УДК 004.043

К.В. САЗОНОВ

ОЦЕНИВАНИЕ СЕМАНТИЧЕСКОГО СОДЕРЖАНИЯ СООБЩЕНИЙ НА ОСНОВЕ ПОТЕНЦИАЛЬНОЙ ИНФОРМАТИВНОСТИ

Сазонов К.В. **Оценивание семантического содержания сообщений на основе потенциальной информативности.**

Аннотация. В настоящей статье представлен анализ существующих подходов к оцениванию количества информации на различных уровнях ее представления. Введены и математически описаны понятия информационного потока и потенциальной информативности сообщения на синтактическом уровне представления. Сформулированы и доказаны теоремы, которые позволяют выполнить количественную оценку потенциальной информации. Предложен подход к оцениванию количества потенциальной информативности.

Ключевые слова: информация, неопределенность, прогнозирование, синтаксис, семантика, потенциальная информативность.

Sazonov K.V. **Evaluation of Semantic Content of Message based on Potential Informativeness.**

Abstract. This paper presents an analysis of the existing approaches to the estimation of the amount of information at different levels of its submission. The concepts of information flow and potential informative messages on the syntactic level of representation are described. Theorems that allow a quantitative estimation of the potential information are formulated and proved. An approach to the estimation of the number of potentially informative is proposed.

Keywords: information, uncertainty, forecasting, syntax, semantics, potential informativeness.

1. Введение. Одним из основных проблемных понятий в современной науке является понятие "информация". В настоящее время существует множество подходов к оцениванию количества информации:

– принцип неопределенности Гейзенберга [1];

– информация Фишера;

– информация и энтропия Шеннона [2].

В результате определить все подходы формально в универсальном смысле чрезвычайно сложно, что подчеркивает актуальность проблемы построения единой теории, призванной формализовать понятие информации и информационных процессов, а также описать превращения информации в процессах разной природы.

В теории передачи информации под формой представления информации подразумеваются сведения, являющиеся объектом некоторых операций, а именно: передачи, распределения, преобразования, хранения или непосредственного использования [2]. Особый интерес для систем обработки информации представляет процесс информационного взаимодействия, который состоит в передаче информации и

предполагает наличие источника (передатчика) S и ее получателя (интерпретанта) Pr.

Описание состояний источника S и получателя информации Pr осуществляется с помощью соответствующих им множеств параметров, мощности которых определяются числом возможных состояний, свойств и целей объектов информационного взаимодействия.

Процесс восприятия информации на уровне получателя связан с некоторым набором таких субъективных свойств информации, как важность, достоверность, своевременность, доступность, возможность ее измерения или количественного соотнесения и т.д. Свойства информации влияют на свойства сообщений и проявляются во взаимосвязях их элементов.

Под сообщением длиной m следует понимать совокупность $< s_j >_m$, $j = 0(1)m-1$ символов (знаков) источника (генеральной совокупности сообщений) S, определенных на множестве (алфавите) $S_{<N>}$, находящихся в определенных отношениях и связях друг с другом и образующих определенную целостность.

Таким образом, исходное сообщение в общем смысле есть форма представления информации в виде последовательности длиной m взаимосвязанных символов $s_j \in S_{<N>}$.

Анализируя количество, содержание и ценность информации в сообщениях, следует исходить из возможностей соответствующего анализа знаковых структур. Изучение связей между символами сообщения может быть реализовано путем измерения количества информации, содержащейся в сообщениях [3, 4].

При измерении количества информации ее свойства во внимание не принимаются. Кроме того, не вся информация имеет объективно измеряемое количество.

В результате многообразия и неоднозначности термина «информация» существует множество подходов к измерению количества информации. Однако, многообразие способов оценивания информации и неоднозначности субъективных оценок обуславливает необходимость разработки и внедрения новой концепции оценивания количества информации. При этом следует учитывать тот факт, что процесс информационного взаимодействия объектов с помощью телекоммуникационных систем предусматривает применение системы преобразований формы представления информации, отличающихся по сложности и структуре, а именно: кодирование источника, помехоустойчивое кодирование, модуляция и др., что обусловливает появление априорной неопределенности различных уровней относительно параметров этих преобразований.

2. Концептуальное описание понятия потенциальная информация. Обсуждения термина «информация» продолжаются до сих пор, однако часто носят не познавательный, а терминологический характер. Действительно, давно известно, что в реальном мире все материальное взаимодействует друг с другом (обменивается информацией), а проявлением этого взаимодействия является отражение, которое зачастую можно интерпретировать как сообщение, содержащее информацию об объекте. Отражение всякого материального объекта представляет собой некоторую упорядоченную вдоль оси времени структуру, которая характеризуется совокупностью связей между элементами, присутствующими в ней, распространяется в пространстве и изменяется во времени в том случае, когда объект является нестационарным.

Понятие «знаний» применимо не только к материальным, но и к идеальным объектам. В силу этого оно является более абстрактным, чем понятие информации, но понятие знаний представляется более конкретным, ориентирующим субъекта на совершение определенных действий. Понятие информации увязывается со знаниями конкретных предметов, их свойств, сторон и пр., что в философии называют предметом познания.

Предмет познания, по определению, является материальным, в силу чего информация действительно связана с отражением, которое проявляется как в физической, так и знаковой форме. Все знаковые и физические формы информации содержатся на каких-либо материальных предметах (носителях). Выявить смысл физической формы отражения без знаковой формы возможно, но описать без знаний нельзя. Указанное обстоятельство определяет более общий характер понятия знания по сравнению с понятием информации [5].

Таким образом, можно сформулировать такие важные понятия, как [6]:

– пространственный информационный поток – это конечное множество отражений материальных объектов, распространяющееся в пространстве, изменяющееся во времени и одновременно характеризующее его;

– сообщение информационного потока – это форма представления информации об окружающем пространстве, удобная для регистрирующей системы, выраженная в виде отражения материального объекта, содержащего данные о его структуре, свойствах и параметрах их изменения во времени;

– потенциальная информация как мера количественного описания воздействия сообщения информационного потока на рецепторы субъекта познания посредством отражений – количественная мера информации, которая содержится в сообщении информационного потока,

ограниченного на определенном временном интервале, и взаимосвязана с неопределенностью присутствующей в его структуре.

Опираясь на концептуальное понятие потенциальной информации можно сформулировать понятие потенциальной информативности сообщения.

Потенциальная информативность представляет собой среднее количество потенциальной информации, которая содержится в сообщении информационного потока, ограниченного на определенном временном интервале, и характеризует среднее значение неопределенности в связях между его элементами.

3. Концептуальное описание контента сообщения. Анализ и оценивание информации, содержащейся в различных типах сообщений семантических форматов (текстовых, звуковых, неподвижных и подвижных графических сообщениях), проводится на трех основных уровнях представления информации (рисунок 1), а именно: синтаксическом, семантическом и прагматическом [4].

Синтаксический уровень представления информации используется для описания комбинаторики символов (знаков) без учета значения и ценности, которую представляют эти символы и их сочетания как для субъекта или устройства, передающего информацию (передатчика), так и для потребителя информации (интерпретанта). Иными словами, на синтаксическом уровне описываются структурные свойства знаковых систем безотносительно к каким-либо их интерпретациям (составляющим предмет интересов семантики) и возможным интерпретаторам (рассматриваемым прагматикой).

