УДК 004.041

К ВОПРОСУ О СИСТЕМНОЙ ИНФОРМАЦИИ

Болбаков Р.Г., к.т.н., МГТУ МИРЭА, E-mail: bolbakov@mirea.ru, Москва, Россия

Аннотация. Статья проводит обзор особенностей состояния и развития информационной энтропии и системной информации. За основу оценки системной информации берутся чистые и смешанные состояния, в которых может находится составная система. Анализируется понятие информационной энтропии.
Ключевые слова: Информация, информатика, информационные ресурсы, информационная энтропия, информационные модели.

TO THE PROBLEM OF SYSTEM INFORMATION

Bolbakov R.G., PhD., MSTU MIREA, E-mail: bolbakov@mirea.ru, Moskow, Russia

Abstact. The article reviews the characteristics of the state and development of the informational entropy and system information. As a basis for evaluation of the system information is taken pure and mixed states, which can contain a composite system. Is analyzed the concept of information entropy.
Keywords: Information, informatics, information resources, informational entropy, information models

Введение. Последние десятилетия характеризуются широким применением информации и информационных систем[1, 2]. Развитие и использование информационных технологий ведут к тому, что жизненные циклы информационных систем становятся все более короткими. Существующие тенденции развития общества требуют эффективного использования информационных ресурсов и систем. Современные информационные ресурсы и технологии связывают с понятием «цифровые» [3, 4].Основные свойства информационных ресурсов включают: кодифицируемость, индентифицируемость, стандартизованность, измеримость [5]. Среди этих свойств следует выделить измеримость, как основу количественной оценки информационных процессов и информационных взаимодействий[6].

Анализ информационных систем на основе измеримости можно осуществлять разными методами, к числу которых относится и метод системной информации. Поскольку информационные системы носят относительно перманентный характер и в то же время легко изменяются под влиянием новой информации, можно утверждать, что важную роль для них играет теория информации[7]. Большую роль в статистической теории информации играет энтропия [8]. В современном понимании это полисемическое понятие, имеющее множество значений.

Энтропия (в естественных науках) –мера беспорядка системы, состоящей из

многих элементов [9];

Энтропия (в статистической физике) – мера вероятности осуществления какого–либо макроскопического состояния [9];

Энтропия (в теории информации) – мера неопределённости какого–либо опыта (испытания), который может иметь разные исходы, а значит, и разное количество информации [9];

Энтропия (в исторической науке) – экспликация феномена альтернативности истории (инвариантности и вариативности исторического процесса) [9];

Энтропия (в информатике) – степень неполноты, неопределённости знаний.

Энтропия (в корпоративных информационных системах) – степень неполноты, неопределённости управленческой информации [10].

Таким образом, энтропия – один из подходов анализа информации и информационных систем.

Процессы обработки информации составляют суть умственной деятельности человека. Человек думает, вычисляет, говорит, слушает, читает, пишет и печатает. При этом он всегда имеет дело с информацией.

Получаемая потребителем информация всегда поступает из некоторого источника. В этом случае говорят о передаче информации. Информация передается по каналу передачи, направляясь от источника к приемнику. Канал передачи – это некоторое среда, которая осуществляет доставку информации [11]. Природа информационных каналов – колебательные движения среды: звуковые, световые, электромагнитные волны и пр. С открытием радиоволн и созданием устройств, их генерирующих и улавливающих в процессах передачи информации произошли революционные изменения. Возникли и быстрыми темпами развиваются инфокоммуникационные технологии.

Информация передается в виде последовательности сигналов, составляющих информационное сообщение [12]. Физический смысл сигнала, с помощью которого передается информация, может не совпадать со смыслом передаваемой информации. Восприятие информации немыслимо без определенных предварительных соглашений и знаний, без которых сигнал будет восприниматься лишь как сообщение о некотором факте, который непонятно как интерпретировать. Для достижения взаимопонимания необходима предварительная договоренность о значениях сигналов.

