TOWARD FRAMEWORK FOR DEVELOPMENT OF SPREADSHEET DATA EXTRACTION SYSTEMS

Alexey O. Shigarov[(1)], Vasiliy V. Khristyuk [(1,2)], Viacheslav V. Paramonov[(1)], Alexandr Yu. Yurin[(1)], Nikita O. Dorodnykh[(1)]

[(1)] Matrosov Institute for System Dynamics and Control Theory SB RAS, Irkutsk, Russia
[(2)] Irkutsk national research technical university, Irkutsk, Russia

The paper presents a problem formulation for the development of a theoretical and software framework for creating systems of data extraction from arbitrary spreadsheet tables. The problem covers the tasks of the automatic recovering semantic markup of tables, conceptualization of their natural-language content, data cleaning and lineage, generating relational and linked data, as well as a synthesis of tabular data transformation systems based on table analysis and interpretation rules. We consider the state of the art methods and discuss some perspective techniques for the development of a consistent solution to this problem.

Keywords: information extraction, data integration, unstructured data management, rule-based programming.

КОНЦЕПЦИЯ ПЛАТФОРМЫ ИЗВЛЕЧЕНИЯ ДАННЫХ ИЗ ПРОИЗВОЛЬНЫХ ЭЛЕКТРОННЫХ ТАБЛИЦ

Шигаров А.О.[(1)], Христюк В.В. [(1, 2)], Парамонов В.В.[(1)], Юрин А.Ю.[(1)], Дородных Н.О.[(1)]

[(1)] Институт динамики систем и теории управления имени В. М. Матросова Сибирского отделения Российской академии наук, г. Иркутск
[(2)] Иркутский национальный исследовательский технический университет, г. Иркутск

В работе рассматриваются вопросы создания теоретической и инструментальной платформы для ускоренной разработки программных систем извлечения данных из произвольных электронных таблиц. Данная проблема охватывает задачи автоматического восстановления семантической разметки таблиц, концептуализации их естественно-языкового содержания, очистки и отслеживания происхождения табличных данных, генерации реляционных и связных данных, а также синтеза исходного кода программ трансформации табличных данных. В работе рассматривается современное состояние исследований в данной области и обсуждаются некоторые перспективные подходы к решению поставленной проблемы.

Ключевые слова: извлечение информации, интеграция данных, управление неструктурированными данными, программирование на основе правил.

Введение. Произвольные таблицы являются распространённым способом представления информации. Они повсеместно используются в документах и веб-пространстве в форматах HTML, PDF, EXCEL, CSV. Современные оценки, сделанные на основе изучения экспериментальных срезов содержания Веба (COMMONCRAWL и CLUEWEB), показывают, что их количество исчисляется сотнями миллионов (WEB TABLE CORPORA, DRESDEN WEB

TABLE CORPUS). Предположительно они содержат сотни миллиардов фактов. Такие таблицы характеризуются большим разнообразием и разнородностью компоновок, стилей и содержания, а также высокой скоростью роста их объема.

Большой объем и свойства структуры таких таблиц делают их ценным источником в приложениях науки о данных и бизнес аналитики. Однако, как правило, они не сопровождаются явной семантикой необходимой для машинной интерпретации своего содержания так, как задумано их автором. Накапливаемая в них информация часто является неструктурированной и не стандартизированной. Анализ этих данных нуждается в их предварительном извлечении и трансформации к структурированному представлению с заданной формальной моделью.

Сегодня, сталкиваясь с перечисленными задачами, исследователи и разработчики прибегают к инструментальным средствам общего назначения, часто предлагая собственные реализации однотипных задач. В сравнении с последними, специализированные инструментальные средства могут позволить сократить время разработки целевого программного обеспечения, скрывая несущественные детали и фокусируясь на обозначенной проблемной области. Это особенно важно в тех случаях, когда необходимо в короткие сроки и при недостатке ресурсов разработать заказное или исследовательское программное обеспечение для массовой обработки слабоструктурированных данных из разнотипных произвольных таблиц.

Проблема управления и интеграции неструктурированной информации включает вопросы извлечения данных из таблиц произвольной формы, представленных в веб-страницах, текстовых и табличных документах. Извлечения табличных данных состоит в восстановлении недостающих структурных и содержательных метаданных (семантики) произвольной таблицы. Как правило произвольная электронная таблица не включает метаданные, описывающие ее структуру и содержание. Неизвестно то, какие роли играют её ячейки (например, содержат ли они данные или атрибуты), как они связаны между собой (например, какие заголовки адресуют значение данных), какими предметными понятиями (категориями) описывается её содержание. Решение обозначенных вопросов позволяет перейти от произвольной (слабоструктурированной) формы таблиц к некоторому структурированному представлению данных (например, реляционной базе данных).

