ИДЕНТИФИКАЦИЯ ВЫЧИСЛЯЕМЫХ ЗНАЧЕНИЙ В СЛАБОСТРУКТУРИРОВАННЫХ ТАБЛИЧНЫХ ДОКУМЕНТАХ

Парамонов В. В., Шигаров А. О.
Институт динамики систем и теории управления им. В. М. Матросова Сибирского отделения Российской академии наук
Иркутск, Российская Федерация
E-mail: {slv, shigarov}@icc.ru

В настоящее время большое внимание уделяется созданию инструментальных средств для управления базами данных, позволяющих интегрировать информацию принадлежащей одной предметной области, но получаемую из различных источников. Часто интегрируемые данные представлены в виде слабоструктурированных электронных таблиц с различной компоновкой. В ряде случаев данные формируются из открытых источников и не являются стандартизованными. Это приводит к тому, что над ними требуется проведение операций очистки. Один из аспектов очистки – идентификация вычисляемых значений, т.к. они являются избыточными при интеграции данных, а также могут содержать ошибки вычислений. В работе предлагаются подходы для поиска таких значений.

Введение

Для принятия обоснованных решений необходима надежная система данных. В связи с этим возрастает потребность в новых методах и технологиях для организации процессов интеграции данных [1]. Интегрированные данные имеют некий единый интерфейс, что упрощает доступ к ним. Благодаря большому объему и наличием расширенного числа показателей представляется возможность проведения их анализа с учетом множества критериев, что повышает ценность данных. Также интеграция предоставляет более широкие возможности для дальнейшего проведения совместных научных исследований. Ввиду того, что информация может быть получена из различных источников, данные могут значительно отличаться по формату своей организации и представлению. В первичных документах, являющихся основой для формирования массива данных, встречаются вычислимые данные. Как правило, они представляют собой различные агрегированные значения, полученные в результате математических операций (сумма, среднее значение, максимальное, минимальное значения, нарастающий итог и пр.) над каким-либо набором данных. Т.е. использованы функции обрабатывающие набор значений для подсчета и возвращающих одного значение.

В случае слияния данных, подобные значения являются избыточными, т.к. их всегда можно получить по имеющейся информации. К тому же вычисляемые значения, представленные в электронных таблицах, часто содержат ошибки [2], что в итоге отражается на качестве интегрированной информации. Таким образом, в рамках вопросов очистки интегрируемой информации представляется важным проведение работ по идентификация вычисляемых значений и их коррекции или экстрагирования для повышения качества данных.

В работе не рассмотрены вопросы извлечения слабоструктурированных данных из табличных документов [3]. В данном исследовании, считается, что структура и характеристики таблицы уже определены.

I. Представление данных

Данные, используемые для интеграции, как правило, представлены в виде таблиц, но в документах различных форматов и, соответственно, имеют отличную структуру. Связано это с тем, что источниками являются не только выполненные в строгом, определенном формате документы, сформированные, например, подразделениями службы Государственной статистики, различными министерствами и ведомствами, но и предоставленные заинтересованными в обмене, накоплении, обработке информации организациями и частными лицами. Иными словами, данные могут собираться по принципу краудсорсинга. В следствии этого интегрируемая информация представляются в форматах и структурах удобных именно владельцам данных. В следствии этого даже однотипные данные могут быть организованы совершенно по–разному. Как правило такие данные представлены в виде документов табличных процессоров (Excel, Calc), форматах CSV (Comma Separated Values), PDF (Portable Document Format). Не всегда из названия полей возможно узнать о наличии каких-либо агрегированных значений в ячейках таблицы. Подобное разноформатное представление приводит к сложности идентификации вычисляемых значений.

II. Идентификация искомых значений

Рассмотрим 3 основных случая представления вычислимых значений в документах:

– есть конкретная формула;

– есть ключевые слова, идентифицирующие вероятное вычислимое значение;
– требуется анализ близлежащих ячеек.