Обработка информации – процесс преобразования уже имеющейся информации в информационном поле [13]. Преобразование информации может быть связано с изменением ее содержания или формы представления. В последнем случае говорят о кодировании информации. Например, к обработке информации могут быть отнесены шифрование информации или перевод текстов на другой язык.

Информационная энтропия. Информационная энтропия — мера информации приходящейся на одно элементарное сообщение источника, вырабатывающего статистически независимые сообщения.

Информационная энтропия для независимых случайных событий x с n возможными состояниями (от 1 до n) рассчитывается по формуле:

$$H(x) = -\sum_{i=1}^{n} p(i)\log_2 p(i)$$

Эта величина также называется средней энтропией сообщения. Величину

$$\log_2 \frac{1}{p(i)}$$

называют частной энтропией, характеризующей только i-е состояние, исследуемой системы.

Таким образом, энтропия события «x» является суммой с противоположным знаком всех произведений относительных частот появления события i, умноженных на их же двоичные логарифмы (основание 2 выбрано только для удобства работы с информацией, представленной в двоичной форме). Это определение для дискретных случайных событий можно расширить для функции распределения вероятностей.

Шеннон [7] предположил, что прирост информации равен утраченной неопределённости, и задал требования к её измерению:

- мера должна быть непрерывной, т.е. изменение значения величины вероятности на малую величину должно вызывать малое результирующее изменение функции;
- в случае, когда все варианты, например, буквы в приведённом примере равновероятны, увеличение количества вариантов (букв) должно всегда увеличивать значение функции;
- должна быть возможность сделать выбор (в нашем примере букв) в два шага, в которых значение функции конечного результата должно являться суммой функций промежуточных результатов.

В связи с этим, функция энтропии *H* должна удовлетворять условиям:

$H(p_1,...,p_n)$ определена и непрерывна для всех $p_1,...,p_n$, где $p_1 \in [0,1]$ для всех i=1, …, n и $p_1+...+p_n = 1$. Нетрудно видеть, что эта функция зависит только от распределения вероятностей, но не от алфавита.

Для целых положительных *n*, должно выполняться следующее неравенство:

$$H\underbrace{\left(\frac{1}{n},...,\frac{1}{n}\right)}_{n} < H\underbrace{\left(\frac{1}{n+1},...,\frac{1}{n+1}\right)}_{n+1}$$

Для целых положительных b_i, где $b_1 + \ldots + b_k = n$, должно выполняться равенство:

$$H\underbrace{\left(\frac{1}{n},...,\frac{1}{n}\right)}_{n} = H\left(\frac{b_1}{n},...,\frac{b_k}{n}\right) + \sum_{i=1}^{k}\frac{b_i}{n}H\underbrace{\left(\frac{1}{b_i},...,\frac{1}{b_i}\right)}_{b_i}$$

Шеннон показал, что единственная функция, удовлетворяющая этим требованиям, имеет вид:

$$-K\sum_{i=1}^{n} p(i)\log_2 p(i)$$

где K – константа (и в действительности нужна только для выбора единиц измерения). Он определил, что измерение энтропии применимое к источнику информации,

$$H = -p_1\log_2 p_1 - ... - p_n\log_2 p_n,$$

может определить требования к минимальной пропускной способности канала, требуемой для надёжной передачи информации в виде закодированных двоичных чисел.

Для вывода формулы Шеннона необходимо вычислить математическое ожидание «количества информации», содержащегося в цифре из источника информации. Мера энтропии Шеннона выражает неуверенность реализации случайной переменной.

Таким образом, информационная энтропия является разницей между информацией, содержащейся в сообщении, и той частью информации, которая точно известна (или хорошо предсказуема) в сообщении.

Примером этого является избыточность языка – имеются явные статистические закономерности в появлении букв, пар последовательных букв, троек и т. д.

