Проблемы анализа таблично структурированной информации

А. А. Михайлов, Д. В. Полевой

Работа посвящена вопросам разбора, преобразований и использования таблично структурированной информации. Дан обзор мировой практики в этом направлении. Описана концептуальная модель таблицы. Выделены и систематизированы ключевые проблемы в обработке таблиц.

Введение

Таблицы являются удобной и компактной формой [1] представления информации на бумаге или экране. Бухгалтерские, финансовые и статистические документы, научная и справочная литература — все они содержат таблицы. Правильно построенная таблица является эффективным инструментом анализа информации, поэтому их используют для выявления тенденций и принятия решений. В то же время неправильная интерпретация таблицы или ошибка могут иметь высокую стоимость.

Данная работа посвящена возможностям и проблемам использования табличного структурированной информации. Особое внимание уделено электронным представлениям табличной информации, которые предоставляют широкие возможности по манипулированию данными и их представлениями.

1. Обзор особенностей электронных таблиц

Из исследований психологов [2] известно, что человеческий мозг может оперировать одновременно с ограниченным числом (до 5–9) единиц информации. Такие единицы могут быть составными, и тогда человек способен обрабатывать больший объем информации. Таблица является способом визуального группирования единиц информации. Элементы таблицы могут быть расположены значительно более плотно и структурировано, чем это возможно при размещении тех же данных в линейном тексте. Хорошо составленная таблица позволяет буквально одним взглядом охватить всю информацию.

Удобство использования конкретной таблицы определяется именно тем, что все необходимые для решения задачи данные представлены

в удобной для восприятия и обработки форме. Решение других задач с помощью этой же таблицы может оказаться неудобным, и тогда требуется изменить представление таблицы.

Таблицы часто используются для быстрого поиска. Заголовки строк и столбцов содержат значения аргументов, а в ячейке на пересечении строки и столбца лежит соответствующее заголовкам значение. Выбор строки и столбца по заголовкам позволяют быстро найти нужное значение. С помощью такой таблицы можно решать обратную задачу — поиск значений параметров для заданного значения функции. Просмотр значений ячеек тела таблицы вдоль строки или столбца позволяет сравнить значения функции при одном фиксированном параметре. Подходящая сортировка строк/столбцов может сильно упростить решение перечисленных задач.

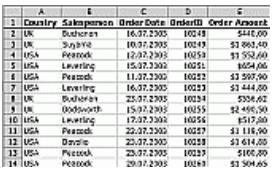
Размещение данных разного уровня детализации рядом в таблице позволяет выборочно просматривать факты в выбранном масштабе, а при необходимости быстро переходить к уточняющей информации. Подобное агрегирование также выполняет важную контрольную функцию, что наиболее ярко выражается в бухгалтерских и финансовых таблицах.

Табличные процессоры или электронные таблицы (Lotus 1–2–3, Excel, Open Office Calc) являются результатом естественного процесса переноса привычных объектов физического мира в мир информационных объектов. При наличии всех необходимых инструментов создание и изменение таблицы в виде информационного объекта несоизмеримо проще, чем проведение тех же операций в физическом мире. Развитые технологии построения изображений (WYZIWIG) на экране и при печати позволяют легко получать бумажные аналоги экранных представлений таблиц.

Табличный процессор предоставляет дополнительные возможности по сортировке и фильтрации данных. Вычисляемые формульные ячейки составляют основу электронной таблицы и позволяют воспользоваться вычислительными возможностями компьютеров для упрощения рутинных вычислений. Экранное представление облегчает оперирование таблицами больших размеров, которые при печати формируют сложные отчеты.

Современные электронные таблицы позволяют манипулировать данными в стиле реляционных БД. Данные в соответствие с реляционной моделью размещаются на листе: каждый столбец соответствует атрибуту, а строка — кортежу отношения (рис. 1 *а*).

