклость зависимости $I[x, y]$ от аргумента ω, т.е. единственность решения уравнения

$$\partial I[x,\ y] / \partial\omega = 0.$$

Более подробный анализ связей между параметрами, характеризующими энтропийную и корреляционную модели, можно найти в [2].

5.6. МАТРИЧНЫЕ МОДЕЛИ ДОКУМЕНТАЛЬНОГО ПОИСКА

Первые публикации по матричным моделям документального поиска восходят к шестидесятым годам нашего столетия. Повышенный интерес к разработке и совершенствованию этих моделей отчасти был обусловлен появлением возможности создания специализированных матричных процессоров, способных выполнять матричные операции существенно быстрее обычных процессоров общего назначения.

Пусть рассматривается некоторое множество из n документов. На основе этого множества можно построить множество всех m терминов, которые хоть раз встречались в каком-либо одном или более документах. При наличии этих двух множеств можно говорить по крайней мере о трех типах сопряженности:

сопряженность типа "документ – документ";

сопряженность типа "термин – термин";

сопряженность типа "документ – термин".

Если первые два типа сопряженностей можно выразить квадратными матрицами порядков соответственно n и m, то сопряженность типа "документ – термин" выражается в общем случае прямоугольной матрицей C $[c_{ij}]$ размерности $n \times m$. В простейшем случае это матрица, где значение c_{ij} равно единице, если j-й термин содержится в i-документе, и нулю – в противном случае. При этом каждому документу ставится в соответствие некоторый m-мерный бинарный вектор (строка), т.е. некоторая из 2^m вершин m-мерного единичного куба пространства терминов. Аналогично, каждому термину ставится в соответствие некоторый n-мерный бинарный вектор (столбец), т.е. некоторая из 2^n вершин n-мерного единичного куба пространства документов.

Простейшим примером матрицы сопряженности типа "документ – документ" может служить квадратная матрица D $[d_{ij}]$ порядка n, где значение d_{ij} равно единице, если существует хоть один термин, одновременно содержащийся как в i-м, так и в j-м документах, и нулю – в противном случае.

В качестве же простейшего примера матрицы сопряженности типа "термин – термин" может служить квадратная матрица T $[t_{ij}]$ порядка m, где значение t_{ij} равно единице, если существует хоть один документ, где одновременно содержатся как i-й, так и j-й термины, и нулю – в противном случае.

Продолжая рассматривать лишь простейшие случаи, положим, что пользовательские запросы также представлены соответствующими бинарными векторами, а именно:

n-мерным бинарным вектором, значение i-й координаты которого равно единице, если i-й документ пользователем включен в список документов, представляющий его запрос, и нулю – в противном случае;

m-мерным бинарным вектором, значение i-й координаты которого равно единице, если i-й термин пользователем включен в список терминов, представляющий его запрос, и нулю – в противном случае.

Далее векторы Q $[q_i]$, представляющие пользовательские запросы, будем рассматривать (в зависимости от формы их представления) либо как матрицы-строки, либо же как матрицы-столбцы.

В случае, когда Q является бинарным вектором размерности m, можно говорить об n-мерном векторе A $[a_i]$ – реакции системы на запрос Q:

$$A = C \cdot Q. \tag{5.55}$$

Значение i-й координаты n-мерного вектора A $[a_i]$ при этом оказывается равным числу терминов запроса, оказавшихся в i-м документе. Естественно (хотя и небезупречно) полагать, что чем больше значение a_i, тем больше вероятность того, что i-й документ релевантен пользовательскому запросу Q.

Можно условиться, например, выдать пользователю только те документы, для которых имеет место $a_i > \beta$, где β – некоторое пороговое значение.

Заметим, что даже тогда, когда все строки и столбцы матрицы $C[c_{ij}]$, а также вектор Q представлены бинарными векторами, вектор A в общем случае не получается бинарным и значениями a_i могут служить произвольные натуральные числа.

На практике чаще всего приходится иметь дело со случаями, когда d_{ij}, t_{ij}, c_{ij}, q_i и a_i могут принимать произвольные действительные значения.

Линейный ассоциативный поиск.

Рассмотрим матрицу сопряженности $C[c_{ij}]$ "документ – термин", где c_{ij} принимают произвольные действительные значения, указывающие на значимость j-го термина при описании i-го документа. Например, c_{ij} могут принимать значения, равные числу встречаемости j-го термина в тексте i-го документа. В другом случае c_{ij} могут принимать значения, равные отношению числа встречаемости j-го термина в i-м документе к общему числу терминов в этом документе. В ряде случаев значения c_{ij} приписываются индексаторами, которые после ознакомления с текстами документов сами определяют по своему усмотрению значимость j-го термина при описании i-го документа.

Аналогично, пользовательский запрос также может быть представлен вектором $Q[q_j]$ размерности m, где q_j принимают произвольные действительные значения, указывающие на значимость j-го термина при описании данного запроса. Степень значимости j-го термина, т.е. значение q_j определяется самим пользователем либо самостоятельно, либо же с помощью индексаторов.

Как и в случае бинарных векторов, будем рассматривать матричное произведение

$$A^{(0)} = CQ^{(0)}, \tag{5.56}$$

где верхний индекс "0" при векторе $Q^{(0)}\left[q_j^{(0)}\right]$ указывает на то, что речь идет именно о первоначально сформулированном пользователем запросе.

Из (5.56) непосредственно следует, что

$$a_i^{(0)} = \sum_{j=1}^{m} c_{ij} q_j^{(0)}. \tag{5.57}$$

Руководствуясь рядом соображений (в том числе и интуитивными), ряд авторов считает, что значения $a_i^{(0)}$ можно принимать за формальную меру релевантности i-го документа пользовательскому запросу $Q^{(0)}$. На этом, собственно, и базируются различные критерии оценки семантической близости, оперирующие методом весовых коэф-

фициентов. В сущности, значение $a_i^{(0)}$ равно скалярному произведению i-го вектора – строки матрицы C на вектор $Q^{(0)}$. О правомочности принятия скалярного произведения двух векторов за меру их близости (подобия) будем говорить ниже. А сейчас рассмотрим матричное произведение

$$Q^{(1)} = C^T A^{(0)}, \tag{5.58}$$

где C^T – матрица, транспонированная относительно матрицы C.

Руководствуясь теми же соображениями, что и при рассмотрении формулы (5.57), можно было значения

$$q_j^{(1)} = \sum_{i=1}^{n} c_{ji}^T a_i^{(0)} \tag{5.59}$$

прокомментировать как уточненные значения $q_j^{(0)}$, т.е. уточненные значения значимости j-го термина при описании пользовательского запроса. Тогда стало бы актуальным рассматривать матричное произведение

$$A^{(1)} = CQ^{(1)}, \tag{5.60}$$

приняв значения

$$a_i^{(1)} = \sum_{j=1}^{m} c_{ij} q_j^{(1)} \tag{5.61}$$

за уточненную формальную меру релевантности i-го документа пользовательскому запросу.

Продолжая "в том же духе", мы придем к рассмотрению бесконечного процесса

$$\begin{aligned} A^{(0)} &= CQ^{(0)} \\ Q^{(1)} &= C^T A^{(0)} \\ A^{(1)} &= CQ^{(1)} \\ &\cdots\cdots\cdots\cdots \\ A^{(t)} &= CQ^{(t)} \\ Q^{(t+1)} &= C^T A^{(t)} \end{aligned} \tag{5.62}$$

характер поведения которого при достаточно больших t и будет предметом нашего рассмотрения.

Из (5.62) легко обнаружить, что

$$Q^{(t)} = \left(C^T C\right)^t Q^{(0)}, \tag{5.63}$$

$$A^{(t)} = \left(CC^T\right)^t A^{(0)}, \tag{5.64}$$

где $\left(C^T C\right)^t$ и $\left(CC^T\right)^t$ – это матрицы соответственно $C^T C$ и CC^T, возведенные в степень t.

Можно показать, что если $F_2(\lambda)$ и $F_1(\lambda)$ являются характеристическими многочленами соответственно матриц C^TC и CC^T, где C – произвольная матрица размерности $n \times m$, то справедлива формула (см. приложение 1):

$$F_1(\lambda) = \lambda^{n-m} F_2(\lambda). \tag{5.65}$$

Пусть среди корней характеристического многочлена $F_2(\lambda)$ имеется старшее по модулю собственное значение λ_0 матрицы C^TC. Тогда из (5.65) следует, что среди корней характеристического многочлена $F_1(\lambda)$ также имеется старший по модулю корень – собственное значение матрицы CC^T. Более того, значение этого собственного значения также равно λ_0. Но из теоремы Сильвестра следует, что наличие у матрицы C^TC старшего по модулю собственного значения λ_0 влечет справедливость при достаточно больших t приближенной формулы [3, 9]:

$$\left(C^TC\right)^{t+1} = \lambda_0 \left(C^TC\right)^t, \tag{5.66}$$

с учетом которой из (5.63) имеем:

$$Q^{(t+1)} = \left(C^TC\right)Q^{(t)} = \lambda_0 Q^{(t)}. \tag{5.67}$$

Аналогично, из (5.64) имеем

$$A^{(t+1)} = \left(CC^T\right)A^{(t)} = \lambda_0 A^{(t)}. \tag{5.68}$$

Из (5.67) и (5.68) следует, что с увеличением значения t векторы $Q^{(t)}$ и $A^{(t)}$ стремятся принимать направления собственных векторов матриц C^TC и CC^T, соответствующих собственным значениям этих матриц, равным λ_0. Иными словами, при произвольном ненулевом векторе $Q^{(0)}$ чем больше значение индекса t, тем в меньшей степени векторы $Q^{(t)}$ и $A^{(t)}$ зависят от вектора $Q^{(0)}$, а в пределе, когда $t \to \infty$, эти векторы и вовсе перестают зависеть от $Q^{(0)}$.

Образно говоря, если вектор $Q^{(0)}$ вообще не учитывает свойства поисковой среды (выразителем которого является матрица C), то при формировании вектора $Q^{(1)}$ фактор среды уже учитывается. Еще в большей степени фактор среды учитывается при формировании вектора $Q^{(2)}$ и далее, чем больше значение индекса t, тем в большей степени при формировании вектора $Q^{(t)}$ учитывается фактор среды и тем в меньшей степени – пользовательский запрос, т.е. вектор $Q^{(0)}$. В результате при достаточно больших значениях индекса t при формировании вектора $Q^{(t)}$ пользовательский запрос $Q^{(0)}$ вовсе предается забвению и вектор $Q^{(t)}$ становится своеобразным выразителем свойств самой поисковой среды. Аналогично обстоит дело также с вектором $A^{(t)}$.

Таким образом, если в нескольких первых тактах, при не очень больших значениях индекса t процесс (5.62) действительно улучшает качество поиска (так как наряду с пользовательским запросом учитывается также фактор среды), то дальнейшее продолжение этого процесса при чрезмерно больших значениях индекса t приводит к резкому ухудшению качества поиска, так как результаты поиска при этом перестают зависеть от пользовательского запроса.

Чтобы не предать забвению пользовательский запрос $Q^{(0)}$, можно было рассматривать процесс

$$\begin{aligned} A^{(0)} &= CQ^{(0)} \\ Q^{(1)} &= C^T A^{(0)} + Q^{(0)} \\ A^{(1)} &= CQ^{(1)} \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ A^{(t)} &= CQ^{(t)}, \\ Q^{(t+1)} &= C^T A^{(t)} + Q^{(0)} \end{aligned} \tag{5.69}$$

который отличается от процесса (5.62) наличием слагаемого $Q^{(0)}$, что, собственно, и предотвращает забвение этого вектора, т.е. пользовательского запроса.

Из (5.69) следует, что

$$Q^{(t)} = \left[1 + C^T C + \left(C^T C\right)^2 + \ \ldots\ + \left(C^T C\right)^t\right] Q^{(0)}, \tag{5.70}$$

$$A^{(t)} = C\left[1 + C^T C + \left(C^T C\right)^2 + \ \ldots\ + \left(C^T C\right)^t\right] Q^{(0)}. \tag{5.71}$$

Формулу (5.71) можно переписать также как

$$A^{(t)} = \left[1 + CC^T + \left(CC^T\right)^2 + \ \ldots\ + \left(CC^T\right)^t\right] A^{(0)}. \tag{5.72}$$

Рассмотрим функцию $f(x) = (1-x)^{-1}$. Для этой функции ряд Тейлора в окрестности точки $x = 0$ (ряд Маклорена) имеет вид бесконечной суммы геометрической прогрессии:

$$P(x) = 1 + x + x^2 + \ldots, \tag{5.73}$$

которая сходится к функции $f(x)$ тогда и только тогда, когда имеет место $|x| < 1$. В силу теоремы Кэлли – Гамильтона отсюда следует, что если в (5.73) скалярную величину x заменить матрицей Z, то для сходимости уже матричного ряда

$$P(Z) = 1 + Z + Z^2 + \ldots \tag{5.74}$$

к функции $f(Z) = \left(1 - Z\right)^{-1}$ необходимо и достаточно, чтобы все корни характеристического многочлена матрицы Z по модулю были меньше единицы [3, 9].

Таким образом, соблюдение условия

$$|\lambda_0| < 1 \tag{5.75}$$

является необходимым и достаточным для того, чтобы из формул (5.70) и (5.71) следовало

$$Q = \lim_{t\to\infty} Q^{(t)} = \left(1 - C^T C\right)^{-1} Q^{(0)}, \tag{5.76}$$

$$A = \lim_{t\to\infty} A^{(t)} = C\left(1 - C^T C\right)^{-1} Q^{(0)}. \tag{5.77}$$

Обратим внимание, что соблюдение условия (5.75) одновременно гарантирует отсутствие у матрицы $(1 - C^T C)$ нулевого собственного значения, т.е. обратимость этой матрицы. Заметим также, что при соблюдении (5.75) из (5.72) следует

$$A = \lim_{t\to\infty} A^{(t)} = \left(1 - CC^T\right)^{-1} CQ^{(0)} \tag{5.78}$$

Из (5.76) ÷ (5.77) следует, что при достаточно больших значениях t матрицы Q и A являются решением системы уравнений

$$\begin{aligned} A &= CQ \\ Q &= C^T A + Q^{(0)}, \end{aligned} \tag{5.79}$$

которую можно переписать в виде матричного уравнения

$$\begin{pmatrix} A \\ Q \end{pmatrix} = \begin{pmatrix} 0 & C \\ C^T & 0 \end{pmatrix} \begin{pmatrix} A \\ Q \end{pmatrix} + \begin{pmatrix} 0 \\ Q^{(0)} \end{pmatrix}. \tag{5.80}$$

К этому уравнению мы пришли путем изучения поведения процесса (5.69) при достаточно больших t. Именно такое уравнение постулируется Сэлтоном Г. при рассмотрении линейного ассоциативного поиска, где в качестве предположения приводится матричное уравнение [14]

$$\begin{pmatrix} A \\ Q \end{pmatrix} = \begin{pmatrix} D & C \\ C^T & T \end{pmatrix} \begin{pmatrix} A \\ Q \end{pmatrix} + \begin{pmatrix} 0 \\ Q^{(0)} \end{pmatrix}. \tag{5.81}$$

Здесь D и T матрицы, учитывающие ассоциации типа "документ – документ" и "термин – термин". Если в (5.81) принять $D = 0$ и $T = 0$, т.е. априори пренебречь ассоциациями этих типов, то придем к полученной нами системе (5.80).