Определение энтропии по Шеннону связано с понятием термодинамической энтропии. Работы Больцмана и Гиббса выполнили большую работу по статистической термодинамике, которая способствовала принятию термина «энтропия» в информационной теории. Существует связь между термодинамической и информационной энтропией.

Определить энтропию случайной величины можно введя предварительно понятия распределения случайной величины X, имеющей конечное число значений:

$$P_X(x_i) = p_i,$$

$$P_X(x_i) = p_i,\ p_i \geq 0, i = 1,2,...,n_1$$

$$\sum_{i=1}^{n} p_i = 1$$

и собственной информации:

$$I(X) = \log_2 P_x(X).$$

Тогда энтропия определяется как:

$$H(X) = E(I(X)) = -\sum_{i=1}^{n} p(i) \log_2 p(i)$$

От основания логарифма зависит единица измерения информации и энтропии: бит, нат или хартли.

Рассмотрим свойства энтропии в семантическом представлении. Энтропия является количеством, определённым в контексте вероятностной модели для источника данных.

Например, кидание монеты «как принятие решения» имеет энтропию $-2(0{,}5 \log_2 0{,}5) = 1$ бит на одно действие (при условии его независимости). У источника, который генерирует строку, состоящую только из букв «А», энтропия равна нулю:

$$-\sum_{i=1}^{\infty} \log_2 1 = 0.$$

Так, например, опытным путём можно установить, что энтропия английского текста равна 1,5 бит на символ, что конечно будет варьироваться для разных текстов [14]. Степень энтропии источника данных означает среднее число битов на элемент данных, требуемых для её зашифровки без потери информации, при оптимальном кодировании.

Некоторые биты данных могут не нести информации. Например, структуры данных часто хранят избыточную информацию или имеют идентичные секции независимо от информации в структуре данных.Количество энтропии не всегда выражается целым числом бит.

Энтропийные характеристики и математические свойства в семантике текстового анализа заключаются в следующем:

Неотрицательность:$H(X) \geq 0$

Ограниченность: $H(X) \leq \log_2 [X]$. Равенство, если все элементы из *X* равновероятны.

Если X, Yнезависимы, то$H(XY) \leq H(X) + H(Y)$.

Энтропия– выпуклая вверх функция распределения вероятностей элементов.

Если X,Yимеют одинаковое распределение вероятностей элементов, то$H(X) = H(Y)$.

Алфавит может иметь вероятностное распределение далекое от равномерного. Если исходный алфавит содержит *n* символов, тогда его можно сравнить с «оптимизированным

алфавитом», вероятностное распределение которого равномерное.

Соотношение энтропии исходного и оптимизированного алфавита – это эффективность исходного алфавита, которая может быть выражена в процентах. Эффективность исходного алфавита с *n* символами может быть также определена как его *n*–арная энтропия.

Энтропия ограничивает максимально возможное сжатие без потерь (или почти без потерь), которое может быть реализовано при использовании теоретически типичного набора или, на практике,– кодирования Хаффмана, кодирования Лемпеля – Зива – Велча или арифметического кодирования.

Анализ информационных систем. Аддитивные ИС[15, 16] (от лат. *additivus, additio* – «прибавляю») – системы, которые являются аддитивными по отношению друг к другу.Если они содержат разностороннюю информацию об одних объектах, то множества идентифицированных классов и атрибутов в таких системах пересекаются. Это позволяет провести идентификацию объекта в разных системах, а множества функциональных и дополнительных классов атрибутов различаются.

Аддитивные системы отвечают принципу сложения энтропийных вкладов и реализуют передаточную функцию в виде комплексного переменного, где действительная часть – энтальпия, мнимая – вероятностная энтропийная характеристика качества ИС, т.е. отвечают принципам эмерджентности, эргодичности и аддитивности, значения функции всегда не выходят из поля (пространства) Лебега [15].

Наиболее простые в управлении аддитивные системы (Рис. 1) называются транзакционными [15], что обусловлено однозначной предсказуемой и всегда повторяемой их реакцией на воздействие одиночных транзакций.

