ИНТЕГРАТИВНЫЙ ПОДХОД ПРИ СОЗДАНИИ ИНФОРМАЦИОННО-АНАЛИТИЧЕСКОЙ СИСТЕМЫ ЭНТРОПИЙНОГО ОЦЕНИВАНИЯ ДАННЫХ

Русяева Е. Ю.[1], Полтавский А. В.[2]
(ФГБУН *Институт проблем управления им. В.А. Трапезникова РАН, Москва*)

Представлен подход и алгоритмы к анализу информации на основе интегративной оценки для мультимедийных потоков данных в компьютерной сети с расчетом энтропийного коэффициента конкордации. Приведены примеры использования интегративного подхода на базисе формул теории информации, теории алгоритмов и технической кибернетики к анализу потоков данных в компьютерной сети. Новизна предложенного подхода в синтезе формализованных моделей подсчета энтропийных характеристик оцениваемого мультимедийного объекта и расчета энтропийных оценок мнений экспертов. Представленный интегративный подход в перспективе нацелен на создание частных методов и моделей для анализа, энтропийного оценивания мультимедийной информации, что в итоге, при наработке достаточной базы знаний, позволит определять также и авторство создаваемого контента (информации, знаний) как естественного (созданного субъектом, человеком), так и искусственного (автоматически созданного) происхождения. В среде конструируемой информационно-аналитической системы (ИАС) возможно не только устанавливать соответствие авторства имеющегося в интернете контента, но и по определенным критериям оценивать качество передаваемой мультимедийной информации. В среде ИАС сопоставляются энтропийные характеристики и оценивается уровень переводов имеющихся естественных и искусственных (машинных) переводчиков из текстов. Создаваемая система позволяет формализовать информационные процессы идентификации данных, формировать адекватную выборку актуальной информации для поддержки принятия решений субъектами управления.

Ключевые слова: информационно-аналитическая система, нейросеть, алгоритм, коэффициент конкордации

[1] *Русяева Елена Юрьевна, к.филос.н., с.н.с.* (eur5@mail.ru).
[2] *Полтавский Александр Васильевич, д.т.н., профессор (avp57avp@yandex.ru).*

1. Введение

На фоне современных тенденций импортозамещения в России, одним из наиболее важных и «чувствительных» к заменам стал сегмент информационных технологий. Сейчас как никогда востребованы отечественные разработки информационных систем. Отметим, что в российской науке традиции автоматизации информационных процессов для задач управления актуализированы уже давно, например, в работах В.А. Трапезникова [14]. Напомним, что практически вместе с разработками первых электронных цифровых компьютеров появились и первые программы и системы советчиков и машинных переводчиков для них, работающих с символьной информацией. Наряду с первыми разработками аппаратно-программных средств для цифровых информационно-вычислительных систем (ИВС) развивались и цифровые сети передачи информации. Например, появились такие сети как ISDN (Integrated Services Digital Network — цифровая сеть с интеграцией услуг), которые относятся к классу сетей, изначально предназначенных для передачи как данных, так и голоса.

По мере развития ИВС и компьютерных сетей эволюционировали программно-аппаратные средства для передачи информации. Сегодня информация в большинстве своем мультимедийная, а качество ее требует адекватной оценки с использованием накопленных знаний, методов, моделей и алгоритмов ее получения [7,8].

Одна из важных задач исследования заключается в разработке компьютерной среды и методов, алгоритмов для реализации программ в формируемой информационно-аналитической системе (ИАС). Во многих современных ИАС не учитываются в полной мере энтропийные характеристики для основных показателей по идентификации потоков мультимедийной информации в цифровой сети.

Например, во многих нейросетях весовые коэффициенты используются в управлении самообучением самой сети. По существу – это программное обеспечение как правило, настроенное как некий эталон, куда внесены желаемые (требуемые) характеристики информационных процессов. В подобных разработках [2] как в черных ящиках не видны четко сами действующие алгоритмы, что затрудняет давать интегративную оценку об истинности и комплексности анализа для всего процесса (это «зашито» в черном ящике разработок).

Очевидно, что во всем мире пакетные разработки в нейросетях рождаются в условиях конкурентной среды и защиты интеллектуальной собственности, а теперь еще и условий санкционного режима. Ни фирмам разработчикам, ни отдельным программистам не выгодно делиться методами, подходами и алгоритмами, которые используются в создаваемых пакетах.

Научному сообществу (как, впрочем, и всем пользователям в сети) остается верить «вслепую» и как бы условно соглашаться на те решения нейросетевого программного обеспечения, которые предлагают в готовом виде. К сожалению, часто в красивой обертке из фраз (нейросеть настроена на то-то и то-то) кроется выгода от продвижения данного программного продукта на рынок (как маркетинговый ход). Причем, зачастую этот маркетинговый ход далек от научно-прикладных задач, причем, как в военной области, так и в народном хозяйстве (двойные технологии). Так что сами эти программные продукты (искусственные нейросети) требуют экспертизы и построения соответствующих информационно-аналитических систем (ИАС) для оценки мультимедийных данных. Вот почему предлагаемый авторами интегративный подход не просто соответствует духу времени, но и вписывается в актуальный тренд современности комплексной деятельности [3].

В данном исследовании представлен подход, разработанный в целях повышения информативности данных для решения задач идентификации мультимедийной информации. При построении алгоритма интегративной оценки мультимедийных данных и расчета энтропийного коэффициента конкордации использован принцип открытой архитектуры, применяемый для современных ИТ с использованием мультимедийной информации. Разработанный частный метод позволяет звуки, буквы, речь, символы, тексты и любые другие мультмедийные данные привести к энтропийной оценке. Далее с учетом получаемых измерений потоков данных для анализа абсолютной погрешности с учетом получаемых энтропийных измерений, с получением коэффициента подобия, энтропийного коэффициента конкордации и других подсчетов можно проводить идентификацию данных. Интегративность подхода в том, что к методам подсчета энтропийных характеристик самого оцениваемого мультимедийного объекта, исходных данных (текста, символа, знака, потока чисел и др. мультимедийных данных), добавлен метод подсчета энтропийных оценок мнений экспертов. Это исследование продолжает авторские разработки в русле интегративного подхода построения информационных систем [9].