Следует особо подчеркнуть, что переход от системы (5.69) к (5.79) (как и, впрочем, от (5.79) к (5.69)) возможен лишь при соблюдении условия (5.75).

В заключение настоящего параграфа попробуем оценить правомочность использования матричных моделей документального поиска в том виде, в каком они существуют в настоящее время.

Выше мы уже говорили о том, что в матричных моделях документального поиска, включающих операцию умножения матриц, в качестве критерия подобия (близости) двух векторов фактически принимается значение их скалярного произведения. Например, для оценки степени близости двух m-мерных векторов, один из которых представляет пользовательский запрос, а другой – некоторый документ, по величине их скалярного произведения судят о степени соответствия (релевантности) данного документа пользовательскому запросу. Здесь m – число элементов во множестве терминов, т.е. рассматриваемые векторы по сути представляют соответствующие (в общем случае нечеткие) подмножества этого множества.

В другом случае речь идет об оценке степени близости двух n-мерных векторов, представляющих соответствующие подмножества множества из n документов. Например, речь может идти о скалярном произведении двух векторов, один из которых представляет пользовательский запрос, сформулированный как нечеткое подмножество множества документов, а другой – некоторый термин, также представленный как некоторое подмножество этого множества.

Таким образом, когда говорят о степени близости двух векторов размерности z, подразумевается, что речь идет о близости двух подмножеств некоторого множества из z элементов. Этим обстоятельством мы будем пользоваться для оценки (там, где это возможно) правомерности, корректности тех или иных интуитивных соображений, лежащих в основе построения различных моделей документального поиска. В простейшем случае, когда осуществляется оценка близости двух бинарных z-мерных векторов, речь идет о степени близости двух обычных подмножеств, определенных на множестве из z элементов. Данный случай полностью характеризуется матрицей сопряженности этих подмножеств (рис. 5.3), где

		первое подмножество	
		X_1	$\overline{X}_1$
второе подмножество	X_2	a	b
	$\overline{X}_2$	c	d

Рис. 5.3. Матрица, характеризующая степень близости подмножеств X_1 и X_2

a – число элементов исходного множества, одновременно принадлежащих обоим подмножествам;

b – число элементов исходного множества, которые принадлежат второму подмножеству и не принадлежат первому;

c – число элементов исходного множества, которые принадлежат первому и не принадлежат второму;

d – число элементов исходного множества, которые не принадлежат ни первому, ни второму подмножествам.

Очевидно, имеет место $a + b + c + d = z$.

Пусть, например, речь идет об оценке степени близости двух обычных подмножеств, определенных на множестве терминов и представляющих пользовательский запрос и очередной документ. Представляется вполне естественным, чтобы документы, которым при фиксированных значениях a и z соответствуют большие значения b и c, признались бы менее релевантными пользовательскому запросу, чем документы, которым соответствуют меньшие значения b и c. Скалярное же произведение двух векторов, равное в данном случае величине a, не учитывает этого обстоятельства, и поэтому его применение нельзя признать правомочным, корректным. В то же время, практически во всех существующих матричных моделях документального поиска фигурирует операция умножения матриц, сводящаяся, как известно, к вычислению скалярных произведений соответствующих векторов. Это обстоятельство ставит под сомнение правомочность, корректность использования всех моделей, где используется операция умножения матриц. С другой стороны, при рассмотрении корреляционных моделей документального поиска мы неоднократно убедились в правомерности, корректности использования для оценки степени близости двух векторов (или, что то же самое, двух подмножеств) формулы коэффициента линейной корреляции (см. например, формулу 5.28). В рассматриваемом смысле представляется заманчивым разработать модели, где корректность формулы коэффициента корреляции сочеталась бы с изяществом инструментария матричного исчисления (собственные векторы, собственные значения и т.п.). Наши исследования показали, что этого можно достичь путем введения в рассмотрение операции R-произведения матриц.

В разделе, где приведена теория динамического взаимодействия различных стратегий анализа, аргументируется использование R-произведения матриц вместо обычной операции умножения матриц. Свидетельством же практической целесообразности использования R-произведения матриц могут служить результаты промышленной эксплуатации аналитических систем, функционирующих по принципу динамического взаимодействия различных стратегий анализа.

5.7. ЭФФЕКТИВНОСТЬ ДОКУМЕНТАЛЬНОГО ПОИСКА И КРИТЕРИИ ЕЕ ОЦЕНКИ

Одно из центральных мест в общей проблематике информационного поиска занимает проблема оценки эффективности АСДП. Оценка эффективности осуществляется не только для пассивной констатации преимуществ или недостатков уже существующих АСДП, но и для выбора оптимальных решений из альтернативных вариантов на этапе проектирования. Оценке могут подвергаться как АСДП в целом, так и отдельные их компоненты (подсистемы). При этом необходимо исходить из того, что целевая функция каждого из компонентов должна быть подчинена целевой функции АСДП в целом.

Будем различать *технико-экономическую и функциональную эффективность АСДП.*

Под технико-экономической эффективностью обычно понимают совокупность таких факторов, как быстродействие АСДП, полнота охвата документов при комплектовании баз данных, себестоимость поиска, оснащенность системы современной множительной аппаратурой, возможность ее эксплуатации в сетевом режиме, оснащенность различными средствами защиты информации, минимальная конфигурация, комфортность и т.п.

Под функциональной эффективностью будем понимать способность системы извлечь из базы данных и выдать пользователю как можно большее число релевантных документов и как можно меньшее число нерелевантных.

Долгие годы в отечественной практике фактически отсутствовала ценовая политика на информационные услуги. О таких важных показателях, как себестоимость и цена информационного обслуживания, обычно умалчивалось. В наше же время, в период рыночной экономики, именно за этими показателями зачастую остается последнее слово при решении вопроса о том, "быть или не быть" данному автоматизированному центру информации. Поэтому представляется вполне естественным желание ряда исследователей разработать комплексные критерии для одновременной оценки функциональной и технико-экономической эффективности АСДП. Способ взвешенного суммирования (или умножения) значений отдельных критериев – составляющих, учитывающих различные аспекты функциональной и технико-экономической эффективности АСДП, вряд ли можно признать перспективным, так как в сущности своей этот способ искусственный, не отражающий природы информационного поиска. Неудачи в области синтеза комплексных критериев, одновременно учитывающих функциональную и технико-экономическую эффективность АСДП, объясняются еще и отсутствием общепринятых количественных характеристик, устанавливающих зависимость стоимости информации от оперативности оповещения. В сложившейся обстановке порою приходится соглашаться с

авторами, рекомендующими при комплексной оценке функциональной и технико-экономической эффективности АСДП пользоваться не метрическими стоимостными критериями, а анализом результатов сравнительной оценки альтернативных вариантов.

Представляется, что разработку комплексных критериев оценки технико-экономической и функциональной эффективности информационных систем следует начинать с четкого разграничения функций собственно информационной системы и систем (отдельных индивидуумов, общества в целом), призванных использовать найденную информацию. Несмотря на кажущуюся простоту, это является довольно сложной задачей, так как на практике такие системы оказываются до такой степени взаимно сроснимися, что оказывается чрезвычайно трудно провести четкую грань между ними. Тем не менее, эти трудности следует преодолеть хотя бы для того, чтобы избавиться от необходимости рассматривать чрезвычайно сложный, трудно поддающийся формальной количественной оценке фактор "полезности" информации. Ведь эффективность использования результатов работы не является характеристикой собственно информационной системы. Работа любой информационной системы, будь то система поиска информации, система дезинформации, система защиты информации и т. п., направлена на то, чтобы уменьшить, увеличить или предотвратить приращение в нежелательную сторону предварительно имеющуюся у других систем неопределенность (энтропию). И именно это приращение энтропии, фактически имевшее место или предотвращенное, вкупе с теми расходами, в которые обошлась работа информационной системы, должны служить отправной точкой при разработке комбинированных критериев. При анализе, например, информационно-поисковой системы речь может идти о том, во сколько рублей в среднем обходится работа системы по ликвидации у пользователя одного килобайта неопределенности.

Далее сконцентрируем наше внимание на вопросах оценки лишь функциональной эффективности АСДП.

Еще в 1953 г. Дж. Перри, А. Кент и М. Берри для оценки функциональной эффективности АСДП предложили использовать коэффициенты полноты и точности. В 1964 г. У. Гоффман и В. Ньюилл для этих же целей ввели в рассмотрение коэффициент специфичности. В предыдущих разделах мы уже обратили внимание на то, что коэффициенты полноты и специфичности действительно являются характеристиками АСДП, тогда как коэффициент точности при заданных коэффициентах полноты и специфичности зависит только от статистической характеристики поисковой среды, т. е. от параметра ω:

$$\omega_1 = \omega\lambda_1/(1 - \lambda_2 + \omega(\lambda_1 + \lambda_2 - 1)).$$

В рассматриваемом смысле представляется разумным оперировать не коэффициентом точности, а непосредственно параметром ω – концентрацией, долей релевантных документов в исходном множестве документов N.

Поскольку пара значений λ_1 и λ_2 (а если имеется возможность их изменения, то зависимость $\lambda_1 = \lambda_1(\lambda_2)$) является исчерпывающей характеристикой собственно АСДП, то можно было бы ограничиться только их рассмотрением. При этом, однако, мы не смогли бы судить о реальном качестве информационного поиска, так как при заданной зависимости $\lambda_1 = \lambda_1(\lambda_2)$ фактическая эффективность работы АСДП в значительной мере зависит от степени согласованности поисковой среды с этой зависимостью. При анализе работы каналов связи К. Шеннон особо обратил внимание на это обстоятельство, что, собственно, и привело его к понятию пропускной способности каналов связи. К сожалению, нам не известны работы, где бы столь же обстоятельно и всесторонне рассматривались вопросы информационного поиска, как это осуществил К. Шеннон применительно к каналам связи.

Путем установления соответствующих аналогий мы разработали энтропийную модель информационного поиска и соответствующие критерии оценки эффективности АСДП. Наряду с понятием количества информации мы ввели в рассмотрение понятие коэффициента проводимости. Именно эти две величины мы и рекомендуем использовать в качестве критерия для оценки функциональной эффективности АСДП.

В рамках корреляционной модели мы разработали корреляционный критерий оценки функциональной эффективности АСДП. Как и критерии $I[x, y]$ (количество информации) и $\varkappa[x, y]$ (коэффициент проводимости), критерий r_{xy} наряду со свойствами собственно АСДП учитывает также степень согласованности поисковой среды с этими свойствами. Представляется заслуживающим внимание также вариант использования для оценки функциональной эффективности АСДП параметра $I_R[x, y]$, рассмотренного в разделе 5.5.

Важным преимуществом критерия r_{xy} является то обстоятельство, что этот критерий может применяться не только к бинарным ситуациям, но и к системам, оперирующим размытыми подмножествами.

Из других критериев, предложенных для оценки функциональной эффективности АСДП, следует упомянуть разработанный в рамках американского проекта "СМАРТ" критерий "косинус". На недостатки этого критерия, равно как и критерия "скалярное произведение", нами уже указывалось в разделе 5.4.

Более подробный анализ различных критериев оценки функциональной и технико-экономической эффективности можно найти в [2].

ЛИТЕРАТУРА К ГЛАВЕ 5

1. *Аветисян Д.О.* О вероятностном подходе к построению интеллектуальных систем, Ч. 1. Теория // Математические вопросы кибернетики и вычислительной техники: Сб. науч. тр. / Вычисл. центр. АН Арм. ССР, Ереванск. гос. уиив-т. – Ереван, 1984. – Т. 13.

2. *Аветисян Д.О.* Проблемы информационного поиска. – М.: Финансы и статистика, 1981.
3. *Анго Андре.* Математика для электро- и радиоинженеров. – М.: Наука, 1967.
4. *Березин Ф.М.* История лингвистических учений. – М.: Высшая школа, 1975.
5. *Брутян Г.А.* Гипотеза Сепира – Уорфа. – Ереван: Луйс, 1968.
6. *Гегель.* Сочинения. Т. 4. – М.: Соцэкгиз, 1959.
7. *Жданова Г.С., Колобродова Е.С., Полушкин В.А., Черный А.И.* Словарь терминов по информатике на русском и английском языках. – М.: Наука, 1971.
8. *Заде Л.* Понятие лнигвистической переменной и его применение к принятию приближенных решений. – М.: Мир, 1976.
9. *Корн Г., Корн Т.* Справочник по математике для научных работников и инженеров. – М.: Наука, 1968.
10. *Лифшиц Н.А., Пугачев В.Н.* Вероятностный анализ систем автоматического управления. – М.: Советское радио, 1963.
11. *Михайлов А.И., Черный А.И., Гиляревский Р.С.* Основы информатики. – М.: Наука, 1968.
12. *Панфилов В.З.* Взаимоотношения языка и мышления. – М.: Наука, 1971.
13. *Рассел Б.* Человеческое познание. – М.: Изд-во ин. лит., 1957.
14. *Сэлтон Г.* Автоматическая обработка, хранение и поиск информации. – М.: Советское радио, 1973.
15. *Шеннон К.* Работы по теории информации и кибернетике. – М.: Изд-во ин. лит., 1963.
16. *Bar-Hillel Y.* Alogician's reaction to theorizing on information search systems // American Documentation. – 1957. – Vol. 8, № 2. – P. 105.
17. *Bar-Hillel Y.* Some theoretical aspects of mechanization of literature searching: Technical report № 3 / Hebrew University. – Jerusalem, 1960. – P. 42–44.
18. *Doyle L.B.* Is relevance an adequate criterion in information system evaluation? // Automation and scientific communication. Pt. 2. – D.C., Washington: American Documentation Institute, 1963. – P. 200.
19. *Taube M.* A note on the pseudo-mathematics of relevance // American Documentation. – 1965. – Vol. 16, № 2. – P. 71.

