Двухсоставность феномена информации и анализ данных (с примерами из когнитивного анализа)

С.В. Смирнов

Институт проблем управления сложными системами РАН, 443020, ул.Садовая, 61, Самара, Россия

Аннотация

Рассматривается принцип двухсоставности информации как единства ее «носителя» и несомых «признаков» (например, в оптике для выделения информации необходимо знание об опорном и отраженном лучах). В анализе данных предлагается считать «носителем» процедуру измерения, а измеренное значение - «признаком». Констатируется существование риска получения некорректного результата обработки данных измерений из-за неполноты знаний о в общем случае сложно устроенном механизме информирования. Предложена обобщенная форма таблицы «объекты-свойства» для отражения реалий накопления эмпирических данных. Исследован один из фундаментальных факторов риска – наличие зависимостей между результатами работы процедур измерения. Обоснована необходимость введения в контекст задач анализа данных так называемых «ограничений существования свойств», что иллюстрируется примерами из когнитивного анализа данных.

Ключевые слова: анализ данных; информация; принцип двухсоставности; процедура измерения; таблица «объекты-свойства»; ограничения существования свойств; когнитивный анализ

1. Введение

В познавательной деятельности велико значение метафор – весьма полезного инструмента вербализации, фиксирования в языке (и сознании) эмпирического опыта и гипотез о предметах познания. Поначалу интуитивное усмотрение сходства с известным, нередко выражаемое в речи как раз в форме скрытых сравнений, способно дать старт формированию новых научных парадигм [1], а позднее - раскрепощать умственные способности человека в выстраивании конкретных методологий [2] как внутри отдельных направлений, так и на междисциплинарном поприще. Вместе с тем, справедливо и обратное, когда поддержанный выразительной метафорой аспект предмета познания может до некоторой степени «затмевать» другие его существенные стороны, мешать появлению адекватных и эффективных теорий и технологий.

В тезаурусе анализа данных (АД) изначально большую роль играют метафоры поиска (*search, exploration*), извлечения (*retrieval, extraction*), добычи полезных ископаемых (*mining*) и т.п. Представляется, что индуцируемые этими референциями аллюзии, как собственно и этимология слова «данные», уводят исследователя от общенаучной концепции «двухсоставности» (этот малоупотребительный, но, несомненно, лапидарный термин почерпнут в [3]) феномена информации, смысл получения которой составляет суть АД.

Предлагаемое сообщение ставит своей задачей привлечь (скорее всего, не впервые, но с новыми идеями и на материале когнитивного анализа) внимание исследователей к нередко упускаемому ходу, направлению мыслей при разработке проблематики АД.

2. Информация и аспект измерения в анализе данных

Принято считать, что фундаментальную роль в постижении субъектом действительности играют две когнитивные способности человеческого сознания: различение в окружающем мире обособленных объектов и обнаружение связей между ними. При этом имманентное качество любого объекта – его «неотделимая от бытия существенная определенность» [4] - может проявляться лишь отдельными сторонами - свойствами (признаками), - и только в результате взаимодействия с объектом. Содержание понятия «двухсоставности» феномена информации связано именно с непреложностью такого взаимодействия для получения сведений об объекте, т.е. с обязательным осуществлением некоторой акции, которая в интересующем нас контексте известна как «процедура измерения». Одна составляющая этого феномена сосредоточивает знание о процедуре измерения, другая отражает собственно результаты измерения.

Например, изображение в оптике – это не только зафиксированные параметры фронта отраженной от объекта световой волны. Адекватная интерпретация субъектом растрового изображения, когда данные представляют координаты, амплитуды и частоты точек фронта, возможна лишь при знании каким светом освещался объект. В случае голограммы, сохраняющей сведения об амплитуде и фазе фронта, интерпретация невозможна без знания характеристик опорного когерентного освещения объекта.

Таким образом, в АД под «данными» резонно понимать не только результаты измерения, зафиксированные в общепринятом протоколе «объекты-свойства» и его экстенсивных вариациях [5, 6], но и ту или иную совокупность априорных знаний, связанных с процедурами измерений, а также с условиями их выполнения.