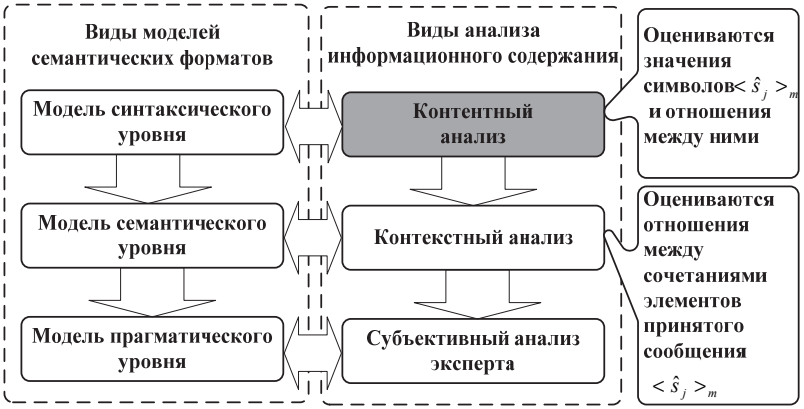

Рис. 1. Уровни представления информации

Исследование информационного содержания на синтаксическом уровне подразумевает оценивание значения символов $< s_j >_m$ и отношения между ними. Анализ такого типа позволяет выявить абсолютно любое информационно значимое либо содержательное наполнение сообщение, представленное в виде совокупности символов алфавита, а следовательно представляет собой ни что иное, как контентный анализ сообщения, а выявленное информационное содержание – контент сообщения. Таким образом, под содержанием сообщения подразумевается множество информационных символов, объединенных в единую структуру, с ярко выраженными особенностями в рамках одного сообщения (файла).

Исходя из выше сказанного, можно сделать вывод, что у сообщений с разным содержанием различная комбинаторика и различные отношения между символами, а следовательно можно предположить и различное количество информации на синтаксическом уровне, и данное количество информации может выступать в качестве информативного признака содержания сообщения.

Модель семантического уровня позволяет учитывать взаимосвязанность между символами с их содержанием (в терминах семиотики – между означающим и означаемым) без учета состояния как источника, так и приемника (интерпретанта) этих символов.

Предметом исследования и описания на этом уровне являются отношения между сочетаниями элементов принятого сообщения $S_{<m>}$ и понятиями, которые образуются в процессе обратной интерпретации сообщения $S_{<m>}$.

Анализ информационного содержания на семантическом уровне представления информации подразумевает оценивание отношений между сочетаниями элементов принятого сообщения. Представленный тип анализа позволяет оценить контекст сообщения (контекстный анализ) – относительно законченный по смыслу отрывок сообщения, в пределах которого наиболее точно и конкретно выявляется смысл и значение отдельного входящего символа или совокупности символов.

При исследовании и моделировании свойств сообщения на прагматическом уровне анализируется его смысловое содержание и отношение к источнику информации. При этом в качестве предмета анализа рассматриваются отношения между понятиями внутри некоторой системы или между понятиями, принадлежащими различным системам. Одна из таких систем образуется в процессе обратной интерпретации принятого сообщения $S_{<m>}$, вторая система отражает знания интерпретанта сообщения. В этом случае прагматика исследует

символы сообщения $S_{<m>}$ с точки зрения их ценности для интерпретанта, а иногда и для источника информации.

4. Математическая модель пространственного информационного потока. Пусть в некоторой области пространства присутствует множество процессов, доступных для наблюдения $\hat{S}^{*}_{<n>} \subseteq S_{<N>}$ и развивающихся во времени согласно законам теории вероятностей. Каждое сообщение $\hat{S}^{(i)}_{<m>} \in \hat{S}^{*}_{<n>}$ представляет собой набор данных:

$$\hat{S}^{*}_{<n>} = \{\hat{S}^{(i)}_{<m>}\}_n = \{<\hat{s}_j>^{(i)}_m\}_n = \{<\hat{s}_0, \hat{s}_1, ..., \hat{s}_j, ..., \hat{s}_{m-1}>^{(i)}\}_n,\ j = 0(1)m-1,\ i = 1(1)n, \quad (1)$$

где $i = 1(1)n$ – пространственная координата, определяющая положение сообщения в пространстве в фиксированный отсчет времени;

$j = 0(1)m-1$ – координата во временной области (последовательность дискретных отсчетов времени $t_j = j\Delta t$).

Данные (1) упорядочены вдоль оси аргумента – времени и пространственной координаты i, то есть представляют собой сообщение информационного потока, а множество $\hat{S}^{*}_{<n>}$ – непосредственно информационный поток.

Описанное множество сообщений имеет стохастическую природу и динамически изменяется с течением времени. Пусть $T = \{t_j \in T \mid t_0 \le t_j < t_0 + m\}$ – фрагмент оси времени, в пределах которого определяется наблюдаемый информационный поток $\hat{S}^{*}_{<n>}$. В этом случае каждое сообщение информационного потока может быть представлено как случайный процесс $\hat{S}^{(i)}_{<m>} \Leftrightarrow \hat{S}^{(i)}(t_j)$, а каждый элемент сообщения может быть описан в виде сечения случайного процесса в дискретные моменты времени как:

$$\hat{S}^{(i)}(t_j) = <\hat{s}^{(i)}_0, \hat{s}^{(i)}_1, ..., \hat{s}^{(i)}_{t_j}, ..., \hat{s}^{(i)}_{m\text{-}1}>, t_j \in T,\ j = 0(1)m-1,\ i = 1(1)n, \quad (2)$$

а в целом упорядоченная двумерная структура может быть представлена как множество временных рядов $\left\{\hat{S}^{(i)}(t_j)\right\}_n$, $j = 0(1)m-1,\ i = 1(1)n$.

Модель любого стохастического процесса может быть описана в следующем виде:

$$\hat{S}^{(i)}(t_j) = f^{(i)}(t_j) + \hat{a}^{(i)}(t_j),\ j = 0(1)m-1,\ i = 1(1)n, \quad (3)$$

где $\hat{a}^{(i)}(t_j)$ – случайная величина (шум, погрешность измерения), характеризуемая нормальным законом распределения с математическим ожиданием $M[\hat{a}]=0$ и дисперсией $D[\hat{a}]=const$;

$f^{(i)}(t_j)$ – функциональная зависимость, представляющая собой информационную составляющую сообщения;

$i=1(1)n$ – пространственная координата, определяющая пространственное положение стохастического процесса в фиксированный отсчет времени t_j, $j=0(1)m-1$.

Значения функций $f^{(i)}(t_j)$, $j=0(1)m-1$, $i=1(1)n$, соответствующие одному моменту времени и описывающие смежные векторы, могут быть связанны между собой в рамках плоскости, соответствующей сечению процесса во времени, и формируют единый образ, параметры которого изменяются во времени.

Таким образом, неопределенность, присутствующая в информационном потоке, может быть описана как множество функций времени $\{f^{(i)}(t_j)\}_n$, $j=0(1)m-1$, $i=1(1)n$, что эквивалентно множеству функций пространственных координат $\{f^{(i)}(j)\}_n$, $j=0(1)m-1$, $i=1(1)n$, в соответствии с описанием (1), где значения $i=1(1)n$ определяют положение вектора в плоскости сечения, а $j=0(1)m-1$ – положение символа в сообщении (векторе).