В некоторых случаях, например при обработке электронных таблицы, таких как Excel или Calc, возможно получить сведения как о структуре документа, опираясь на его объектную модель. Через объектную модель можно идентифицировать формулу, используемую для ячейки, если таковая существует. Извлеченную формулу следует сопоставить с остальным значениями ячеек исследуемой строки (столбца). Это позволяет проверить похожесть используемых формул с учетом смещения. Однако данный подход имеет достаточно ограниченную сферу применения, т.к. зависит от формата документов и использования в них формул. В ряде случаев формулы не используются или используются не корректно, например для одной ячейки применена функция «SUM», для другой – сложение значений из других ячеек или число. Поэтому представляет интерес анализ текстовых значений, описывающих структурные элементы таблицы - категории, метки, дочерние метки [4]. Пример возможной структуры таблицы приведен на рис. 1.

Категория	Метка	Метка
Метка	Вхождение	Вхождение
Дочерняя метка	Вхождение	Вхождение
Метка	Вхождение	Вхождение

Рис. 1 – Пример структуры таблицы

Анализируя структурные элементы таблицы выделяются диапазоны значений находящихся между словами соответствующим операциям агрегации: «минимум», «максимум», «среднее», «сумма», «нарастающий итог», «за период» и т.п. Список возможным операций агрегации и действия над предполагаемым набором данных содержится в специальном глоссарии. После вычисления значения выбранного множества данных проводится сверка со значением фактически содержащимся в ячейке. Если возможно экстрагировать формулу, то проверяется на сколько проводимые операции и набор данных соответствуют имеющейся формуле. В случае, если вычисленное и фактическое значения различаются, то проводится анализ других вычисляемых значений, относящихся к данной области, т.к. требуется установить является некорректным значение в ячейке или операция вычисления выполнена не корректно.

Наиболее сложным представляется способ идентификации вычисляемых данных, когда не доступна (или нет возможности однозначно её интерпретировать) информация, их характеризующая. В таком случае проводится анализ значений содержащихся в соседних ячейках. И, в случае определения, отклонения проводится перебор возможных формул для идентификации значений. При этом для анализируемых в области ячеек ведутся вычисления агрегированных значений, т.е. для каждой ячейки проверяется ряд гипотез. Число гипотез зависит от числа вхождений в глоссарий. Как только гипотеза подтвердилась, проводится её проверка для смежного строки (столбца). Если для соседнего набора данных гипотеза не подтверждается, то исследуемая область расширяется. Операция продолжается пока не будет гипотеза не подтверждена либо пока не достигнут конец данных по строке (столбцу).

III. Выводы

Поиск вычисляемых значений – одна из значимых задач по очистке данных, что в свою очередь является частью работ по интеграции данных.

Предложенные методы апробированы при интеграции данных в рамках информационного сервиса по слежению за природными очагами клещевых инфекций и гельминтозов в туристско-рекреационных зонах Республики Бурятия, а также для предварительной обработки данных используемых для поддержки управления социально-экономическим развитием территорий. В случае, когда нет информации описывающих данные в ячейки происходит полный перебор возможных возможных значений. Для больших таблиц это может привести к снижению быстродействия. Для повышения быстродействия и эффективности планируется применять методы машинного обучения для поиска вычисляемых значений.

Работа выполнена при частичной финансовой поддержке РФФИ гранты № 16-57-44034, 16-07-00411, РФФИ и Правительства Республики Бурятия в рамках научного проекта № 15-47-04348, программы Президиума РАН № 27 «Фундаментальные проблемы решения сложных практических задач с помощью суперкомпьютеров».

1. Намиот Д. Е. [и др.] стандарты в области больших даных / Д. Е. Намиот, В. П. Куприяновский, Д. Е. Николаев, Е. В. Зубарева // International Journal of Open Information Technologies. Vol. 6. № 11 — 2016. — pp 12–17.
2. Raymond R. Panko. What We Know About Spreadsheet Errors // The Journal of End User Computing's Special issue on Scaling Up End User Development Vol. 10, No 2. Spring 1998, pp. 15–21
3. Шигаров А. О., Бычков И. В., Парамонов В. В., Белых П. В. Анализ и интерпретация произвольных таблиц на основе исполнения CRL правил // Вычислительные технологии. — 2015. — т. 20, № 6. — С. 87 – 112.
4. Alexey O. Shigarov, Viacheslav V. Paramonov, Polina V. Belykh, Alexander I. Bondarev. Rule-Based Canonicalization of Arbitrary Tables in Spreadsheets // Communications in Computer and Information Science. — 2016. — Vol. 639, — pp. 78-–91.