ГЛАВА 6

ЭЛЕМЕНТЫ ТЕОРИИ ДИНАМИЧЕСКОГО ВЗАИМОДЕЙСТВИЯ РАЗЛИЧНЫХ СТРАТЕГИЙ ПОИСКА

ПРИ РАССМОТРЕНИИ математических моделей документального поиска мы уже говорили о возможности представления n-мерными векторами различных подмножеств (в том числе нечетких), определенных на произвольных множествах из n элементов. В частном случае, когда эти векторы являются бинарными, они представляют обычные (четкие) подмножества соответствующих множеств. Говорилось также об операциях центрирования и нормирования этих векторов как необходимых этапах при вычислении коэффициента линейной корреляции между соответствующими векторами, или, что то же самое, соответствующими подмножествами. Динамическое взаимодействие различных стратегий поиска (анализа) фактически сводится к реализации последовательности операций, приводящих к достижению максимальных значений тех или иных параметров оптимизации. Таковыми служат параметры, устанавливающие меру близости (подобия) двух векторов. От корректности выбора этих параметров в основном и зависит эффективность динамического взаимодействия различных стратегий анализа. Проанализировав различные критерии, призванные оценить меру подобия двух векторов, мы остановили свой выбор на энтропийном и корреляционном критериях, значения которых и были использованы нами в качестве параметров оптимизации при построении различных алгоритмов динамического взаимодействия.

Для простоты и не в ущерб общности изложения в рамках настоящего пособия мы сконцентрируем наше внимание лишь на корреляционных моделях динамического взаимодействия, не забывая, однако, что многое из нижесказанного после определенных модификаций вычислительного (а не идеологического) характера остается справедливым для энтропийных моделей динамического взаимодействия [2].

Идеологической основой теории динамического взаимодействия различных стратегий является теорема транзитивности. Ниже приводится доказательство этой теоремы для случаев n-мерной сферы и

n-мерного куба. Мы не сочли излишним привести по ходу изложения также ряд сопутствующих результатов (включая теорему синонимии), знание которых, на наш взгляд, будет способствовать лучшему пониманию теоремы. Мы надеемся также, что читатели сами оценят значимость этих результатов для возможного самостоятельного их использования вне предмета нашего рассмотрения [2].

Пусть рассматривается универсальное множество U из n элементов, на котором определены различные размытые подмножества X_i с функцией принадлежности $\mu_i(u_t)$, задающей меру принадлежности t-го элемента универсального множества U подмножеству X_i. Линейным преобразованием размытого подмножества X_i будем называть произвольное размытое подмножество X_j, для которого имеет место

$$\mu_j(u_t) = \beta_{ij} + \alpha_{ij}\mu_i(u_t) \ (\alpha_{ij} \neq 0, \ t = 1, 2, \ldots, n). \tag{6.1}$$

При $\alpha_{ij} > 0$ будем говорить также, что подмножества X_i и X_j являются линейными повторениями друг друга и, наоборот, когда $\alpha_{ij} < 0$ будем говорить, что X_i и X_j являются линейными дополнениями друг друга. Из основных свойств коэффициента линейной корреляции непосредственно следует, что в первом случае имеет место

$$r(X_i, X_j) = 1, \tag{6.2}$$

а во втором

$$r(X_i, X_j) = -1. \tag{6.3}$$

Во всех остальных случаях имеет место

$$|r(X_i, X_j)| < 1. \tag{6.4}$$

Далее через X_i будем обозначать также точки n-мерного пространства, радиусы-векторы которых представляют соответствующие размытые подмножества. Легко убедиться, что все координаты вектора X_{iE} – геометрической проекции вектора X_i на направление вектора $E(1, 1, \ldots, 1)$ – равны m_i, где

$$m_i = \frac{1}{n}\sum_{t=1}^{n}\mu_i(u_t). \tag{6.5}$$

При этом

$$r(X_i, X_j) = \cos(X_i^*, X_j^*), \tag{6.6}$$

где вектор X_i^* получается из вектора X_i путем его центрирования:

$$X_i^* = X_i - X_{iE}. \tag{6.7}$$

В свою очередь, путем операции нормирования вектора X_i^* можно

получить вектор

$$x_i = X_i^* / |X_i^*|. \tag{6.8}$$

Геометрическим местом точек, имеющих в качестве своих радиусов-векторов векторы типа x_i, является единичная сфера с центром в начале координат, все радиусы-векторы которой перпендикулярны вектору $E(1, 1,\ldots, 1)$. Фактическая размерность этой сферы равна $n - 1$. Преобразования (6.8) осуществляют однозначное отображение произвольного размытого подмножества X_i в соответствующую точку x_i этой сферы. Обратное отображение не является однозначным, так как при прямом отображении в эту же точку x_i отображаются все другие подмножества – линейные повторения размытого подмножества X_i. В точку же $-x_i$ при этом отображаются все размытые подмножества – линейные дополнения размытого подмножества X_i.

6.1. ТЕОРЕМЫ ТРАНЗИТИВНОСТИ И СИНОНИМИИ (СЛУЧАЙ n-МЕРНОЙ СФЕРЫ)

Утверждение 1.

Пусть x – случайная точка, имеющая равномерное распределение на n-мерной единичной сфере с центром в начале координат, а y_0 – произвольная фиксированная точка. Тогда имеют место:

$$\text{а) } M(\cos(xy_0)) = 0, \tag{6.9}$$

$$\text{б) } M(\cos^2(xy_0)) = D(\cos(xy_0)) = 1/n, \tag{6.10}$$

где через $M(z)$ и $D(z)$ обозначены соответственно математическое ожидание и дисперсия случайной величины z.

Справедливость (6.9) непосредственно следует из симметричности распределения случайной величины $\cos(xy_0)$ относительно оси ординат. Из условия принадлежности точки x рассматриваемой сфере имеем $\sum_{k=1}^{n} x_k^2 = 1$, т. е. $\sum_{k=1}^{n} M(x_k^2) = 1$. Отсюда из соображений симметрии получим $M(x_k^2) = 1/n$, т. е. $M(\cos^2(xy_0)) = 1/n$. С учетом (6.9) отсюда следует (6.10).

В частном случае, когда $n = 2$, формулировка пункта (б) приводит к известной формуле

$$\int_0^{2\pi} \cos^2 \varphi d\varphi = \pi.$$

Замечание к утверждению 1.

Из соображений симметрии легко убедиться, что (6.9) и (6.10) остаются в силе при замене фиксированной точки y_0 случайной точкой y, имеющей произвольное независимое от x распределение.

Утверждение 2.

Пусть x и y – независимые случайные точки, имеющие равномерные распределения на соответствующих множествах точек рассматриваемой сферы, определенных фиксированными значениями $\cos(xz_0)$ и $\cos(yz_0)$, где $z_0(z_{01}, z_{02},\ldots, z_{0n})$ – произвольная фиксированная точка. Тогда имеют место:

$$\text{а)}\ M(\cos(xy)) = \cos(xz_0)\cos(yz_0), \tag{6.11}$$

$$\text{б)}\ D(\cos(xy)) = \frac{1}{n-1}\sin^2(xz_0)\sin^2(yz_0), \tag{6.12}$$

в) в области $\cos^2(xz_0) + \cos^2(yz_0) < 1$ величина

$$p(\cos(xy) > 0) - 0{,}5$$

является монотонно возрастающей нечетной функцией аргумента

$$\cos(xz_0)\cos(yz_0),$$

г) в области $\cos^2(xz_0) + \cos^2(yz_0) \geqslant 1$ имеет место

$$p(\cos(xy) > 0) - 0{,}5 = 0{,}5\,\mathrm{sgn}(\cos(xz_0)\cdot\cos(yz_0)). \tag{6.13}$$

Действительно, приняв в качестве базиса n-мерного евклидова пространства систему векторов, один из которых (например, с индексом 1) проходит через точку z_0, получим:

$$\cos(xy) = \cos(xz_0)\cos(yz_0) + \sum_{k=2}^{n} x_k y_k. \tag{6.14}$$

Легко показать, что закон распределения случайной величины $\sum_{k=1}^{n} x_k y_k$ совпадает с законом распределения случайной величины $R_1R_2\cos(uv)$, где

$$R_1 = \sqrt{1-\cos^2(xz_0)},\quad R_2 = \sqrt{1-\cos^2(yz_0)},$$

а $u(u_1, u_2,\ldots, u_{n-1})$ и $v(v_1, v_2,\ldots, v_{n-1})$ – радиусы-векторы независимых случайных точек, имеющих равномерные распределения на $(n-1)$-мерной единичной сфере с центром в начале координат. Отсюда, с учетом (6.9), (6.10) и замечания к утверждению 1, легко убедиться в справедливости всех пунктов утверждения 2. Из (6.11), (6.12) легко получить выражение для

$$M(\cos^2(xy)) = \frac{1}{n-1}\sin^2(xz_0)\sin^2(yz_0) + \cos(xz_0)\cos(yz_0). \tag{6.15}$$

Заметим, что формула (6.13) остается в силе при произвольных законах распределений случайных точек x и y на упомянутых выше множествах точек сферы.

В частном случае, когда $n = 2$, (6.11) и (6.15) приводят к известным формулам:

$$\cos(\alpha+\beta)+\cos(\alpha-\beta)=2\cos\alpha\cdot\cos\beta,$$

$$\cos(\alpha+\beta)\cdot\cos(\alpha-\beta)=\cos^2\alpha\cdot\cos^2\beta-\sin^2\alpha\cdot\sin^2\beta.$$

Замечание к утверждению 2.

Утверждение 2 остается справедливым при замене случайной точки y произвольной фиксированной точкой $y_0(y_{01}, y_{02},\ldots, y_{0n})$ с заданным значением $\cos(y_0z_0)$.

Действительно, при этом в выражении $R_1R_2\cos(uv)$ вместо случайной точки $v(v_1, v_2,\ldots, v_{n-1})$ приходится рассматривать фиксированную точку $v_0(v_{01}, v_{02},\ldots, v_{0(n-1)})$ $(n-1)$-мерной единичной сферы, что не сказывается на распределении случайной величины $R_1R_2\cos(uv)$ (см. замечание к утверждению 1).

Теорема транзитивности для n-мерной единичной сферы (обобщение пунктов(а) и (б) утверждения 2)

Пусть x и y – независимые случайные точки, имеющие равномерные распределения на соответствующих множествах точек рассматриваемой сферы, определенных фиксированными значениями $\cos(xz_0)$ и $\cos(yq_0)$, где z_0 и q_0 произвольные фиксированные точки. Тогда имеют место:

$$\text{а) } M(\cos(xy))=\cos(xz_0)\cos(z_0q_0)\cos(q_0y), \tag{6.16}$$

$$\text{б) } D(\cos(xy))=\frac{1}{(n-1)^2}(n-2+\cos^2(xz_0)+$$

$$+\cos^2(yq_0)+\cos^2(z_0q_0)-n(\cos^2(xz_0)\cos^2(yq_0)+$$

$$+\cos^2(xz_0)\cos^2(z_0q_0)+\cos^2(yq_0)\cos^2(z_0q_0))+$$

$$+(2n-1)\cos^2(xz_0)\cos^2(yq_0)\cos^2(z_0q_0)). \tag{6.17}$$

Действительно, приняв в качестве базиса n-мерного пространства систему векторов, один которых (например, с индексом 1) проходит через точку z_0, получим:

$$x_1=\cos(xz_0),\ \ y_1=\cos(yz_0),\ \ M(x_2)=M(x_3)=\ldots=M(x_n)=0.$$

Отсюда, в силу независимости случайных точек x и y, а также с учетом (6.11) и замечания к утверждению 2, получим:

$$M(\cos(xy))=\sum_{k=1}^{n}M(x_k)M(y_k)=\cos(xz_0)M(\cos(yz_0))=$$

$$= \cos(xz_0)\cos(yq_0)\cos(z_0q_0).$$

Из соотношения $\cos(xy) = \sum_{k=1}^{n} x_k y_k$ имеем:

$$\cos^2(xy) = \sum_{k=1}^{n} x_k^2 y_k^2 + 2\sum_{k=1}^{n-1} (x_k y_k)(x_t y_t)$$

$$(n \geqslant t > k),$$

или

$$M(\cos^2(xy)) = M\left(\sum_{k=1}^{n} x_k^2 y_k^2\right) + 2M\left(\sum_{k=1}^{n-1} (x_k y_k)(x_t y_t)\right). \qquad (6.18)$$

$$(n \geqslant t > k)$$

В силу независимости случайных точек x и y, с учетом симметрии получим:

$$M\left(\sum_{k=1}^{n-1} (x_k y_k)(x_t y_t)\right) = \sum_{k=1}^{n-1} M(x_k x_t)M(y_k y_t) = 0.$$

$(n \geqslant t > k)$ $\qquad$ $(n \geqslant t > k)$

Пользуясь соотношениями

$$x_1 = \cos(xz_0), \quad \sum_{k=1}^{n} M(x_k^2) = 1, \quad \sum_{k=1}^{n} M(y_k^2) = 1,$$

выражением (6.15) и замечанием к утверждению 2, с учетом симметрии из (6.18) получим:

$$M(\cos^2(xy)) = \sum_{k=1}^{n} M(x_k^2)M(y_k^2) = \cos^2(xz_0) \times$$

$$\times\left(\frac{1}{n-1}\sin^2(yq_0)\sin^2(z_0q_0) + \cos^2(yq_0)\cos^2(z_0q_0)\right) +$$

$$+\frac{1}{n-1}\sin^2(xz_0)\left(1 - \frac{1}{n-1}\sin^2(yq_0)\sin^2(z_0q_0) -\right.$$

$$\left. -\cos^2(yq_0)\cos^2(z_0q_0)\right) = \frac{n-2}{(n-1)^2} + \frac{1}{(n-1)^2}(\cos^2(yq_0) + \qquad (6.19)$$

$$+\cos^2(xz_0) + \cos^2(z_0q_0)) - \frac{n}{(n-1)^2}(\cos^2(xz_0)\cos^2(yq_0) +$$

$$+\cos^2(xz_0)\cos^2(z_0q_0) + \cos^2(yq_0)\cos^2(z_0q_0)) +$$

$$+\frac{n^2}{(n-1)^2}\cos^2(xz_0)\cos^2(yq_0)\cos^2(z_0q_0).$$

Отсюда с учетом (6.16) легко получить (6.17).

В частном случае, когда векторы z_0 и q_0 совпадают, т.е. $\cos(z_0q_0)=1$, формулы (6.16), (6.17) и (6.19) совпадают соответственно с формулами (6.11), (6.12) и (6.15).

