

ОБ ЭФФЕКТИВНОСТИ АСОД И ПРОБЛЕМЕ ИЗМЕРЕНИЯ ЦЕННОСТИ ДАННЫХ

Е. Г. ЯСИН

(Москва)

В настоящее время разрабатывается и внедряется большое число автоматизированных систем обработки данных (АСОД) и управления (АСУ). Масштабы работ и затраты ресурсов столь велики, что мы вправе говорить о формировании новой отрасли — индустрии информации. С ростом вложений все острее становится проблема их эффективности. Между тем ей уделяется мало внимания практически, и она весьма слабо разработана теоретически. Нередко у создателей АСОД нет четкого представления о целях и критериях ее создания. Иногда цель видят просто во внедрении ЭВМ самом по себе, в использовании их способности к обработке больших массивов данных. В результате возникают системы, выдающие груды цифр, которым трудно или вообще невозможно найти применение.

Очевидно, необходимо научиться более корректно определять задачи АСОД, найти критерии отбора данных, которые должны войти в программу выпуска системы, иными словами, разработать методы отбора наиболее ценных данных. Состав и объем выходных данных должен формироваться таким образом, чтобы их суммарная ценность при заданной величине затрат была максимальной. Тогда относительно просто решается и проблема определения эффективности АСОД.

Теоретически задача ставится следующим образом. Пусть ценность данных для потребителей определяется рядом факторов, полнотой, достоверностью, своевременностью и т. п. — и измеряется в рублях. Тогда отношение между ценностью данных и затратами на их получение можно представить графиком, изображенным на рис. 1.

Последовательное расширение перечня наблюдаемых показателей, повышение их достоверности и своевременности, очевидно, дают при данных технических средствах и технологии лишь убывающий темп роста суммарной ценности, тогда как затраты имеют тенденцию расти при этом с ускорением. Наибольшая эффективность будет достигаться при максимальной разности между суммарной ценностью и затратами, которой на оси абсцисс соответствует точка А, определяющая оптимальные харак-

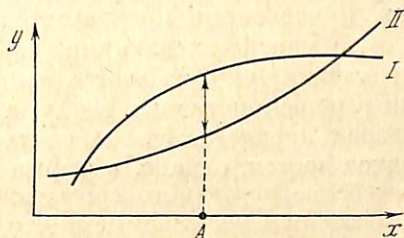


Рис. 1. Зависимость ценности данных и затрат на их получение от их полноты, достоверности и своевременности: x — мера полноты, достоверности, своевременности; y — мера ценности и затрат (в рублях); I — затраты на получение данных; II — ценность данных

теристики номенклатуры показателей, их достоверности и своевременности.

Подобное теоретическое представление могло бы приобрести и практическое значение, если бы удалось найти методы измерения ценности данных.

К настоящему времени имеется несколько подходов к измерению ценности данных, первый из которых основан на предположении, что мерой ценности данных может быть количество содержащейся в них информации в смысле Винера—Шеннона. Это допущение положено в основу работ [1, 2]. В [3, стр. 23] высказывается мнение, что количество информации может быть мерой ценности данных, но лишь в известном отношении, когда речь идет о качественно сравнимых событиях. Поскольку шенноновская теория информации учитывает лишь фактор вероятности наступления событий, отвлекаясь от содержания данных и их отношения к потребителю, этот подход не удовлетворяет многих исследователей. В связи с этим в [4] выдвигается теория семантической информации, в [5] предлагается метод измерения количества алгоритмической информации и т. п. Однако все эти попытки, интересные сами по себе, не дали решения проблемы ценности данных, так как в них учитывались не все факторы, влияющие на ценность данных. Подход, описанный в [6], выгодно отличается от других тем, что автор его пытается выяснить и принять в расчет весь комплекс факторов, характеризующих ценность данных. В частности, предлагается учитывать *значимость* данных, определяемую мерой их влияния на качество решения той или иной задачи управления; *употребимость* данных как показатель частоты их применения в решении разных задач, а также степень важности этих задач, по которой предлагается определять ранг информации. Этот интересный подход дается, однако, в форме довольно произвольной конструкции: требует более детального обоснования выбор факторов, нет уверенности в том, что они независимы и что балльные оценки являются для них наилучшей мерой.