Сводные таблицы [3] (pivot-tables) табличных процессоров используется для просмотра комплексных данных путем вложения нескольких измерений по осям и отображения данных на нескольких страницах с соответствующими итоговыми результатами (рис. 1 *б*).



	A	B	C	D	E
1	Country	Salesperson	OrderDate	OrderID	Order Amount
2	UK	Buchanan	16.07.2003	10248	$440,00
3	UK	Suyama	10.07.2003	10249	$1 863,40
4	USA	Peacock	12.07.2003	10250	$1 552,60
5	USA	Leverling	15.07.2003	10251	$654,06
6	USA	Peacock	11.07.2003	10252	$3 597,90
7	USA	Leverling	16.07.2003	10253	$1 444,80
8	UK	Buchanan	23.07.2003	10254	$556,62
9	UK	Dodsworth	15.07.2003	10255	$2 490,50
10	USA	Leverling	17.07.2003	10256	$517,80
11	USA	Peacock	22.07.2003	10257	$1 119,90
12	USA	Davolio	23.07.2003	10258	$1 614,88
13	USA	Peacock	25.07.2003	10259	$100,80
14	USA	Peacock	29.07.2003	10263	$1 504,65

а)

	A	B	C	D	E	F
1	Country	UK				
2						
3	Sum of Order Amount	Quarters				
4	Salesperson	Qtr1	Qtr2	Qtr3	Qtr4	Общий итог
5	Buchanan	$22 719	$6 858	$16 035	$23 181	$68 792
6	Dodsworth	$32 480	$14 920	$9 649	$17 999	$75 048
7	King	$34 866	$38 584	$21 950	$21 563	$116 963
8	Suyama	$18 903	$18 107	$14 276	$21 241	$72 528
9	Общий итог	$108 968	$78 468	$61 910	$83 984	$333 331

б)

Рис. 1. Пример *а*) фрагмента исходной и *б*) сводной таблицы

Такие таблицы поддерживают итеративный выбор подмножеств данных и изменение отображаемого уровня детализации, однако они слишком жестко связывают данные с их внешним видом, не отделяя структурную информацию от желаемого визуального представления. Сводные таблицы не предоставляют возможности прямого редактирования данных, а добавление новых измерений или группировка данных требует дополнительной настройки.

Многие авторы считают появление табличных процессоров революцией, которая во многом определила рост популярности использовании персональных компьютеров для решения задач бизнеса. Согласно исследованиям середины и конца 90-х годов [4–6] 60–70 % опрошенных менеджеров использовали табличные процессоры. Это сравнимо с процентом использовавших электронную почту менеджеров ($\sim$ 60 %) и существенно превышает число пользовавшихся веб-браузерами ($\sim$ 25 %). Электронные таблицы являются важным средством коммуникации между дочерними и родительскими компаниями [7], они широко применяются для подготовки налоговых и других отчетов в федеральные органы [8, 9].

Некоторые авторы говорят о табличной парадигме и парадигме табличных вычислений [10]. Простота и интуитивная понятность модели табличных вычислений делает табличные процессоры популярными прикладными программами, но имеет и обратную сторону. Несовпадение модели электронной таблицы и предметной области, а также легкость визуального манипулирования данными и построения вычислительных таблиц являются причиной большого числа скрытых ошибок [11]. Для уменьшения числа ошибок в электронных таблицах и снижения связанных с ними рисков разрабатываются различные методы [7–11] визуализации скрытых структур данных и вычислений.

Появившиеся как инструмент автоматизации бухгалтерских расчетов, табличные процессоры сейчас являются одними из самых распространенных офисных и бизнес-приложений. Помимо широких возможностей для реализации вычислений в матричной модели, табличные процессоры позволяют готовить документы к публикации и использовать

электронные документы. При этом использующаяся базовая матричная модель сильно ограничивает возможности по манипулированию данными и их представлению.

2. Концептуальная модель таблицы

Таблица — способ описания, хранения, манипулирования и представления (вывода на носитель) фактов предметной области. Правильнее говорить о некоторой модели предметной области или реального объекта, которая выделяет значимые параметры описываемого явления (рис. 2).

