АЛГОРИТМИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ДЛЯ ИЗВЛЕЧЕНИЯ ТАБЛИЧНЫХ ДАННЫХ ИЗ HTML-СТРАНИЦ

Алакшин М.С., Парамонов В.В.

Ключевые слова: HTML таблицы, реляционные данные, классификация, обнаружение, извлечение.

В веб-среде содержится большое количество таблиц, часть из которых является ценным источником информации [3]. Извлечение, обобщение данных, содержащихся в подобного виде таблицах, имеет высокую значимость и может быть использовано в различных областях, таких как бизнес-аналитика, научные исследования и т.п. Для интеграции, анализа табличных данных одной из первичных задач является обнаружение и извлечение таблиц.

Один из основных форматов представления документов в веб-среде – HTML [2]. Табличные данные, организованные в таком формате, в первую очередь ориентированы на удобное восприятие человеком. Это влечет разнообразие схем организации данных, что, порождает трудности для автоматизированного понимания HTML-таблиц.

В работе [3] предлагается деление всех веб-таблиц на два класса – подлинные и неподлинные. Под подлинными понимаются таблицы, которые могут быть представлены в реляционном виде. Под неподлинными – те, которые используются для разметки непосредственно страницы.

Следует отметить, что подлинные таблицы, ввиду разнообразия организации в них данных можно разбить на насколько классов. Однако, проведенный анализ показал, что таксономии, предложенные различными исследователями, содержат различные классы таблиц. Также отмечается противоречивость и неполнота классов в имеющихся таксономиях [1, 6].

В данной работе за основу взята таксономия Е. Крестана [4, 5]. Однако в предоставленной классификации не рассматривается очень широкий класс таблиц – таблицы с иерархическим заголовками. Данный класс использован как расширение таксономии.

По результатам исследования был разработан алгоритм, позволяющий идентифицировать и извлечь подлинные таблицы, с соответствия с организацией данных, описанных в [4, 5] и с учетом возможной иерархии заголовков.

Данный алгоритм был использован для создания инструментального программного средства, HTML-Tables-Extractor[1], обеспечивающего извлечение подлинных таблиц, содержащихся на HTML страницах. Приложение позволяет проводить обнаружение подлинных таблиц, соответствующих таксономии и экспортировать их в документ формата Excel (файл *.xlsx). Каждая выгруженная таблица при этом представляется на одном листе электронной книги. Разработанное инструментальное программное средство обеспечивает подготовку данных для дальнейшего понимания таблиц.

Список литературы

1. Парамонов, В. В., Шигаров, А. О. Обзор таксономий веб-таблиц // Материалы 6-й Международной конференции «Динамические системы и компьютерные науки: теория и приложения (DYSC 2024)». 2024. С. 180–183.
2. Рудакова, Е. А., Тулупова, А. А. Создание и оформление web-сайта // Научно-образовательный потенциал как фактор национальной безопасности. 2021. С. 43–47.
3. Cafarella M.J., Halevy A., Wang D.Z., Wu E., Zhang Y. WebTables: Exploring the Power of Tables on the Web // Proceedings of the VLDB Endowment. 2008. Т. 1, №1. С. 538–549.
4. Crestan E., Pantel P. A fine-grained taxonomy of tables on the Web // Proceedings of the 19th ACM International Conference on Information and Knowledge Management. 2010. С. 1405–1408.
5. Crestan E., Pantel P. Web-scale table census and classification // Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. 2011. С. 545–554.
6. Kostyleva O., Paramonov V., Shigarov A., Vetrova V. Towards Comparison of Table Type Taxonomies // Proceedings of the 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO). IEEE. 2022. С. 1461–1465.Communication and Electronic Technology (MIPRO). IEEE, May 2022, pp. 1461–1465.

[1] https://github.com/Marks473/HTML-Tables-Extractor