Наиболее плодотворным, на наш взгляд, является рассмотрение необходимых для управления данных как продукта определенной отрасли производства. Как и другие продукты, они имеют потребительную стоимость, которая реализуется при потреблении их в процессе решения задач управления и зависит от совокупности их потребительских свойств. Процесс их сбора, обработки, хранения и передачи аналогичен процессу производства других продуктов. Данные производятся в определенной номенклатуре (состав показателей) и количестве. В процессе производства им придаются именно те потребительские свойства, которые делают их пригодными для удовлетворения потребностей органов управления. Соответствием совокупности потребительских свойств этим потребностям определяется их качество. Очевидно, при таком подходе суммарная ценность производимых в АСОД данных будет зависеть от номенклатуры выдаваемых показателей, количества их (определяемого с учетом частоты выдачи различных показателей за определенный период) и качества.

Большое значение имеет определение номенклатуры показателей, выдаваемых АСОД, или, иными словами, отбор показателей для программы выпуска. При этом, несомненно, необходимо принять во внимание такие упомянутые в [6] факторы, как значимость данных, их употребимость, а также важность задач, решаемых с их применением. На наш взгляд, количественная оценка этих факторов в принципе может быть получена методами теории информации.

Пусть y_1 — решение задачи 1, причем его получение ликвидирует неопределенность, измеряемую энтропией $H(y_1)$. Величину $H(y_1)$ можно

рассматривать как меру важности задачи: по крайней мере интуитивно ясно, что составление народнохозяйственного плана устраняет большую неопределенность, чем составление плана отдельного предприятия, если в понятие народнохозяйственного плана включать совокупность взаимно согласованных планов всех элементов народного хозяйства. В противном случае, если, например, составление народнохозяйственного плана мыслится как решение макроэкономической модели, учитывающей ограниченное число агрегированных факторов, то такой этап будет устранять меньшую энтропию, но при этом и полученное решение будет менее важным. Пусть x_1 — показатель, необходимый для решения задачи 1. Мерой его значимости, т. е. степени влияния на получаемое решение, может быть условная энтропия y_1 при условии x_1 , т. е. величина $H_{x_1}(y_1)$. Действительно, если изменение x_1 мало влияет на изменение y_1 , то $H_{x_1}(y_1)$ будет мала, и наоборот*. Между тем, среднее количество информации, содержащееся в x_1 относительно y_1 ,

$$I_{11} = I(x_1, y_1) = H(y_1) - H_{x_1}(y_1). \quad (1)$$

Из (1) следует, что величина I_{11} , характеризующая информативность показателя x_1 для решения задачи 1, дает оценку совместного влияния на его ценность факторов важности задачи и значимости показателя.

Пусть далее x_1 применяется для решения M независимых задач. Тогда фактор утилиты можно учесть простой суммой

$$I_1 = \sum_{j=1}^M I_{1j}, \quad j = 1, 2, \dots, M.$$

Если для решения M задач АСОД производит N независимых показателей, то они будут содержать количество информации, измеряемое величиной

$$I = \sum_{i=1}^N \sum_{j=1}^M I_{ij}, \quad i = 1, 2, \dots, N. \quad (2)$$

Очевидно, суммарная ценность данных, выдаваемых АСОД, будет тем больше, чем больше I . Учитывая это, в принципе можно не только отбирать номенклатуру показателей, но и определять необходимую частоту (периодичность) их получения. Для этого необходимо принять во внимание то обстоятельство, что x_1 может принимать значения в некотором интервале, причем вероятность, что он примет данное значение, зависит от того, каково было его значение в предшествующий отрезок времени. Иными словами,