Теорема синонимии для n-мерной единичной сферы (обобщение пункта (б) утверждения 1)

Пусть x – случайная точка, имеющая равномерное распределение на n-мерной единичной сфере с центром в начале координат, а z_0 и q_0 – произвольные фиксированные точки. Тогда имеет место:

$$M(\cos(xz_0)\cos(xq_0))=\frac{1}{n}\cos(z_0q_0). \qquad (6.20)$$

Действительно, приняв в качестве базиса n-мерного пространства систему векторов, один из которых (например, с индексом 1) проходит через точку z_0, а другой (например, с индексом 2) принадлежит плоскости, содержащей радиусы-векторы точек z_0 и q_0, получим

$$\cos(xz_0)=x_1,\ \cos(z_0q_0)=q_{01},\ \ q_{03}=q_{04}=\ldots=q_{0n}=0,$$

$$\cos(xq_0)=x_1q_{01}+x_2q_{02},\ \cos(xz_0)\cos(xq_0)=q_{01}x_1^2+q_{02}x_1x_2,$$

$$M(\cos(xz_0)\cos(xq_0))=\cos(z_0q_0)M(x_1^2)+q_{02}M(x_1x_2).$$

В силу симметрии и с учетом (6.10) имеем $M(x_1x_2)=0$, $M(x_1^2)=1/n$, откуда следует справедливость (6.20).

В частном случае, когда векторы z_0 и q_0 совпадают, т.е. $\cos(z_0q_0)=1$, выражение (6.20) совпадает с выражением (6.10).

В случае, когда $n=2$, формулировка теоремы синонимии приводит к известной формуле:

$$\int_0^{2\pi}\cos(\alpha-\varphi)\cos\varphi d\varphi=\pi\cos\alpha.$$

Выше мы условились использовать в качестве формальной меры степени подобия двух размытых подмножеств X_i и Y_i значение критерия $r(X_i, X_j)$, численно равного величине $\cos(X_i^*, X_j^*)$ или, что то же самое, скалярному произведению x_ix_j, где x_i и x_j – отображения соответствующих размытых подмножеств на единичную сферу с центром в начале координат, все радиусы-векторы которой перпендикулярны

вектору $E(1, 1,\ldots,1)$. В рамках настоящего раздела рассматривались случаи, когда интересущие нас случайные точки имели непрерывные распределения на единичной сфере с центром в начале координат или на соответствующих участках ее поверхности. При решении же ряда прикладных задач, связанных с построением систем динамического взаимодействия различных стратегий анализа, приходится иметь дело с размытыми подмножествами, соответствующие которым случайные точки имеют дискретные распределения на соответствующих точках единичной сферы. В некоторых из этих случаев, представляющих определенный практический интерес, имеют место теоремы – дискретные аналоги рассмотренных только что теорем транзитивности и синонимии. В частности, такие аналоги имеют место при рассмотрении важнейшего класса обычных (четких) подмножеств, или соответствующих им бинарных векторов.

6.2. ТЕОРЕМЫ ТРАНЗИТИВНОСТИ И СИНОНИМИИ (СЛУЧАЙ *n*-МЕРНОГО КУБА)

Задание каждого обычного подмножества X_i, определенного на универсальном множестве U из n элементов, сводится к заданию функции принадлежности $\mu_i(u_t)$, которая выделяет из n элементов множества U те элементы, которым соответствуют значения $\mu_i(u_t) = 1$, т.е. те, которые принадлежат подмножеству X_i. Тем самым выделяются также те элементы, которые не принадлежат подмножеству X_i, т.е. те, которым соответствуют значения $\mu_i(u_t) = 0$. Элементы второго типа в совокупности представляют подмножество $\overline{X}_i$ – дополнение подмножества X_i до универсального множества U. Каждому обыкновенному подмножеству X_i при этом ставится в соответствие n-мерный вектор $X_i(X_{i1}, X_{i2}, \ldots, X_{in})$, t-я координата которого равна единице или нулю в зависимости от того, принадлежит или нет t-й элемент универсального множества U данному подмножеству X_i. Число различных обычных подмножеств, определенных на универсальном множестве U из n элементов, равно 2^n. Геометрическим местом точек X_i, радиусы-векторы которых представляют эти подмножества, являются 2^n вершин n-мерного куба. Точки $0(0, 0, \ldots, 0)$ и $E(1, 1, \ldots, 1)$ и соответствующие им подмножества – нулевое подмножество и его дополнение, совпадающее с универсальным множеством U, в дальнейшем не будем рассматривать, так как, во-первых, они не представляют практического интереса, а во-вторых, на них не определены значения критерия $r(X_i, X_j)$. Остальные $2^n - 2$ подмножества называются собственными подмножествами множества U. Из основного определения значения критерия $r(X_i, X_j)$ (см. предыдущие разделы) легко убедиться, что для произвольных двух

собственных подмножеств X_i и X_j множества U имеет место

$$r(X_i, X_j) = \frac{n a_{x_i x_j} - a_{x_i} a_{x_j}}{\sqrt{a_{x_i} a_{x_j} (n - a_{x_i})(n - a_{x_j})}}, \tag{6.21}$$

где a_{x_i} – число элементов подмножества X_i, $a_{x_i x_j}$ – число элементов подмножества $X_i \cap X_j$.

Легко убедиться также, что имеют место $r(X_i, X_i) = 1$ и

$$r(X_i, X_j) = r(X_j, X_i) = -r(X_i, \overline{X}_j), \tag{6.22}$$

т.е. $r(X_i, X_j)$ является симметричной мерой подобия подмножеств X_i и X_j. Заметим, что линейное повторение произвольного обыкновенного подмножества X_i совпадает с самим этим подмножеством, а линейное дополнение – с обычным его дополнением $\overline{X}_i$. При этом наличие элемента x_i во множестве радиусов-векторов типа x предопределяет наличие в этом же множестве радиуса-вектора – x_i, соответствующего подмножеству $\overline{X}_i$.

Прежде чем приступить к формулировке и доказательству дискретных аналогов теорем транзитивности и синонимии, приведем доказательство одного соотношения, являющегося обобщением теоремы сложения, которой, как и ее обобщением, будем пользоваться ниже.

Обобщение теоремы сложения

Известная теорема сложения гласит [6]:

$$\sum_{q=0}^{\alpha} C_{\alpha}^{q} C_{n-\alpha}^{\beta-q} = C_{n}^{\beta}. \tag{6.23}$$

Покажем сначала, что имеет место

$$\sum_{q=0}^{\alpha} q^k C_{\alpha}^{q} C_{n-\alpha}^{\beta-q} = \sum_{i=1}^{k} C_{\alpha}^{i} C_{n-i}^{\beta-t} A_k(i), \tag{6.24}$$

где

$$A_k(i) = \sum_{t=1}^{i} C_i^t t^k (-1)^{i-t}.$$

Представим выражение q^k в виде суммы

$$q^k = \sum_{i=1}^{k} a_k(i) \prod_{j=0}^{i-1} (q - j). \tag{6.25}$$

Подставляя здесь поочередно $q = 1, 2, \ldots, k$, относительно $a_k(i)$ полу-

чим систему линейных уравнений, решение которой приводит к

$$a_k(i) = \frac{1}{i!}\sum_{t=1}^{i}(-1)^{i+t}C_i^t t^k. \qquad (6.26)$$

Путем очевидных преобразований отсюда, с учетом (6.25), получим:

$$q^k C_\alpha^q = \sum_{i=1}^{k} a_k(i) C_\alpha^q \prod_{j=0}^{i-1}(q-j) = \sum_{i=1}^{k} C_\alpha^i C_{\alpha-i}^{q-i} A_k(i), \qquad (6.27)$$

где

$$A_k(i) = a_k(i) i! = \sum_{t=1}^{i} C_i^t t^k (-1)^{i+t}. \qquad (6.28)$$

С учетом (6.27) имеем:

$$\sum_{q=0}^{\alpha} q^k C_\alpha^q C_{n-\alpha}^{\beta-q} = \sum_{q=0}^{\alpha}\sum_{i=1}^{k} C_\alpha^i C_{\alpha-i}^{q-i} C_{n-\alpha}^{\beta-q} A_k(i) =$$

$$= \sum_{i=1}^{k} C_\alpha^i A_k(i) \sum_{q=0}^{\alpha} C_{\alpha-i}^{q-i} C_{n-\alpha}^{\beta-q}. \qquad (6.29)$$

Введя новую переменную $z = q - i$ и пользуясь теоремой сложения (6.23), получим:

$$\sum_{q=0}^{\alpha} C_{\alpha-i}^{q-i} C_{n-\alpha}^{\beta-q} = \sum_{z=-i}^{\alpha-i} C_{\alpha-i}^{z} C_{(n-i)-(\alpha-i)}^{(\beta-i)-z} = C_{n-i}^{\beta-i}. \qquad (6.30)$$

Подставляя этот результат в (6.29), приходим к (6.24). Из (6.24), (6.25) и (6.28) легко получить соотношение

$$\sum_{q=i}^{\alpha} C_\alpha^q C_{n-\alpha}^{\beta-q} C_q^i = C_\alpha^i C_{n-i}^{\beta-i}, \qquad (6.24а)$$

которое при $i = 0$ совпадает с (6.23). Отсюда, как и, впрочем, из (6.24), легко получить:

$$\sum_{q=0}^{\alpha} q C_\alpha^q C_{n-\alpha}^{\beta-q} = \alpha C_{n-1}^{\beta-1}, \qquad (6.31)$$

$$\sum_{q=0}^{\alpha} q^2 C_\alpha^q C_{n-\alpha}^{\beta-q} = \frac{\alpha(n-\alpha-\beta+\alpha\beta)}{\beta-1} C_{n-2}^{\beta-2}. \qquad (6.32)$$

Перейдем к доказательству дискретных аналогов утверждений и теорем, рассмотренных в предыдущем разделе. Будем рассматривать множество V, элементами которого служат $2^n - 2$ точки – все вершины n-мерного единичного куба за исключением точек $0(0, 0, \ldots, 0)$ и $E(1, 1, \ldots, 1)$. Напомним, что это точки – отображения в n-мерном пространстве всех возможных собственных подмножеств X_i универсального множества U.

Утверждение 3 (дискретный аналог утверждения 1).

Пусть X – случайная точка, имеющая равномерное распределение на элементах множества V, а Y_0 – произвольный фиксированный элемент этого множества. Тогда имеют место:

$$\text{а) } M(r(X, Y_0)) = 0 \tag{6.33}$$

$$\text{б) } M(r^2(X, Y_0)) = D(r(X, Y_0)) = 1/(n-1), \tag{6.34}$$

или для радиусов-векторов $x = X^* / |X^*|$ и $y_0 = Y_0^* / |Y_0^*|$

$$\text{в) } M(\cos(xy_0)) = 0, \tag{6.33а}$$

$$\text{г) } M(\cos^2(xy_0)) = D(\cos(xy_0)) = 1/(n-1). \tag{6.34а}$$

Справедливость пункта (а) или, что то же самое, пункта (в) непосредственно следует из симметричности распределения случайной величины $r(X, Y_0) = \cos(xy_0)$, в чем легко убедиться из (6.22).

Для доказательства пункта (б) или, что то же самое, пункта (г), пользуемся соотношениями (6.23), (6.31) и (6.32). Покажем, что (6.34) имеет место при произвольном фиксированном значении a_x, что, очевидно, будет достаточным для утверждения (6.34) в общем случае, т.е. при равномерном распределении случайной точки X на всех элементах множества V.

Из (6.21) с учетом (6.23), (6.31), (6.32) и (6.33) получим:

$$M(r^2(X, Y_0)) = D(r(X, Y_0)) =$$

$$= \frac{1}{C_n^{a_x}} \sum_{a_{xy_0}=0}^{a_{y_0}} \frac{(a_{xy_0} n - a_{x_0} a_{y_0})^2}{a_x a_{y_0} (n-a_x)(n-a_{y_0})} C_{a_{y_0}}^{a_{xy_0}} C_{n-a_{y_0}}^{a_x - a_{xy_0}} =$$

$$= \frac{1}{a_x a_{y_0} (n - a_x (n - a_{y_0}) C_n^{a_x}} \left(n^2 \sum_{a_{xy_0}=0}^{a_{y_0}} a_{xy_0}^2 C_{a_{y_0}}^{a_{xy_0}} C_{n-a_{y_0}}^{a_x - a_{xy_0}} - \right.$$

$$\left. -2n a_x a_{y_0} \sum_{a_{xy_0}=0}^{a_{y_0}} a_{xy_0} C_{a_{y_0}}^{a_{xy_0}} C_{n-a_{y_0}}^{a_x - a_{xy_0}} + a_x^2 a_{y_0}^2 \sum_{a_{xy_0}=0}^{a_{y_0}} C_{a_{y_0}}^{a_{xy_0}} C_{n-a_{y_0}}^{a_x - a_{xy_0}} \right) = \frac{1}{n-1}.$$

Замечание к утверждению 3.

Из соображений симметрии легко убедиться, что утверждение 3 остается в силе при замене фиксированной точки Y_0 случайной точкой Y, имеющей произвольное независимое от X распределение на элементах множества V.

Утверждение 4.

Пусть X – случайная точка, имеющая равномерное распределение на элементах множества V, определенных заданным значением $r(X, Z_0)$, где Z_0 – произвольный фиксированный элемент этого множества. Тогда для точек x и z_0, соответствующих X и Z_0, имеет место

$$M(x - \Pr_{z_0} x) = 0. \tag{6.35}$$

Из (6.21) следует, что заданное значение $r(X, Z_0)$ в общем случае может быть достигнуто при различных парах значений a_{xz_0} и a_x. Покажем, что для произвольной фиксированной пары значений a_{xz_0} и a_x имеет место $M(x - \Pr_{z_0} x) = 0$, что, очевидно, является достаточным условием для утверждения 4 в общем случае. Для удобства изложения в дальнейшем координатные представления n-мерных векторов будем рассматривать как соответствующие n-разрядные коды.