Научная значимость решения этой проблемы определяется расширением знаний о природе произвольных электронных таблиц и процессах их преобразования в структурированную форму. Решение проблемы вносит важный вклад в ряд научных направлений, где произвольные таблицы являются естественным объектом исследований. К таким направления относятся: управление неструктурированными данными (вопросы интеграции табличной информации), информационный поиск (вопросы извлечения данных из произвольных таблиц), а также семантический веб (вопросы формирования открытых связных данных).

В работе предлагается создать теоретическую и инструментальную платформы ускоренной разработки систем извлечения данных из произвольных электронных таблиц. Поставленная задача охватывает исследование и разработку следующих вопросов: (1) алгоритмов ролевого и структурного анализа таблиц (автоматического восстановления их синтаксической и семантической разметки); (2) алгоритмов интерпретации таблиц (очистки и концептуализации их естественно-языкового содержания); (3) объектной модели и формального языка трансформации табличных данных; (4) инструментов синтез исходного кода программ

трансформации табличных данных; (5) инструментов генерации реляционных и связных данных из произвольных таблиц.

Современное состояние исследований. Два последних десятилетия активно развиваются методологические основы извлечения и трансформации данных из таблиц, представленных в веб-страницах, текстовых и табличных документах. В работах [1, 6, 8, 11, 16, 22, 24, 27, 29] предлагаются методы ролевого и структурного анализа таблиц (восстановления отношений между ячейками). Каждый из этих методов ограничен обработкой небольшого количества (1-5) широко распространенных видов табличных компоновок. Однако, многие виды таблиц специфичных для некоторых областей (например, паспорта безопасности продуктов в химии или паспорта электротехнического оборудования в энергетике) остаются неохваченными.

Методы интерпретации таблиц [5, 9, 23, 26, 30, 31] предназначены для связывания табличного содержания с внешними концепциями (предметными онтологиями, глобальными таксономиями, открытыми связными данными). В основном, они полагаются на анализ естественно-языкового содержания таблиц и их контекста, пренебрегая их компоновочными и стилевыми особенностями. На практике, это бывает недостаточно, например, часть информации может выражаться через различные шрифтовые или цветовые свойства.

Продолжают развиваться специализированные инструменты для извлечения и трансформации данных из произвольных электронных таблиц в структурированную форму, в т. ч., системы трансформации табличных данных [2, 3, 7, 16–21], системы извлечения связанных данных [13–15, 25, 28]; TANGO — система трансформации произвольных таблиц к реляционной форме на основе поиска критических ячеек [11, 27]; SENBAZURU — система извлечения реляционных данных из таблиц с иерархиями заголовков [6], DEEXCELERATOR — система нормализации данных электронных таблиц к реляционной форме [10]. Инструменты TANGO, SENBAZURU и DEEXCELERATOR преследуют схожие цели с нашими — преобразовать таблицы от произвольной к реляционной форме. Однако, они используют заданные модели исходных таблиц, при которых смешивается их физическая и логическая компоновка. Это ограничивает возможности их применения сводными таблицами характерными для статистических отчетов.

Анализ современного состояния исследований в рассматриваемой области показывает, что на сегодняшний день не существует специализированных библиотек моделей и платформ для разработки систем извлечения и трансформации табличных данных. Кроме того, нет инструментальных средств, в которых извлечение и трансформация данных выражаются как последовательность известных операций автоматического анализа и интерпретации таблиц. Ценность извлекаемых данных может быть повышена путем их связывания со свободными глобальными таксономиями понятий общего назначения, например, DBPEDIA, YAGO, WIKIDATA, WIKITOLOGY. Согласованные, непротиворечивые и обогащенные URI данные могут быть использованы при дальнейшей их интеллектуальной обработке или построении на их основе интеллектуальных систем. Исследования в данной области представлены работами [12, 17, 25] и др. Однако, перспективным является создание не только связанных RDF и OWL онтологий, но создание узкоспециализированных проблемно-ориентированных таксономий и онтологических паттернов.

Предлагаемые подходы. Процесс извлечение данных из произвольной таблицы включает следующие этапы: (i) ролевой анализ (извлечение единиц данных из табличного содер-

жания и сопоставление их с функциональными ролями); (ii) структурный анализ (восстановление связей между единицами табличных данных); (iii) интерпретация (ассоциирование единиц табличных данных с категориями внешних словарей).