2. *Потоки данных и их поуровневая обработка в информационно-экспертных системах*

Как известно, само понятие «информация» часто понимают в очень широком смысле, используют его как зонтичный бренд для разных целей. Для данного исследования достаточно и классического понимания информации как некого объекта, как «обозначение содержания, полученное нами из внешнего мира в процессе приспосабливания к нему нас и наших чувств» [4]. Нам важнее указать на уровни информации как некого объекта изучения, обозначить стадии трансформаций, когда из инфор-

мационного шума (данных) путем естественной и/или искусственной переработки, структуризации, и интеллектуальной обработки возникает новый уровень смысла для субъекта. Мы начнем изучение с самого нижнего уровня, с потока данных, согласно пирамиде знаний Акоффа [1]. То есть, в перспективе нас интересует вопрос о том, как из нынешней многоплановой мультимедийной информации, поступающей из разных источников, включая звуки, музыку, картинки, видео, числа, символы, буквы, иероглифы, знаки, слова, тексты и т.д., в итоге субъект создает знания. Но пока мы лишь формируем подход, в котором интегрируются формальные оценки исследуемого объекта.

Поясним, что, как известно, и в кибернетике, информатике, чтобы моделировать алгоритмы принятия решений приходится вводить определенные ограничения. Это делать все труднее в силу непредсказуемости, многофакторности и неизбежной субъективности при выделении исходных предпосылок о факторах, мотивах в контурах управления. Результаты моделирования управленческих воздействий существенно зависят от логики исходных предположений о поведении управляемых агентов, заложенных исследователем в модель. Здесь приходится соприкасаться с социально-культурными, психологическими, философскими аспектами, исследования часто носят релятивный характер, а их результаты зависят от системы представлений и предпочтений самих исследователей. Мы моделируем информационно-аналитическую систему - ИАС, в среде которой оцениваем потоки данных как некий объект, используемый для принятия решений субъектом управления. Разрабатываем также алгоритм идентификации потоков исходных данных для формирования контекстной информации в системе управления.

Мы будем отталкиваться от представлений классиков науки. Весь общий неструктурированный поток информации (понимаемый как общий информационный шум, общие сведения) будем называть данными. Поток данных, поступающих в среду (см. Рис.1) информационно-аналитических, экспертных систем, требует операций распараллеливания в современной компьютерной среде. Далее следуют процессы упорядочивания

данных, их хранения, а затем и выдачи в виде уже структурированной информации с целью дальнейшего принятия управленческих решений.

По факту, принятие решения по-прежнему остается за субъектом управления, ведь только человек может интеллектуально обрабатывать поступающую информацию, формируя знания (см. пирамиду знаний Акоффа, Рис.1). Искусственные экспертные системы могут использоваться лишь для информационной поддержки принятия решений.

Мы формируем уровни ИАС, опираясь на базу так называемой пирамиды знаний (эксперта) Р. Акоффа[1] [1]. Моделируем

---

[1] Напомним, Расселу Акоффу условно приписывается разработка уровней информационной иерархии восприятия потоков информации естественным интеллектом (человеком). Здесь каждый уровень добавляет определённые свойства к предыдущему, почему условно эта модель и была названа пирамидой. По сути, каждый этаж пирамиды характеризуется бо́льшим уровнем зрелости (пригодностью к продолжению жизни) и кратно меньшим объёмом сведений.

Термин «пирамида знаний» стал популярным после речи Р. Акоффа в 1989. В основании этой, сокращенно, DIKW-пирамиды, располагаются данные, на следующем «этаже» находится информация - она к данным добавляет контекст. Этажом выше в пирамиде располагается знание, оно добавляет ответ на вопрос «как», то есть, по сути, определяет механизм использования информации. Иногда, в частности, сам Р.Акофф, добавляют уровень понимания после этажа «знаний». На верхнем этаже располагается мудрость, она добавляет ответ на вопрос «когда», то есть определяет условия использования знания. Считается, что по мере движения от нижнего этажа пирамиды к верхнему увеличивается степени полезности [4] переработки информации для принятия решений ЛПР.

Таким образом, согласно Р. Акоффу, знание – это данные вместе с контекстом и механизмом их применения. Что же такое тогда неявное знание, если понимать знание вообще именно так? Неявные знания – это такие данные вместе с контекстом, механизм применения которых существует и эффективно работает «в голове» у эксперта, но в явном виде этот механизм не представлен ни вербально, ни математически, ни как-либо ещё.

алгоритм анализа потока мультимедийных данных в ИАС для системы управления и решения задач по идентификации потоков данных в новой структуре. Состав этой структуры представляют современные сетевые и облачные технологии с распределенными базами мультимедийных данных, базами знаний и алгоритмов для обработки информации, составляющей ядро для компьютеризированной среды и системы ИАС в целях принятия управленческих решений.

А) Рис.1 (а) Пирамида знаний эксперта

Б) Рис.1 (б) Модель искусственной ИС с элементами ИИ

Рис. 1. Схематизация процесса обмена мультимедийными данными, информационными и знаниевыми потоками между естественными и искусственными информационными системами.

Данная схема лишь условно отображает особенности к восприятию и переработке мультимедийных данных, информации человеком-экспертом (естественной системой) и созданными человеком, искусственными информационными системами – ИИС, причем они сейчас не способны генерировать объекты выше уровня информации. Слева (см. Рис.1 (а)) представлена уже достаточно известная пирамида знаний эксперта (восприятия данных человеческим, естественным интеллектом) по Р. Акоффу. Справа, на Рис.1(б) в виде перевернутого конуса изображена модель обработки данных, информации, и баз знаний ИИС.

Подчеркнем, ИИС не создают новое знание! Пока ИИС не могут достигнуть даже уровня понимания ребенка, для этого требуется слишком много энергии для простраивания всех возможных связей в системе. У человека же (субъекта управления), в его нейросетке, эти связи выстроены в ходе эволюции естественным (природным) путем.

Люди в современном мире погружены в такой огромный поток информации, всевозможных мультимедийных данных, что в одиночку уже ни один человеческий, естественный интеллект не в силах с ним справиться. Человечество за свою историю уже создало огромную базу знаний. Создание новых знаний – прерогатива людей, причем, современное знание все чаще создается путем комплексирования и интеграции уже имеющейся информации, знаний благодаря какой-то новой комбинации смыслов. Это отдельная тема, выходящая за рамки данного исследования. Здесь мы лишь подчеркнем, что человек использует ИИС с целью автоматизированной обработки данных, информации, применяет ИИС и для поддержки принятии решений.