Будем различать так называемые зоны нулей и единиц, включающие разряды, где код вектора Z_0 содержит соответственно нули и единицы. Значения каждого разряда кода вектора z_0 в этих зонах соответственно равны

$$-\frac{\sqrt{a_{z_0}}}{\sqrt{n(n-a_{z_0})}} \quad \text{и} \quad \frac{\sqrt{n-a_{z_0}}}{\sqrt{na_{z_0}}},$$

т.е. соответствующие разряды кода вектора $\Pr_{z_0} x$ равны

$$\frac{\sqrt{a_{z_0}}\, r(X, Z_0)}{\sqrt{n(n-a_{z_0})}} \quad \text{и} \quad \frac{\sqrt{n-a_{z_0}}\, r(X, Z_0)}{\sqrt{na_{z_0}}}. \tag{6.36}$$

Легко показать, что при фиксированных значениях a_{xz_0} и a_x средние значения каждого разряда кода вектора x в зонах нулей и единиц соответственно равны:

$$\frac{a_x a_{z_0} - n a_{xz_0}}{(n-a_{z_0})\sqrt{na_x(n-a_x)}} \quad \text{и} \quad \frac{na_{xz_0} - a_x a_{z_0}}{a_{z_0}\sqrt{na_x(n-a_x)}},$$

откуда, с учетом (6.21) и (6.36), непосредственно следует справедливость (6.35).

Утверждение 5 (дискретный аналог утверждения 2).

Пусть X и Y – независимые случайные точки, имеющие равномерные распределения на соответствующих элементах множества V, определенных заданными значениями $r(X, Z_0)$ и $r(Y, Z_0)$, где Z_0 – произвольный фиксированный элемент этого множества. Тогда имеет место дискрет-

ный аналог пункта (а) утверждения 2:

$$\text{а) } M(r(X, Y)) = r(X, Z_0)r(Y, Z_0), \tag{6.37}$$

или для радиусов-векторов x, y и z_0:

$$\text{б) } M(\cos(xy)) = \cos(xz_0)\cos(yz_0). \tag{6.37а}$$

Из уравнения

$$\cos(xy) = \cos(xz_0)\cos(yz_0) + (x - \Pr_{z_0} x)(y - \Pr_{z_0} y)$$

с учетом независимости случайных точек X и Y имеем

$$M(\cos(xy)) = \cos(xz_0)\cos(yz_0) + M(x - \Pr_{z_0} x)M(y - \Pr_{z_0} y).$$

Здесь в силу утверждения 4 как $M(x - \Pr_{z_0} x)$, так и $M(y - \Pr_{z_0} y)$ равны нулю, что свидетельствует о справедливости (6.37а) или, что то же самое, (6.37).

Дискретный аналог пункта (б) утверждения 2, а именно:

$$\text{в) } D(r(X,Y)) = \frac{1}{n-2}(1 - r^2(X, Z_0))(1 - r^2(Y, Z_0)), \tag{6.38}$$

или для радиусов-векторов x, y и z_0:

$$\text{г) } D(\cos(xy)) = \frac{1}{n-2}\sin^2(xz_0)\sin^2(yz_0). \tag{6.38а}$$

в общем случае не имеет места. Эти соотношения справедливы лишь в отдельных частных случаях, например, в тривиальных случаях, когда имеет место какое-либо одно из следующих условий:

$$r(X, Y) = \pm\, r(X, Z_0),\ \ r(X, Y) = \pm\, r(Y, Z_0)$$

Во всех этих случаях $(1 - r^2(X, Z_0))(1 - r^2(Y, Z_0)) = 0$, т.е. левая и правая части (6.38) одновременно равны нулю.

Пользуясь (6.23), (6.31) и (6.32) можно показать, что (6.38) имеет место также в ряде нетривиальных случаев, например, когда $a_{z_0} = 1$ или $a_{z_0} = n - 1$.

Замечание к утверждению 5.

Легко заметить, что (6.37) и (6.37а) остаются в силе при замене случайной точки Y произвольным фиксированным элементом Y_0 с заданным значением $r(Y, Z_0)$.

Действительно, при этом

$$M(\cos(xy)) = \cos(xz_0)\cos(yz_0) + (x - \Pr_{z_0} x)(y_0 - \Pr_{z_0} y_0),$$

где в силу утверждения 4 имеет место $M(x - \Pr_{z_0} x) = 0$.

Можно показать, что по крайней мере для рассмотренных нами случаев $a_{z_0} = 1$ и $a_{z_0} = n - 1$ при такой замене остаются в силе также (6.38) и (6.38а).

Теорема транзитивности для n-мерного единичного куба (дискретный аналог теоремы транзитивности для n-мерной сферы, обобщение утверждения 5)

Пусть X и Y – независимые случайные точки, имеющие равномерные распределения на соответствующих элементах множества V, определенных заданными значениями $r(X, Z_0)$ и $r(Y, Q_0)$, где Z_0 и Q_0 произвольные фиксированные элементы множества V. Тогда имеет место дискретный аналог пункта (а) теоремы транзитивности, доказанной для n-мерной сферы:

$$\text{а)}\quad M(r(X,Y)) = r(X,Z_0)r(Y,Q_0)r(Z_0,Q_0), \tag{6.39}$$

или для радиусов-векторов x, y, z_0 и q_0:

$$\text{б)}\quad M(\cos(xy)) = \cos(xz_0)\cos(yq_0)\cos(z_0q_0). \tag{6.39а}$$

Действительно, из уравнения

$$\cos(xy) = \cos(xz_0)\cos(yz_0) + (x - \Pi\text{р}_{z_0}x)(y - \Pi\text{р}_{z_0}y)$$

в силу независимости случайных точек X и Y получим:

$$M(\cos(xy)) = \cos(xz_0)M(\cos(xz_0)) + M(x - \Pi\text{р}_{z_0}x)M(y - \Pi\text{р}_{z_0}y),$$

где в силу утверждения 4 имеет место $M(x - \Pi\text{р}_{z_0}x) = 0$, а в силу замечания к утверждению 5

$$M(\cos(yz_0)) = \cos(yq_0)\cos(q_0z_0),$$

т.е. имеем:

$$M(\cos(xy)) = \cos(xz_0)\cos(yq_0)\cos(q_0z_0),$$

или, что то же самое,

$$M(r(X,Y)) = r(X,Z_0)r(Y,Q_0)r(Q_0,Z_0).$$

В частном случае, когда векторы Z_0 и Q_0 совпадают, т.е. $r(Z_0, Q_0) = 1$, выражение (6.39) совпадает с (6.37). Дискретный аналог пункта (б) теоремы транзитивности для сферы в общем случае не имеет места. Эти соотношения справедливы лишь в отдельных частных случаях. Можно показать, например, что они имеют место в случаях, когда $a_{z_0} = a_{q_0} = 1$, т.е. когда коды векторов Z_0 и Q_0 содержат по одной единице.

Теорема синонимии для n-мерного единичного куба (дискретный аналог теоремы синонимии для n-мерной сферы, обобщение пункта (б) утверждения 3)

Пусть X – случайная точка, имеющая равномерное распределение на элементах множества V, а Z_0 и Q_0 – произвольные фиксированные элементы этого множества. Тогда имеет место:

а) $$M(r(X,Z_0)r(X,Q_0)) = \frac{1}{n-1} r(Z_0,Q_0), \qquad (6.40)$$

или для радиус-векторов x, q_0 и z_0:

б) $$M(\cos(xz_0)\cos(xq_0)) = \frac{1}{n-1} r(z_0q_0). \qquad (6.40а)$$

Приняв в качестве базиса $(n-1)$-мерного пространства систему векторов, один из которых (например, с индексом 1) проходит через точку z_0, а другой (например, с индексом 2) принадлежит плоскости, проходящей через радиусы-векторы точек z_0 и q_0, получим:

$$\cos(xz_0) = x_1,\ \cos(z_0q_0) = q_{01},\ \cos(xq_0) = x_1q_{01} + x_2q_{02},$$

$$\cos(xz_0)\cos(xq_0) = q_{01}x_1^2 + q_{02}x_1x_2, \qquad (6.41)$$

$$M(\cos(xz_0)\cos(xq_0)) = q_{01}M(x_1^2) + q_{02}M(x_1x_2). \qquad (6.42)$$

В силу (6.34а) имеет место

$$M(x_1^2) = M(\cos^2(xz_0)) = \frac{1}{n-1}.$$

Величина $M(x_1x_2)$ равна нулю, в чем легко убедиться, рассматривая группы векторов X, характеризующиеся постоянством значений $x_1 = r(X, Z_0)$. Для каждой из этих групп в силу утверждения 4 имеет место $M(x_2) = 0$, т.е. $M(x_1x_2) = 0$.

Подставляя значения $M(x_1^2) = 1/(n-1)$ и $M(x_1x_2) = 0$ в (6.42), получим (6.40а) или, что то же самое, (6.40).

В частном случае, когда векторы z_0 и q_0 совпадают, т.е. $r(z_0, q_0) = 1$, выражение (6.40) совпадает с (6.34).

Прежде чем перейти к лексико-семантической интерпретации полученных результатов, отметим следующее:

Из всех утверждений и теорем, рассмотренных в настоящем разделе, нас будут интересовать лишь теоремы транзитивности и синонимии, доказанные для случая n-мерного куба. Ниже мы займемся их интерпретацией и укажем пути их конкретного применения при построении аналитических систем различного назначения. Необходимость же в рассмотрении утверждений и теорем, относящихся к n-мерной сфере, была обусловлена тем, что задачи, касающиеся n-мерного куба, путем операций центрирования и нормирования сводятся к решению соответствующих задач, определенных на n-мерной сфере, и в этом смысле предварительное рассмотрение утверждений и теорем, касающихся этой сферы, как бы подготовило почву для лучшего понимания и восприятия результатов, касающихся n-мерного куба.

Представленные здесь утверждения и теоремы (касающиеся как сферы, так и куба) могут быть использованы также самостоятельно,

вне предмета нашего рассмотрения, например, при необходимости вычисления различных кратных сумм или интегралов, обладающих определенными свойствами симметрии.

6.3. ЛЕКСИКО-СЕМАНТИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ И ПУТИ ПРАКТИЧЕСКОГО ПРИМЕНЕНИЯ ТЕОРЕМ ТРАНЗИТИВНОСТИ И СИНОНИМИИ

В этом разделе приводятся результаты практического истолкования теорем транзитивности и синонимии, приведенного в работе [3].

Теорема транзитивности

В подавляющем большинстве современных баз данных, предназначенных для промышленной эксплуатации, их элементы – конкретные документы, – представлены не одним, а группой описателей их семантики. В качестве таких описателей могут выступать языковые средства различных естественных и/или искусственных языков. Так, в библиографических базах данных, кроме совокупности слов базового естественного языка (приведенных в виде кратких аннотаций, рефератов, ключевых слов и т.п.), каждый вторичный документ может снабжаться также набором индексов – лексических единиц тех или иных классификационных языков, например, набором индексов (рубрик) УДК и/или ББК, и/или ГРНТИ и др. Более того, в ряде случаев во вторичных документах приводятся их заглавия на языках первоисточников, т.е. в них присутствуют также некоторые наборы слов на иностранных языках.

Присутствие во вторичных документах одновременно нескольких логических полей, отводимых под те или иные его описатели, способствует тому, чтобы у пользователя, получившего распечатку вторичных документов, создавалось более полное представление об их первоисточниках. Если к тому же эти поля "поисково активизированы", т.е. их содержимое включено в поисковые образы документов (ПОД), то создается возможность организовать поиск по совокупности поисковых реквизитов, принадлежащих в общем случае различным поисковым полям. Многие из современных информационно-поисковых систем (поисковых оболочек) располагают этой возможностью, что, естественно, способствует некоторому повышению их функциональной эффективности. Вместе с тем, даже при наличии такой возможности, существующие информационно-поисковые системы остаются лишенными "внутреннего интеллекта": их работа сводится к тривиальной проверке документов на наличие в них тех или иных конкретных контекстных ситуаций, сформулированных пользователями в своих запро-

сах. Документы, не содержащие этих контекстных ситуаций, не будут выданы пользователю вне зависимости от того, соответствует ли их семантика семантике пользовательского запроса или нет. Чтобы уменьшить негативные последствия, обусловленные отсутствием взаимно однозначного соответствия между семантическими категориями (идей, понятий, информационных потребностей) и их языковыми формулировками, используются различные предварительно подготовленные пособия, призванные осуществить терминологический контроль и наращивание запросов, сформулированных пользователями. В ряде случаев наличие таких пособий (различных тезаурусов, дескрипторных или иных словарей и т.д.) действительно приводит к некоторому повышению функциональной эффективности автоматизированных систем документального поиска (АСДП). Однако эти пособия, как бы тщательно они ни были составлены, не могут учитывать все многообразие поисковых ситуаций, которые могут возникать при реальной работе различных пользователей с системой. Потерминное наращивание пользовательских запросов (если даже под словом "термины" понимать также отдельные словосочетания) не может учитывать те изменения в семантике одного и того же термина, которые претерпевает она (эта семантика) в зависимости от того, в каком терминологическом или контекстном (лексико-семантическом) окружении фигурирует рассматриваемый термин в конкретных запросах. Иными словами, наличие этих пособий также не внесло ничего принципиально нового в современные информационные технологии. Если в их отсутствие вся ответственность за качество поиска была возложена на самих пользователей, то при их наличии часть этой ответственности берут на себя составители словарей. Что же касается самих АСДП, "напичканных" огромным количеством вторичных документов – ценнейшим интеллектуальным материалом, то, как это ни странно, на них возложены лишь функции пассивного контроля наличия или отсутствия в очередных документах тех или иных контекстных ситуаций. В процессе поиска остается неиспользованным огромный интеллектуальный потенциал, в явном или неявном виде содержащийся в самих базах данных.

Интуиция же подсказывает, что наличие в памяти ЭВМ достаточно большого числа вторичных документов – носителей определенных интеллектуальных усилий их авторов, индексаторов, аналитиков и т.п. – создает реальные предпосылки для использования в процессе поиска интеллектуального потенциала, содержащегося в базах данных. Естественно, что чем большее число документов содержится в базе данных, тем, при прочих равных условиях, большими должны быть аналитические возможности этих систем.

Представляется, например, вполне резонным поручить ЭВМ, чтобы она, после обычного поиска документов, релевантных пользовательскому запросу (независимо от того, подвергся ли этот запрос предварительному терминологическому наращиванию или нет), сама извлекла бы из различных поисковых полей лексические единицы, осуществила

бы их ранжирование по степени соответствия смыслу запроса и на основе такого анализа скомпоновала бы своеобразный микротезаурус, ориентированный на обслуживание конкретно сложившейся поисковой ситуации "конкретный запрос – конкретная база данных". При такой постановке вопроса лексические единицы могут быть извлечены из логических полей, о существовании которых пользователь может и не знать. Например, в этом микротезаурусе могут фигурировать слова английского языка, независимо от того, владеет ли пользователь английским языком или нет. В другом случае в этом микротезаурусе могут фигурировать индексы УДК, независимо от того, знает или нет пользователь о существовании такого информационно-поискового языка. С некоторыми оговорками составление этих микротезаурусов может рассматриваться как некий эквивалент автоматического перевода пользовательского запроса на те или иные языки, в том числе и иностранные. Следует особо подчеркнуть, что такой перевод будет осуществляться без помощи каких-либо словарей (например, англо-русских или русско-английских), а лишь на основе той информации, которая содержится в самой базе данных. Естественно, что степень соответствия этих микротезаурусов, их адекватность конкретно сложившимся поисковым ситуациям будет находиться в прямой зависимости от того, насколько удачно будет осуществлено ранжирование различных лексических единиц по степени их соответствия пользовательским запросам. Для выработки формальной количественной меры для такого ранжирования будем пользоваться доказанной нами теоремой транзитивности (случай n-мерного единичного куба).