Анализ таблиц может базироваться на следующих решениях: (i) алгоритмах извлечения единиц данных (вхождений, меток, ссылок) из естественно-языкового содержания ячеек на основе восстановления литеральных типов данных и распознавания именованных сущностей; (ii) алгоритмах выделения типичных функциональных областей ячеек (боковик, шапка, тело, подвал, итог) на основе выделения кластеров по восстановленным литеральным типам (распознанным именованным сущностям) и эвристик поиска критических ячеек типа MIPS (Minimum Indexing Point Search); (ii) алгоритмах восстановления иерархических отношений между единицами данных в типичных функциональных областях на основе обучения по прецедентам (предполагается создать модели выбора таких отношений среди кандидатов; сформировать обучающие выборки на основе разметки существующих корпусов таблиц; рассматривать, как компоновочные, так и стилевые признаки, связанные с извлечёнными единицами данных).

Многие ошибки этой фазы возникают из-за того, что произвольные таблицы ориентированы прежде всего на восприятие человеком. Их физическая (синтаксическая) разметка часто не совпадает с визуальной разметкой (передаваемой прочерченными границами), которою читает человек. Например, одна визуальная ячейка может быть составлена из нескольких физических. Для уменьшения ошибок такого рода планируется исследовать возможность реконструкции физической (машиночитаемой) структуры ячеек по их визуальной (человекочитаемой) структуре. Предлагается разработать алгоритмы реконструкции на основе анализа визуальных границ ячеек.

Интерпретация таблиц обеспечивается за счет применения свободных глобальных таксономий понятий общего назначения (DBPEDIA, YAGO, WIKIDATA, WIKITOLOGY) для понимания и концептуализации естественно-языкового содержания произвольных таблиц. Благодаря свойствам табличной компоновки, возможно выделить каждый набор единиц данных, относящихся к одной неизвестной категории. Каждая единица (гипоним) может быть ассоциирована с одной или несколькими категориями (гиперонимами) из заданной глобальной таксономии. Его окружение (гипонимы из того же набора) может использоваться, чтобы уточнить его категорию. Кроме того, предлагается задействовать техники очистки данных для удаления знаков-заполнителей, ссылок на сноски, "мусорных" слов ("из них", "всего" и др.), вычислимых (агрегированных) значений.

Язык правил трансформации произвольных таблиц CRL предлагается усовершенствовать на основе расширения его функциональных возможностей алгоритмами, реализованными в рамках проекта. Планируется разработать специализированный диалект этого языка для выражения лаконичных программ трансформации данных электронных таблиц в виде наборов правил. При этом одна программа может охватывать широкий диапазон произвольных таблиц, разных на просвет, но разделяющих общий набор компоновочных, стилевых и содержательных свойств. В процессе исполнения правила трансформации табличных данных должны отображать факты о физической структуре таблицы (данные о компоновке, стиле и содержании) в факты о ее логической структуре (функциональные единицы и их семантические связи). Для представления фактов предлагается дополнить язык объектной моделью таблицы со строгим разделением физической и логический структуры.

Синтез исходного кода программ трансформации табличных данных. Планируется также дополнить развиваемый язык специализированным диалектом для генерации исходного кода программ трансформации табличных данных в формате языка общего назначения. Предлагается разработать формальную (ANTLR) грамматику предлагаемого языка, объектную модель и интерпретатор данного языка. Это позволит реализовать синтаксический разбор наборов правил, восстановления объектной модели программы и её интерпретацию для синтеза целевого исходного кода. Предполагается, что порожденная программа не должна требовать доработки для компиляции и запуска, но может быть расширена, если это необходимо.

Генерация реляционных и связных данных из произвольных таблиц. Предлагается обобщить опыт исследователей в данной области, в частности [12, 17, 25], и рассмотреть процесс генерации связанных данных как некую последовательность этапов: выявление классов-кандидатов из открытой таксономии (далее, LOD), соответствующих столбцам; соотнесение значений ячеек таблиц с соответствующим объектам LOD (если это возможно); выявление отношений между столбцами таблицы и их связь со свойствами LOD; генерация связанного представления данных с указанием URI в формате RDF. Достаточно интересными подзадачами в рамках подхода являются выбор метрик для вычисления близости между понятиями, а также подбор релевантных узлов.