На Рис.1. стрелками указаны потоки входа и переработки данных из внешней среды по уровням и обмен данными, информацией между экспертами и ИИС. В первую очередь – это уровни восприятия и переработки данных экспертом и перевернутая архитектура последовательности ввода и переработки данных в информационной системе.

Напомним, что ранее, во времена Р. Акоффа, для создания пирамиды знаний еще не существовало аппаратно-программных

средств с элементами так называемого искусственного интеллекта - ИИ[1], которые могли бы реализовать идентификацию потоков данных на всех этажах пирамиды. Но теперь, в XXI-й информационный век обойтись без структурирования и координации процессов прохождения больших потоков данных в информационных системах (ИС) невозможно в принципе.

Как правило, в любом проекте и системе управления (технической, социально-экономической, биологической и др.) имеется своя ИС. Это особенно заметно в ныне создаваемых ИС (документальных, экспертных, информационно-аналитических, фактографических и пр.) в новом поколении аппаратно-программных средств, роботизированных комплексов, систем.

Все они могут быть настроены на оценки и энтропийные характеристики данных. Сами же эти характеристики связаны непосредственно с вероятностными (стохастическими) процессами, без которых современные информационно-аналитические

---

[1] Полностью разделяем мнение академика Д.А. Новикова, высказанного им в интервью по поводу ИИ, публикация [5]. Цитируем: «искусственный интеллект является очень вредным термином, потому что для обывателей он создает иллюзию действительно чего-то сравнимого с человеком». «А что такое искусственный интеллект, по-вашему? Можно сказать, что с точки зрения теории управления – это раздел математики. Традиционно есть несколько областей науки, как правило – методов прикладной математики, которые считают относящимися к искусственному интеллекту. Они возникли в разное время, и некоторые из них появились еще до 1956 года, когда этот термин возник. Те же столь безумно модные сейчас искусственные нейронные сети берут начало от математической модели искусственного нейрона Мак-Каллока–Питтса – а это 1943 год. Первая практическая реализация нейронной сети – это перцептрон Розенблатта (1957 год). Традиционно к методам искусственного интеллекта, помимо сетей, относят: нечеткие множества (которыми начали заниматься в 1965 году), эволюционные алгоритмы и экспертные системы (возникшие в 1960-70-е годы), мультиагентные системы (1974 год), различные логические средства и многое другое» [5].

(ИАС) с элементами ИИ, в том числе и нейросети, представляются нам в сложившихся структурах ИАС проблематичными. Требуются новые решения и научно-обоснованные подходы к разработкам таких ИАС.

3. Интегративный подход, схематизация информационных процессов с прохождением данных по каналам связи в среде ИАС

Как правило, информационной системой (ИС), либо автоматизированной ИС, АИС, часто называют цифровую программно-аппаратную систему, предназначенную для автоматизации целенаправленной деятельности конечных пользователей, обеспечивающую, в соответствие с заложенной в неё логикой обработки, возможность получения, модификации и хранения информации. В разработанной информационной модели для ИАС производится анализ из естественных и искусственных источников поступления информации.

Наш подход мы назвали интегративным, так как добавляем в ИАС блоки для выбора экспертов и «фильтрацию» их оценок в отношении объектов, в частности, мультимедийных данных, поступающих на дальнейшую экспертизу. Как схематично показано на Рис. 1, на всех уровнях пирамиды ведется обработка данных, информации. Упрощенную схему цифровой трансформации из потоков данных покажем на Рис. 2.

Новизна исследования заключается в интеграции, адекватном комбинировании научных достижений в компьютерной технике с сочетанием программных средств поддержки управленческих решений, на базе использования фундаментальных теорем, методов и моделей, рабочих формул известных ученых - Кл. Шеннона [17], А.Н. Колмогорова, В.С. Пугачева [12], Э.А. Трахтенгерца [15] и др.

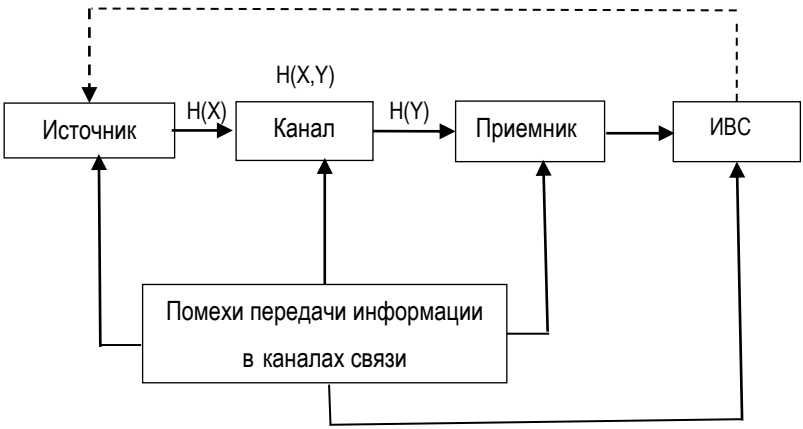


Рис. 2. Схематизация передачи мультимедийной информации по каналам связи и управления в компьютеризированной среде ИАС

На Рис. 2. видно, что при наличии помех (естественной и искусственной природы) в канале связи имеется некоторая неопределенность, связанная с передачей данных. Степень неопределенности (как информационная энтропия) будет оценивается разностью с общепринятой формулой следующего вида:

$I(Y,X) = H(X) – H(X/Y)$, где

H(x) – энтропия входных источников в компьютеризированной среде ИАС

H(X/Y) – величина для энтропии, характеризующей меру неопределенности до приема мультимедийного сигнала (потоков данных, передаваемых сообщений и т.д.)

I(YX) – энтропийная мера снятия уровня для неопределенности после приема и обработки информации в среде ИАС. В данном случае, в контуре управления и обработки информации для такой системы является создаваемая среда ИАС. Датчиками

служат источники приема мультимедийных сообщений как потоков данных на входе и с обратной связью после их обработки.