Пусть на множестве документов N осуществляется обычный поиск документов, релевантных некоторому запросу. Обозначим искомое подмножество документов через Z. Естественно, что при работе в реальных условиях АСДП выдаст пользователю некоторое подмножество документов Z_0, в общем случае не совпадающее с подмножеством Z. С каждой i-й лексической единицей, подлежащей ранжированию по степени соответствия ее семантики семантике запроса, будем связывать по два подмножества, а именно, подмножество документов I_0, содержащих данную лексическую единицу, и подмножество документов I, связанных с i-м термином на семантическом уровне, вне зависимости от факта присутствия в этих документах рассматриваемого термина.

Если бы имело место взаимно однозначное соответствие семантических категорий их языковым формулировкам, то подмножества Z_0 и I_0 в точности совпали бы с подмножествами соответственно Z и I, а значения коэффициентов корреляций $r(Z_0, Z)$ и $r(I_0, I)$ между этими подмножествами оказались бы равными единице. В реальных же условиях такое соответствие не имеет места, эти подмножества не совпадают, а значения $r(Z_0, Z)$ и $r(I_0, I)$ оказываются меньшими единицы. Поскольку в процессе ранжировки терминов речь идет о конкретном запросе, то

можно считать, что подмножество Z_0 нам задано. Известно также, что существует некоторое конкретное подмножество Z с конкретным значением $r(Z_0, Z)$, но само это подмножество Z, как и значение $r(Z_0, Z)$, нам не заданы. Из определения коэффициента линейной корреляции можно судить о положительности значения $r(Z_0, Z)$, так как отрицательные его значения означали бы, что качество работы АСДП хуже случайной выборки. Также к абсурдному результату мы пришли бы, допустив возможность отрицательных значений $r(I_0, I)$.

Пусть требуется определить, какому из терминов I и J следует отдать предпочтение при их ранжировке по степени соответствия смыслу запроса. Поскольку подмножества Z_0, I_0 и J_0 считаются заданными, то согласно теореме транзитивности имеем:

$$M(r(I,Z)) = r(I_0, I)r(I_0, Z_0)r(Z_0, Z), \tag{6.43}$$

$$M(r(J,Z)) = r(J, J_0)r(J_0, Z_0)r(Z_0, Z). \tag{6.43а}$$

В приведенных формулах нам известны лишь значения $r(I_0, Z_0)$ и $r(J_0, Z_0)$, а в качестве формальной меры соответствия данного термина пользовательскому запросу следовало бы располагать если не конкретными значениями $r(I, Z)$ и $r(J, Z)$, то хотя бы значениями их математических ожиданий, т.е. $M(r(I, Z))$ и $M(r(J, Z))$.

Априори мы не располагаем никакой информацией не только о конкретных значениях $r(I_0, I)$ и $r(J_0, J)$, но и о характере распределения этих случайных величин. Естественно, однако, полагать, что эти значения меняются в довольно узком диапазоне вокруг некоего среднего значения r_L, характерного для каждого естественного и/или искусственного языка. Сами значения r_L, по крайней мере для естественных языков, весьма близки к единице. Исходя из вышеизложенного, с определенными оговорками можно принять, что

$$r(I_0, I) = r(J_0, J) = r_L. \tag{6.44}$$

Из (6.43) и (6.43а) с учетом (6.44) приходим к

$$M(r(I,Z) - r(J,Z)) = \beta(r(I_0, Z_0) - r(J_0, Z_0)), \tag{6.45}$$

где коэффициент

$$\beta = r_L r(Z_0, Z), \tag{6.46}$$

согласно приведенным выше соображениям, всегда положителен.

Исходя из положительности коэффициента β и с учетом (6.45) легко заключить, что количественной мерой соответствия того или иного термина пользовательскому запросу может служить соответствующее ему значение $r(I_0, Z_0)$. Таким образом, если не в абсолютных метрических единицах, то хотя бы в плане сравнительной оценки эти значения являются своеобразной опосредствованной (транзитивной) оценкой значений $r(I, Z)$.

Не следует забывать, что формула (6.44) верна лишь в контексте конкретно рассматриваемой нами задачи. На самом же деле при переходе в рамках фиксированных естественных и/или искусственных языков от одних терминов к другим конкретные значения $r(I_0, I)$ могут отклоняться от среднего значения r_L, присущего данному языку.

Вместе с тем, поскольку нас интересуют не конкретные значения тех или иных рассматриваемых нами величин, а лишь сравнительная их оценка, то в силу взаимных компенсаций упомянутых выше отклонений их наличие не приводит к сколь-либо серьезным негативным последствиям при рассмотрении значений $r(I_0, Z_0)$ и $r(J_0, Z_0)$ вместо значений $r(I, Z)$ и $r(J, Z)$. По крайней мере, об этом свидетельствуют результаты промышленной эксплуатации микротезаурусов, где термины ранжируются с использованием описанного выше механизма.

Теорема синонимии

Выше уже говорилось о роли предварительно подготовленных дескрипторных словарей, тезаурусов при эксплуатации автоматизированных систем документального поиска. Важное место при подготовке этих словарей занимает задача формирования синонимических рядов, с помощью которых осуществляется терминологическое наращивание пользовательских запросов.

Традиционно терминологические пособия составлялись до начала эксплуатации АСДП, без учета того, с какими именно базами данных приходилось работать пользователям. Со временем специалисты убедились в необходимости адаптации этих пособий к конкретным предметным областям. Появились так называемые отраслевые словари, в процессе эксплуатации которых специалисты обнаружили необходимость дальнейшей их корректировки с учетом специфики конкретных баз данных. Составление этих пособий неразрывно связано с огромными затратами интеллектуального труда специалистов-лингвистов. Что же касается адаптации этих пособий к тем или иным поисковым средам, то эту работу нельзя выполнять без помощи специалистов по рассматриваемым предметным областям. Непременное присутствие при этом элементов субъективизма, с одной стороны, и большие затраты интеллектуального труда – с другой, привели к необходимости формализовать процедуру подготовки синонимических рядов с тем, чтобы частично или полностью поручить эту работу ЭВМ. Возникла необходимость разработки формальных количественных мер по оценке степени синонимичности различных терминов. Это довольно трудная задача, если учесть, что само определение "слово, совпадающее или близкое по значению с другим словом" носит подчеркнуто семантический характер, трудно поддающийся формализации. Для решения этой задачи введем в рассмотрение несколько иное определение синонимии, которое вкупе с доказанной нами теоремой синонимии для

n-мерного куба и будем использовать при разработке формальной количественной меры синонимичности слов.

Пусть мы располагаем некоторой конкретной базой данных и применительно к этой базе требуется оценить степень синонимичности терминов i и j. Подмножества документов, содержащие эти термины, обозначим соответственно через I_0 и J_0, а формальную количественную меру степени их синонимичности – через $s(i, j)$. При определении последней будем исходить из соображений, на которые ранее уже обратили внимание специалисты-лингвисты:

1) шансов на то, что в рамках одного и того же вторичного документа встретятся два слова-синонима, меньше обычного, так как при подготовке этих документов их авторы или референты обычно придерживаются какого-либо одного из терминов данного синонимичного ряда.

2) если некоторая пара терминов i и x встречается в документах чаще обычного, а термин j является синонимом термина i, то следует ожидать, что чаще обычного будут встречаться также термины j и x. Действительно, если, например, термины i и j выражены именами существительными и прилагательное x является характерным описателем термина i, то оно будет также характерным описателем термина j.

Формальную количественную меру синонимичности $s(i, j)$ терминов i и j определим как

$$s(i,j)=\frac{1}{m}\sum_{k=1}^{m} r(X_k, I_0)r(X_k, J_0)-\frac{1}{n-1}r(I_0, J_0), \qquad (6.47)$$

где

n – число документов в рассматриваемой базе данных;

m – число терминов в инверсном списке, т.е. число терминов, которые хоть раз встретились в каком-либо одном или более документах рассматриваемой базы данных;

X_k – подмножество документов, содержащих k-й термин инверсного списка.

Из теоремы синонимии следует, что в гипотетическом случае, когда случайная точка X имеет равномерное распределение на элементах множества V, для любой фиксированной пары терминов i и j величина $s(i, j)$ равна нулю. В реальных же условиях, в силу наличия определенных лексико-семантических корреляционных связей, значения $s(i, j)$ могут отличаться от нуля. Так, если I_0 и J_0 случайные точки, имеющие равномерные распределения на элементах множества V, то в силу замечания к утверждению 3 имело бы место $M(r(I_0, J_0)) = 0$. В реальных же условиях, в силу соображений, сформулированных в пункте (1), можно утверждать, что если термины i и j синонимы, то значения $r(I_0, J_0)$ склоняются к отрицательным величинам, т.е. $M(r(I_0, J_0)) < 0$. В силу же

же соображений, сформулированных в пункте (2), можно утверждать, что если рассматриваемые термины i и j являются синонимами, то величина

$$\frac{1}{m}\sum_{k=1}^{m} r(X_k,\ I_0)r(X_k,\ J_0)$$

окажется больше величины

$$M(r(X,\ I_0)r(X,\ J_0)),$$

вычисленной из расчета равномерного распределения случайной точки X на элементах множества V.

Таким образом, мы приходим к заключению, что чем сильнее выражены свойства синонимичности терминов i и j, тем, при прочих равных условиях, большим становится в формуле (6.47) значение уменьшаемой (соображение 2) и тем меньшим – значение вычитаемой (соображение 1). В результате можно утверждать, что большие значения $s(i, j)$ свидетельствуют о том, что свойства синонимичности между терминами i и j применительно к данной базе данных выражены сильнее, и наоборот.

6.4. R-ПРОИЗВЕДЕНИЕ МАТРИЦ. ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

В предыдущем разделе мы убедились в том,что значения $r(I_0, Z_0)$ могут быть использованы в качестве формальной количественной меры при ранжировке различных лексических единиц по степени их семантического родства с пользовательскими запросами. Как мы убедимся ниже, процедуру такой ранжировки можно свести к одной матричной операции, а именно, R-произведению матрицы размерности $m \times n$ на вектор-столбец размерности n, где m – число терминов, подлежащих ранжировке, n – число документов в базе данных. Эта операция является частным случаем операции R-произведения матрицы размерности $m \times n$ на матрицу размерности $n \times q$. В настоящем разделе мы приведем основное определение R-произведения матриц и укажем на ряд специфических свойств этой операции.

Пусть $a(a_1,\ a_2,\ldots,\ a_n)$ – произвольный n-мерный вектор-строка и требуется осуществить его центрирование, т.е. требуется от этого вектора перейти к вектору $a_0(a_{01},\ a_{02},\ldots,\ a_{0n})$, где $a_{0k} = a_k - m_a$, а $m_a = \frac{1}{n}\sum_{k=1}^{n} a_k$ (см. предыдущие разделы).

Введем в рассмотрение квадратную матрицу $S_n[s_{n_{ij}}]$ порядка n, где

$$s_{n_{ij}} = \begin{cases} \dfrac{n-1}{n} & \text{при } j = i \\ -\dfrac{1}{n} & \text{при } j \neq i. \end{cases} \qquad (6.48)$$

Легко убедиться, что для центрирования произвольного вектора a размерности n достаточно умножить этот вектор на матрицу S_n, т.е. $a_0 = aS_n$.

Произведение же произвольной прямоугольной матрицы $A[a_{ij}]$ размерности $m \times n$ на матрицу S_n осуществляет центрирование всех m векторов-строк матрицы A. Так, все векторы-строки матрицы $A_0 = AS_n$ суть центрированные векторы-строки соответствующих строк матрицы A.

Матрица S_n обладает рядом замечательных свойств. Можно показать, например (см. приложение 2), что характеристический многочлен этой матрицы определяется формулой

$$S_n(\lambda) = \lambda(\lambda - 1)^{n-1}, \tag{6.49}$$

что свидетельствует о ее вырожденности. Ее ранг равен $n - 1$. Для произвольного натурального k имеет место

$$S_n^k = S_n. \tag{6.50}$$

Пусть теперь рассматривается вектор-столбец $b(b_1, b_2, \ldots, b_n)$ и требуется осуществить его центрирование. Легко убедиться, что для этого достаточно матрицу S_n умножить на этот вектор, т.е. вектор-столбец

$$b_0 = S_n b \tag{6.51}$$

и есть результат центрирования вектора-столбца b.

Произведение же матрицы S_n на произвольную прямоугольную матрицу $B[b_{ij}]$ размерности $n \times q$ осуществляет центрирование всех q векторов-столбцов матрицы B. Так, все векторы-столбцы матрицы S_nB суть центрированные векторы-столбцы соответствующих столбцов матрицы B.

Пусть теперь задана матрица A размерности $m \times n$, не содержащая нулевых строк, и требуется осуществить нормирование всех ее m строк. Очевидно, этого можно добиться в результате матричного умножения

$$A_w = H_A A, \tag{6.52}$$

где $H_A[h_{a_{ij}}]$ диагональная матрица порядка m с элементами

$$h_{a_{ii}} = 1 \Big/ \sqrt{\sum_{k=1}^{n} a_{ik}^2}. \tag{6.53}$$

Все векторы-строки матрицы A_w суть нормированные векторы-строки соответствующих строк матрицы A.