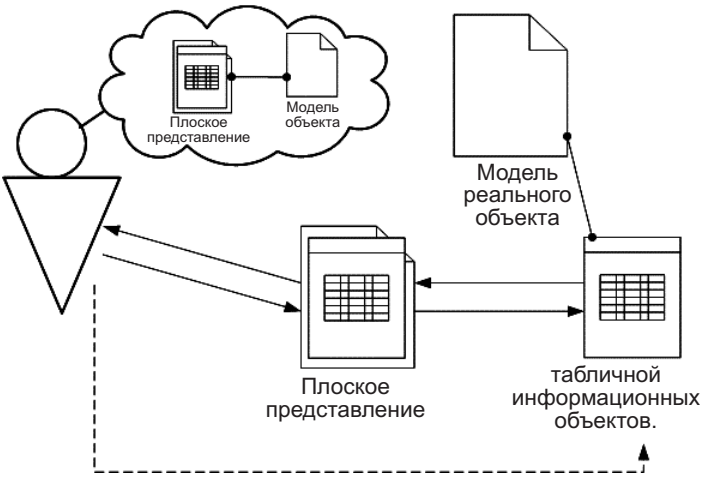


Рис. 2. Концептуальная модель таблицы

Информационный объект содержит набор значений параметров, описывающих состояние предметной области в соответствии с моделью. Плоское представление таблицы используется для визуализации информационного объекта. Возможности плоского представления для передачи моделей информационных объектов ограничены. Представимый в виде одной таблицы на плоскости информационный объект будем называть табличным информационным объектом.

При формировании плоского представления в нем фиксируется часть информации о самой модели предметной области. При работе с табличным представлением человек использует собственную модель описываемого в таблице реального объекта, которая есть у него изначально или формируется на основании интерпретации плоского представления. Таким образом, плоское представление используется для

передачи информации о модели реального объекта. Удобство использования таблицы зависит от возможности правильной интерпретации и восстановления той части модели предметной области, которая необходима для решения текущей задачи.

Отражение различных аспектов одного табличного ИО и разные постановки задач могут требовать различных плоских представлений, при этом между способами построения плоских представлений и классом информационных объектов допускающих табличное представление существует непосредственная связь. Знание структуры табличного ИО позволяет манипулировать им непосредственно. Однако визуализация результатов и визуальное манипулирование возможно только через плоское представление. Для построения эффективного пользовательского интерфейса требуется определить допустимые операций над плоским представлением и организовать отображение действий пользователя на табличный ИО при контроле на соответствие модели.

Будучи моделью некоторого объекта или явления, таблица предоставляет возможности для вычислений и прогнозирования в рамках этой модели. Значения параметров могут быть связаны некоторыми функциями, и табличный ИО должен фиксировать наличие этих зависимостей. Функциональные связи параметров могут использоваться для анализа данных или агрегирования. Использование вычислительных возможностей расширяет возможности контроля. Инструменты манипулирования таблицами должны реализовывать вычисления и контроль данных по правилам в рамках табличного ИО.

В реальном мире носителем плоского представления является физический объект — бумага или экран. Учет возможностей носителей задает набор ограничений, которые должны описываться и использоваться при формировании плоского представления. Примерами практических задач являются размещение нескольких таблиц на странице, разбиение одной таблицы на несколько страниц, автоматический расчет размера ячеек для заданного размера страницы.

Задача распознавания таблиц является обратной к задаче формирования представления и не может решаться без модели плоского представления. Произвольную конфигурацию графических примитивов и текстовых фрагментов не всегда можно интерпретировать как табличный ИО, при этом существуют конструкции, которые визуально похожи на таблицы, но таковыми не являются.

Комплексная модель таблицы состоит из следующих согласованных моделей:

1. Модель табличного информационного объекта.
2. Модель плоского представления (включая модель носителя).