При этом множество всех функций времени $\{f^{(i)}(t_j)\}_n$, каждая из которых представляется в дискретные моменты времени t_j, образует собой информационный поток.

В свою очередь, информационный поток состоит из конечного множества элементов (символов алфавита) $\hat{s}_j^{(i)}$, $t_j \in T$, $j=0(1)m-1$, $i=1(1)n$. Наименьшим носителем неопределенности в структуре информационного потока являются значения двух соседних элементов, то есть элементом неопределенности предлагается считать пару смежных символов $\hat{s}_j^{(i)}$ и $\hat{s}_{j\pm1}^{(i\pm1)}$. В результате минимальный элемент неопределенности определяется как некоторая функция двух переменных $I(\hat{s}_j^{(i)},\hat{s}_{j+1}^{(i+1)})$.

5. Математическое описание количественной меры потенциальной информации. В соответствии с основными постулатами

теории информации синтаксическую информацию сообщения можно выразить как неопределенность его структуры.

Впервые термин неопределенность был выдвинут Гейзенбергом, в соответствии с его принципом неопределенность пропорциональна произведению приращения импульса элементарной частицы Δp к приращению ее координаты Δx.

Применительно к информационному потоку $\left\{\hat{S}^{(i)}(t_j)\right\}_n$ неопределенность и информация, которая в нем содержится, пропорциональна функции $I(\hat{s}_j^{(i)}, \hat{s}_{j+1}^{(i+1)}) \sim h$, с той лишь разницей, что в качестве приращения импульса выступает приращение значения случайной величины $\Delta\hat{s}_{j+1}^{(i+1)}$, а в качестве приращения координаты Δx – приращение одной или нескольких координат $i \pm 1$ и $j \pm 1$ ($t_j \pm \Delta t$) пространственного информационного потока.

Каждая отдельно взятая реализация $s_j^{(i)}$ случайной величины $\hat{s}_j^{(i)}$ представляет собой численное значение, определяемое функцией $f^{(i)}(t_j)$. В соответствии с принципом неопределенности Гейзенберга в фиксированный момент времени можно наблюдать только одну реализацию $\hat{s}_j^{(i)}$, в результате чего возникает неопределенность значений $\hat{s}_{j+1}^{(i+1)}$, смещенных на один отсчет по координатам i или j:

$$h \sim \Delta x \Delta p \Rightarrow \Delta p \sim \Delta\hat{s}_{j+1}^{(i+1)}, \Delta x \sim \Delta t_j \Rightarrow h \sim \Delta\hat{s}_{j+1}^{(i+1)} \Delta t_{j+1} \sim I(\hat{s}_{j+1}^{(i+1)}, \hat{s}_j^{(i)}). \quad (4)$$

Указанную неопределенность можно устранить только в том случае, когда известна реализация случайной величины, характеризующей приращение $\Delta\hat{s}_{j+1}^{(i+1)}$ значения $\hat{s}_j^{(i)}$ по координатам i и j относительно $\hat{s}_{j+1}^{(i+1)}$, при условии, что значение $\hat{s}_{j+1}^{(i+1)}$ неизвестно.

Согласно теории Шеннона количество информации, содержащееся в сообщении, зависит от степени неопределенности этого сообщения, которая характеризуется вероятностью его появления. Количество информации тем больше, чем оно менее вероятно. В результате количество информации, содержащееся в одном символе сообщения $s_j^{(i)}$, целесообразно определить как функцию вероятности появления этого символа $P(s_j^{(i)})$.

Понятие неопределенности неотъемлемо связано с понятием субъекта, воспринимающего информационный поток, ибо с точки зрения различных субъектов в одном и том же потоке может присутствовать различное количество информации (неопределенности).

Оценивание контента сообщения информационного потока становится возможным только после восприятия данного сообщения некоторой регистрирующей системой.

Неопределенность, присутствующая в структуре информационного потока, не позволяет субъекту, воспринимающему данные, безошибочно предсказывать значение последующего элемента. Таким образом, процедуру восприятия сообщения субъектом можно описать в виде следующей структурной схемы (рисунок 2).

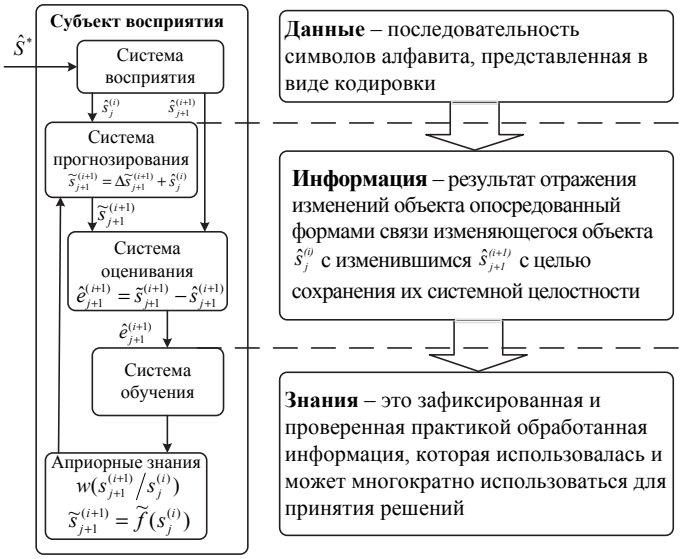

Рис. 2. Модель, поясняющая процесс восприятия сообщения субъектом

Пусть информационный поток $\left\{\hat{S}^{(i)}(t_j)\right\}_n$, $j = 0(1)m-1$, $i = 1(1)n$ доступен для наблюдения на интервале времени T ($t_j \in T$). Тогда субъекту в момент времени t_j становится доступно значение элемента $s_j^{(i)}$, принадлежащего сообщению $\hat{S}^{(i)}(t_j)$. На основе множества априорных данных о вероятностных характеристиках случайной величины $\hat{s}_{j+1}^{(i+1)}$, которые могут быть описаны в виде условного закона распреде-

ления $w(s_{j+1}^{(i+1)} / s_j^{(i)})$, либо на основе сформированного представления о функциональной зависимости $\tilde{s}_{j+1}^{(i+1)} = \tilde{f}(s_j^{(i)}) = \tilde{f}^{(i)}(t_j)$, субъект может осуществлять оценивание значения последующего элемента информационного потока.