3. Представление знаний о процедурах измерения

3.1. Обобщенная таблица «объекты-свойства»

Стандартным протоколом представления эмпирических данных в АД служит таблица «объекты-свойства» (ТОС):

$$(G^*, M, V, \boldsymbol{I}), \qquad (1)$$

где $G^* = \{g_i\}_{i=1,\dots,r}$, $r = |G^*_k| \geq 1$ - множество наблюдавшихся объектов: $G^* \subseteq G$, где G – всё гипотетически мыслимое множество объектов исследуемой предметной области (ПрО); $M = \{m_j\}_{j=1,\dots,s}$, $s = |M| \geq 1$ - множество измеренных у объектов свойств; V - множество значений свойств; $\boldsymbol{I}$ - тернарное отношение между G^*, M и V ($\boldsymbol{I} \subseteq G^* \times M \times V$), определённое для всех пар из $G^* \times M$.

Фундаментальная классификация задач АД отталкивается от традиционного представления о «данных» и указывает задачи двух сопряженных между собой направлений: обнаружение закономерных связей между элементами ТОС и использование обнаруженных закономерностей для прогнозирования одних элементов ТОС по известным значениям других ее элементов [6]. Признавая конструктивность этой классификации и незыблемость ее основы – ТОС, – уместно, тем не менее, поставить вопрос об отражении в сложившихся структурных рамках представления данных обсуждавшегося выше «аспекта измерения» в АД. При этом следует ориентироваться не на создание универсальной методики имплементации в ТОС возможности описания необъятного разнообразия способов и условий проведения измерений, а на отражение наиболее общих реалий накопления эмпирической информации.

Отвечая на этот вызов естественно опереться на то, что номинально «аспект измерения» в ТОС все же представлен: набору столбцов таблицы биективно соответствует множество использованных при формировании ТОС процедур измерения. Дальнейший морфологический анализ ТОС позволяет легко найти способы отражения следующих фундаментальных факторов накопления эмпирической информации:

- выполнение, как правило, многократных независимых измерений (серий измерений) свойства $m_j \in M$ у объекта $g_i \in G^*$;
- дифференциация доверия к результатам разных серий измерений одного и того же объекта (например, по причине различия внешних условий при выполнении серий измерений);
- использование для измерения одного и того же свойства m_j нескольких различных процедур (конгруэнтных источников информации);
- дифференциация доверия к различным процедурам измерения.

Протокол «объекты-свойства» (1) в этом случае заменяет кортеж [7, 8]

$$(G^*, M, Se, Pr, A),$$

описывающий обобщенную ТОС, где (см. рис. 1):

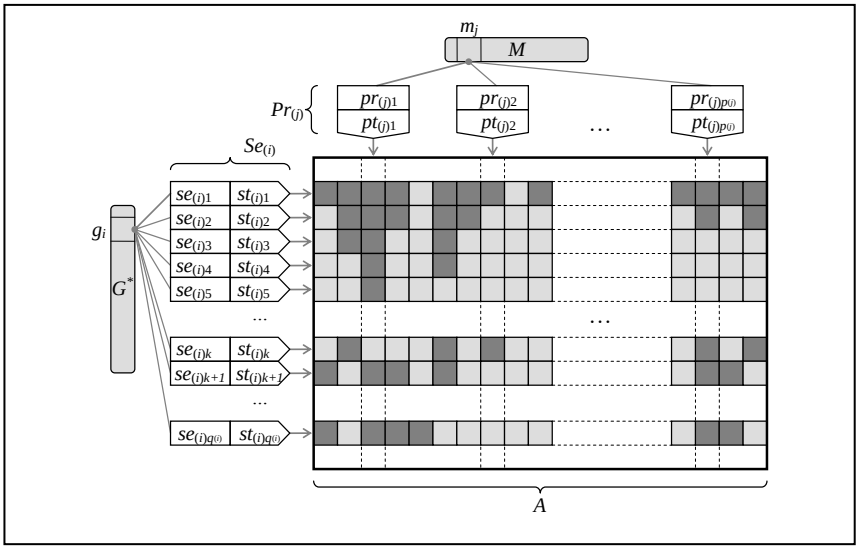

Рис. 1. Структура обобщенной таблицы «объекты-свойства» на основе двумерной матрицы (для приведенных серий измерений темные ячейки матрицы соответствуют результатам из V, а светлые – результату **NM**).