Субтрактивные ИС (Рис 2) – системы, содержащие общую объединенную информацию о разноплановых или разнородных объектах. Их можно рассматривать как множества идентифицированных классов атрибутов в таких системах пересекаются, что позволяет провести идентификацию объекта в разных системах, а множества функциональных и дополнительных классов атрибутов – различаются, например, системы мультимедиа и компьютерной графики (обучающие и дескриптивные, геоинформационные системы, коннекторы и т.п.)[16].

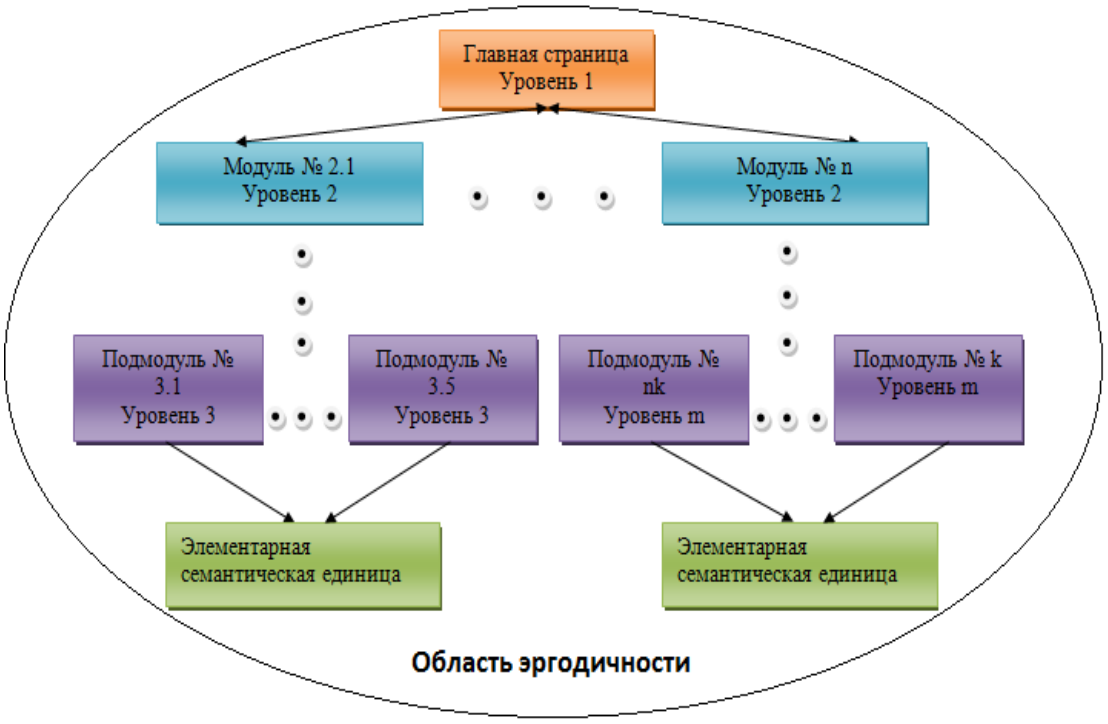

Рис. 1. **Архитектура аддитивных портальных систем.**

Рис. 2. **Архитектура субтрактивных информационных систем**

Мультипликативные ИС (Рис. 3) – системы, обладающие как субтрактивным так и аддитивными свойствами, приобретаемыми в результате взаимодействия в процессе интеграции систем. Для адаптивных интегрированных структур признаки

мультипликативности весьма ожидаемы [15].