Исходя из выше сказанного, в результате оценивания $s_j^{(i)}$ субъект прогнозирует наиболее вероятное, с его точки зрения, изменение элемента $s_j^{(i)}$, то есть $\Delta\tilde{s}_{j+1}^{(i+1)} = \tilde{s}_{j+1}^{(i+1)} - s_j^{(i)}$, в результате этого могут возникнуть следующие ситуации:

– для каждого элемента информационного потока $s_j^{(i)}$ соответствующее значение оценки $\Delta\tilde{s}_{j+1}^{(i+1)}$ совпадает с действительным значением $\Delta s_{j+1}^{(i+1)}$, в этом случае путем измерения любого $s_j^{(i)}$ можно вычислить значения всех элементов информационного потока. В таких условиях неопределенность в информационном потоке полностью отсутствует, в силу чего отсутствует и получаемая информация;

– субъект прогнозирует отличающееся значение элемента $\tilde{s}_{j+1}^{(i+1)} \neq s_{j+1}^{(i+1)}$, тогда после наблюдения реализации $\hat{s}_{j+1}^{(i+1)}$ определяется ошибка прогнозирования $\hat{e}_{j+1}^{(i+1)}$. Величина ошибки пропорциональна неопределенности значения элемента информационного потока $\hat{s}_{j+1}^{(i+1)}$ относительно значения элемента $\hat{s}_j^{(i)}$ и зависит от полноты описания условного закона распределения случайных величин $w(s_{j+1}^{(i+1)}/s_j^{(i)})$ для данного субъекта, либо от точности описания функциональной зависимости $\tilde{s}_{j+1}^{(i+1)} = \tilde{f}(s_j^{(i)})$. В том случае, когда между случайными величинами $\hat{s}_{j+1}^{(i+1)}$ и $\hat{s}_j^{(i)}$ прослеживается тесная взаимосвязь, у субъекта формируется более полное описание данной зависимости, что соответствует меньшим значениям ошибок прогнозирования $\hat{e}_{j+1}^{(i+1)}$;

– если априорные данные у субъекта относительно значения $\hat{s}_{j+1}^{(i+1)}$ полностью отсутствуют (совместное распределение $w(s_{j+1}^{(i+1)}/s_j^{(i)})$ и функция $\tilde{s}_{j+1}^{(i+1)} = \tilde{f}(s_j^{(i)})$ неизвестны, что соответствует полной неопределенности относительно структуры информационного потока) и реализация случайной величины $\hat{s}_j^{(i)}$ наблюдалась только один раз, то

в качестве наиболее ожидаемого значения предсказывается $\widetilde{s}_{j+1}^{(i+1)} = s_j^{(i)}$, в силу чего значение $\Delta\widetilde{s}_{j+1}^{(i+1)} = 0$, в этом случае возможны две ситуации:

а) все элементы сообщения действительно равны друг другу $s_{j+1}^{(i+1)} = s_j^{(i)} \Rightarrow \Delta s_{j+1}^{(i+1)} = 0$, прогноз окажется верным, и ошибки предсказания в этом случае отсутствуют, в силу чего информация так же отсутствует;

в) элементы информационного потока изменяются $s_{j+1}^{(i+1)} \neq s_j^{(i)} \Rightarrow \Delta s_{j+1}^{(i+1)} \neq 0$, ошибки предсказания принимают максимальные значения, что соответствует максимальной информативности потока.

После того, как субъектом определено значение $e_{j+1}^{(i+1)}$, он дополняет априорные данные об информационном потоке и тем самым корректирует совместное распределение случайных величин $w(s_{j+1}^{(i+1)}/s_j^{(i)})$ и параметры функциональной зависимости $\widetilde{s}_{j+1}^{(i+1)} = \widetilde{f}(s_j^{(i)})$ соседних элементов информационного потока (шаг оценивания). В результате после повторного наблюдения значение ошибки при оценивании $\Delta\widetilde{s}_{j+1}^{(i+1)}$ становится меньше, в силу чего количество неопределенности в связях смежных случайных величин уменьшается.

В результате определенного числа L наблюдений субъект может полностью скорректировать свои априорные данные о связанности элементов информационного потока, в результате чего значения оценок будут соответствовать результатам наблюдений $\Delta\widetilde{s}_{j+1}^{(i+1)} = \Delta s_{j+1}^{(i+1)}$, неопределенность в таком потоке устраняется.

Изложенный принцип составляет основу функционирования всех существующих систем распознавания. Число повторных наблюдений L интерпретируется как объем эталонных описаний, от которых зависит вероятность правильного распознавания.

Можно сделать вывод о том, что информативность потока принимает максимальные значения в том случае, когда поток воспринимается субъектом впервые. Воспринимая информационный поток как ранее не оцениваемый субъектом (распределение $w(s_{j+1}^{(i+1)}/s_j^{(i)})$ и $\widetilde{s}_{j+1}^{(i+1)} = \widetilde{f}(s_j^{(i)})$ неизвестны), можно определить количество информации, независимое от субъекта восприятия, то есть объективное количество информации – потенциальную (объективную) информативность.

Субъект оценивает информационный поток, полагаясь на свой априорные знания о данном информационном потоке и ошибку предсказания, на основании которой корректирует свою систему прогнозирования (посредством приобретения новых знаний), то есть обучается. В результате, посредством данной схемы возможно связать три основных понятия теории информации (рисунок 2).

Данные – последовательность символов алфавита, представленная в виде кодировки.

Информация – результат отражения изменений объекта опосредованный формами связи изменяющегося объекта $\hat{s}_{\mathrm{j}}^{(\mathrm{i})}$ с изменившимся $\hat{s}_{\mathrm{j+1}}^{(\mathrm{i+1})}$ с целью сохранения их системной целостности.

Знание – это зафиксированная и проверенная практикой обработанная информация, которая использовалась и может многократно использоваться для принятия решений.

Посредством получения из данных информации субъект пополняет свои знания.

6. Ошибка прогнозирования как количественная характеристика потенциальной информации. Взаимную связь двух смежных элементов информационного потока можно описать в виде некоторой функциональной зависимости двух случайных величин. Известный элемент в этом случае представляет собой объясняющую переменную $\hat{s}_{\mathrm{j}}^{(i)}$, а прогнозируемый – зависимую $\hat{s}_{\mathrm{j+1}}^{(\mathrm{i+1})}$. Такая односторонняя стохастическая зависимость называется простой регрессией (зависимость результата только от одной объясняющей переменной) $\widetilde{s}_{\mathrm{j+1}}^{(\mathrm{i+1})} = f(s_{\mathrm{j}}^{(\mathrm{i})})$. Значение регрессии определяет оценку наиболее вероятного значения зависимой случайной величины $\hat{s}_{\mathrm{j+1}}^{(\mathrm{i+1})}$ при заданном значении реализации $s_j^{(i)}$ случайной величины $\hat{s}_j^{(i)}$. Разброс значений случайной величины $\hat{s}_{\mathrm{j+1}}^{(\mathrm{i+1})}$ вокруг $\widetilde{s}_{\mathrm{j+1}}^{(\mathrm{i+1})}$ обусловлен влиянием неопределенности (энтропии), присутствующей в информационном потоке. Разность между эмпирическими значениями $s_{\mathrm{j+1}}^{(\mathrm{i+1})}$ и расчетным значением $\widetilde{s}_{\mathrm{j+1}}^{(\mathrm{i+1})}$ позволяет получить количественную оценку неопределенности как ошибки предсказания:

$$\tilde{s}_{\mathrm{j+1}}^{(\mathrm{i+1})} - \hat{s}_{\mathrm{j+1}}^{(\mathrm{i+1})} = \hat{e}_{\mathrm{j+1}}^{(\mathrm{i+1})}. \qquad (5)$$

При отсутствии неопределенности в информационном потоке расчетные значения равны эмпирическим $\tilde{s}_{j+1}^{(i+1)} = s_{j+1}^{(i+1)}$, в силу чего между значениями $\hat{s}_j^{(i)}$ и $\hat{s}_{j+1}^{(i+1)}$ существует жесткая функциональная зависимость, неподверженная влиянию случайных факторов (зависимость полностью детерминирована). В том случае, когда какие-либо взаимные связи между случайными величинами $\hat{s}_j^{(i)}$ и $\hat{s}_{j+1}^{(i+1)}$ отсутствуют, определить их функциональную зависимость $\tilde{s}_{j+1}^{(i+1)} = f(s_j^{(i)})$ не представляется возможным, и разница между расчетным значением $\tilde{s}_{j+1}^{(i+1)}$ и эмпирическим максимальна. В этом случае и прогнозы и ошибки не зависят от субъекта, воспринимающего поток, в результате неопределенность при восприятии в таких условиях объективна и принимает значения, непосредственно зависящие только от структуры сообщения. Таким образом, в случае отсутствия знаний у оценивающего субъекта неопределенность в структуре потока максимальна и принимает значение, объективно характеризующее отношение между элементами информационного потока, а следовательно и синтаксическое содержание сообщения, то есть является информативным признаком содержания сообщения.