Модель табличного ИО описывает классы допускающих представление на носителе в соответствии с моделью плоского представления ИО. Эта модель содержит:

- Способ описания структуры данных экземпляра таблицы.
- Способ описания внеструктурных зависимостей и ограничений данных.
- Формализацию допустимых операций над экземпляром таблицы.

Модель плоского представления определяет:

1. Алгоритмы построения представлений экземпляра таблицы.
2. Способ описания выбора конкретного варианта алгоритма и набора его параметров.

Части модели таблицы должны быть согласованы в том смысле, что

1. Любой экземпляр таблицы содержит факты предметной области в соответствии с моделью предметной области.
2. Для любого экземпляра таблицы можно построить плоское представление, в соответствие с алгоритмом построения плоского представления и моделью данных.
3. Операции над таблицей определяются так, чтобы при любом переходе между экземплярами таблицы выполнялись предыдущие пункты.

Представленный в соответствие с моделью плоского представления табличный объект может быть выделен, идентифицирован и разобран в соответствие с моделью распознавания. Полное или частичное восстановление табличного ИО возможно по его представлению.

3. Ключевые проблемы в обработке таблиц

Все основные проблемы в современной обработке таблиц относятся к одной из трех категорий:

— размытость предметной области;

— невозможность создания технологичных электронных таблиц по заданной логической структуре;

— отсутствие разборщика распространенных электронных таблиц.

Первая категория представляет собой единую фундаментальную проблему, видимо, не имеющую точного решения. Проблема же приближенного решения распадается на две — эффективное определение верхней и нижней граней для предметной области.

Вторая категория охватывает большой набор проблем разных типов и сложности — от отсутствующих сейчас, но относительно легко

разрабатываемых средств и методов до тяжелых, порой кажущихся неразрешимыми концептуальных проблем.

Третья категория, напротив, содержит, по-сути, единственную проблему, себя саму. Речь идет фактически о недостающей программе распознавания. И как любую задачу автоматического распознавания, решить ее можно, но со значительными усилиями и с ограниченным качеством решения.

3.1. Проблемы фиксации предметной области

Строго формализовать понятие таблицы, по всей видимости, невозможно. Можно построить последовательность информационных объектов с плавно ухудшающимися табличными свойствами так, что первый объект будет таблицей, а последний ею не будет. В ряде случаев установить объективно точный момент качественного перехода из таблицы в «не таблицу» невозможно.

Тем не менее, представляется принципиально возможным дать **конструктивное определение точной верхней грани для множества табличных объектов**. Но для этого требуется не математический, а естественный язык. Такие категории, как «примерно», «значительное количество», «намного больше», вполне конструктивны, несмотря на формальную неопределенность, и обеспечивают большую мощность естественного языка, по сравнению с алгоритмическим.

Данная проблема представляется краеугольной во всей табличной тематике. Ее решение необходимо и достаточно для трансформации локальных практик и выводов в научную дисциплину.

Для оценки существующих и разработки новых систем обработки табличной информации необходима также **эффективная нижняя грань для множества табличных объектов**. Под этим авторы подразумевают определение

а) формализованное;

б) покрывающее подавляющую часть таблиц;

в) признанное общественным мнением.

3.2. Проблемы создания технологичных таблиц

1. Отсутствие технологичного редактора для нормальных двумерных таблиц. Представляется откровенным упущением современных производителей программного обеспечения.
2. Отсутствие идеи для технологичного редактирования нормальных трехмерных таблиц. Очень интересное место для изобретателей. Такой идеи вполне может и не быть, но и уверенности, что здесь не произойдет открытие, тоже нет.

3. Отсутствие идеи технологичного сворачивания таблиц по измерениям после двух для отчасти приемлемой работы с многомерными таблицами.
4. Отсутствие эстетичного способа записи полной табличной информации.
5. Отсутствие единого формата для записи полной табличной информации.