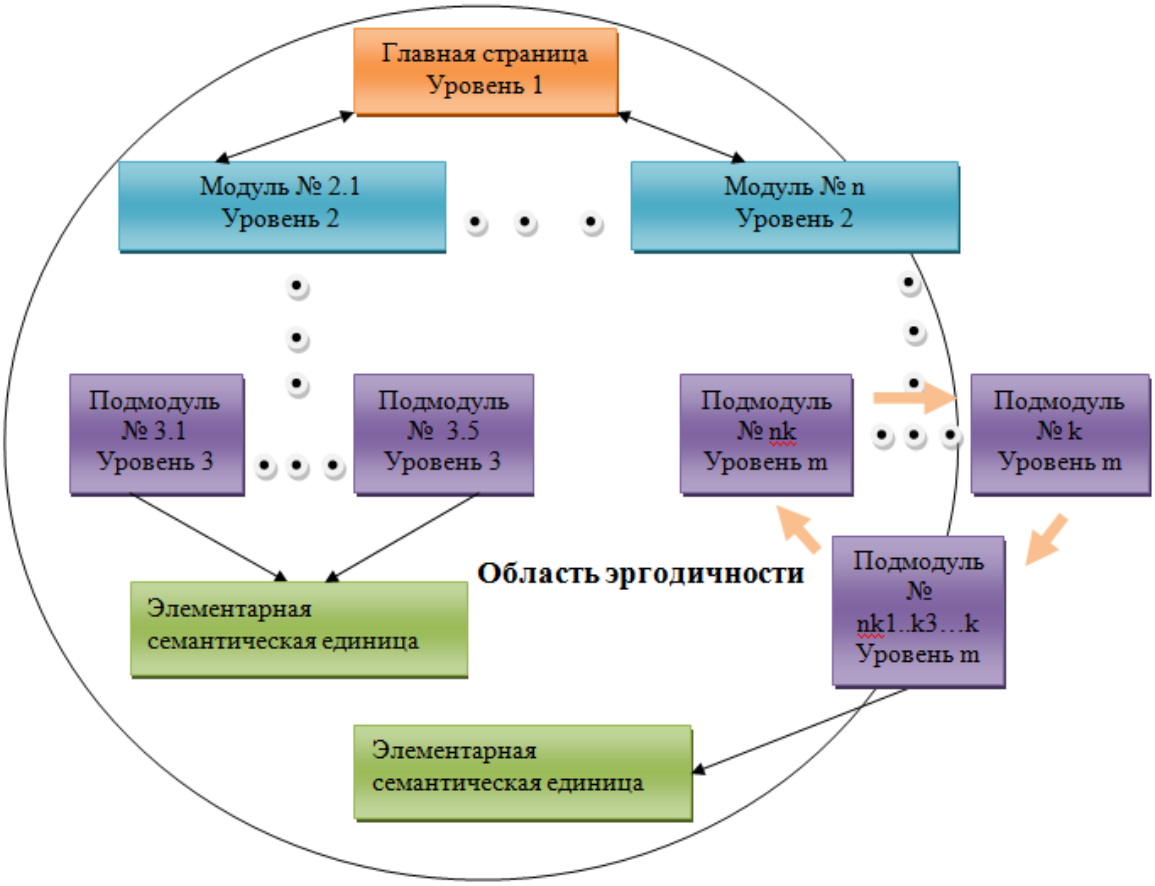

Рис. 1.3. **Архитектура мультипликативных информационных систем**

О системной информации. Введем понятие чистые состояния системы *W,* под которым будем понимать возможные дискретные состояния, в которых может находиться исследуемая система или объект. Через *I*обозначим количество информации по К.Э. Шеннону [7]. Согласно работам Луценко [17, 18] системное обобщение формулы Хартли [7] для равновероятных состояний объекта управления представляется в виде последовательности:

$$I = \log_2 W \tag{1}$$

$$I = \log_2 W^{\varphi} \tag{2}$$

$$I = \log_2 \sum_{m=1}^{M} C_W^m \tag{3}$$

$$I = \log_2 (C_W^1 + C_W^2 + ... + C_W^M), \tag{4}$$

$$при \quad M = W: \quad \sum_{m=1}^{M} C_W^m = 2^W - 1. \tag{5}$$

$$I = \log_2 (2^W - 1) \approx W \tag{6}$$

при W>> 1; I ≈ W с очень малой и быстро уменьшающейся погрешностью. Здесь W – количество чистых (классических) состояний системы;

φ – коэффициент эмерджентности Хартли (уровень системной организации объекта, имеющего W чистых состояний). Он также называется уровнем системности.