Предлагаемые подходы соответствуют современному мировому уровню исследований. Они во многом преодолевают ограничения присущие мировым конкурентным проектам (TANGO, SENBAZURU и DEEXCELERATOR). В частности, предлагается реализовать принципиально новую объектную модель и формальный язык трансформации табличных данных. По сравнению с конкурентными решениями это позволит работать не только с широко-распространенными, но и со специфичными типами произвольных таблиц. В отличие от конкурентных языков, где табличная компоновка задается строго (например, в абсолютных координатах), нами впервые предлагается язык для выражения неизменяемых компоновочных, стилевых и содержательных свойств, разделяемых некоторым набором таблиц. Таким образом, одна программа может обрабатывать существенно более широкий диапазон произвольных таблиц. Кроме того, впервые предлагается разработать инструменты синтеза программных систем трансформации табличных данных от произвольной к реляционной форме.

Заключение. Проектируемая платформа открывают новые возможности интеллектуализации для ускоренной разработки программного обеспечения извлечения информации в научных и индустриальных приложениях с интенсивным использованием разноструктурированных табличных данных. Планируемые результаты расширяет теоретические знания в области разработки программного обеспечения систем интеграции разнородных табличных данных больших объемов. Результаты проекта внесут вклад в решение следующих проблем: интеграция слабоструктурированной табличной информации, извлечение данных из произвольных таблиц, а также формирование открытых связных данных.

Результаты могут использоваться на практике в сфере бизнес аналитики и науки о данных. Они могут быть положены в основу создания новых наукоёмких технологий, продуктов и услуг интеграции слабоструктурированной табличной информации в приложениях аналитики данных. Особенный интерес для их применения представляет аналитика данных в сферах с интенсивным использованием электронных таблиц (например, финансы, экономика,

государственное и бизнес управление). В качестве апробации разрабатываемой платформы предлагается рассмотреть задачи автоматизированного заполнения баз данных: опасных химических веществ, на основе анализа их деклараций, а также результатов диагностирования в рамках проведения экспертизе промышленной безопасности опасных объектов в нефтехимии [4] на основе анализа электронных отчетов.

Работа выполнена при финансовой поддержке Российского научного фонда (грант № 18-71-10001). Результаты получены при использовании ЦКП «Интегрированная информационно-вычислительная сеть Иркутского научно-образовательного комплекса» (http://net.icc.ru).

ЛИТЕРАТУРА

[1] *Adelfio M., Samet H.* Schema extraction for tabular data on the web // Proc. VLDB Endow, 2013, vol 6, iss. 6, pp. 421-432.

[2] *Astrakhantsev N., Turdakov D., Vassilieva N.* Semi-automatic data extraction from tables // Proc. 15th All-Russian Conf. Digital Libraries, 2013.

[3] *Barowy D.W., Gulwani S., Hart T., Zorn B.* FlashRelate: Extracting relational data from semi-structured spreadsheets using examples // Proc. of the 36th ACM SIGPLAN Conference on Programming Language Design and Implementation, 2015, vol. 50, iss. 6, pp. 218-228.

[4] *Berman A.F., Nikolaichuk O.A., Yurin A.Y. and Kuznetsov K.A.* Support of Decision-Making Based on a Production Approach in the Performance of an Industrial Safety Review // Chemical and Petroleum Engineering, 2015, vol. 50 (11-12), pp. 730-738.

[5] *Cao T.D., Manolescu I., Tannier X.* Extracting linked data from statistic spreadsheets // Proc. Int. Workshop Semantic Big Data, 2017.

[6] *Chen Z., Cafarella M.* Integrating spreadsheet data via accurate and low-effort extraction // Proc. 20th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2014. pp. 1126-1135.

[7] *Cunha J., Erwig M., Mendes M.,Saraiva J.* Model inference for spreadsheets // Autom Softw Eng. 2016, vol. 23. pp. 361-392.

[8] *de Vos M., Wielemaker J., Rijgersberg H. et al*. Combining information on structure and content to automatically annotate natural science spreadsheets. // Int. J. of Human-Computer Studies, 2017, vol. 130. pp. 63-76.

[9] *Deng D., Jiang Y., Li G., Li J., Yu C.* Scalable column concept determination for web tables using large knowledge bases // Proc. VLDB Endow. 2013, vol. 6 (13). pp. 1606-1617.

[10] *Eberius J., Werner C., Thiele M. et al*. DeExcelerator: a framework for extracting relational data from partially structured documents // Proc. 22nd ACM Int. Conf. on Information & Knowledge Management, 2013.