3.3. Перспективы создания разборщика таблиц

Выше отмечалась принадлежность этой задачи к семейству задач автоматического распознавания и, как следствие, общие свойства решения. Здесь мы подчеркнем особенности верхнего уровня именно этой задачи.

1. Разнообразие распространенных электронных форматов. Это не очень существенно, основных есть и будут единицы (xls, doc, xml, html). Тем не менее, высококачественная программа должна будет содержать две компоненты — собственно распознаватель и требующий постоянной актуализации чтец разных форматов.
2. Разнообразие моделей редуцированного представления. Это сердцевинная проблема для распознавателя. Программа всегда будет покрывать это разнообразие не полностью. Обычные требования к автоматическому распознаванию — промышленно высокий процент правильно обрабатываемых объектов и надежная диагностика сомнительных ситуаций — обещают значительные трудности, не меньшие, чем в других хрестоматийных задачах распознавания.
3. Потенциальная ценность конвертора таблиц велика, но жестко ограничена уровнем современных ему средств извлечения выгоды из знания логики таблицы. Т. е. ценность разборщика сильно снижается в отсутствие решения проблем первой категории.

Заключение

Обработка таблиц — динамично развивающаяся область информационных технологий. Новые открытия и изобретения представляются здесь вполне возможными и несущими отличные коммерческие перспективы. Многочисленные частные задачи представляют несомненный научный интерес в области разработки искусственного интеллекта.

Однако современное развитие систем обработки таблиц принципиально ограничено размытостью предметной области. Возможно, обработка таблиц станет первой областью, где будет разработан подлинный искусственный интеллект, т. е. программа сможет решать задачи, определенные на естественном, а не на алгоритмическом языке.

Литература

1. *Арлазаров В. Л.*, *Емельянов Н. Е.* От баз данных к базам знаний (объекты, формы, содержание) // Сборник трудов ИСА РАН, М.: КомКнига/URSS, 2006. С. 6–17.
2. *Миллер Дж. А.* Магическое число семь плюс или минус два. О некоторых пределах нашей способности перерабатывать информацию // Психология памяти / Под ред. Ю. Б. Гиппенрейтер и В. Я. Романова. М.: ЧеРо, 2000. С. 564–582.
3. *Dalgleish D.* Excel Pivot Tables Recipe Book: A Problem-Solution Approach. APRESS, 2006. 335 p.
4. *Gerald J. O'Brien and David Wilde W.* Australian managers' perceptions, attitudes and use of information technology // Information and Software Technology, 1996. V. 38. P. 783–789.
5. *Vlahos G. E.*, *Ferratt T. W.* The use of information technology by managers of corporations in greece to support decision making. In Proceedings of the conference on Computer Personal Research. ACM, 1992. P. 136–151.
6. *Vlahos G. E.*, *Ferratt T. W.*, *Knoepfle G.* Use and perceived value of computer-based information systems in supporting the decision making of german managers. In Proceedings of the conference on Computer Personal Research. ACM, 2000. P. 111–123.
7. *Clermont M.*, *Hanin C.*, *Mittermeir R.* A Spreadsheet Auditing Tool Evaluated in an Industrial Context // In Spreadsheet Risks. Audit and Development Methods. 2002. V. 3. P. 35–46.
8. *Clermont M.* A Scalable Approach to Spreadsheet Visualization. PhD thesis, Universität Klagenfurt, Austria, 2003. 202 p.
9. *Butler R.* Is This Spreadsheet a Tax Evader? How H. M. Customs & Excise Test Spreadsheet Applications // In Proceedings of the 33rd Hawaii International Conference on System Sciences. 2000. V. 4. P. 400407
10. *Nunez F.* An extended spreadsheet paradigm for data visualization systems and its implementation. M. Sc. dissertation, University of Cape Town, 2000. 156 p.
11. *Tukiainen M.* Developing a New Model of Spreadsheet Calculation: A Goals and Plans Approach. PhD dissertation, University of Joensuu, 2001. 121 p.