Практические аспекты формализации сложных информационных объектов на примере задачи анализа плоского представления табличной информации

Д. Я. Слободецкий, Д. В. Полевой

Работа посвящена проблемам формализации сложных информационных объектов в контексте информационного компьютерного моделирования. В качестве объекта моделирования рассматривается табличное представление информации.

Введение

Невозможно представить себе современное развитие науки и техники без широкого использования методов математического и компьютерного моделирования. Сущность этой методологии состоит в замене исходного объекта его аналогом — математической моделью — и оперированию ею при помощи реализуемых на компьютерах алгоритмических средств.

Важным широко распространенным случаем является информационное моделирование, когда моделируются не физические, а информационные объекты. В этом случае модель используется для реализации методов манипулирования информацией и извлечения знаний из данных [1].

1. Требования к моделям

При разработке модели существует ряд требований, которым она должна удовлетворять. Помимо вытекающих из конкретной прикладной задачи частных требований, существуют некоторые общие для всех моделей требования, которые можно подразделить на два основных типа:

- необходимые;
- качественные.

Выполнение необходимых требований является обязательным для правильного описания объекта. К таким требованиям относятся: адекватность и непротиворечивость.

Степень выполнение качественных требований определяет границы применимости модели и удобство использования последней. К таким требованиям относятся: полнота, минимальность внутреннего описания, минимальность покрытия неправильных объектов.

Дальше расшифруем каждое из требований:

- *Адекватность* гарантирует правильное и достаточное для решения исходной прикладной задачи описание объекта. Выполнено это требование или нет можно определить только в контексте конкретной прикладной задачи.
- *Непротиворечивость* обеспечивает однозначную интерпретацию результатов моделирования. Это свойство является основополагающим, поскольку исходный объект, по своей природе, является непротиворечивой сущностью, в которой все свойства согласованы.
- *Полнота* показывает степень охвата предметной области моделью. Полная модель описывает все рассматриваемые объекты.
- *Минимальность внутреннего описания* отражает степень дублирования информации в разных частях программной реализации модели. В идеальном случае описание является минимальным, то есть не содержит дублирующихся фрагментов. Минимизация внутреннего описания уменьшает затраты на поддержание непротиворечивости модели при реализации или развитии последней.
- *Минимальность множества неправильных объектов* позволяет учитывать тот факт, что часть соответствующих разработанной модели объектов не входят в множество исходно моделируемых. Другими словами, модель не позволяет отделить «правильные» объекты от «неправильных». В такой ситуации либо модель сужается дополнительными ограничениями, либо используются внемодельные инструменты контроля, обеспечивающие корректность объектов.

2. Задача анализа плоского представления табличной информации

В своем плоском представлении таблица фиксирует два вида информации: данные о некоторых параметрах предметной области и некоторую

информацию о структуре предметной области [2]. Удобство использования таблицы зависит от возможности правильного представления той части предметной области, которая существенная для решения текущей задачи. Электронные формы таблиц предоставляют некоторые средства манипулирования, но все эти средства реализуют модель представления таблицы в конкретном редакторе или формате, а не связаны со структурой данных.

Так возникает задача создания средств обработки табличной информации, опирающихся на более сложные модели, чем модели представлений [3]. При этом одной из первых задач в цепочке обработки является *разбор таблицы* (восстановление структуры информационного объекта таблицы по плоскому представлению). Для таблицы в некотором электронном виде необходимо отделить ключевую информацию от производной (выделить заголовки), разобрать структуру ключевой информации и сформировать информационный объект (восстановить исходный информационный объект).

Для решения задачи разбора таблицы необходима модель представления исходных, промежуточных и конечных данных. Первым, наиболее очевидным, источником таких моделей являются уже реализованные программные проекты.

3. Сложности переноса моделей между прикладными задачи

Современная теория описания таблиц ([2–5]) выделяет не менее двух уровней описания: физический и логический. Физическая структура таблицы описывает расположение частей таблицы на носителе. Логическая структура таблицы описывает типы этих частей, их назначения и соотношения. Рассмотрим несколько моделей, использующихся в электронных представлениях таблиц.

Современные табличные процессоры (MS Excel, OO Calc) используют модель, которую условно можно назвать матричной [2]. Плюсами такого представления являются:

- точная информация о значениях и типах данных;
- описание форматирования ячеек и оформления границ;
- определяющие зависимости между ячейками формулы.

Однако возможность произвольного объединения ячеек в прямоугольные группы сильно расширяет множество описываемых объектов, давая возможность формировать «нетабличные» образования (рис. 1).