В целях обоснования потенциальной неопределенности в качестве информативного признака требуется описать количественную меру неопределенности, заключенной в смежных элементах потока.

Теорема 1. Пусть случайные величины $\hat{s}_{t_j}^{(i)}$ и $\hat{s}_{j+1}^{(i+1)}$ являются соседними элементами одного информационного потока. Тогда количественно величину, характеризующую их неопределенность, можно представить как меру разброса значений элементов потока вокруг некоторой функциональной зависимости их значений – регрессией.

Доказательство. В теории математической статистики разброс наблюдаемых значений случайной величины $\hat{s}_{j+1}^{(i+1)}$ около ее математического ожидания $M[\hat{s}_{j+1}^{(i+1)}]$ характеризуется оценкой дисперсии вида:

$$\tilde{D}[\hat{s}_{j+1}^{(i+1)}] = \frac{\sum_{\gamma=1}^{N} (s_{j+1,\gamma}^{(i+1)} - \tilde{M}[\hat{s}_{j+1}^{(i+1)}])^2}{N-1}. \qquad (6)$$

Оценка дисперсии (6) является общей, и ее значение обусловлено изменениями объясняющих переменных, в силу чего можно выпол-

нить разложение дисперсии. Отклонение γ-го результата наблюдения от общего среднего можно представить в следующем виде:

$$s_{j+1,\gamma}^{(i+1)} - \tilde{M}[\hat{s}_{j+1}^{(i+1)}] = (s_{j+1,\gamma}^{(i+1)} - \tilde{s}_{j+1,\gamma}^{(i+1)}) + (\tilde{s}_{j+1,\gamma}^{(i+1)} - \tilde{M}[\hat{s}_{j+1}^{(i+1)}]). \quad (7)$$

После возведения в квадрат обеих частей соотношения (7) и суммирования по γ имеет место следующее равенство:

$$\sum_{\gamma=1}^{N}(s_{j+1,\gamma}^{(i+1)} - \tilde{M}[\hat{s}_{j+1}^{(i+1)})^2 = \sum_{\gamma=1}^{N}(s_{j+1,\gamma}^{(i+1)} - \tilde{s}_{j+1,\gamma}^{(i+1)})^2 + \\ +2\sum_{\gamma=1}^{N}(s_{j+1,\gamma}^{(i+1)} - \tilde{s}_{j+1,\gamma}^{(i+1)})(\tilde{s}_{j+1,\gamma}^{(i+1)} - \tilde{M}[\hat{s}_{j+1}^{(i+1)}]) + \sum_{\gamma=1}^{N}(\tilde{s}_{j+1,\gamma}^{(i+1)} - \tilde{M}[\hat{s}_{j+1}^{(i+1)}])^2. \quad (8)$$

С учетом выражения (6) и $\sum_{\gamma=1}^{N}\tilde{s}_{j+1,\gamma}^{(i+1)}\hat{e}_{j+1,\gamma}^{(i+1)} = 0$ выражение (8) принимает вид:

$$\sum_{\gamma=1}^{N}(s_{j+1,\gamma}^{(i+1)} - \tilde{M}[\hat{s}_{j+1}^{(i+1)}])^2 = \sum_{\gamma=1}^{N}(s_{j+1,\gamma}^{(i+1)} - \tilde{s}_{j+1,\gamma}^{(i+1)})^2 + \sum_{\gamma=1}^{N}(\tilde{s}_{j+1,\gamma}^{(i+1)} - \tilde{M}[\hat{s}_{j+1}^{(i+1)}])^2. \quad (9)$$

Разделив соотношение (9) на $N-1$, получим:

$$\frac{\sum_{\gamma=1}^{N}(s_{j+1,\gamma}^{(i+1)} - \tilde{M}[\hat{s}_{j+1}^{(i+1)}])^2}{N-1} = \frac{\sum_{\gamma=1}^{N}(s_{j+1,\gamma}^{(i+1)} - \tilde{s}_{j+1,\gamma}^{(i+1)})^2}{N-1} + \frac{\sum_{\gamma=1}^{N}(\tilde{s}_{j+1,\gamma}^{(i+1)} - \tilde{M}[\hat{s}_{j+1}^{(i+1)}])^2}{N-1} = \\ = \frac{\sum_{\gamma=1}^{N}(e_{j+1,\gamma}^{(i+1)})^2}{N-1} + \frac{\sum_{\gamma=1}^{N}(\tilde{s}_{j+1,\gamma}^{(i+1)} - \tilde{M}[\hat{s}_{j+1}^{(i+1)}])^2}{N-1}. \quad (10)$$

Выражение (10) представляет собой разложение оценки общей дисперсии на две составляющие:

$$\tilde{D}[\hat{s}_{j+1}^{(i+1)}] = \tilde{D}[\hat{e}_{j+1}^{(i+1)}] + \tilde{D}[\hat{\tilde{s}}_{j+1}^{(i+1)}], \quad (11)$$

где $\tilde{D}[\hat{\tilde{s}}_{j+1}^{(i+1)}]$ – оценка дисперсии значений регрессии, представляющая собой ту часть общей дисперсии $\tilde{D}[\hat{s}_{j+1}^{(i+1)}]$, которая обусловлена влиянием случайной величины $\hat{s}_{j+1}^{(i+1)}$ («объясненная» дисперсия, или дисперсия, обусловленная регрессией); $\tilde{D}[\hat{e}_{j+1}^{(i+1)}]$ – оценка дисперсии зна-

чений ошибки прогнозирования, представляющая собой ту часть общей дисперсии $\widetilde{D}[\hat{s}_{j+1}^{(i+1)}]$, которая не объясняется функцией регрессии («случайная», или остаточная дисперсия) и отражает независимость $\hat{s}_{j+1}^{(i+1)}$ от значения $\hat{s}_j^{(i)}$.

Из выражения (11) следует, что чем ближе оценка $\widetilde{D}[\hat{e}_{j+1}^{(i+1)}]$ приближается к нулю, тем меньше эмпирические значения случайной величины $\hat{s}_{j+1}^{(i+1)}$ отклоняются от значения регрессии $\widetilde{s}_{j+1}^{(i+1)}$. Иными словами, чем больше оценка дисперсии $\widetilde{D}[\hat{s}_{j+1}^{(i+1)}]$ по сравнению с $\widetilde{D}[\hat{e}_{j+1}^{(i+1)}]$, тем больше общая дисперсия формируется за счет объясняющей величины $\hat{s}_j^{(i)}$, в силу чего связь между значениями $\hat{s}_{j+1}^{(i+1)}$ и $\hat{s}_j^{(i)}$ более интенсивная.