$$I(x_{i,t-1}, x_{i,t}) = H(x_{i,t}) - H_{x_{i,t-1}}(x_{i,t}). \quad (3)$$

Показатель $x_{i,t-1}$ будет содержать тем больше информации относительно $x_{i,t}$, чем больше $H_{x_{i,t-1}}(x_{i,t})$ в (3), а она будет тем больше, чем меньше отрезок времени между моментами t и $t-1$ и чем меньше колеблемость показателя x_i . Но с ростом этой величины падает информативность показателя $x_{i,t}$, и получать его имеет смысл только тогда, когда он может доставить определенное минимальное количество информации. Зная величину допустимой ошибки x_i , можно так выбрать периодичность его получения, чтобы предупредить появление чрезмерных ошибок и в то же время иметь данные с достаточной информативностью.

* Например, при исследовании линейной модели межотраслевого баланса в НИЭИ Госплана СССР выяснилось, что из всех показателей норм материальных затрат только по 500 нормам изменение в 1,5 раза меняет итоговые показатели больше чем на 1% [7, стр. 40]. Условная энтропия по этим нормам будет больше, чем по другим.

Таким образом, методы теории информации позволяют оценить некоторые факторы, а именно те, влияние которых на ценность данных суммируется в информативности сообщения. Однако это не все. Необходимо оценить еще качество данных. В соответствии со сказанным выше, качество показателя K можно характеризовать выражением [8, стр. 19]

$$K = c_1 m_1 + c_2 m_2 + \dots + c_l m_l, \quad (4)$$

где $m_1, m_2, \dots, m_l$ — измеримые характеристики потребительских свойств 1, 2, ..., l ; $c_1, c_2, \dots, c_l$ — оценки важности соответствующих потребительских свойств определенных групп задач: $c_1 + c_2 + \dots + c_l = 1$.

Чтобы оценить качество показателей, необходимо решить три трудные проблемы: отобрать перечень потребительских свойств, найти для них адекватные меры, найти способы определения оценок важности для разных групп задач. Ниже предпринимается попытка наметить пути решения этих проблем.

В первом приближении для включения в перечень потребительских свойств данных можно предложить следующие свойства: а) достоверность; б) актуальность; в) своевременность; г) удобство восприятия; д) содержательность. К этому перечню можно было бы добавить рассмотренные уже ранее значимость данных, употребимость их и важность задачи или суммирующую их информативность. Однако последняя может рассматриваться двояко: как потребительское свойство *данных* (т. е. как элемент качества) и как количество информации, т. е. объемная характеристика информационных процессов. *Мерой достоверности* данных может быть оценка вероятности того, что искомая величина находится в интервале, определяемом величиной допустимой ошибки, которую в принципе можно вычислить по интегральной формуле Лапласа при условии, что для этого будут получены необходимые данные наблюдения. Вообще же теория измерения достоверности экономических данных еще нуждается в серьезном развитии. *Актуальность и своевременность* — характеристики весьма близкие по содержанию, но все же различные. Актуальность данных зависит от их возрастов, т. е. от времени между моментами наблюдения и использования информации, а также от скорости изменения значений данного показателя. Своевременность же отражает степень соответствия времени поступления данных моменту, когда должна решаться задача. Ее можно характеризовать временем от поступления данных до решения задачи — знак минус при запаздывании и плюс при опережении. Большое опережение тоже нежелательно, так как связано с потерей актуальности. Здесь мы пока не будем различать эти свойства, полагая их тождественными. Зная содержание этих свойств, можно подобрать такие функции, значения которых могли бы использоваться как их количественные величины. Мы пока этого делать не будем, исходя в дальнейших рассуждениях из наличия обобщенного свойства данных — *срочности*, измеряемой в единицах времени. Это допущение делается из практических соображений.

Удобство восприятия — свойство данных, зависящее от времени их восприятия человеком или машиной. Имеются исследования, раскрывающие содержание удобства восприятия и факторы, от которых оно зависит [9], или обращающие внимание на связь его с другими потребительскими свойствами данных [10].