Если же задана матрица B размерности $n \times q$, не содержащая нуле-

вых столбцов, и требуется осуществить нормирование всех ее q столбцов, то этого можно добиться операцией

$$B_w = BF_B, \tag{6.54}$$

где $F_B[f_{b_{ij}}]$ диагональная матрица порядка q с элементами

$$f_{b_{ii}} = 1\Big/\sqrt{\sum_{k=1}^{n} b_{ki}^2}. \tag{6.55}$$

Рассмотрим пару матриц A и B размерностей соответственно $m \times n$ и $n \times q$, таких, что ни одна из строк матрицы A и ни один из столбцов матрицы B не является коллинеарным вектору E (1, 1,..., 1). Это означает, что все векторы-строки матрицы $A_0 = AS_n$ суть центрированные векторы-строки соответствующих строк матрицы A, причем матрица A_0 не содержит нулевых строк. Аналогично, все векторы-столбцы матрицы $B_0 = S_nB$ суть центрированные векторы-столбцы соответствующих столбцов матрицы B, причем матрица B_0 не содержит нулевых столбцов. Отсутствие нулевых строк у матрицы A_0 позволяет подвергать ее операции построчной нормировки, т.е. рассматривать матрицу

$$A_{0w} = H_{AS_n} AS_n = H_{A_0} A_0. \tag{6.56}$$

Аналогично, отсутствие нулевых столбцов у матрицы B_0 позволяет подвергать ее операции постолбцовой нормировки, т.е. рассматривать матрицу

$$B_{0w} = S_n BF_{S_nB} = B_0 F_{B_0}. \tag{6.57}$$

Определение

R-произведением матриц A и B будем называть матрицу

$$A * B = H_{A_0} AS_n \cdot S_n BF_{B_0}. \tag{6.58}$$

Из этого определения следует, что R-произведение матриц A и B отличается от обычного матричного их произведения лишь тем, что вместо значений скалярных произведений i-х строк матрицы A с j-ми столбцами матрицы B берутся значения коэффициентов линейной корреляции этих векторов.

Очевидно, из существования матрицы $A*B$ вовсе не следует существование матрицы $B*A$. Независимо от того, каковыми являются элементы этих матриц a_{ij} и b_{ij}, если матрица $A*B$ существует, то все ее элементы нормированы в интервале [–1, 1]. Как и в случае обычного произведения матриц $A \cdot B$, если A и B матрицы размерностей $m \times n$ и $n \times q$, то матрица $A*B$ имеет размерность $m \times q$.

В частном случае, когда матрица A имеет размерность $1 \times n$, то в формуле (6.58) вместо матрицы H_{A_0} фигурирует обычный множитель, равный $h_{A_0} = 1/|A_0|$. Аналогично, когда матрица B имеет размерность $n \times 1$, то в формуле (6.58) вместо матрицы F_{B_0} фигурирует обычный множитель, равный $f_{B_0} = 1/|B_0|$.

При рассмотрении матричных моделей документального поиска (раздел 5.6) мы обратили внимание на неправомочность использования моделей, оперирующих обычным произведением матриц. Напомним, что это было обусловлено использованием при обычном умножении матриц скалярного произведения соответствующих векторов, которое не может служить корректной мерой оценки степени подобия этих векторов. Модели же, где используется R-произведения матриц, свободны от этого недостатка,и поэтому представляется резонным вернуться к рассмотрению процессов (5.62) и (5.69),заменив в последних обычные операции умножения матриц на их R-произведение:

$$\begin{aligned} A^{(0)} &= C * Q^{(0)} \\ Q^{(1)} &= C^T * A^{(0)} \\ A^{(1)} &= C * Q^{(1)} \\ &\cdots \\ A^{(t)} &= C * Q^{(t)} \\ Q^{(t+1)} &= C^T * A^{(t)} \\ &\cdots \end{aligned} \qquad (6.59)$$

Здесь $A^{(0)}$ – n-мерный вектор, представляющий подмножество (в общем случае нечеткое) документов, выданных системой в ответ на первоначально сформулированный пользовательский запрос $Q^{(0)}$. Каждая i-я строка матрицы C^T представляет размытое подмножество $I_{(i)}$, соответствующее i-му термину инверсного списка. Руководствуясь соображениями, изложенными в разделе 6.3, легко заключить, что m-мерный вектор $Q^{(1)}$ и есть ранжированный по величине $r(I_{(i)}, A^{(0)})$ инверсный список терминов. Таким образом, при заданной матрице C^T размерности $m \times n$, операция $Q^{(1)} = C^T * A^{(0)}$ осуществляет R-отображение на множество терминов подмножества $A^{(0)}$, определенного на множестве документов. В свою очередь, операция $A^{(1)} = C * Q^{(1)}$ осуществляет R-отображение на множество документов подмножества $Q^{(1)}$, определенного на множестве терминов. В рассматриваемом смысле подмножество $A^{(1)}$ является однократным R-отражением подмножества $A^{(0)}$, определенного на множестве документов, на это же множество. Далее, однократное R-отражение подмножества $A^{(t-1)}$, т.е подмножество $A^{(t)}$, будем называть t-кратным отражением подмножества $A^{(0)}$.

Аналогично вышеизложенному, будем говорить, что операция $A^{(0)} = C^* Q^{(0)}$ осуществляет R-отображение на множество документов подмножества $Q^{(0)}$, определенного на множестве терминов. Подмножество же $Q^{(1)} = C^T * A^{(0)}$ может рассматриваться как однократное R-отражение подмножества $Q^{(0)}$, определенного на множестве терминов. Далее, однократное R-отражение подмножества $Q^{(t-1)}$, т.е. подмножество $Q^{(0)}$, будем называть t-кратным отражением подмножества $Q^{(0)}$.

Нас будет интересовать поведение подмножеств (векторов) $A^{(t)}$ и $Q^{(t)}$при $t \to \infty$. Поскольку во всех уравнениях системы (6.59) мы имеем дело с R-произведениями матриц на соответствующие векторы-столбцы, то вместо постолбцово нормирующих матриц у нас будут фигурировать обычные множители, равные обратным величинам модулей этих векторов после их центрирования. Как и раньше, эти множители мы обозначим буквой f, но снабдим их только именем вектора-столбца, опустив индекс "0". Этот индекс мы опустим также в выражении построчно-нормирующей матрицы. Там, где это не может привести к недоразумениям, мы опустим также индекс, указывающий на порядок центрирующей матрицы S. И, наконец, исходя из свойства матрицы S_n, приведенной в (6.50), везде, где фигурирует произведение $S_n \cdot S_n$, его заменим на S_n.

Из (6.59) имеем:

$$
\begin{aligned}
A^{(0)} &= H_C CSQ^{(0)} f_{Q^{(0)}} \\
Q^{(1)} &= H_{C^T} C^T SA^{(0)} f_{A^{(0)}} \\
A^{(1)} &= H_c CSQ^{(1)} f_{Q^{(1)}} \\
&\cdots \\
A^{(t)} &= H_c CSQ^{(t)} f_{Q^{(t)}} \\
A^{(t+1)} &= H_{C^T} C^T SA^{(t)} f_{A^{(t)}} \\
&\cdots
\end{aligned}
\qquad (6.60)
$$

Введем обозначения:

$$G = H_{C^T} C^T SH_C CS, \qquad (6.61)$$

$$L = H_C CSH_{C^T} C^T S, \qquad (6.62)$$

$$\gamma_t = f_{Q^{(0)}} f_{A^{(0)}} f_{Q^{(1)}} f_{A^{(1)}} \cdots f_{Q^{(t-1)}} f_{A^{(t-1)}} \qquad (6.63)$$

при $t > 0$ и $\gamma_t = 1$ при $t = 0$.

Тогда из (6.60) легко установить, что

$$Q^{(t)} = G^t \gamma_t Q^{(0)}. \tag{6.64}$$

Примем, что в спектре собственных значений матрицы G содержится старшее по модулю собственное значение λ_0. Тогда, согласно теореме Сильвестра из (6.64), следует, что при достаточно больших t имеет место [4, 5, 6]:

$$Q^{(t+1)} = GG^t \cdot \gamma_{t+1} Q^{(0)} = \lambda_0 \cdot G^t \cdot \gamma_{t+1} Q^{(0)}, \tag{6.65}$$

или, пользуясь (6.63):

$$GG^t \gamma_t Q^{(0)} = \lambda_0 G^t \gamma_t Q^0, \tag{6.65а}$$

т.е.

$$GQ^{(t)} = \lambda_0 Q^{(t)}. \tag{6.65б}$$

Отсюда следует, что при $t \to \infty$ вектор $Q^{(t)}$ стремится принимать направление собственного вектора матрицы G, соответствующего ее старшему собственному значению λ_0. В рассматриваемом смысле при $t \to \infty$ поведение вектора $Q^{(t)}$ в процессе (6.60) аналогично поведению этого вектора в процессе (5.62), с той лишь разницей, что при рассмотрении процесса (6.60) вместо матрицы C^TC приходится иметь дело с матрицей

$$G = H_{C^T} C^T S H_C C S,$$

т.е. с матрицей-произведением тех же матриц C^T и C, но лишь после их предварительного построчного центрирования и нормирования.

Из (6.60) и (6.64) с учетом (6.62) и (6.63) легко получить выражение для $A^{(t)}$:

$$A^{(t)} = L^t A^{(0)} \gamma_t \cdot f_{Q^{(t)}} / f_{Q^{(0)}}. \tag{6.66}$$

Исходя из (6.61) и (6.62) можно утверждать, что по крайней мере для ненулевых значений λ_0 спектры собственных значений матриц G и L совпадают. Т.е. если в спектре матрицы G собственное значение λ_0 является старшим по модулю, то это же значение будет старшим по модулю в спектре собственных значений матрицы L. И тогда из (6.66) легко установить, что

$$LA^{(t)} = \lambda_0 A^{(t)}, \tag{6.67}$$

т.е. при $t \to \infty$ вектор $A^{(t)}$ стремится принимать направление собственного вектора матрицы L, соответствующего ее старшему собственному значению λ_0.

Из (6.65) и (6.66) легко установить, что при $t \to \infty$

$$Q^{(t+1)} = \lambda_0 f_{Q^{(t)}} f_{A^{(t)}} Q^{(t)}, \tag{6.68}$$

$$A^{(t+1)} = \lambda_0 f_{A^{(t)}} f_{Q^{(t+1)}} A^{(t)}. \tag{6.69}$$

Из основного определения коэффициента корреляции нетрудно убедиться в справедливости

$$r(k_1 x,\ k_2 y) = r(x,\ y)\,\mathrm{sgn}(k_1 \cdot k_2). \tag{6.70}$$

Из (6.59), (6.68) и (6.70) с учетом положительности всех f получим

$$Q^{(t+1)} = Q^{(t)}\,\mathrm{sgn}(\lambda_0), \tag{6.71}$$

$$A^{(t+1)} = A^{(t)}\,\mathrm{sgn}(\lambda_0), \tag{6.72}$$

т.е. при $t \to \infty$ векторы $Q^{(t)}$ и $A^{(t)}$ сходятся не только по направлению, но и по модулю. Так, если $\lambda_0 > 0$, то имеет место

$$Q^{(t+1)} = Q^{(t)}, \tag{6.73}$$

$$A^{(t+1)} = A^{(t)}, \tag{6.74}$$

если же $\lambda_0 < 0$, то имеют место

$$Q^{(t+1)} = -Q^{(t)}, \tag{6.75}$$

$$A^{(t+1)} = -A^{(t)}. \tag{6.76}$$

Из (6.68) и (6.69) с учетом (6.71) и (6.72) можно заключить также, что при достаточно больших t имеет место

$$|Q^{(t)}||A^{(t)}| = |\lambda_0|. \tag{6.77}$$

Замена в процессе (5.62) операции обычного умножения матриц на их R-произведение делает этот процесс более корректным, что, естественно, приводит к заметному повышению функциональной эффективности АСДП, по крайней мере, в начальных тактах ее функционирования. Вместе с тем, процесс (6.59) также, как и процесс (5.62), "грешит" забвением при достаточно больших значениях t первоначально сформулированного пользователем запроса, т.е. вектора $Q^{(0)}$. Это приводит к тому, что повышение качества поиска наблюдается лишь в первых нескольких тактах функционирования АСДП, после чего дальнейшее продолжение процесса (6.59) приводит к резкому ухудшению качества поиска. В рассматриваемом смысле представляет интерес R-аналог

процесса (5.69), а именно, процесс

$$
\begin{aligned}
A^{(0)} &= C * Q^{(0)} \\
Q^{(1)} &= C^T * A^{(0)} + Q^{(0)} \\
A^{(1)} &= C * Q^{(1)} \\
&\cdots \\
A^{(t)} &= C * Q^{(t)} \\
Q^{(t+1)} &= C^T * A^{(t)} + Q^{(0)} \\
&\cdots
\end{aligned}
\qquad (6.78)
$$

Отсюда следует

$$
\begin{aligned}
A^{(0)} &= H_C CSQ^{(0)} f_{Q^{(0)}} \\
Q^{(1)} &= H_{C^T} C^T SA^{(0)} f_{A^{(0)}} + Q^{(0)} \\
A^{(1)} &= H_C CSQ^{(1)} f_{Q^{(1)}} \\
&\cdots \\
A^{(t)} &= H_C CSQ^{(t)} f_{Q^{(t)}} \\
Q^{(t+1)} &= H_{C^T} C^T SA^{(t)} f_{A^{(t)}} + Q^{(0)} \\
&\cdots
\end{aligned}
\qquad (6.79)
$$

Пользуясь введенными выше обозначениями, получим:

$$
Q^{(t)} = \gamma_t \left(\sum_{p=0}^{t} G^p / \gamma_{t-p} \right) Q^{(0)}. \qquad (6.80)
$$

Из (6.61) и (6.62) следуют

$$
H_C CSG^p = L^p H_C CS, \qquad (6.81)
$$

$$
H_{C^T} C^T L^p = G^p H_{C^T} C^T S, \qquad (6.81а)
$$

с учетом которых из (6.79) и (6.80) легко получить:

$$
A^{(t)} = \gamma_t \frac{f_{Q^{(t)}}}{f_{Q^{(0)}}} \left(\sum_{p=0}^{t} L^p / \gamma_{t-p} \right) A^{(0)}. \qquad (6.82)
$$

Из сопоставления (6.80) и (6.64) легко обнаружить, что в отличие от процесса (6.59), в процессе (6.78), благодаря наличию слагаемого $Q^{(0)}$, вектор $Q^{(t)}$ зависит от вектора $Q^{(0)}$ при произвольных сколь угодно больших значениях t.

