УДК 004.056:004.75

М.Р. Шабан, м.Київ

ПРЕДВАРИТЕЛЬНАЯ ОЦЕНКА ЭНТРОПИИ ЭЛЕКТРОННЫХ ДОКУМЕНТОВ ПРИ ПРОВЕДЕНИИ ГОСУДАРСТВЕННОЙ ЭКСПЕРТИЗЫ АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ

Abstract. In this regard, it is advisable to stay on the received wide popularity Shannon model. The model shows that for the description of information exchange essential quantitative assessment of participating information can be obtained by introducing a discrete probability measure space communications.

Различные научные направления предполагают различные подходы к определению информации. Существует множество определении [1]. Наличие этого множества подтверждает отсутствие универсального подхода к определению информации, поэтому новая область применения требует нового подхода к определению информации и требует разработки и формированию исходных понятий. Для анализа информации в случае проведения государственной экспертизы автоматизированной системы определяющим является наличие нескольких групп документов семантически и причинно-следственно связанных между собой, а существенным, в этом случае, является отображение информации в виде дискретного множества элементов (информационных символов или даже констант). Современный тренд развития математической логики и теории моделирования позволяет охватывать все больше объем знаний, которые можно описать формально.

В связи с этим, целесообразно остановиться на получившей широкую известность модели К.Шенонна [2]. В модели показано, что для описания информационного взаимодействия существенные количественные оценки участвующем в нем информации можно получить, вводя на дискретном пространстве сообщений вероятностную меру. Согласно Шеннону, дискретный источник информации представляется как некоторый эргодический марковский процесс. Исходя из вероятности генерации источником того или иного сообщения, Шеннон вводит показатель энтропии информации, характеризующую количество генерируемой информации. Модель Шеннона можно усовершенствовать применительно к информационным потокам электронного взаимодействия в документообороте.

Дискретный вероятностный процесс является марковским, если некоторая система в дискретные моменты времени может находиться в одном из конечного числа состояний S_1, S_2, … S_n , и для любой пары состояний i, j задана вероятность p_i (j) перехода системы за один шаг из состояния S_i в состояние S_j . Подчеркнем, что вероятность зависит только от номеров

состояний i и j и не зависит от того, каким образом система попала в состояние i, т.е. не зависит от момента времени (номера шага). Марковский процесс является эргодическим, если при стремлении времени (числа шагов) у бесконечности, вероятность нахождения системы в состоянии j стремится к некоторому пределу.

Пусть имеется некоторое множество возможных событий, вероятности осуществлений которых p_1, p_2, … p_n . Эти вероятности известны, но это все, что нам известно относительно того, какое событие произойдет. Шеннон строит на множестве событий меру H(p_1, p_2, … p_n), называемую им энтропией и показывающую, насколько велик «выбор» из такого набора событий или сколь неопределен для нас его исход. В приложении к информации такая мера эквивалентна оценке значимости сообщения о том, что произошло некоторое событие из n возможных и, фактически, предполагается количественная оценка информации.

Естественно в таком случае потребовать, чтобы мера (энтропия) H = H(p_1, p_2, … p_n), удовлетворяла следующим условиям:

- энтропия H должна быть непрерывной относительно;
- если все равны, то энтропия H должна быть монотонно возрастающей функцией от n. При увеличении неопределенности число n возможных событий, значимость информации об осуществлении конкретного события должна увеличиваться;
- если бы выбор распадался на два последовательных выбора, то первоначальная энтропия H должна быть взвешенной суммой энтропии индивидуальных выборов.

Существует единственная функция H, удовлетворяющая перечисленным свойствам. При этом энтропия H имеет вид:

$$H = -K\sum_{t=1}^{n} p_t \log p_t ,$$

где K – некоторая положительная константа, определяющая выбор единицы измерения. Не уменьшая общности, можно считать, что K = 1. Тогда энтропией множества вероятности $\{p_1, p_2, \dots p_n\}$ называют величину:

$$H = -\sum_{t=1}^{n} p_t \log p_t . \quad (1)$$

Пусть имеются два события x и y с m исходами для первого и n исходами для второго. Пусть p(i,j) означает вероятность осуществления исхода i для x и исхода j для y. Если x случайная величина, то обозначим ее энтропию через H(x); таким образом, x – не аргумент функции , а лишь знак, отличающий ее, скажем, от H(y) – энтропии случайной величины y. Энтропия совместного события $H(x, y)$ равна:

$$H(x,y) = -\sum_{i,j}^{m,n} p(i,j) \log p(i,j).$$

Используя предикативную форму, энтропия совместного события $H(x,y)$ равна

$$\forall H(x,y) \sim F_{m,n}(x,y) \exists p(m,n) = - \sum_{i,j}^{m,n} p(i,j) \log p(i,j),$$