- $Se = \cup^r_{i=1} Se_{(i)}$ - множество всех выполненных при зондировании ПрО серий измерений, $|Se| = \Sigma^r_{i=1}|Se_{(i)}| = m$, и $Se_{(i)} = \{se_{(i)k}\}_{k=1,\dots,q(i)}$, $q_{(i)} \geq 1$, $i = 1,\dots, r$ - множество серий измерений, которым подвергнут объект $g_i \in G^*$ причем всякая серия $se_{(i)k}$ характеризуется степенью доверия к ее результатам $st_{(i)k}$;
- $Pr = \cup^s_{j=1} Pr_{(j)}$ - арсенал всех используемых при зондировании ПрО процедур измерения, $|Pr| = \Sigma^s_{j=1}|Pr_{(j)}| = n$, и $Pr_{(j)} = \{pr_{(j)k}\}_{k=1,\dots,p(j)}$, $p_{(j)} \geq 1$, $j = 1,\dots, s$ - множество конгруэнтных процедур измерения свойства $m_j \in M$, причем всякая процедура $pr_{(j)k}$ характеризуется степенью доверия к ее результатам $pt_{(j)k}$;

- $A = (a_{ij})_{i=1,\ldots,m;\ j=1,\ldots,n}$ – матрица результатов серий измерений *Se* свойств *M* у объектов из выборки G^*, выполненных с помощью процедур измерения *Pr*, $a_{ij} \in V \cup$ **NM**, **NM** (*not measured*) – константа, указывающая, что в действительности в *i*-й серии измерение *j*-й процедурой не производилось (введение этого формального результата помимо прочего весьма полезно для сохранения двумерного характера обобщенной ТОС).

3.2. Ограничения существования свойств

Разумеется, обобщение ТОС, выполненное на высоком уровне абстрагирования, не в состоянии отразить все возможное разнообразие априорных сведений о процедурах измерения и условиях их выполнения, необходимое для осуществления результативного анализа данных. Для описания и использования этих знаний в настоящее время используются самые различные методики, и их обобщение – актуальная научная задача. Вместе с тем уже сегодня здесь можно указать ряд полезных идей, и, в частности, предложение фиксировать в задаче анализа данных ограничения существования свойств у объектов исследуемой ПрО.

Термин «ограничения существования свойств» (*properties existence constraints*) предложен в работе [9], где априорные знания процедурах измерения (в [9] это были процедуры анализа текстов) конвертированы в ограниченное число необходимых бинарных отношений на множестве измеряемых свойств. Этот подход обеспечил авторам [9] эффективное решение задачи концептуального анализа корпуса текстов и был применен и развит в других работах [8, 10].

4. Примеры вовлечение в анализ данных знаний о процедурах измерения

4.1. Онтологический анализ данных

В онтологическом анализе данных [7, 8] обрабатываются эмпирические свидетельства вида «объект g_i обладает свойством m_j». Обобщенная ТОС вполне отражает условия получения подобных семантических суждений об исследуемой ПрО, однако при консолидации данных возникает необходимость использования многозначных логик, например, векторных [11]. Другой необходимостью становится построение специальной модели ограничения существования свойств, поскольку возникновение таких ограничений в задачах когнитивного анализа органично связано с применением фундаментальной процедуры концептуального шкалирования свойств [8, 12].

4.2. Формирование коллективных когнитивных карт

Когнитивные карты – широко применяемый на практике и активно развиваемый комплекс моделей, методов и компьютерных средств для формализации экспертных знаний при управлении слабоструктурированными ситуациями (см., например, обзоры [13, 14]). Однако этап построения когнитивной карты динамической ситуации, к чему, как правило, привлекается большая группа экспертов, слабо поддержан формальными моделями и методами. Представляется, что для увязки понимания разными экспертами анализируемой проблемной ситуации могут быть применены очерченные выше подходы [15].