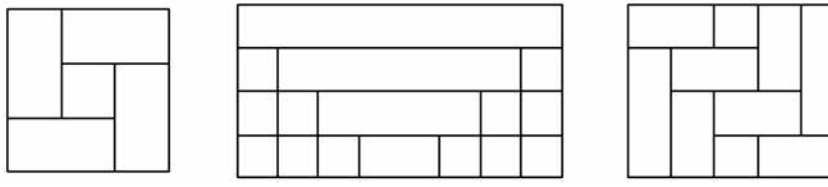

Рис. 1. Примеры порождаемых объединением ячеек матрицы «нетабличных» объектов, все границы ячеек видимы

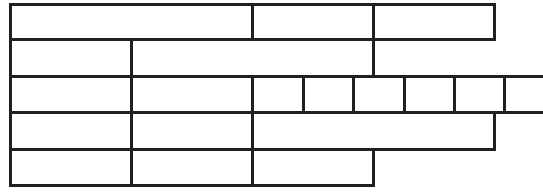

Рис. 2. Пример порождаемого независимым редактированием строк «нетабличного» объекта, все границы ячеек видимы

Распространенные текстовые процессоры (например, MS Word) предоставляют меньше информации о данных таблицы, по сравнению с табличными процессорами, но обладают не менее выразительными графическими средствами оформления таблиц (рис. 2).

Независимость строк таблицы позволяет произвольным образом формировать ячейки в конце строки (например, «откусывать» последние). Другой особенностью является независимость параметров границ каждой ячейки. Это означает, что для смежных ячеек общая граница может быть определена с разными параметрами для каждой из ячеек. Алгоритм прорисовки таблицы скрывает от пользователя такие детали внутреннего представления, но с точки зрения разбора и анализа таблицы подобная модель является противоречивой и избыточной.

Модели таблиц для вычислений или формирования печатных документов успешно обеспечивают решение задач миллионов пользователей в различных приложениях, но является некорректными для задачи разбора. Критическое рассмотрение ряда других современных моделей ([6–8]) показало их частичную или полную непригодность для реализации решения задачи разбора. Далее будет рассмотрен процесс построения модели таблицы для решения задачи восстановление структуры информационного объекта таблицы по плоскому представлению.

Размытость предметной области

Иногда тяжело, а порой почти невозможно точно сказать принадлежит ли конкретный объект к множеству описываемых нами объектов, что становится ощутимым препятствием для моделирования. Многие объекты даже в нечетких терминах естественного языка описываются недостаточно строго, что лишь усугубляется при попытке математической формализации.

Проблема размытости может быть решена выделением существенной и в то же время более простой для формализации части предметной области. Таким образом, строится модель, которая точно не покрывает все возможные случаи, но предоставляет возможность решить ощутимую часть пользовательских задач. Когда построенная модель становится недостаточной, из неформализованной части предметной области выделяется следующее по значимости подмножество объектов, а модель расширяется для описания последних. Сознательный отказ от полной и исчерпывающей единственной модели позволяет строить ряд ограниченных по выразительным возможностям моделей с известными характеристиками.

Недостаток структурной информации в моделях таблиц, по всей видимости, является не случайностью, а результатом принципиальной невозможности строго формализовать понятие таблицы [3]. Таблицы, как способ представления информации, используют возможности человеческого мозга по интерпретации визуальных зависимостей, при этом формализация разницы между допускающими табличную интерпретацию объектами и объектами, не допускающими такой интерпретации, затруднительна (рис. 3).

В качестве базового множества табличных объектов для решения задачи разбора структуры были выбраны таблицы с двумя заголовочными зонами без «подвала» (footnote). Последние, по оценкам исследователей, составляют 56 % всех таблиц в печатных источниках, при этом расширение

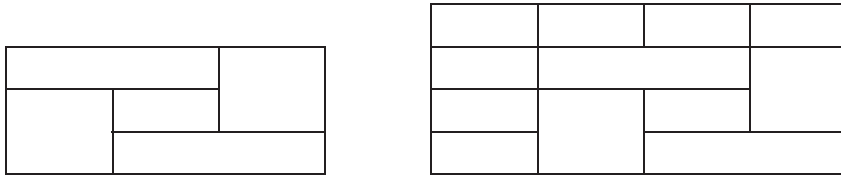

Рис. 3. Пример близких геометрических объектов с различиями в возможности табличной интерпретации

Таблица 1

	Европеоиды			Негроиды			Монголоиды		
	2000	**2001**	**2002**	**2000**	**2001**	**2002**	**2000**	**2001**	**2002**
Мужской	65	60	75	55	50	65	75	70	85
Женский	70	65	80	60	55	70	80	75	90

модели за счет «подвала» позволяет увеличить охват модели до 95 %–97 % всех таблиц [5].