Гипотезао законе возрастания эмерджентности. Исследование математических выражений системной теории информации позволило сформулировать гипотезу о существовании "Закона возрастания эмерджентности"[17, 18]. Суть этой гипотезы состоит в том, что в самих элементах системы содержится сравнительно небольшая доля всей записанной в ней информации, а основной ее объем составляет системная информация подсистем различного уровня иерархии.

Различие между классическим и предложенным системными понятиями информации соответствует различию между понятиями множества и системы, на основе которых они сформированы.

$$I = \log_2 W^{\varphi} = \log_2 \sum_{m=1}^{M} C_W^m \tag{7}$$

$$\varphi = \frac{\log_2 \sum_{m=1}^{M} C_W^m}{\log_2 W} \tag{8}$$

$$I(W,M) = \log_2 W^{\frac{\log_2 \sum_{m=1}^{M} C_W^m}{\log_2 W}} \tag{9}$$

$$I(W,M) \approx \log_2 W^{\frac{W}{\log_2 W}} = W \tag{10}$$

$$I_{SYSTEM} \approx W - \log_2 W \tag{11}$$

$$I(W,M) = \log_2 W + \log_2 W^{\varphi - 1} \tag{12}$$

$$\varphi = \frac{\log_2 \sum_{m=1}^{M} C_W^m}{\log_2 W} \approx \frac{W}{\log_2 W} \tag{13}$$

$$I_{SYSTEM} \approx W - \log_2 W \tag{14}$$

Гипотеза 1: «О природе сложности системы». Сложность системы определяется разными факторами, одним из которых является количество, содержащейся в ней информации. В этой гипотезе открытым остается вопрос о мере количества информации. Например, количество бит может описывать текстовую структурированную систему, текстовую неструктурированную систему, иерархическую систему и топологическую

систему

Гипотеза 2: «О видах системной информации». Системная информация включает две составляющие:зависящую от количества элементов системы; зависящую также от характера взаимосвязей между элементами

Лемма 1:При увеличении количества элементов в системе доля системной информации в ней возрастает нелинейно с тенденцией к уменьшению.

Закон возрастания эмерджентности:Чем больше элементов в системе, тем большую долю от всей содержащейся в ней информации составляет системная информация, содержащаяся не в элементах, а в подсистемах различной сложности и уровнях иерархии.

Следствие:Увеличение уровня системности влияет на объект аналогично повышению уровня детерминированности: понижение уровня системности, также как и степени детерминированности системы приводит к ослаблению влияния факторов на поведение системы, т.е. к понижению управляемости системы за счет своего рода «инфляции факторов»

Лемма 2:Чем выше уровень системности, тем большая доля информации системы содержится во взаимосвязях ее элементов

Лемма 3:Чем меньше элементов в системе, тем быстрее возрастает доля информации, содержащейся во взаимосвязях элементов при возрастании уровня системности.

Классическая формула Харкевичаимеет вид.

$$I_{ij} = \log_2 \frac{P_{ij}}{P_j} \qquad (15)$$

где P_{ij}– вероятность перехода объекта управления в *j*-е состояние в условиях действия *i*-го фактора; P_j– вероятность самопроизвольного перехода объекта управления в *j*-е состояние, т.е. в условиях отсутствия действия *i*-го фактора или в среднем.

Известно, что корреляция не является мерой причинно-следственных связей. Если значение корреляции между действием некоторого фактора и переходом объекта управления в определенное состояние высокое, то это не значит, что данный фактор является причиной этого перехода. Для того чтобы по корреляции можно было судить о наличии причинно-следственной связи, необходимо сравнить исследуемую группу с контрольной группой, в которой данный фактор не действовал.