Содержательность — специфическое свойство данных, удовлетворительную меру которого сегодня еще не удалось найти. Ее обычно связывают с возможностью интерпретации одних данных в связи с другими. Отдельный показатель оказывается беднее содержанием при сопоставлении его

с взаимосвязанными показателями. Поскольку экономико-математические модели можно рассматривать как форму организации знаний о взаимосвязях между показателями, то включенный в модель показатель имеет большую содержательность, чем отдельно взятый. Повышение содержательности достигается также обеспечением сопоставимости показателей по объему и методологии расчета. Очевидно, возможности интерпретации зависят также от получателей и центров обработки, от массы накопленных ими знаний и уровня их организации. В этом смысле говорят обычно о богатстве тезауруса приемника. Проблема содержательности данных требует дальнейших исследований.

Перечисленные потребительские свойства данных должны быть изучены с точки зрения их взаимосвязи. Можно выдвинуть гипотезу, что число независимых потребительских свойств может быть сведено к двум — достоверности и своевременности (срочности), причем требования к другим свойствам данных для разных классов задач будут изменяться примерно по тому же закону, что и требования к одному из этих свойств. Например, повышению требований к достоверности будет соответствовать повышение требований к содержательности. Тогда (4) можно упростить

$$K = c_a d + c_t t^{-1}, \quad (5)$$

где d — мера достоверности; t^{-1} — мера срочности; c_a и c_t — соответствующие оценки важности.

Выбор линейной формы функции диктуется необходимостью максимального упрощения процедуры получения оценок важности. Можно было бы в принципе выбрать для качества данных функцию вида $K = d^{\alpha} t^{\beta-1}$, где α и β характеризуют эластичность требований, предъявляемых к достоверности и срочности данных для группы задач. Получение их было бы связано со значительными трудностями, тогда как коэффициенты важности c_a и c_t возможно получить путем опроса работников управления и статистической обработки полученных результатов.

Однако (5) все же нуждается в корректировке с целью приведения к единой размерности. Достаточно умножить t^{-1} на нормирующий коэффициент r_t , такой, чтобы произведение $r_t t^{-1}$ было безразмерной величиной. Тогда $K = c_a d + c_t r_t t^{-1}$ будет тоже безразмерным коэффициентом.

Зная параметры качества, производственную программу АСОД можно определить номенклатурой показателей, нормами их периодичности, достоверности и срочности. Эти сведения необходимы и для определения суммарной ценности данных.

Учитывая разнообразие задач и предъявляемых ими требований к информации, целесообразно как-то сгруппировать ее по признаку этих требований. Так, экономико-статистическую информацию можно условно разделить на три крупных группы: а) *аналитическая информация* — для задач комплексного анализа, долгосрочного прогнозирования и планирования; б) *текущая (конъюнктурная) информация* — для текущего (квартального, годового) анализа и планирования; в) *оперативная информация* — для задач оперативного контроля и регулирования.

Существуют, конечно, примеры информации, аналитической по своему назначению, требования к которой в отношении достоверности будут не очень жесткими. Однако в целом такая группировка, видимо, отображает характерные особенности задач управления.

Для каждой группы оценки важности потребительских свойств будут изменяться в определенном интервале. При этом оценка достоверности для аналитической информации будет выше, чем для оперативной, но для последней будет особенно важна срочность. В конкретных условиях повышение достоверности требует больших затрат времени и ухудшает по-

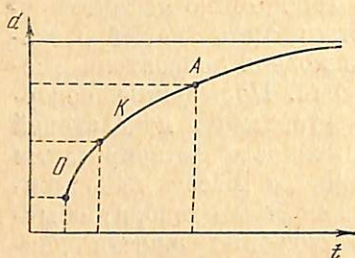


Рис. 2. Зависимость между срочностью и достоверностью данных: d — мера достоверности, t — срочность (в ед. времени)

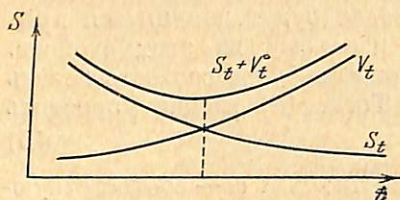


Рис. 3. Изменение затрат и потерь вследствие запаздывания данных в зависимости от срочности: S — затраты и потери