Пусть для некоторого значения $t = t_0$ имело место

$$Q^{(t_0+1)} = Q^{(t_0)}. \tag{6.83}$$

Очевидно, это повлекло бы

$$A^{(t_0+1)} = A^{(t_0)} \tag{6.83а}$$

и далее

$$Q^{(t_0+p)} = Q^{(t_0)}, \tag{6.84}$$

$$A^{(t_0+p)} = A^{(t_0)} \tag{6.84а}$$

для любого $p = 0, 1, 2,\ldots$

Из (6.79) следует, что

$$Q^{(t+1)} - Q^{(t)} = (Gf_{Q^{(t)}} f_{A^{(t)}} - 1)Q^{(t)} + Q^{(0)}, \tag{6.85}$$

$$A^{(t+1)} - A^{(t)} = (Lf_{A^{(t)}} f_{Q^{(t+1)}} - 1)A^{(t)} + H_C CSf_{Q^{(t+1)}} Q^{(0)}, \tag{6.86}$$

т.е. соблюдение условия (6.83) повлекло бы:

$$(1 - Gf_{Q^{(t_0)}} f_{A^{(t_0)}})Q^{(t_0)} = Q^{(0)}, \tag{6.87}$$

$$(1 - Lf_{Q^{(t_0)}} f_{A^{(t_0)}})A^{(t_0)} = H_C CSf_{Q^{(t_0)}} Q^{(0)}. \tag{6.88}$$

Но для того, чтобы существовали $Q^{(t_0)}$ и $A^{(t_0)}$, необходимо и достаточно, чтобы матрицы $(1 - Gf_{Q^{(t_0)}} f_{A^{(t_0)}})$ и $(1 - Lf_{Q^{(t_0)}} f_{A^{(t_0)}})$ были обратимы, т.е. чтобы все собственные значения матриц G и L удовлетворяли условию:

$$\lambda \neq 1 / f_{Q^{(t_0)}} f_{A^{(t_0)}}. \tag{6.89}$$

При соблюдении же этого условия, введя обозначения

$$Q = Q^{(t_0+p)} \quad p = 0,1,2,\ldots, \tag{6.90}$$

$$A = A^{(t_0+p)} \quad p = 0,1,2,\ldots, \tag{6.91}$$

получим

$$Q = (1 - Gf_Q f_A)^{-1} Q^{(0)}, \tag{6.92}$$

$$A = (1 - Lf_Q f_A)^{-1} H_C CSf_Q Q^{(0)}. \tag{6.93}$$

Если для всего спектра собственных значений матриц G и L соблюдено условие

$$|\lambda| f_Q f_A < 1, \tag{6.94}$$

то матрицы $(1 - Gf_Q f_A)^{-1}$ и $(1 - Lf_Q f_A)^{-1}$ можно представить в виде

соответствующих бесконечных сумм, т.е. из (6.92) и (6.93) будем иметь [5]:

$$Q = \left(\sum_{p=0}^{\infty} (Gf_Q f_A)^p \right) Q^{(0)}, \tag{6.95}$$

$$A = \left(\sum_{p=0}^{\infty} (Lf_Q f_A)^p \right) H_C C S f_Q Q^{(0)}. \tag{6.96}$$

Таким образом, если процесс (6.79) сходится к паре векторов Q и A, то они могут быть найдены с помощью (6.92) и (6.93). Если к тому же имеет место (6.94), то формулы (6.92) и (6.93) можно переписать в виде (6.95) и (6.96).

Если сравнивать процессы (6.59) и (6.78), то легко обнаружить, что по сравнению с (6.59) процесс (6.78) носит более "консервативный" характер, который обусловлен слагаемым $Q^{(0)}$ в выражении

$$Q^{(t)} = C^T * A^{(t-1)} + Q^{(0)}.$$

Из-за наличия этого слагаемого на каждом этапе отображения вектора $A^{(t-1)}$ на множество терминов фактор "изменчивости", обусловленный слагаемым $C^T * A^{(t-1)}$, "отягощается", затушевывается присутствием слагаемого $Q^{(0)}$, учитывающего фактор преемственности. От этого недостатка свободен процесс (6.59), где слагаемое $Q^{(0)}$ отсутствует. Но отсутствие $Q^{(0)}$ приводит к тому, что процесс в целом оказывается во власти среды и поэтому после нескольких первых тактов работы полностью "забывает" о векторе $Q^{(0)}$ и векторы $Q^{(t)}$ и $A^{(t)}$ принимают собственные направления матриц G и L, т.е. становятся выразителями свойств среды, "забывая" о существовании пользовательского запроса $Q^{(0)}$.

В рассматриваемом смысле представляется целесообразным придерживаться процесса (6.59) с тем, чтобы не "отягощать" фактор изменчивости, но своевременно прекращать процесс, до того, как векторы $Q^{(t)}$ и $A^{(t)}$ "забудут" о векторе $Q^{(0)}$. В случае, когда документы оснащены группой различных описателей их семантики, мы имеем дело с набором различных матриц смежности "документ-термин" и поэтому после нескольких этапов работы с некоторой матрицей C, после получения очередного вектора $A^{(t)}$, процесс можно прекратить и перейти к работе с другой матрицей, соответствующей другому информационно-поисковому языку.

В результате такого перехода от одних матриц к другим получаются новые запросы, эквивалентные (или почти эквивалентные) по смыслу пользовательскому запросу $Q^{(0)}$, но сформулированные средствами (лексическими единицами) других языков – присутствующих в документах описателей их семантики. При этом нет никаких ограничений на

характер этих описателей-языков. Например, запрос $Q^{(0)}$, представленный набором русских слов, может переводиться на запрос, представленный набором английских слов. Особо отметим, что этот перевод осуществляется без помощи каких-либо словарей, так как необходимая для такого перевода информация по крупицам извлекается системой из самой базы данных и уже в рафинированном виде как конечный результат представляется пользователю.

Процесс перевода запросов с одних языков на другие может представлять самостоятельный научно-практический интерес в плане его применимости в ряде других информационных технологий, не имеющих прямого отношения к документальному поиску. Так, описанный выше механизм динамического взаимодействия различных языков может быть применен при установлении диагнозов, лечебных или вредных свойств различных естественных и/или искусственных продуктов, а также в ряде других систем аналитического назначения при необходимости извлечения из большого объема данных качественно новой, до этого не существовавшей (несформулированной) информации.

Представленные в настоящей главе идеи и методы в энтропийном варианте их реализации легли в основу построения автоматизированной информационно-поисковой (аналитической) системы "Бумеранг" [1].

ЛИТЕРАТУРА К ГЛАВЕ 6

1. *Аветисян Д.О., Аветисян Р.Д.* Автоматизированная информационно-поисковая система "Бумеранг" // Информационные ресурсы России. – 1995. – № 2.
2. *Аветисян Д.О.* О вероятностном подходе к построению интеллектуальных систем. Ч. 1. Теория // Математические вопросы кибернетики и вычислительной техники: Сб. науч. тр. / Вычисл. центр АН Арм. ССР, Ереванск. гос. унив-т. – Ереван, 1984. – Т. 13.
3. *Аветисян Р.Д.* Об одном методе построения многоязычного лексико-семантического интерфейса в автоматизированных ИПС // Научно-техническая информация. Серия 2. – 1996. – № 1.
4. *Анго Андре.* Математика для электро- и радиоинженеров. – М.: Наука, 1967.
5. *Гантмахер Ф.Р.* Теория матриц. – М.: Наука, 1988.
6. *Корн Г., Корн Т.* Справочник по математике для научных работников и инженеров. – М.: Наука, 1968.

Утверждение о спектре собственных значений

Пусть X и Y – произвольные матрицы размерностей $n \times m$ и $m \times n$. Примем, что $n \geqslant m$. Нас будут интересовать матрицы $A = X \cdot Y$ и $B = Y \cdot X$. Очевидно, это квадратные матрицы порядков соответственно n и m. Характеристические многочлены этих матриц обозначим соответственно через $A(\lambda)$ и $B(\lambda)$.

Теорема.

$$A(\lambda) = (-\lambda)^{n-m} B(\lambda),$$

т.е. спектры собственных значений матриц A и B совпадают, за исключением, быть может, нулевых собственных значений.

Доказательство.

Характеристический многочлен матрицы A можно представить как

$$A(\lambda) = (-\lambda)^n + \sum_{p=1}^{n} S_p (-\lambda)^{n-p}, \qquad (1)$$

где S_p равно сумме всех возможных главных миноров p-го порядка матрицы A (см., например: Гантмахер Ф.Р. Теория матриц. – М.: Наука, 1988):

$$S_p = \sum_{1 \leqslant i_1 < i_2 < \ldots < i_p \leqslant n} \begin{vmatrix} a_{i_1 i_1} & a_{i_1 i_2} & \cdots & a_{i_1 i_p} \\ a_{i_2 i_1} & a_{i_2 i_2} & \cdots & a_{i_2 i_p} \\ \cdots & \cdots & \cdots & \cdots \\ a_{i_p i_1} & a_{i_p i_2} & \cdots & a_{i_p i_p} \end{vmatrix}.$$

Эту формулу можно переписать в виде:

$$S_p = \sum_{1 \leqslant i_1 < i_2 < \ldots < i_p \leqslant n} \left| \begin{pmatrix} x_{i_1 1} & x_{i_1 2} & \cdots & x_{i_1 m} \\ x_{i_2 1} & x_{i_2 2} & \cdots & x_{i_2 m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{i_p 1} & x_{i_p 2} & \cdots & x_{i_p m} \end{pmatrix} \begin{pmatrix} y_{1 i_1} & y_{1 i_2} & \cdots & y_{1 i_p} \\ y_{2 i_1} & y_{2 i_2} & \cdots & y_{2 i_p} \\ \cdots & \cdots & \cdots & \cdots \\ y_{m i_1} & y_{m i_2} & \cdots & y_{m i_p} \end{pmatrix} \right|.$$

Пользуясь формулой Бине–Коши, отсюда получим:

$$S_p = \sum_{\substack{1\le i_1<i_2<\ldots<i_p\le n \\ 1\le j_1<j_2<\ldots<j_p\le m}} \begin{vmatrix} x_{i_1j_1} & x_{i_1j_2} & \cdots & x_{i_1j_p} \\ x_{i_2j_1} & x_{i_2j_2} & \cdots & x_{i_2j_p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{i_pj_1} & x_{i_pj_2} & \cdots & x_{i_pj_p} \end{vmatrix} \begin{vmatrix} y_{j_1i_1} & y_{j_1i_2} & \cdots & y_{j_1i_p} \\ y_{j_2i_1} & y_{j_2i_2} & \cdots & y_{j_2i_p} \\ \cdots & \cdots & \cdots & \cdots \\ y_{j_pi_1} & y_{j_pi_2} & \cdots & y_{j_pi_p} \end{vmatrix}.$$

Поскольку $m \leqslant n$, то очевидно, что для всех $p > m$ имеет место $S_p \equiv 0$, т.е. выражение (1) можно переписать как

$$A(\lambda) = (-\lambda)^{n-m}\left((-\lambda)^m + \sum_{p=1}^{m} S_p(-\lambda)^{m-p}\right). \qquad (2)$$

Характеристический многочлен матрицы B можно представить как

$$B(\lambda) = (-\lambda)^m + \sum_{p=1}^{m} L_p(-\lambda)^{m-p}, \qquad (3)$$

где

$$L_p = \sum_{\substack{1\le j_1<j_2<\ldots<j_p\le m \\ 1\le i_1<i_2<\ldots<i_p\le n}} \begin{vmatrix} y_{j_1i_1} & y_{j_1i_2} & \cdots & y_{j_1i_p} \\ y_{j_2i_1} & y_{j_2i_2} & \cdots & y_{j_2i_p} \\ \cdots & \cdots & \cdots & \cdots \\ y_{j_pi_1} & y_{j_pi_2} & \cdots & y_{j_pi_p} \end{vmatrix} \begin{vmatrix} x_{i_1j_1} & x_{i_1j_2} & \cdots & x_{i_1j_p} \\ x_{i_2j_1} & x_{i_2j_2} & \cdots & x_{i_2j_p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{i_pj_1} & x_{i_pj_2} & \cdots & x_{i_pj_p} \end{vmatrix}.$$

Отсюда легко установить, что при всех $p \leqslant m$ имеет место $S_p \equiv L_p$, т.е. формулу (2) можно переписать как

$$A(\lambda) = (-\lambda)^{n-m}\left((-\lambda)^m + \sum_{p=1}^{m} L_p(-\lambda)^{m-p}\right).$$

Отсюда, с учетом (3), окончательно получим:

$$A(\lambda) = (-\lambda)^{n-m} B(\lambda).$$

Утверждение о характеристическом уравнении матрицы S_n

Рассмотрим матрицу

$$S_n = \frac{1}{n} A,$$

где $A[a_{ij}]$ – матрица с элементами

$$a_{ij} = \begin{cases} n-1 & \text{при } i = j \\ -1 & \text{при } i \neq j \end{cases}.$$

Очевидно, если в характеристическом уравнении матрицы A заменить ее собственные значения α_i на $\alpha_i = n\lambda_i$, то получим характеристическое уравнение матрицы S_n относительно ее собственных значений λ_i. Характеристическое уравнение матрицы A имеет вид:

$$\begin{vmatrix} n-\alpha-1 & -1 & -1 & \ldots & -1 & -1 \\ -1 & n-\alpha-1 & -1 & \ldots & -1 & -1 \\ -1 & -1 & n-\alpha-1 & \ldots & -1 & -1 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ -1 & -1 & -1 & \ldots & n-\alpha-1 & -1 \\ -1 & -1 & -1 & \ldots & -1 & n-\alpha-1 \end{vmatrix} = 0.$$

Вычитая поочередно из первой строки вторую, из второй строки третью, ..., из $(n-1)$-й строки n-ю, получим

$$\begin{vmatrix} (n-\alpha) & -(n-\alpha) & 0 & \ldots & 0 & 0 \\ 0 & (n-\alpha) & -(n-\alpha) & \ldots & 0 & 0 \\ 0 & 0 & (n-\alpha) & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & (n-\alpha) & -(n-\alpha) \\ -1 & -1 & -1 & \ldots & -1 & n-\alpha-1 \end{vmatrix} = 0,$$

или, что то же самое,

$$\begin{vmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & -1 \\ -1 & -1 & -1 & \dots & -1 & n-\alpha-1 \end{vmatrix} (n-\alpha)^{n-1} = 0.$$

Отсюда, путем поочередного сложения последней (n-й) строки с первой строкой, умноженной на 1, со второй строкой, умноженной на 2,..., с $(n-1)$-й строкой, умноженной на $(n-1)$, получим:

$$\begin{vmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & -1 \\ 0 & 0 & 0 & \dots & 0 & -\alpha \end{vmatrix} (n-\alpha)^{n-1} = 0,$$

т.е. характеристическое уравнение матрицы A относительно ее собственных значений α_i имеет вид

$$\alpha(n-\alpha)^{n-1} = 0.$$

Подставляя сюда $\alpha = n\lambda$, получим характеристическое уравнение матрицы S_n относительно λ_i:

$$\lambda(n-\lambda)^{n-1} = 0.$$

ОГЛАВЛЕНИЕ

Аветисян Рубен Декартович
Аветисян Декарт Овсепович

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ИНФОРМАТИКИ