ОБНАРУЖЕНИЕ И ИЗВЛЕЧЕНИЕ ПОДЛИННЫХ HTML-ТАБЛИЦ В ВЕБ-ДОКУМЕНТАХ НА ОСНОВЕ ЭВРИСТИК И МАШИННОГО ОБУЧЕНИЯ

Алакшин М. С., Парамонов В. В.

Ключевые слова: HTML-таблицы, подлинные таблицы, эвристики, машинное обучение, понимание таблиц

В веб-среде содержится большое количество таблиц, часть из которых является ценным источником данных для аналитики и извлечения знаний [2]. Один из наиболее популярных форматов представления данных в веб-среде – HTML. При этом HTML-таблицы используются как для разметки страниц, так и для организации данных, представимых в реляционном формате. Такое, базовое деление веб-таблиц на подлинные и неподлинные используется во многих работах [1]. К подлинным таблицам, в дальнейшем, применяются различные методы классификации, понимания, позволяющие извлечь данные и представить их в каноническом виде. Это делает данные доступными для анализа и интеграции.

В работе предлагается гибридная стратегия обнаружения подлинных таблиц: (1) эвристическая проверка структурной корректности и согласованности типов данных в ячейках; (2) бинарная классификация таблиц методом машинного обучения по относительным частотам типов содержимого ячеек. На эвристическом этапе учитываются семантические роли ячеек (заголовки/атрибуты/данные/производные), адаптированные для HTML-таблиц [5]. Для второго этапа таблица отображается в вектор признаков из 7 долей типов ячеек («числа», «элемент формы», «строка», «медиа данные», «другое», «ссылка», «нет данных»), после чего обучается классификатор; подобные признаки и постановка задачи бинарной классификации используются в ML-подходах к определению таблиц [6]. Эвристические подходы на правилах и метриках структуры таблицы и согласованности данных содержимого рассматриваются для сегментации структуры таблицы и снижения времени обработки, выявляя заранее неподлинные таблицы [3, 4].

Эксперименты выполнены на наборе из 1000 HTML-таблиц, извлечённых из архивного снимка Common Crawl (июль 2015): 46 подлинных (4.6%) и 954 неподлинных (95.4%). Для сравнения моделей использовалась стратифицированная 5-блочная кросс-валидация; наилучший баланс качества/устойчивости показал Random Forest Оценка эффективности: только эвристики — 74.78%; только машинное обучение — 90.63%; гибрид (эвристики + Random Forest) — 99.28%

Таким образом, предложенный подход обеспечивает высокое качество обнаружения подлинных HTML-таблиц по сравнению с использованием каждого из подходов по отдельности.

Список литературы

1. Парамонов В. В. Обзор таксономий веб-таблиц / В. В. Парамонов, А. О. Шигаров // Динамические системы и компьютерные науки: теория и приложения: тр. 6-й Междунар. конф. DYSC 2024. — 2024.
2. Cafarella M. J. WebTables: Exploring the Power of Tables on the Web / M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, Y. Zhang // Proceedings of the VLDB Endowment. — 2008.
3. Chen H. H. Mining Tables from Large Scale HTML Texts / H. H. Chen, S. C. Tsai, J. H. Tsai // Proceedings of the 18th conference on Computational linguistics (COLING ’00).
4. Kim Y.-S. Detecting Tables in Web Documents using Semantic and Syntactic Coherency / Y.-S. Kim, K.-S. Lee // Proceedings of the 6th International Conference on Web Information Systems Engineering Workshops (WISE’05). — 2005.
5. Koci E. Layout Inference in Spreadsheets / E. Koci, M. Thiele, W. Lehner, J. H. T. // 2017 IEEE 17th International Conference on Data Mining Workshops (ICDMW). — 2017.
6. Wang Y. A Machine Learning Based Approach for Table Detection on The Web / Y. Wang, J. Hu // Proceedings of the 11th International Conference on World Wide Web (WWW ’02). — 2002. — P. 242–250.