[11] *Embley D. W., Krishnamoorthy M. S., Nagy G., Seth S.* Converting heterogeneous statistical tables on the web to searchable databases // Int. J. Document Analysis and Recognition, 2016. vol. 19, iss. 2. pp. 119-138.

[12] *Embley D.W., Tao C., Liddle S.W.* Automating the extraction of data from HTML tables with unknown structure // Data & Knowledge Engineering. 2005. vol. 54 (1). pp.3-28.

[13] *Ermilov I., Ngonga Ngomo A.* TAIPAN: Automatic Property Mapping for Tabular Data // Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016. pp. 163-179.

[14] *Fiorelli F.,Lorenzetti T., Teresa M. et al.* Sheet2RDF: a flexible and dynamic spreadsheet import & lifting framework for RDF // Proc. 28th Int. Conf. Industrial, Eng. and Other Appl. of Applied Intelligent Systems, 2015. pp. 131-140.

[15] *Galkin M., Musomtsev D.* Identifying web tables: Supporting a neglected type of content on the web // Proc. 6th Int. Conf. Know. Eng. and Semantic Web, 2015. vol. 518. pp. 48-62.

[16] *Goto K., Ohta Yu., Inakoshi H., Yugami N.* Extraction algorithms for hierarchical header structures from spreadsheets // Proc. Workshops of the EDBT/ICDT 2016 Joint Conference, 2016. vol. 1558.

[17] *Han L., Finin T., Parr C., Sachs J., Joshi A.* RDF123: From Spreadsheets to RDF // Proceedings of the 7th International Semantic Web Conference, ISWC. 2008. P. 451-466.

[18] *Harris W., Gulwani S.* Spreadsheet table transformations from examples. // SIGPLAN Not., 2011. vol. 46. iss. 6. pp. 317-328.

[19] *Hung V., Benatallah B., Saint-Paul R.* Spreadsheet-based complex data transformation. Proc. 20th ACM Int. Conf. Inf. and Know. Management, 2011. pp. 1749–1754.

[20] *Jin Z., Anderson M. R., Cafarella M., Jagadish H. V.* Foofah: Transforming data by example // Proc. ACM Int. Conf. Management of Data.

[21] *Kandel S., Paepcke A., Hellerstein J., Heer J.* Wrangler: Interactive visual specification of data transformation scripts // Proc. SIGCHI Conf. Human Factors in Computing Systems. pp. 3363-3372.

[22] *Koci E., Thiele M., Romero O., Lehner W.* A machine learning approach for layout inference in spreadsheets // Proc. 8th Int. Joint Conf. Knowledge Discovery, Knowledge Engineering and Knowledge Management, 2016. pp.77-88.

[23] *Limaye G., Sarawagi S., Chakrabarti S.* Annotating and searching web tables using entities, types and relationships // Proc. VLDB Endow, 2010. vol. 3. Iss. 1-2. pp. 1338-1347.

[24] *Mauro N., Esposito F., Ferilli S.* Finding critical cells in web tables with SRL: Trying to uncover the devil's tease // Proc. 12th Int. Conf. on Document Analysis and Recognition, 2013. pp. 882-886.

[25] *Mulwad V., Finin T., Joshi A.* A Domain Independent Framework for Extracting Linked Semantic Data from Tables. // Search Computing, 2012. pp. 16-33.

[26] *Muñoz E., Hogan A., Mileo A.* Using linked data to mine RDF from wikipedia's tables // Proc. 7th ACM Int. Conf. Web Search and Data Mining, 2014. pp. 533-542.

[27] *Nagy G., Seth S.* Table headers: An entrance to the data mine // Proc. 23rd Int. Conf. Pattern Recognition, 2016. pp. 4065-4070.

[28] *O'Connor M. J., Halaschek-Wiener C., Musen M. A.* Mapping Master: A Flexible Approach for Mapping Spreadsheets to OWL // LNCS: The Semantic Web – ISWC 2010, 2010. pp 194-208.

[29] *Rastan R., Paik H., Shepherd J., Haller A.* Automated Table Understanding Using Stub Patterns. // LNCS: Database Systems for Advanced Applications, 2016. vol. 9642. pp 533-548.

[30] *Venetis P., Halevy A., Madhavan J. et al.* Recovering semantics of tables on the web // Proc. VLDB Endow, 2011. vol. 9, iss. 9. pp. 528-538.

[31] *Wang J., Wang H., Wang Z., Zhu K.* Understanding tables on the web // LNCS: Proc. 31st Int. Conf. Conceptual Modeling, 2012. vol. 7532. pp. 141-155.