Интегративный подход анализа потоков данных актуален для повышения уровня оцениваемых характеристик (в условиях помех), прежде всего, вероятностных показателей. Подход служит для повышения функциональных возможностей и «оперативности» обработки информации в ИАС, а также коррекции поступающих данных и оптимизации самого процесса анализа. Сущность подхода заключается в комплексном использовании комбинации алгоритмов из экспертных и формальных оценок. Это важно в целях создания новых возможностей для формирования среды ИС и создания блока для информационных экспертных систем (ИЭС). Этот подход продолжает и развивает тематику исследований [6, 7, 8, 11] и патента [10]. В данном исследовании рассматривается приложение интегративного подхода к энтропийному анализу мультимедийных потоков данных, потоков текстовой (символьной) информации.

Отметим, что ошибки и энтропийная погрешность присутствует в любом процессе по автоматической обработке потоков данных в информационно-вычислительных системах – ИВС. Наверное, больше всего это затрагивает потоки различного рода лингвистической информации. Поиски решения ведутся в рамках прикладных разработок компьютерной, прикладной лингвистики, в машинных переводах, интернет-роботах, видеотрансформерах звук-текст, текст-звук и т.д.

Следует отметить, что исторически сложилась такая ситуация, что под термином «прикладная лингвистика» в разных странах понимают разные дисциплины и делают упор на разные аспекты применения этой науки. Так, например, за рубежом прикладную лингвистику в основном воспринимают как науку, ответственную за обучение иностранным языкам и переводческой деятельности. В то же время в советской науке и российской, её преемнице, в основном прикладная лингвистика синонимична понятиям «компьютерная лингвистика» и «инженерная лингвистика». Это было связано с тем, что термин получил широкое распространение в СССР в 1950-х годах одновременно с появлением первых компьютерных систем для автоматической

обработки именно текстовой информации, в том числе и машинного перевода их числе [7, 8]. Именно использование математических методов при изучении естественных языков способствовало возникновению математической (компьютерной) лингвистики. Сегодня эти методы значительно усовершенствованы.

Покажем сущность интегративного подхода при разработке цифровой среды ИАС.

Во-первых, поясним для начала работу блока выбора «наилучшего» эксперта. Для этого блока воспользуемся общепризнанным понятием о уровне компетентности i-го эксперта в ИАС, обозначим - $К_i$ ($К^э_i$). Он изменяется в пределах от 0 до 1. Как правило, сами эксперты подбираются с большим значением для коэффициентом компетентности, который, в свою очередь, почти пропорционален доверительной вероятности. Для ее подсчета, как правило, оперируют эмпирическими значениями из математических ожиданий и вероятностями, связанных по количественному оцениваемому фактору для i-го эксперта. Методики подсчета коэффициента компетентности i-го эксперта насчитывают множество подходов и алгоритмов, что позволяет решать «гибкую» задачу выбора метода для среды ИАС.

На наш взгляд, среди них предпочтителен алгоритм по выбору более компетентного эксперта, основанный именно на вероятностной оценке. Таким образом, этот процесс для формируемой среды ИАС имеет также случайный характер. Для получения оценки и визуализации информационных процессов в компьютерной среде, по ранее известным методикам из теории вероятностей и математической статистики, строят полигон для вероятностей оцениваемых экспертов. То есть, в среде ИАС отображен подход, заключающийся в энтропийной оценке как самой мультимедийной информации, поступающей в компьютеризированную среду ИАС, так и в учете мнений участвующих при обработке данных экспертов, которые также принимают свои решения в условиях неопределенности для всей системы.

Во-вторых, как отмечено, сама передача мультимедийной информации связана с ее потерей, искажением и др., которые носят также случайный характер. В [10] было показано как

это следует учитывать на примерах энтропийного анализа текстов. Это же происходит и с анализом данных для мультимедиа.

Как видим, имеются две основные компоненты по неопределенности в компьютерной среде ИАС при обработке различной информации, поступающей от источника данных и от самих экспертов. Поэтому подход к «взвешиванию» из потоков мультимедийных данных и «взвешиванию» экспертов назван интегративным в рамках выбранных путей оцениваемых информационных потоков с выбранной метрикой и пространства измерений. Важно и влияние на информационные процессы в среде ИАС различных факторов междисциплинарной природы, но этот аспект требует дальнейших исследований. Реализовать в среде ИАС предлагаемый интегративный подход продемонстрируем на основе разработанного алгоритма.

4. Алгоритмизация оценивания потоков данных и информации конкордации

Отметим, что применимость интегративного подхода как инновационного инструмента обосновывается тем фактом, что он в самой ИАС может быть использован в разных областях (и отраслях) для управления знаниями. Например, как мы указывали, при расчетах, связанных с идентификацией текстовой информации. Это важно в случаях необходимости определения авторства текста с целью принятия объективных решений при защите авторских прав (как антиплагиат), для адекватной идентификации авторства, установления персональной уникальности текста и пр. Применение алгоритма для идентификации лингвистического (например, переводчика) робота также сегодня очень актуально.

Покажем основы цифровой трансформатизации потоков текстов. В этом процессе все большое значение приобретают инновационные (и «цифровые») проекты и модели, направленные, прежде всего, на создание автоматизированных процедур и алгоритмов для сетевых телекоммуникационных ИС, которые, как отмечено, также должны учитывать энтропийную составля-

ющую для передачи и обработки потоков информации по различным коммуникативным каналам связи. Данное обстоятельство, а также и сам такой стохастический процесс связаны с потерей качества передачи информации по каналам связи в сетевых структурах ИС, которая может быть частично утеряна или преобразована (и/или получила недопустимое искажение) с большими ошибками. Это можно показать на примере построения моделей и алгоритмов по идентификации текстовых потоков в экспертной системе, подробнее описано в работе [10].

На практике прикладных задач и в ИАС часто ведется поиск числа повторений той или иной служебной частицы среди потоков данных из многих тысяч слов и отождествлением с известной задачей математической статистики о повторении испытаний. То есть, количество слов текста будем считать числом испытаний n_i, а число m_i повторений для частицы – числом появлений события. Тогда можно ввести понятие для оцениваемой относительной частоты в компьютеризированной среде ИАС, которую находим из рабочей формулы

$$P_i = \frac{m_i}{n_i} , \qquad (1)$$

как отношение указанных чисел.