Теорема доказана. ▲

Показателем интенсивности связи двух случайных величин является коэффициент детерминации, который с учетом разложения (11) представляет собой величину:

$$B(\hat{s}_j^{(i)}, \hat{s}_{j+1}^{(i+1)}) = \frac{\widetilde{D}[\widetilde{s}_{j+1}^{(i+1)}]}{\widetilde{D}[\hat{s}_{j+1}^{(i+1)}]} = 1 - \frac{\widetilde{D}[\hat{e}_{j+1}^{(i+1)}]}{\widetilde{D}[\hat{s}_{j+1}^{(i+1)}]},\ j = 0(1)m-1,\ i = 1(1)n. \qquad (12)$$

В том случае, когда коэффициент детерминации $B(\hat{s}_j^{(i)}, \hat{s}_{j+1}^{(i+1)}) = 1$, все эмпирические значения $\hat{s}_j^{(i)}$ лежат на регрессионной прямой ($\widetilde{D}[\hat{e}_{j+1}^{(i+1)}] = 0$). В свою очередь, при $B(\hat{s}_j^{(i)}, \hat{s}_{j+1}^{(i+1)}) = 0$ линия регрессии параллельна оси абсцисс, а коэффициенты регрессии при этом равны нулю.

Из выражения (12) следует, что отношение $\dfrac{\widetilde{D}[\hat{e}_{j+1}^{(i+1)}]}{\widetilde{D}[\hat{s}_{j+1}^{(i+1)}]}$ представляет собой показатель неопределенности связей и может быть использовано в качестве коэффициента неопределенности:

$$H(\hat{s}_j^{(i)}, \hat{s}_{j+1}^{(i+1)}) = 1 - B(\hat{s}_j^{(i)}, \hat{s}_{j+1}^{(i+1)}) = \frac{\widetilde{D}[\hat{e}_{j+1}^{(i+1)}]}{\widetilde{D}[\hat{s}_{j+1}^{(i+1)}]},\ j = 0(1)m-1,\ i = 1(1)n. \qquad (13)$$

Из выражения (13) следует, что большей связанности элементов соответствует меньшая энтропия и меньшее количество информации, заключенной в них, в силу чего коэффициент неопределенности и ко-

личество информации, содержащейся в близлежащих элементах сообщения, должны быть пропорциональны, т.е.:

$$\forall \hat{s}_{\mathrm{j}} \exists \{\hat{I}_j\}_m, \hat{I}_j = I_j(\hat{s}_{\mathrm{j}}^{(\mathrm{i})}, \hat{s}_{\mathrm{j+1}}^{(\mathrm{i+1})}) \sim H_j(\hat{s}_{\mathrm{j}}^{(\mathrm{i})}, \hat{s}_{\mathrm{j+1}}^{(\mathrm{i+1})}), j = 0(1)m-1, i = 1(1)n, \quad (14)$$

где $I_j(\hat{s}_{\mathrm{j}}^{(\mathrm{i})}, \hat{s}_{\mathrm{j+1}}^{(\mathrm{i+1})})$ – количество информации, содержащейся в значениях двух соседних элементов сообщения информационного потока;

m – число элементов в сообщении $\hat{S}^{(i)}(t_j)$.

Информационный поток можно представить как множество временных рядов $\left\{\hat{S}^{(i)}(t_j)\right\}_n$, $j = 0(1)m-1$, $i = 1(1)n$. Для прогнозирования последующего элемента временного ряда $\hat{S}^{(i)}(t_j)$ при известном значении $\hat{s}_{\mathrm{j}}^{(i)}$ принято использовать два подхода:

– в первом случае прогнозирование последующего значения ряда по одному или нескольким предыдущим значениям осуществляется с помощью известной функциональной зависимости (регрессии) $\tilde{s}_{\mathrm{j+1}}^{(\mathrm{i+1})} = f(\mathrm{s}_{\mathrm{j}}^{(\mathrm{i})})$, в этом случае оценка элемента ряда называется регрессионной средней;

– во втором случае (вид регрессии неизвестен) в качестве оценки принимается наиболее вероятное значение случайной величины $\hat{s}_{\mathrm{j}}^{(\mathrm{i})}$, то есть оценка моды $\tilde{s}_{\mathrm{j+1}}^{(\mathrm{i+1})} = \tilde{M}o[\hat{s}_{\mathrm{j+1}}^{(\mathrm{i+1})}]$, в этом случае оценка элемента ряда называется вариационной средней и обозначается как $\dot{\tilde{s}}_{\mathrm{j+1}}^{(\mathrm{i+1})}$.

Таким образом, процессы прогнозирования, используемые в теории анализа временных рядов, эквивалентны процедуре оценивания субъектом значения случайной величины $\hat{s}_{\mathrm{j+1}}^{(\mathrm{i+1})}$.

Проведенный анализ позволяет сделать вывод, что по аналогии с описанием количества информации в информатике, величина ошибки прогнозирования одной случайной величины относительно другой характеризует количество информации совместно с условным распределением этих случайных величин. При этом количество информации $I_j(\hat{s}_{\mathrm{j}}^{(\mathrm{i})}, \hat{s}_{\mathrm{j+1}}^{(\mathrm{i+1})})$, присутствующее в реализации случайной величины $\hat{s}_{\mathrm{j+1}}^{(\mathrm{i+1})}$ относительно $\hat{s}_{\mathrm{j}}^{(\mathrm{i})}$, можно выразить через ошибку прогнозирования. Вместе с тем, в соответствии с выражением (14) количество информации зависит от коэффициента неопределенности $H_j(\hat{s}_{\mathrm{j}}^{(\mathrm{i})}, \hat{s}_{\mathrm{j+1}}^{(\mathrm{i+1})})$, в силу

чего можно утверждать, что величина коэффициента неопределенности $H_j(\hat{s}_{\text{j}}^{(\text{i})}, \hat{s}_{\text{j+1}}^{(\text{i+1})})$ и реализации ошибки $e_{\text{j+1}}^{(\text{i+1})}$ функционально зависимы.

В целях отыскания количественных характеристик коэффициента неопределенности требуется описать функциональную зависимость коэффициента неопределенности и ошибки оценивания элемента информационного потока.

Теорема 2. Пусть существует величина $H_j(\hat{s}_{\text{t}_\text{j}}^{(\text{i})}, \hat{s}_{\text{j+1}}^{(\text{i+1})})$, представляющая собой показатель неопределенности связей двух элементов информационного потока $\hat{s}_{\text{j+1}}^{(\text{i+1})}$ и $\hat{s}_{\text{j}}^{(\text{i})}$. Тогда значение неопределенности пропорционально величине ошибки прогнозирования значения $\widetilde{s}_{\text{j+1}}^{(\text{i+1})}$ при известном значении $\hat{s}_{\text{j}}^{(\text{i})}$:

$$H_j(\hat{s}_{\text{j}}^{(\text{i})}, \hat{s}_{\text{j+1}}^{(\text{i+1})}) \sim \left| \widetilde{s}_{\text{j+1}}^{(\text{i+1})} \mid \hat{s}_{\text{j}}^{(\text{i})} - \hat{s}_{\text{j+1}}^{(\text{i+1})} \right| \sim e_{\text{j+1}}^{(\text{i+1})} , \tag{15}$$

где $\widetilde{s}_{\text{j+1}}^{(\text{i+1})} \mid \hat{s}_{\text{j}}^{(\text{i})}$ – оценка случайной величины $\hat{s}_{\text{j+1}}^{(\text{i+1})}$ при условии, что известно значение реализации величины $\hat{s}_{\text{j}}^{(\text{i})}$.