Высокая вероятность перехода объекта управления в определенное состояние, так же как и высокая корреляция, в условиях действия некоторого фактора сама по себе не говорит о наличии причинно-следственной связи между ними, т.е. о том, что данный фактор обусловил переход объекта в это состояние. Это связано с тем, что вероятность перехода объекта в это состояние может быть вообще очень высокой и независимо от действия

фактора.

Поэтому в качестве меры силы причинной обусловленности определенного состояния объекта действием некоторого фактора Харкевич предложил логарифм отношения вероятностей перехода объекта в это состояние в условиях действия фактора и при его отсутствии или в среднем (15). Таким образом, семантическая мера информации Харкевича является мерой наличия причинно-следственных связей между факторами и состояниями объекта управления.Выражение классической формулы Харкевича через частоты фактов.

$$P_{ij} = \frac{N_{ij}}{N_i}; \quad P_i = \frac{N_i}{N}; \quad P_j = \frac{N_j}{N} \tag{16}$$

где $N_i = \sum_{j=1}^{W} N_{ij}; \quad N_j = \sum_{i=1}^{M} N_{ij}; \quad N = \sum_{i=1}^{W}\sum_{j=1}^{M} N_{ij}$

$$I_{ij} = \log_2 \frac{N_{ij} N}{N_i N_j} \tag{17}$$

Однако мера Харкевича (15), в отличие от меры Шеннона, не удовлетворяет принципу соответствия с мерой Хартли, т.е. не переходит в меру Хартли в детерминистском случае, когда каждому будущему состоянию объекта управления соответствует единственный уникальный фактор и между факторами и состояниями имеется взаимно однозначное соответствие (19):

$$I_{ij} = \log_2 \left(\frac{N_{ij} N}{N_i N_j} \right)^{\Psi} \tag{18}$$

$$\forall\ N_{ij} = N_i = N_j = 1 \tag{19}$$

Откуда $I_{ij} = \log_2 N^{\Psi} = \log_2 W^{\varphi}$ (20)

$$\Psi = \frac{\log_2 W^{\varphi}}{\log_2 N} \tag{21}$$

Ниже приводится вывод системного обобщения формулы Харкевича:

$$\Psi = \frac{\log_2 W^{\frac{\log_2 \sum_{m=1}^{M} C_W^m}{\log_2 W}}}{\log_2 N} \tag{22}$$

$$I_{ij} = \log_2\left(\frac{N_{ij}N}{N_i N_j}\right)^{\Psi} = \log_2\left(\frac{N_{ij}N}{N_i N_j}\right)^{\frac{\log_2 W^{\varphi}}{\log_2 N}} =$$

$$= \frac{\log_2 W^{\varphi}}{\log_2 N}\left(\log_2\left(\frac{N_{ij}}{N_i N_j}\right) + \log_2 N\right) =$$

$$= \log_2\left(\frac{N_{ij}}{N_i N_j}\right)^{\frac{\log_2 W^{\varphi}}{\log_2 N}} + \log_2 W^{\varphi}$$

Окончательное выражение для системного обобщения формулы Харкевича

$$I_{ij} = \log_2\left(\frac{N_{ij}}{N_i N_j}\right)^{\frac{\log_2 W^{\varphi}}{\log_2 N}} + \log_2 W^{\varphi} \qquad (23)$$

Таким образом, в распоряжении исследователя имеются выражения: для системных обобщений Хартли и Харкевича ; для количества информации и плотности информации Шеннона ; гипотезы о законе возрастания эмерджентности; аналитические выражения для коэффициентов Хартли и Харкевича, которые являются обоснованными в рамках оговоренных условий системной теории информации количественными мерами уровня системности и степени детерминированности систем (23) [07, 18].

Заключение. Использование системной теории информации в ряде случаев, оговоренных в гипотезах, позволяет оценить эффективность такой системы и осуществлять управление такими системами. Использование системной теории информации и информационной энтропии позволяет решать задачи когнитивного моделирования[19, 20] и проводить оценку меры познания, то есть когнитивной энтропии.