Из математической статистики нам известны случаи, когда при увеличении числа испытаний при проведении вычислительного эксперимента числовые значения частот колеблются около некоторой величины и отклонения частот от указанной величины уменьшаются с ростом числа испытаний (наблюдений) в ИАС. Как правило, в качестве таковой величины принимается среднее $P_{ср}$ по частоте P_i. Если в формуле (1) символом i обозначать номер серии испытаний в компьютерной среде ИАС, тогда значение $P_{ср}$ необходимо вычислять:

$$P_{cp} = \frac{\sum_i P_i}{N} , \qquad (2)$$

где N – число серий в испытаниях.

Этот факт в математической статистике (о повторяемости частот) называется законом устойчивости частот, а на основе

теоремы Я. Бернулли, сама величина $P_{ср}$ принимается в качестве оценки для вероятности появления события в вычислительном эксперименте. Таким образом, показаны теоретические основы к разработке алгоритмов потоков данных в ИАС.

Кратко представим описание к алгоритму для формируемой среды ИАС по соответствующей цепочке передачи и обработки данных в виде первичной блок-схемы (Рис. 3):

Блок 1 – вводятся потоки исходных данных для оценивания.

Блок 2 – собираются блоки обработки и структурируются исходные данные.

Блок 3 – производится предварительная обработка данных.

Блок 4 – расчет идентификаторов данных (буквы, числа, идентификаторы звука, видео и т.д.).

Блок 5 – осуществляется отбор данных из блоков оцениваемой информации.

Блок 6 – сравнительного анализа (индикатор), данные сопоставляются.

Блок 7 – информация структурируется и фильтруется по направлениям задач.

Блок 8 – блок учета оценки экспертизы информационных потоков данных.

Блок 9 – отбор результатов оцениваемых энтропийных характеристик.

Блок 10 – блок энтропийных характеристик и конкордации (индикатор ИЛИ).

Блок 11 – блок вывода потоков данных и индикации оцениваемой информации.

Блок 12 – формулируется итоговые оценки для потоков информации в компьютеризированной среды ИАС.

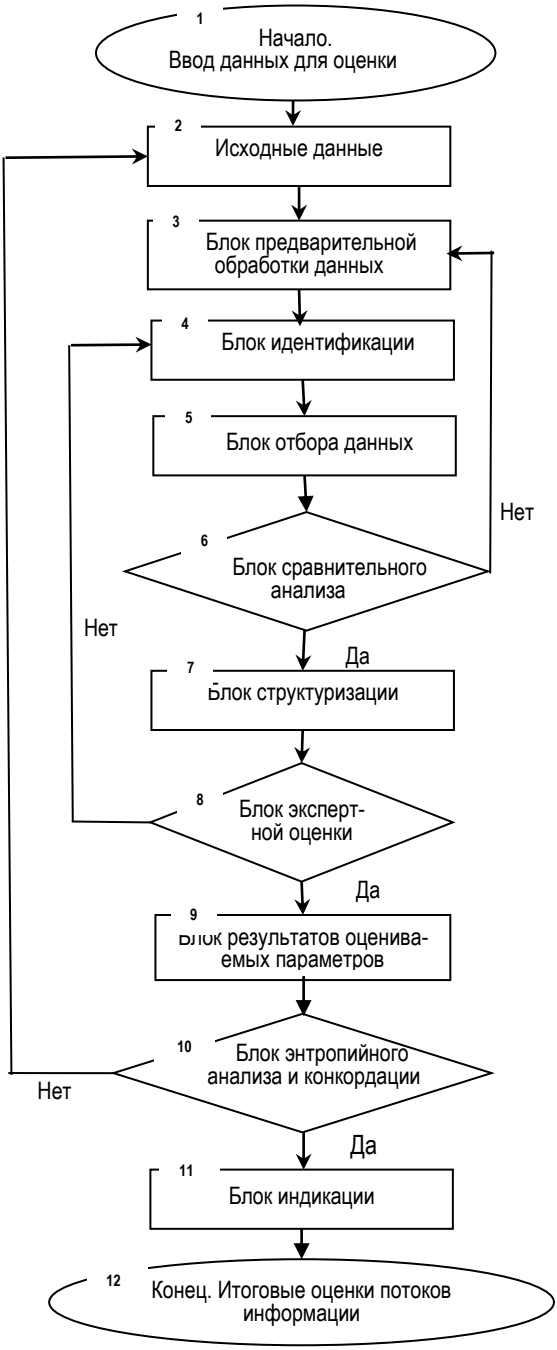

1
Начало.
Ввод данных для оценки
2
Исходные данные
3
Блок предварительной обработки данных
4
Блок идентификации
5
Блок отбора данных
6
Блок сравнительного анализа
Нет
Нет
Да
7
Блок структуризации
8
Блок эксперт-ной оценки
Нет
Да
9
Блок результатов оцениваемых параметров
10
Блок энтропийного анализа и конкордации
Нет
Да
11
Блок индикации
12
Конец. Итоговые оценки потоков информации

Рис. 3. Алгоритм интегративной оценки мультимедийных данных и расчета энтропийного коэффициента конкордации

В рамках стратегии национальной безопасности страны [13] в России одобрена госпрограмма построения системы информационной безопасности [16], которая непосредственно затрагивает эти вопросы. В современных условиях России требуются свои отечественные аппаратно-программные средства, свои информационные ресурсы и алгоритмы для решения этих и многих других проблем.

Рассмотрим детальнее работу блока системы индикации в цифровой среде ИАС

Рис. 4. Схематизация блока индикации среды ИАС

5. Теоретические основы оценивания энтропийных характеристик экспертизы

Разработанный в рамках интегративного метода алгоритм (Рис.3) является цифровой платформой для оценки и мультимедийной информации, поэтому выбираем пока его для ИАС в качестве основного. Далее, в соответствии с замыслом ИАС, для формирования «инструментального ядра» из множества известных методов экспертных оценок следует также выбрать более приемлемый, который предполагает учитывать сам уровень экспертизы, а также случайной составляющей в информационной конструкции ИАС. Покажем один из них, который связан с определением коэффициентов информационной конкордации.

Известный коэффициент конкордации – это число от 0 до 1, показывающее согласованность мнений экспертов при проведении ранжирования каких-то свойств. Чем ближе это значение к 0, тем согласованность считается более низкой. При величине данного коэффициента менее 0,3 мнения экспертов считаются несогласованными. При нахождении величины коэффициента в диапазоне от 0,3 до 0, 7 согласованность считается средней. При величине более 0,7 согласованность экспертов принимается как высокая. Энтропийный коэффициент конкордации (коэффициент согласия экспертов) находим по общепринятой формуле

$$W_э = 1 - \frac{H_э}{H_{max}}$$

*W*э - энтропийный коэффициент конкордации (согласованности) мнений и оценок экспертов по i-му объекту.