Невозможность описание объекта одной моделью

Другой сложностью при построении моделей информационных объектов является необходимость использования нескольких согласованных моделей, отражающих различные аспекты моделируемого явления. Попытка ограничения сложности модели за счет выбора только одного аспекта сокращает область применимости модели (см. п. 3).

Рассмотрим следующую таблицу (1) зависимости некоторой величины от пола, расы и года.

Выделение и анализ заголовочных зон дает следующие три независимых набора ключей: «пол», «раса» и «год», при этом структура данных определяется полным декартовым произведением множеств ключей (рис. 4).

Исходной моделью для разбора была выбрана географическая модель таблицы, которая описывает взаиморасположение ячеек таблицы, без учета содержимого ячеек. Классификация квазитабличных объектов по их географическому представлению выявила четыре основных класса объектов:

- «Соты» — плотное объединение изотетичных прямоугольников.

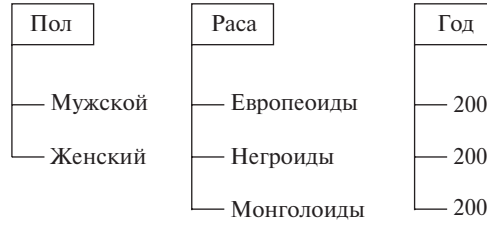

Рис. 4. Пример ключевой информации в таблице

- «Форма» — покрытие прямоугольной области изотетичными прямоугольниками.
- «Таблоид» — форма, у которой любая прямоугольная подобласть допускает сквозные разрезы между ячейками.
- «Таблица» — удовлетворяющий дополнительным ограничениям таблоид.

По определению каждый последующий класс объектов накладывает дополнительные ограничения, сужая множество описываемых объектов. Таким образом, географическая модель непротиворечиво и полно описывает корректные входные данные разборщика таблиц, а многомерный куб определяет ответ разбора.

Для таблиц существует много распространенных форматов и популярных программ обработки, которые используют различные внутренние представления. Способы представления, а также явные и неявные предположения требуют унификации, что и приводит к необходимости использования некоторой промежуточной модели, которой стала географическая модель таблицы. Последняя позволяет отсекать таблицы, допустимые с точки зрения стороннего формата, но не согласованные с выбранной логической моделью.

Заключение

Рассмотренные практические аспекты формализации сложных информационных объектов показывают, что построение модели является индивидуальным для сложных случаев и должно опираться на анализ прикладной задачи.

Формализация плоского представления табличной информации и разработка разборщика структуры информационного объекта таблицы по плоскому представлению является только одним из первых шагов для построения полноценного инструментария обработки таблиц.

Литература

1. *Арлазаров В. Л.*, *Емельянов Н. Е.* От баз данных к базам знаний // Системный подход к управлению информацией: Труды ИСА РАН. Т. 23. М.: КомКнига/URSS, 2006.
2. *Полевой Д. В.* Разработка моделей, методов и средств обработки табличных документов в информационных системах: Дисс. ...канд. тех. наук. М., 2007.
3. *Михайлов А. А.*, *Полевой Д. В.* Проблемы анализа таблично структурированной информации // Информационно-аналитические аспекты в задачах управления. Труды Института системного анализа Российской академии наук (ИСА РАН): Т. 29. М.: Издательство ЛКИ/URSS, 2007.

4. *Haralick R. M.* Document image understanding: geometric and logical layout // In Proceedings of the 1994 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 1994. P. 385–390.
5. *Wang Y.* Document analysis: table structure understanding and zone content classification: Ph. D. Thesis, University of. Washington, Seattle, WA, 2002. 161 p.
6. HTML 4.01 Specification W3C Recommendation 24 December 1999. [Электронный ресурс] http://www.w3.org/TR/1999/REC-html401–19991224
7. OpenDocument v1.1 specification. [Электронный ресурс] http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=office
8. Document Object Model (DOM) Level 3 Core Specification. [Электронный ресурс] http://www.w3.org/TR/2004/REC–DOM–Level-3–Core-20040407
9. Extensible Markup Language (XML) 1.0 (Third Edition) W3C Recommendation, 04 February 2004. [Электронный ресурс] http://www.w3.org/TR/2004/REC-xml-20040204