Доказательство. Пусть имеет место несмещенная оценка дисперсии ошибки прогнозирования величины $\hat{e}_{\text{j+1}}^{(\text{i+1})}$, математическое ожидание которой известно и с учетом ее аналитического описания равно $M[\hat{e}_{\text{j+1}}^{(\text{i+1})}] = 0$. При этом указанная оценка дисперсии может быть представлена в виде:

$$\widetilde{D}[\hat{e}_{\text{j+1}}^{(\text{i+1})}] = \frac{1}{N}\sum_{\gamma=1}^{N}(e_{\text{j+1},\gamma}^{(\text{i+1})} - M[\hat{e}_{\text{j+1}}^{(\text{i+1})}])^2 = \frac{1}{N}\sum_{\gamma=1}^{N}(e_{\text{j+1},\gamma}^{(\text{i+1})})^2 . \tag{16}$$

Тогда коэффициент неопределенности (13) с учетом выражения (16) можно записать следующим образом:

$$H_j(\hat{s}_j^{(i)}, \hat{s}_{j+1}^{(i+1)}) = \frac{\widetilde{D}[\hat{e}_{j+1}^{(i+1)}]}{\widetilde{D}[\hat{s}_{j+1}^{(i+1)}]} = \frac{\sum_{\gamma=1}^{N}(e_{j+1,\gamma}^{(i+1)})^2}{N \cdot \widetilde{D}[\hat{s}_{j+1}^{(i+1)}]} = \frac{\overline{(e_{j+1}^{(i+1)})^2}}{\widetilde{D}[\hat{s}_{j+1}^{(i+1)}]}, \tag{17}$$

где $e_{\text{j+1},\gamma}^{(\text{i+1})}$ – ошибка при прогнозировании γ-го элемента алфавита.

Отношение (17) позволяет выявить, какая часть общего рассеяния значений случайной величины $\hat{s}_{\text{j+1}}^{(\text{i+1})}$ обусловлена факторами, не

зависящими от значения случайной величины $\hat{s}_{\mathrm{j}}^{(\mathrm{i})}$. При этом чем большую долю в общей дисперсии составляет оценка $\widetilde{D}[\hat{e}_{\mathrm{j+1}}^{(\mathrm{i+1})}]$, тем меньшее влияние оказывает значение случайной величины $\hat{s}_{\mathrm{j}}^{(\mathrm{i})}$, то есть связь между $\hat{s}_{\mathrm{j}}^{(\mathrm{i})}$ и $\hat{s}_{\mathrm{j+1}}^{(\mathrm{i+1})}$ менее интенсивная.

Из выражения (17) следует пропорциональность значений коэффициента неопределенности $H_j(\hat{s}_{\mathrm{t_j}}^{(\mathrm{i})}, \hat{s}_{\mathrm{j+1}}^{(\mathrm{i+1})})$ элемента сообщения $\hat{s}_{\mathrm{j+1}}^{(\mathrm{i+1})}$ относительно $\hat{s}_{\mathrm{j}}^{(\mathrm{i})}$ и среднего значения квадрата ошибки прогнозирования $\overline{(e_{\mathrm{j+1}}^{(\mathrm{i+1})})^2}$ (начального момента 2-го порядка $\nu_2[\hat{e}_{\mathrm{j+1},\gamma}^{(\mathrm{i+1})}]$), причем, чем большее значение принимает коэффициент неопределенности, тем ближе среднее значение квадрата ошибки прогнозирования $\overline{(e_{\mathrm{j+1}}^{(\mathrm{i+1})})^2}$ к значению оценки дисперсии $\widetilde{D}[\hat{s}_{\mathrm{j+1}}^{(\mathrm{i+1})}]$. Коэффициент пропорциональности в соответствии с выражением (17) представляет собой величину $1/\widetilde{D}[\hat{s}_{\mathrm{j+1}}^{(\mathrm{i+1})}]$. В том случае, когда неопределенность максимальна, справедливо равенство $\overline{(e_{\mathrm{j+1}}^{(\mathrm{i+1})})^2} = \widetilde{D}[\hat{s}_{\mathrm{j+1}}^{(\mathrm{i+1})}]$, что полностью соответствует процессу с нулевым математическим ожиданием («белый шум»), который в классической теории информации принято воспринимать как максимально информативное сообщение.

Теорема доказана. ▲

Основываясь на выводах, полученных в соответствии с теоремой 2, можно предложить альтернативный вариант вычисления среднего значения квадрата ошибки прогнозирования:

$$\overline{(e_{\mathrm{j+1}}^{(\mathrm{i+1})})^2} = H_{\mathrm{j}}^{(\mathrm{i})}(\hat{s}_{\mathrm{j+1}}^{(\mathrm{i+1})}, \hat{s}_{\mathrm{j}}^{(\mathrm{i})})\widetilde{D}[\hat{s}_{\mathrm{j+1}}^{(\mathrm{i+1})}] \ . \qquad (18)$$

Подставляя в выражение (18) выражение для коэффициента детерминации (13) можно представить выражения для вычисления среднего значения квадрата ошибки оценивания смежных случайных величин по их реализациям:

$$\overline{(e_{j+1}^{(i+1)})^2} = \frac{1}{N(N-1)}\left(N\sum_{\gamma=1}^{N}(s_{j+1,\gamma}^{(i+1)})^2 - \left(\sum_{\gamma=1}^{N}s_{j+1,\gamma}^{(i+1)}\right)^2\right) - \frac{\left(N\sum_{\gamma=1}^{N}s_{j,\gamma}^{(i)}s_{j+1,\gamma}^{(i+1)} - \left(\sum_{\gamma=1}^{N}s_{j,\gamma}^{(i)}\right)\left(\sum_{\gamma=1}^{N}s_{j+1,\gamma}^{(i+1)}\right)\right)^2}{N(N-1)\left(N\sum_{\gamma=1}^{N}(s_{j,\gamma}^{(i)})^2 - \left(\sum_{\gamma=1}^{N}s_{j,\gamma}^{(i)}\right)^2\right)} \cdot \qquad (19)$$

Выражение (19) позволяет отойти от процедур прогнозирования элементов сообщения и использовать для вычисления квадрата среднего значения ошибки предсказания непосредственно значения элементов сообщения.