Интеграция (сущность интеграции [9]) в том, что берутся оценки вероятности *P*ср и энтропийные характеристики для потоков данных и количества информации – *H* от самих экспертов, методы подсчетов показаны в [10] с использованием дополнительного фактора, а именно, учета мнений экспертов, в чем,

собственно, в данном случае и проявляются возможности интегративности. Она заключается, как было указано ранее, в том, что энтропийные характеристики самого оцениваемого объекта (текст, символ, знак, потока чисел и др.) соединяем также и с энтропийными оценками с мнениями экспертов.

6. Примеры информационного моделирования процессов идентификации

Средства сетевых ИВС и сформированная ИАС позволяют провести вычислительный эксперимент для поступающей информации из компьютерной сети. На основе статистического анализа потоков данных из компьютерной сети в самой программе производится подсчет основных вероятностных характеристик.

Представим фрагмент предварительных испытаний из программы по расчетам информационной энтропии для произведений великого поэта А.С. Пушкина (Рис. 5). Полученные в среде ИАС оценки для вероятности $P_{ср}$ к потокам данных известных его произведений практически совпадают по полученным спектрам. Созданные алгоритмы и программные средства для спектрального анализа в ИАС позволяют в короткие сроки осуществлять идентификацию оцениваемых потоков данных. В этих условиях используется теория статистических решений.

В этой теории существуют также два источника неопределенности. Во-первых, неизвестно, какому распределению подчиняются исходные потоки данных. Во-вторых, также не совсем известно, какое распределение имеет то множество (генеральная совокупность), о котором мы хотим сделать выводы по его подмножеству, образующему исходные данные.

Статистические процедуры в компьютерной ИАС это и есть процедуры для принятия решений, снимающих оба эти вида неопределенности. Отметим, что существует целый ряд причин, которые приводят к некорректному применению статистических методов:

- полученные статистические выводы, как и любые другие, всегда имеют некоторые оценки надежности или достоверности. Как правило, оцениваемая достоверность статистических выводов неизвестна, и она определяется в ходе статистического исследования;
- качество для принимаемых решений в среде ИАС, полученных посредством из применения статистических процедур, зависит также от качества потоков исходных данных;
- также, по-видимому, не следует «подвергать» статистической обработке потоков данных, не имеющие явно выраженной статистической природы изначально;
- необходимо в ИАС использовать статистические процедуры, соответствующие уровню для априорной информации исследуемой совокупности данных. Если распределение исходных потоков данных в среде ИАС неизвестно, то надо либо его установить, либо использовать несколько различных методов и моделей, чтобы сравнить результаты. Если они в моделировании значительно будут отличаются по своим статистическим характеристикам, то говорим о неприменимости некоторых из использованных процедур непосредственно в компьютерной среде ИАС.

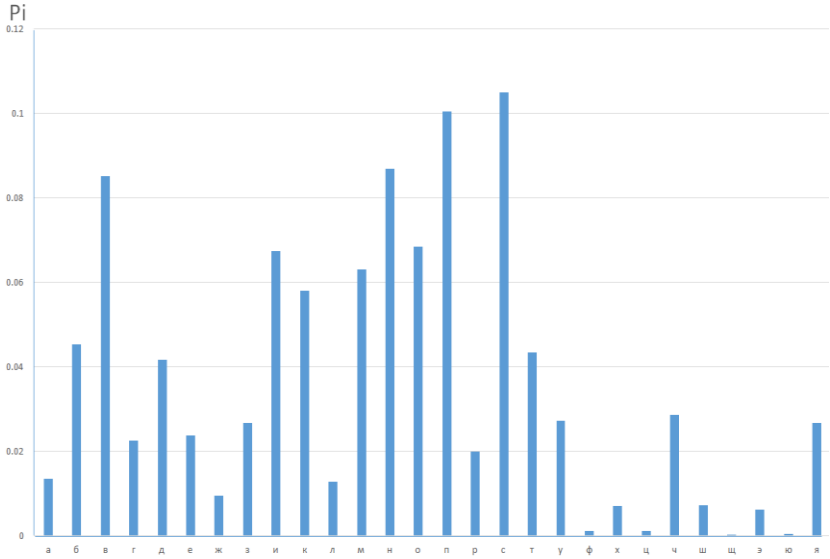


Рис. 5. Схематизация полученных вероятностей и спектров по предлогам в, на и с в произведениях А.С. Пушкина в компьютерной среде ИАС

Программа в ИАС, реализующая расчеты энтропийных характеристик, представляет собой описанные выше действующие алгоритмы. Покажем лишь один фрагмент из этой программы на алгоритмическом языке Питон, сам текст программы непосредственно представлен на экране монитора персонального компьютера (Рис.6 а)). Также покажем некоторые результаты информационного моделирования для машинных переводов известными ИАС и расчета информационной энтропии в созданной компьютеризированной среде ИАС (Рис.6 б), в), г), д)).

In [1]:
```
import pandas as pd
```

Создание списка `RU` возможных букв у русских слов:

In [2]:
```
RU = list('абвгдеёжзийклмнопрстуфхцчшщъыьэюя')
```

Создадим функцию `rf`, которя читает содержимое файла `f` в папке `d` и формирует `pandas` серию из букв русского текста

In [3]:
```
rf = lambda d,f: pd.Series([ l                                          # каждая букв
                             for s in open(d+'/'+f, encoding = 'utf-8') # для каждой
                         if s                                           # если строка
                       for w in s.split()                               # перебираем
                   for l in w.lower()                                   # переводим ,
               if l in RU])                                             # если каждый
```