Простейшая когнитивная карта - знаковый граф - может быть описана кортежем (F, W), где $F = (f_i)_{i=1,\ldots,n}$ – множество вершин-факторов ситуации, $W = (w_{ij})_{i=1,\ldots,n;\ j=1,\ldots,n}$ - матрица смежности графа: $w_{ij} \in \{+1, 0, -1\}$, $w_{ij} \neq 0$ свидетельствует о существовании в графе ребра (i, j), фиксирующего влияние *i*-го фактора на *j*-й: «положительное» при $w_{ij} = +1$ и «отрицательное» при $w_{ij} = -1$, $w_{ii} = 0$ (0 свидетельствует, что влияния не существует).

Результат коллективной работы *k* экспертов над формированием подобной когнитивной карты можно выразить следующим образом:

$$F = \cup_{i=1,\ldots,k} F_i,\ W = \oplus_{i=1,\ldots,k} W_i.$$

Это означает учет в итоговой модели всех факторов, выделяемых каждым экспертом (в первом приближении - путем теоретико-множественного объединения этих факторов F_i, $i = 1,\ldots, k$), и также формирование некоторой композиции описаний межфакторных влияний W_i, для отыскания которой применимы предложенные приемы АД.

В качестве объектов интересующей нас предметной области (ПрО) выделим упорядоченные пары факторов $(f_i, f_j) \in F$, $i \neq j$. Актуальными семантическими суждениями в этой ПрО будут независимые заключения экспертов о «положительном» («+») и «отрицательном» («-») влиянии *i*-го фактора на *j*-й.

Данные, поступившие от экспертов, структурируются в виде обобщенной ТОС (см. таблицу 1). В ней константа **X** указывает на наличие, а **None** – на отсутствие у пары (f_i, f_j) того или иного свойства, константа **NM** указывает, что экспертом оценка соответствующего влияния не проводилась (например, потому, что в его оценке ситуации отсутствует один или оба рассматриваемых в строке ОТОС фактора влияния), константа **Failure** фиксирует случай сомнения и отказа эксперта от определенной оценки влияния *i*-го фактора на *j*-й.

Нетрудно видеть, что данные экспертизы, представленные в таблице 1, вообще говоря, неполны и противоречивы, и также как в онтологическом анализе данных для их совмещения целесообразно применять многозначные логики.

Ограничения существования свойств (т.е. межфакторных влияний) для рассматриваемой ПрО следует искать в фундаментальном определении знакового орграфа [13]. Здесь можно отметить следующее:

- в упорядоченной паре факторов (f_i, f_j), $i \neq j$ фактор f_i влияет на фактор f_j либо «положительно», либо «отрицательно» (т.е. свойства «+» и «-» упорядоченных пар факторов несовместимы);
- ребра знакового орграфа описывают совокупность прямых влияний одного фактора ситуации на другой, а непрямые влияния вычисляются по известному «правилу произведения», но нет ограничения, устанавливающего, что фактор f_i может влиять на фактор f_j, $i \neq j$, либо только прямо, либо только косвенно;
- в общем случае отсутствует запрет на существование в знаковом орграфе петель, хотя такое ограничение может быть легко удовлетворено путем исключения пар факторов вида (f_i, f_i) из анализа влияний – см. таблицу 1.

Таблица 1. Данные экспертизы влияний факторов ситуации

	Эксперт 1		…	Эксперт *k*	
	«+»	«-»		«+»	«-»
(f_1, f_2)	**X**	**None**	…	**NM**	**NM**
(f_1, f_3)	**Failure**	**X**		**NM**	**NM**
…					
(f_n, f_{n-1})	**Failure**	**Failure**		**None**	**None**

5. Заключение

Анализ данных должен выполняться с привлечением априорных знаний о процедурах измерения свойств объектов зондируемой предметной области и об условиях проведения этих измерений. Учету этого положения, прямо вытекающего из «двухсоставности» феномена информации, может мешать устоявшаяся метафоричная терминология анализа данных.

Частично отражения реалий накопления данных для анализа можно достичь определенным обобщением стандартной формы представления эмпирического материала – таблицы «объекты-свойства». Однако, несмотря на фундаментальность факторов, учитываемых таким обобщением, этого, в общем случае, будет недостаточно.