где $F_{m,n}(x,y)$ – функция события x и y для m и n исходов.
В то время как

$$H(x,y) = - \sum_{i}^{m} p(i) \log \sum_{i}^{m} p(i),$$

или в предикативной форме:

$$\forall H(x) \sim F_{m}(x) \exists p(m) = - \sum_{i}^{m} p(i) \log p(i),$$

и

$$H(y) = - \sum_{m}^{n} p(j) \log \sum_{j} p(j),$$

и, соответственно, в предикативной форме:

$$\forall H(y) \sim F_{n}(y) \exists p(n) = - \sum_{j}^{n} p(j) \log p(j),$$

Легко показать, что $H(x,y) \leq H(x) + H(y)$, причем равенство имеет место только в том случае, когда события x и y взаимно независимы (т.е. p(i)* p(j)). Неопределенность совместного события не больше, чем сумма неопределенностей отдельных событий.

Для получителя важно не собственно сообщение, а насколько оно изменяет уровень знаний об источнике, насколько сообщение непредсказуемо: с ростом нетривиальности важность сообщения растет. Нетривиальность можно оценивать условной вероятностью (с позиции ожидания приемника) генерации источником конкретного сообщения. Соответственно, это приводит к понятию условной энтропии источника, отражающей нетривиальность, с точки зрения получателя, информации от источника. Количество информации, полученной приемником от источника, оценивается как разность между энтропией и условной энтропией источника.

Математически указанные соображения интерпретируются следующим образом. События x и y не обязательно независимые. Для каждого частного значения j, которое может принять x, имеется условная вероятность p_i (j) того, что при этом y примет значение j. Она задается выражением:

$$p_i \ (j) = \frac{p_i(j)}{\sum_i p_i(j)}$$

Условная энтропия $H_x(y)$ величины y есть значение, получаемое в результате осреднения энтропии y по всем значениям x с весами, равными вероятности этих значений x:

$$H_x(y) = - \sum_{i,j}^{m,n} p_i \ (i,f) \log p(f).$$

Эта величина показывает, какова в среднем неопределенность значения y, когда известно значение x. Подставляя значение p_i (j), получим $H_x(y) = H(x,y) - H(x)$ или $H(y) = H(x) + H_x(y)$. Энтропия (неопределенность) совместного события $H(x,y)$ равна сумме энтропии события x и условной энтропии события y, когда известно x.

Содержательной точки зрения, энтропия $H(x)$ можно трактовать как некоторый количественный показатель того знания, которое несет информация о событии x. Тогда $H(x)$ знание, обусловленное уже двумя событиями x и y, и $H_x(y)$, есть дополнительное знание ценность информации о реализации события y при базовых знаниях $H(x)$. Опираясь на приведенные соотношения, несложно показать, что $H(y) \geq H_x(y)$ предварительная информация о событиях x «контексте» события y, уменьшает «ценность» информации о событии y.

В реальном мире мы всегда имеем априорный запас знаний. Любое информационное сообщение бывает избыточно. Оно содержит некоторую «сущность», «смысл», «новое знание» и ряд дополнительных, известных сведений, которые, строго говоря, можно было бы не приводить. Поэтому избыточность информации может быть использована для реализации вспомогательных функций, например, для обеспечения юридической силы документа, его защиты. На достаточно строгой основе, приходим к положению, что любой документ характеризуется не только «содержанием», но и рядом вспомогательных «атрибутов».

На основе вышеприведенных рассуждений можно сделать вывод, что суммарное количество информации (I), возникающее в процессах электронного документооборота и энтропии, является постоянным, т.е.

$$\bigcup_{i,j}^{m,n} I(i,j) + \bigcup_{i,j}^{m,n} H(i,j) = const.$$

Вывод: Итак, нами была рассмотрена энтропия электронных документов с учетом двоичного представления документов без учета семантической составляющей. Далее мы будем рассматривать множество входных и выходных документов энтропию которых можно оценить аналогично формуле (1), однако при этом нам необходимо ввести понятие семантических констант и построить математический аппарат учитывающий их влияние.

1. Закон України "Про інформацію" [Електронний ресурс] – Режим доступу: http://zakon0.rada.gov.ua/laws/show/2657-12 – Заголовок з екрану.
2. Г.Г Асеев Электронный документооборот// Г.Г Асеев – К.: Кондор, 2007 – С. 36-42.

Поступила 6.10.2016р.