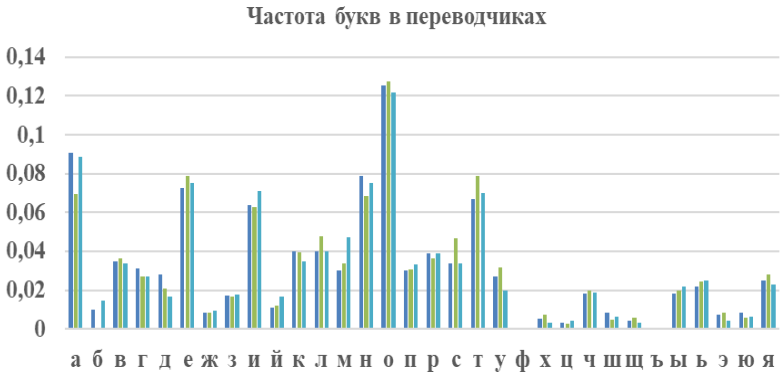
Импортируем из модуля `math` функцию логарифма

In [4]:
```
from math import log
```

Создадим функцию `df` получения `H` - таблицы статистических характеристик символов из множества букв текста

In [5]:
```
def df(d,f):
    exp = rf(d,f)
    Ni = exp.value_counts()                 # число появлений букв
    Pi = exp.value_counts(normalize=True) # вероятность появления каждой буквы

    H = pd.DataFrame( # создание pandas таблицы
        {'No':range(1,len(RU)+1)}, # со столбцом No: 1,2,...,кол-во букв в RU
        index = RU) # и индексами строк RU
    H.index.name = 'Буква' # Назовем индексный столбец 'Буква'

    H['Кол-во'] = H.index.map(                  # Создадим столбец числа появлений
        lambda x: x in Ni.index and Ni[x] or 0 ) # равный Ni, если буква встречалась
    H['Pi'] = H.index.map(                      # Создадим столбец вероятностей поя
        lambda x: x in Ni.index and Pi[x] or 0 ) # равный Pi, если буква встречалась
    H['Hi'] = H.Pi.map(                         # Из Pi создадим столбец для энтроп
        lambda x: x and -x * log(x,2) or 0 )  # равный Pi*Log2(Pi), если буква встре

    # Получим столбцы для букв и их Pi, отсортированных по значению Pi
    sort = H.sort_values(by=['Pi']).Pi
    H['sort'] = sort.index
    H['Psort'] = sort.values

    # Получим столбцы для букв и их Pi, нормализованных по значению Pi
    H['norm'] = H.sort.iloc[1:-1:2].append(H.sort.iloc[::-2]).values
    H['Pnorm'] = H.Psort.iloc[1:-1:2].append(H.Psort.iloc[::-2]).values

    return H
```

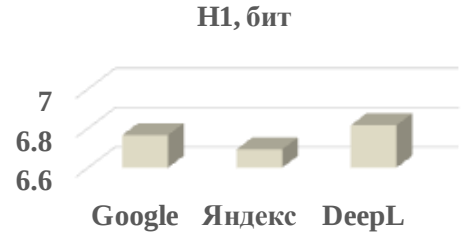
а)



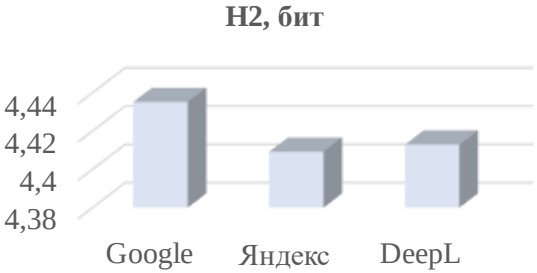
б)

Пере-водчик	Коли-чество слов, N1	H1, бит	Количе-ство букв, N2	H2, бит
Google	169	6,765470691	925	4,435127202
Яндекс	182	6,692949683	1068	4,409139975
DeepL	171	6,815399006	970	4,412856731

в)



г)



д)

Рис. 6. Фрагмент из расчета информационной энтропии в компьютеризированной среде ИАС

Анализ субтитров к полученным потокам данных и фрагменты из видео показали, что у каждого переводчика своя система машинного перевода [18, 19, 20]. Из трёх представленных машинных переводчиков для пары языков «японский-русский» для произведений А.С. Пушкина чаще встречаются буквы во всех трёх текстах субтитров – «о», «е», «т», «а», «и», а наименее частотные буквы – «ф» и «ъ», их показатели равны 0.

Заключение

Как было указано, рассматриваем поуровневое устройство информации как объекта исследования, создания ИАС для поддержки принятия решений субъектами управления. На первом

уровне мы исследуем данные, полученные в среде ИАС, причем отметим, что они не могут рассматриваться как абсолютно точные. Кроме того, очевидно, эти данные могут интересовать не сами по себе, а лишь в качестве некоторых сообщений и сигналов, которые, возможно, несут определенную и значимую информацию о том, что субъекта, принимающего решения в действительности интересует. Реалистичнее считать, что субъект имеет дело не просто с потоками данных, к тому же, вполне вероятно, зашумленными и неточными, но еще и косвенными, а возможно, и не полными изначально. Более того, эти данные касаются, как правило, не всей исследуемой (генеральной) совокупности, а лишь определенного ее подмножества, в котором можно вести обработку потока информации и фактически собрать необходимые оценки в ИАС. Однако при этом часто хочется сделать выводы обо всей совокупности, причем еще и знать степень достоверности этих выводов, что не вполне корректно. Субъект анализирует данные также в соответствии с личными интеллектуальными возможностями, способностями к анализу, синтезу, интеграции, но это тема отдельной работы.

Новизна данного исследования заключается в том, что на основе оценки вероятности и энтропийных характеристик, частные методы, некоторые варианты подсчетов которых были представлены авторами в [10], добавляется важный дополнительный фактор, а именно, учет мнений экспертов в среде ИАС на заключительных этапах. Таким образом, интегративный подход основан на комплексном использовании алгоритмов из экспертных и формальных оценок с целью синтеза и комбинировании новых возможностей для создания среды ИС, формирования автоматизированного блока для ИАС и ИЭС.

Таким образом, интегративный подход заключается в том, что мы соединяем энтропийные характеристики самого оцениваемого мультимедийного объекта (текста, символа, знака, потока чисел и др.) с энтропийными оценками мнений экспертов. Далее интегративный подход позволит, на наш взгляд, не просто идентифицировать и давать энтропийную оценку мультимедийной информации, но и в перспективе, при наработке достаточной базы знаний, поможет идентифицировать авторство ин-

формации как естественных (экспертных), так и искусственных генерируемых (автоматических) источников. Также в созданной среде ИАС можно устанавливать соответствие имеющегося в интернете контента определенным качественным критериям, сопоставлять характеристики и определять адекватность переводов искусственных переводчиков текстов (например, как показано здесь: Google Переводчик, Яндекс Переводчик и DeepL Переводчик [18, 19, 20]). Исследования в этом ключе будут акцентированы по выбору методов экспертизы, действующих алгоритмов анализа потоков мультимедийных данных и продолжены в прикладных задачах.