Список литературы

1. Болбаков Р.Г. Теорема Байеса в когнитивной семантике образовательных информационных систем/ Современные проблемы науки и образования. – 2012. – № 5; URL: www.science–education.ru/105–7074 (дата обращения: 06.11.2012).

2. Болбаков Р.Г., Раев В.К. Моделирование когнитивной семантики образовательных информационных систем //Информатизация образования и науки, 2013. - № 1(17). - с. 91–102.

3. Цветков В.Я. Цифровые карты и цифровые модели // Геодезия и аэрофотосъемка. 2000, №2. с.147-155

4. Цветков В. Я. Клод Элвуд Шеннон, как основоположник цифрового моделирования // Перспективы науки и образования- 2014. - №1. – с44-50

5. Матчин В.Т. Информационные ресурсы как инструмент научного исследования и развития // Вестник МГТУ МИРЭА «MSTUMIREAHERALD» 2014 - № 2 (3) - с.235-256.

6.V. Ya. Tsvetkov.Information Interaction as a Mechanism of Semantic Gap Elimination // European Researcher, 2013, Vol.(45), № 4-1, p.782- 786.

7. Шеннон К. Работы по теории информации и кибернетике. – М.: Изд. иностр. лит., 1963. – 830 с.

8. Волькенштейн М. В. Энтропия и информация. – М.: Наука, 2006., 192 с.

9. Википедия. Свободная энциклопедия [Электронный ресурс]. – Режим доступа: http://ru.wikipedia.org/

10. V. Ya. Tsvetkov, N.V. Azarenkova. Entropy in corporate information systems // European Researcher, 2014, Vol.(70), № 3-1, p.471-477

11. Осин А.В. Мультимедиа в образовании: контекст информатизации. – М.: Агентство "Издательский сервис", 2004. – 320 с.

12. Кудж С.А.,Цветков В. Я. Информационные сообщения. - М.: Московский государственный технический университет радиотехники, электроники и автоматикиМГТУ МИРЭА , 2013.- 142 с., электронное издание рег.свид. №34320 от 24.12.2013. номер гос регистрации 0321305022

13. TsvetkovV.Ya. Informationfield.// LifeScienceJournal2014- 11(5). –pp.551-55.

14. Лийв Э. Х. Инфодинамика. Обобщенная энтропия и негэнтропия // Электронная библиотека Грамотей. 2010. URL: http://www.gramotey.com/books/271133718619.htm (дата обращения: 05.11.2010/

15. Ильин И.В., Петров К.А., Трифонов Л.А., и др. Онтология моделирования и проектирования семантических информационных систем и порталов– 254 с. – ОФАП – МОСКВА, 2008 / Рег.№11389

16.Монахов С.В., Савиных В.П., Цветков В.Я. Методология анализа и проектирования сложных информационных систем. - М.: Просвещение, 2005. - 264с

17. Луценко Е.В. Обобщенный коэффициент эмерджентностиХартли как количественная мера синергетического эффекта объединения булеанов в системном обобщении теории множеств / Политематический сетевой электронный научный журнал кубанского государственного аграрного университета. Издательство: Кубанский государственный аграрный университет (краснодар) ISSN: 1990–4665, 2011, 27–37 с.

18. Луценко Е.В. Количественные меры возрастания эмерджентности в процессе эволюции систем (в рамках системной теории информации) / Политематический сетевой электронный научный журнал кубанского государственного аграрного университета. Издательство: Кубанский государственный аграрный университет (краснодар) ISSN: 1990–4665, 2006, 1–20 с.

19. TsvetkovV.Ya. Cognitive information models. // Life Science Journal -2014; -11(4). - pp468-471.

20. Болбаков Р.Г. Моделирование когнитивной семантики образовательных информационных систем на основе когнитив–энтропии / Деп. в ВИНИТИ № 65–В 2012 от 10.02.2012