Таким образом, если для каждого элемента информационного потока $\hat{s}_{\mathrm{j}}^{(\mathrm{i})}$ определены все ошибки прогнозирования связанных с ним элементов $\hat{s}_{\mathrm{j+1}}^{(\mathrm{i+1})}$, то можно представить неопределенность всего потока в виде множества ошибок прогнозирования:

$$\left\{\sqrt{\overline{(e_{\mathrm{j+1}}^{(\mathrm{i+1})})^2}}\right\}_{\mathrm{mn}}, i = 1(1)n,\ j = 0(1)m-1. \tag{20}$$

Вследствие чего значение потенциальной информативности, которая присутствует в информационном потоке, можно представить в следующем виде:

$$I = \sum_{i=1}^{n}\sum_{j=0}^{m-1}\sqrt{\overline{(e_{\mathrm{j+1}}^{(\mathrm{i+1})})^2}},\ i = 1(1)n,\ j = 0(1)m-1. \tag{21}$$

7. Заключение. В итоге можно сделать вывод, что вся потенциальная информативность, которая содержится в потоке, может быть представлена в виде множества ошибок оценивания смежных случайных величин. Величина I зависит от интенсивности связей между элементами информационного потока и характеризует количество информации сообщения на семантическом уровне. Иными словами, выражение (21) представляет объективно существующую информацию сообщения, которая не зависит от субъекта при условии, что до восприятия первых элементов потока $\hat{s}_{t1}^{(i)}$ любые априорные знания у субъекта отсутствовали.

Количество информации, определяемое выражением (21), соответствует концептуальному понятию потенциальной информации [7]. В результате можно сделать вывод, что величина I является количественной мерой потенциальной информации.

Полученные аналитические выражения, позволяют описать информативные признаки распознавания семантического содержания сообщений информационного потока, в целях обеспечения поиска как потенциально ценной информации, так и вредоносного контента.

Для практического использования аналитических моделей требуется сопоставление различных типов семантических сообщений соответствующим аналитическим моделям и информативным признакам.

Литература

1. *Гейзенберг В.* Избранные философские работы // С.-Пб.: Наука. 2006. 576 с.
2. *Левин В.И.* К.Э. Шеннон и современная наука // Вестник ТГТУ. 2008. Том 14. №3. С. 703–724.
3. *Астахов М.А., Ростовцев Ю.Г., Яфраков М.Ф.* Информационная борьба и знаковые системы // М.: Издательство «ТОМ». 2007. 334 с.
4. *Ростовцев Ю.Г.* Основы построения автоматизированных систем сбора и обработки информации // СПб.: ВИКИ. 1992. 216 с.
5. *Данилин С.Н.* О современном понятии информации // Информационные технологии. 2003. № 11. С. 53–57.
6. *Присяжнюк С.П., Сазонов К.В.* Потенциальная информативность как новая характеристика отражения материального объекта // Информация и космос. 2006. №2. С. 100–105.
7. *Сазонов К.В.* Модели оценивания потенциальной информативности потоков сообщений // Наукоемкие технологии. 2009. Т. 10. № 12. С. 63–69.

References

1. Gejzenberg V. *Izbrannye filosofskie raboty* [Selected philosophical works]. M.: Nauka, 2006. 576 p. (In Russ.)
2. Levin V.I. [C.E. Shannon and modern science]. *Vestnik TGTU – Bulletin of the TGTU*. 2008. vol. 14. no. 3. pp. 703–724 (In Russ.)
3. Astakhov M.A., Rostovtcev Y.G., Yafrakov M.F. *Informacionaia borba i znakovie sistemy* [Information warfare and sign systems]. M .: Publisher “TOM”. 2007. 334 p. (In Russ.)
4. Rostovtcev Y.G. *Osnovy postroeniya avtomatizirovannyh sistem sbora i obrabotki informacii* [Fundamentals of automated systems for the collection and processing of information]. SPb .: Vicki. 1992. 216 p. (In Russ.)
5. Danilin S.N. [On the current concept of information]. *Informacionnye tekhnologii – Information Technology*. 2003. vol. 11. pp. 53–57. (In Russ.)
6. Prysiazhnyuk S.P., Sazonov K.V. [Potential as a new informative reflection characteristics of the material object]. *Informaciya i kosmos – Information and Space.* 2006. vol. 2. pp. 100–105.
7. Sazonov K.V. [Models estimating potential information content of message flows]. *Naukoemkie tekhnologii – High Tech.* 2009. vol. 10. no. 12, pp. 63–69.

Сазонов Константин Викторович — д-р техн. наук, начальник кафедры инженерного анализа, Военно-космическая академия имени А. Ф. Можайского. Область научных интересов: системы сбора и обработки информации, обратное проектирование современных телекоммуникационных систем. Число научных публикаций — 65. Staffa78@mail.ru; ул. Ждановская д. 13, г. Санкт-Петербург, 199178, РФ; р.т.: (812) 230-28-15, Факс: (812)237-12-49.

Sazonov Konstantin Viktorovich — Ph.D., Dr. Sci., head of engineering analysis department, Mozhaisky Military Space Academy. Research interests: the system of data collection and processing, reverse engineering of modern telecommunication systems. The number of publications — 65. Staffa78@mail.ru; st. Zhdanovskaya d. 13, St. Petersburg, 199178, Russian Federation; office phone: (812) 230-28-15, Fax: (812)237-12-49.

РЕФЕРАТ

Сазонов К.В. **Оценивание семантического содержания сообщения на основе потенциальной информативности.**

Любое сообщение информационного потока может быть представлено в виде некоторой упорядоченной структуры элементов и связей между ними. В том случае, когда для определенной интеллектуальной системы, воспринимающей сообщение, структура и связи между элементами известны, для данной системы такое сообщение полностью детерминировано, и неопределенность в его структуре полностью отсутствует. Если предсказание элемента выполняется с ошибкой, система оценивает ее значение для адаптации и корректировки процедуры прогнозирования, в результате чего на следующем шаге ошибка прогнозирования снижается. Иными словами интеллектуальная система получает некоторый объем информации, которая содержится в смежных элементах, на основании этого происходит самообучение системы с целью минимизации ошибки следующих прогнозов.

При условии отсутствия у системы, воспринимающей сообщение, априорных данных о сообщении, процедура обучения выполняется впервые, и как следствие, образуются максимальные ошибки прогнозирования. Эффективное значение ошибок прогнозирования представляет собой объективную оценку количества информации, не зависящую от воспринимающей системы и получившую название потенциальной информации сообщения.

Содержание сообщения на семантическом уровне зависит от структурных особенностей сообщения, а также от плотности (скачков связанности) неопределенности в структуре сообщения. Таким образом, для различных классов контента величина потенциальной информации принимает различные значения, а эффективное значение ошибки прогнозирования предлагается использовать в качестве информативного признака распознавания семантического содержания сообщений.

SUMMARY

Sazonov K.V. **Evaluation of Semantic Content of Message based on Potential Informativeness.**

Any message of information flow can be represented in the form of an ordered structure of the elements and the relationships between them. In that case, when for a certain intelligent system that receives a message, the structure and relationships between elements are known to the system a message is completely determined, and uncertainty in its structure completely absent. If the prediction element fails, the system evaluates its importance for adapting and correcting the prediction procedure, whereby in the next step the prediction error decreases. An intelligent system receives a certain amount of information that is contained in the adjacent cell. Then the system is self-learned in order to minimize errors following forecasts.

In the absence of a system that receives the message, a priori data about the message the process of learning for the first time, and as a result, the maximum prediction error is formed. The effective value of prediction errors is an objective assessment of the amount of information that is independent of the receiving system and received the name of the potential of information messages.

The content of posts on the semantic level depends on the structural features of messages, and the density (jumps connectivity) uncertainty in the structure of the message. Thus, for various classes of potential information content value takes different values, and the effective value of the prediction error is proposed as an informative feature recognition semantic content of messages.