Литература

1. АКОФФ Р. Л. *Менеджмент в 21 веке. Трансформация корпорации*. - Томск: ТГУ, 2006. - 418 с.
2. БЕЛОВ В.С. *Информационно-аналитические системы. Основы проектирования и применения.* – М.: Издательский Центр ЕАОИ, 2008. – 111 с.
3. БЕЛОВ М.В., НОВИКОВ Д.А. *Методология комплексной деятельности*. М.: ЛЕНАНД, 2018. – 320 с.
4. ВИНЕР Н. *Кибернетика, или управление и связь в животном и машине; или Кибернетика и общество*. — М.: Наука; Главная редакция изданий для зарубежных стран, 1983. — 344 с.
5. НОВИКОВ Д.А. *Вокруг искусственного интеллекта складывается очень тревожная структура знаний и компетенций,* – академик Новиков. [Электронный ресурс] –https://new.ras.ru/mir-nauky/news/vokrug-iskusstvennogo-intellekta-skladyvaetsya-ochen-trevozhnaya-struktura-znaniy-i-kompetentsiy-aka/.
6. ПОЛТАВСКИЙ А.В. *Основы математической обработки информации вычислительных систем.* Учебное пособие. – М.: МГПУ, 2017 – 97 с.

7. ПОЛТАВСКИЙ А.В. *Программные средства вычислительных систем. Часть I. ЭВМ первых поколений*. Учебное пособие. – М.: МГПУ, 2014 – 87 с.
8. ПОЛТАВСКИЙ А.В. Программные средства вычислительных систем. Часть II. ЭВМ третьего и четвертого поколений. Учебное пособие. – М.: МГПУ, 2016 – 96 с.
9. ПОЛТАВСКИЙ А.В., РУСЯЕВА Е.Ю. *Интегративный подход к построению информационной системы мониторинга для комплексов беспилотных летательных аппаратов* / Труды 14-й Международной конференции «Управление развитием крупномасштабных систем» (MLSD-2021). – М.: ИПУ РАН, 2021. С. 1670-1677.
10. ПОЛТАВСКИЙ А.В., РУСЯЕВА Е.Ю., БУРБА А.А. *Устройство для содержательного анализа текстовой информации*: Патент на изобретение № 2568272 РФ; Дата публикации 27.10.2015. Бюл. №30
11. ПОЛТАВСКИЙ А.В., ФЕДЯНИНА В.А., СКОТЧЕНКО А.С. *Информационные и телекоммуникационные технологии в исследованиях* / Компьютерный практикум /– Москва, 2020.
12. ПУГАЧЕВ В.С. *Теория случайных функций и ее применение к задачам автоматического управления.*- М., «Наука», 1962.
13. СТРАТЕГИЯ НАЦИОНАЛЬНОЙ БЕЗОПАСНОСТИ РФ. Утверждена указом Президента РФ от 02.07.2021 №400.
14. ТРАПЕЗНИКОВ В.А. *Управление и научно-технический прогресс*. М.: Наука, 1983. – 223 с.
15. ТРАХТЕНГЕРЦ Э.А., ИВАНИЛОВ Е.Л., ЮРКЕВИЧ Е.В. *Современные компьютерные технологии управления информационно-аналитической деятельностью*. – М.: СИНТГЕГ, 2007. – 372 с.
16. УКАЗ *«О мерах по обеспечению технологической независимости и безопасности критической информационной инфраструктуры Российской Федерации».* Подписан Президентом 30 марта 2022 года. [Электронный ресурс] – URL: http://kremlin.ru/catalog/keywords/98/events/68090

17. ШЕННОН К.Э. *Работы по теории информации и кибернетике* (с предисловием академика А.Н. Колмогорова). Издательство иностранной литературы, 1963 г. – М.: – 824 с.

18. ЯНДЕКС ПЕРЕВОДЧИК [Электронный ресурс] – URL: https://translate.yandex.ru/;
19. DEEPL Переводчик [Электронный ресурс] – URL: https://www.deepl.com/ru/translator
20. GOOGLE Переводчик [Электронный ресурс] – URL: https://translate.google.com/?hl=ru;

INTEGRATIVE APPROACH IN CREATING INFORMATION-ANALYTICAL SYSTEM OF ENTROPY EVALUATION OF DATA

Elena Rusyaeva, Institute of Control Sciences of RAS, Moscow, Cand.Sc., (eur5@mail.ru).
Alexander Poltavsky, Institute of Control Sciences of RAS, Moscow, Doctor of Science, professor (avp57avp@yandex.ru).

Abstract: An algorithm for the identification of multimedia information based on the integrative evaluation of multimedia data and the calculation of the entropy concordance coefficient is presented. Examples of using an integrative approach based on formulas from information theory, theory of algorithms and technical cybernetics to the analysis of data flows in a computer network are given. The novelty of the proposed approach is in the synthesis of formalized models for calculating the entropy characteristics of the evaluated multimedia object itself and for calculating the entropy estimates of expert opinions. The presented integrative approach in the future is aimed at creating particular methods of identification, entropy evaluation of multimedia information, which, as a result, with the development of a sufficient knowledge base, will allow determining the authorship of the created content (information, knowledge) as natural (created by the subject, person) and artificial (automatically created) origin. In the environment of the constructed information-analytical system (IAS), it is possible not only to establish the correspondence of the authorship of the content available on the Internet, but also to evaluate the quality of the transmitted multimedia information according to certain criteria. In IAS, entropy characteristics are compared and the adequacy of translations of existing natural and artificial translators of texts is determined. The created system makes it possible to formalize the processes of data identification, to form an

adequate sample of relevant information to support decision-making by management subjects.

Keywords: information-analytical system, neural network, algorithm, concordance coefficient

УДК 004.021

ББК 16.33

Статья представлена к публикации
членом редакционной коллегии …заполняется редактором…

Поступила в редакцию …заполняется редактором…
Опубликована …заполняется редактором…