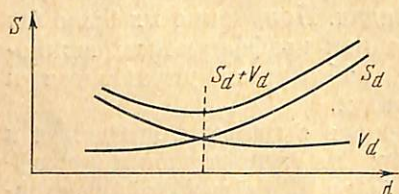


Рис. 4. Изменение затрат и потерь вследствие недостоверности данных в зависимости от меры достоверности данных

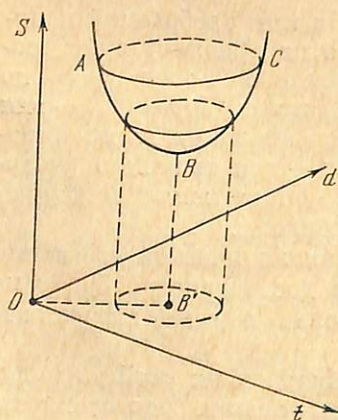


Рис. 5. Изменение суммы затрат и потерь в зависимости от достоверности и срочности

казатели срочности. Наоборот, повышение срочности часто достижимо лишь в ущерб достоверности. Зависимость между этими характеристиками можно отобразить графиком, изображенным на рис. 2.

Форма кривой на рис. 2, не имеет значения: важно лишь, чтобы она отражала увеличение времени прохождения данных при росте достоверности. Участок A на кривой показывает интервал изменения характеристики достоверности и срочности для аналитической информации; K и O — для конъюнктивной и оперативной информации.

Численное определение этих интервалов с целью установления соответствующих нормативов требует вовлечения в анализ показателей затрат на получение данных и потерь от их запаздывания и недостоверности.

Изменение затрат и потерь в зависимости от разных характеристик срочности показано на рис. 3.

Подобным же образом на рис. 4 показано их изменение при разных значениях достоверности.

Здесь S_t — затраты на достижение заданной срочности, растущие с сокращением времени, выделяемого на подготовку данных; V_t — потери от запаздывания данных; S_d — затраты на достижение заданной достоверности; V_d — потери вследствие недостоверности данных.

Кривые $S_t + V_t$ и $S_d + V_d$ на рис. 3 и 4 показывают сумму затрат и потерь, которая должна быть минимизирована.

Для оперативной информации кривая V_t на рис. 3 будет круче, для аналитической — более пологой, в связи с чем точка на оси абсцисс, отражающая характеристику срочности для $S_t + V_t = \min$, будет смещаться к началу координат для оперативной информации и удаляться от него для аналитической. Кривая V_d на рис. 4 будет иметь большую крутизну для оперативной и меньшую для аналитической, в силу чего и здесь точка, соответствующая $S_d + V_d = \min$, будет смещаться для оперативной информации к началу координат, а для аналитической — в противоположном направлении.

Таким образом, при выборе оптимальных значения d и t для разных видов информации может быть применен критерий $S = S_t + V_t + S_d + V_d = \min$.

Точки на поверхности ABC (рис. 5) представляют значения критерия S , достигающего минимума в точке B . Положение ее проекции B' на плоскость dOt зависит от угла наклона и кривизны кривой, образующей поверхность ABC в каждой точке ее сечения.

Точка B' будет смещаться к началу координат для оперативной информации и от него — для аналитической. Тогда кривую OKA на рис. 2 можно интерпретировать как геометрическое место точек минимума критерия S для разных видов информации, которым соответствуют оптимальные для каждого вида значения d и t . Ясно, что сводная характеристика качества в (5) при этих значениях d и t будет максимальной, а оценки важности d и t не субъективны: их объективной основой являются возможные потери от запаздывания и недостоверности данных при решении задач управления.

В конечном счете приведенные выкладки могут быть теоретическим обоснованием при разработке методов отбора данных по ценности, формирования оптимальной программы выпуска АСОД, при разработке рациональных схем потоков информации.