Приемы введения в контекст задач анализа данных априорных знаний, связанных с процедурами измерений, могут быть самыми разными, что и составляет суть конкретных методик анализа. Вместе с тем, целесообразен отбор наиболее общих идей и подходов, к числу которых, несомненно, следует отнести модель ограничений существования свойств.

Благодарности

Статья подготовлена по материалам научных исследований в рамках субсидированного государственного задания Институту проблем управления сложными системами РАН на 2017 год по Программам фундаментальных научных исследований Президиума РАН: Программа III.4, подпрограмма «Комплексные системы управления», проект «Модели и методы формирования когнитивных карт в условиях неполноты информации о проблемной ситуации».

Литература

[1] Кун, Т. Структура научных революций / Т. Кун. – М.: Прогресс, 1977. – 300 с.
[2] Новиков, А.М. Методология / А.М. Новиков, Д.А. Новиков. – М.: СИНТЕГ, 2007. – 668 с.
[3] Вихнин, А.Г. Штурм четвертого мегапроекта: кто будет новым Биллом Гейтсом? Системный анализ и выбор стратегии / А.Г. Вихнин, Н.З. Сакипов. – М.: Изд-во «Диалог-МИФИ», 2008. – 288 с.
[4] Жилин, Д.М. Теория систем: опыт построения курса / Д.М. Жилин. - 4-е изд., испр. – М.: Изд-во ЛКИ, 2007. – 184 с.
[5] Барсегян, А.А. Анализ данных и процессов / А.А. Барсегян, М.С. Куприянов, И.И. Холод, М.Д. Тесс, С.И. Елизаров. - 3-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2009. – 512 с.
[6] Загоруйко, Н.Г. Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. – Новосибирск: Институт математики СО РАН, 1999. – 270 с.
[7] Semenova, V.A. Intelligent analysis of incomplete data to building formal ontologies / V.A. Semenova, S.V. Smirnov // CEUR Workshop Proceedings, 2016; 1638: 796-805. DOI: 10.18287/1613-0073-2016-1638-796-805.
[8] Самойлов, Д.Е. Анализ неполных данных в задачах построения формальных онтологий / Д.Е. Самойлов, В.А. Семенова, С.В. Смирнов // Онтология проектирования. – 2016. – Т. 6, №3(21). - С. 317-339. DOI: 10.18287/2223-9537-2016-6-3-317-339.
[9] Lammari N., Metais E. Building and maintaining ontologies: a set of algorithms // Data & Knowledge Engineering. – 2004. - Vol. 48(2). - P. 155-176.
[10] Пронина, В.А. Использование отношений между атрибутами для построения онтологии предметной области / В.А. Пронина, Л.Б. Шипилина // Проблемы управления. - 2009. - №1. - С. 27-32.
[11] Аршинский, Л.В. Применение векторного формализма в логике и логико-математическом моделировании / Л.В. Аршинский. - Онтология проектирования. – 2016. – Т. 6, №4(22). - С. 436-451. DOI: 10.18287/2223-9537-2016-6-4-436-451.
[12] Samoilov, D.E. Data formation and processing in formal concept analysis: subjective aspects / D.E. Samoilov, S.V. Smirnov // CEUR Workshop Proceedings, 2016; 1638: 806-812. DOI: 10.18287/1613-0073-2016-1638-806-812.
[13] Кузнецов, О.П. Анализ влияний при управлении слабоструктурированными ситуациями на основе когнитивных карт / О.П. Кузнецов, А.А. Кулинич, А.В. Марковский // Человеческий фактор в управлении / Под ред. Н.А. Абрамовой, К.С. Гинсберга, Д.А. Новикова. — М.: КомКнига, 2006. - С. 313-344.
[14] Кулинич, А.А. Компьютерные системы моделирования когнитивных карт: подходы и методы / А.А. Кулинич // Проблемы управления. - 2010. - №3. – С. 2-16.
[15] Смирнов, С.В. Модели и методы формирования когнитивных карт при их коллективной разработке / С.В. Смирнов // Информационные технологии и системы: Труды Шестой международной науч. конф. ИТиС-2017 (1-5 марта 2017 г., Банное, Россия) / Отв. ред.: Ю.С. Попков, А.В. Мельников. - Челябинск: Изд-во Челяб. гос. ун-та, 2017. – С. 281-283.