Выбор схемы потоков информации тесно связан с выбором технологии обработки данных. Оптимальный вариант находится где-то между двумя крайними альтернативами — полной централизацией и полной децентрализацией. Первая предполагает максимальную концентрацию всех технологических процессов обработки данных в одном центре, в рамках единой системы. В качестве таковой можно представить себе единую систему экономической информации страны, построенную на технической базе единой государственной сети вычислительных центров. Вторая схема представляет разрозненное размещение технологических процессов обработки данных и соответствующих технических средств у источников информации и ее потребителей. Пробообразом этого варианта является создание в каждом районе многочисленных вычислительных центров, АСОД и АСУ в министерствах, ведомствах, других органах управления, на предприятиях, в сущности, по тому же принципу, по которому всегда размещалась настольная счетная техника.

Современные тенденции развития техники и технологии обработки данных перемещают оптимальный вариант в сторону большей централизации. При этом совершенно ясно ставится цель значительного повышения комплексности обработки, многократного использования ограниченного объема данных для решения различных задач. Важно заметить, что предметом комплексной обработки по преимуществу могут быть данные аналитического и конъюнктурного характера, предъявляющие высокие требования в отношении достоверности и содержательности. Централизованная схема наилучшим образом приспособлена к удовлетворению этих требований, так как в ней обеспечиваются наибольшие возможности для сопоставлений различных показателей, их взаимного контроля и т. д. В известном смысле централизация есть предпосылка и элемент интеграции обработки данных. Вместе с тем большая централизация обычно требует увеличения времени на прохождение данных. Это обуславливает определенные преимущества децентрализованной схемы, которая проявляет большую приспособляемость к экстренным режимам работы и легче добывается повышения срочности, хотя в других отношениях она почти всегда уступает централизованной схеме. Поскольку оперативная информация характеризуется высокой оценкой срочности, для ее обработки целесообразно строить системы с относительно высокой степенью децентрализации, направляя данные непосредственно от источников к потребителям, например с предприятий в министерства. Отсюда следует, что для разных видов информации целесообразно создавать раз-

личные АСОД в соответствии с предъявляемыми требованиями в отношении достоверности и срочности. Например, отраслевые системы следовало бы специализировать на обработке оперативной информации, а также на решения специальных отраслевых задач. Аналитическая же и конъюнктурная информация должны храниться и обрабатываться преимущественно в системе с высокой степенью централизации, интегрированной в масштабе народного хозяйства на единой технической базе государственной сети вычислительных центров. Такая система могла бы наилучшим образом удовлетворить потребности управления в экономических данных оптимального качества при минимальных затратах на заданную программу выпуска.

ЛИТЕРАТУРА

1. А. А. Харкевич. О ценности информации. В сб. Проблемы кибернетики. Вып. 4. М., Физматгиз, 1960.
2. М. М. Бонгард. О понятии «полезная информация». В сб. Проблемы кибернетики. Вып. 9. М., Физматгиз, 1963.
3. Я. Н. Ханелис. Торгово-экономическая информация. М., «Статистика», 1967.
4. R. Carnap, Y. Bar-Hillel. An Outline of Theory of Semantic Information. M. I. T. Research Lab. Electronic, Techn., Rept., 1953, 247.
5. А. Н. Колмогоров. Три подхода к определению понятия «количества информации». Проблемы передачи информации, 1965, т. 1, вып. 1.
6. В. М. Жеребин. Язык экономической системы и оценка информации. Экономика и матем. методы, 1968, т. IV, вып. 5.
7. Проблемы народнохозяйственного оптимума. М., «Экономика», 1969.
8. Дж. ван Эттингер, Дж. Ситтиг. Больше ... через качество. М., Изд-во стандартов, 1968.
9. Г. И. Покровский, В. Н. Тростников. Организация информационного потока. Будущее науки. Международный ежегодник. Вып. 2, М., «Знание», 1968.
10. Экономическая семиотика. М., «Наука», 1970.

Поступила в редакцию
18 VI 1970