ARIA
APPLIED RESEARCH IN ACTION

Pytheas: Pattern-Based Table Discovery in CSV Files

We present Pytheas, our supervised learning approach for accurately discovering tables in Open CSV files, with a confidence metric for efficiently labeling data.

Christina Christodoulakis

Angela Demke Brown and Moshe Gabel
ACADEMIC SUPERVISORS

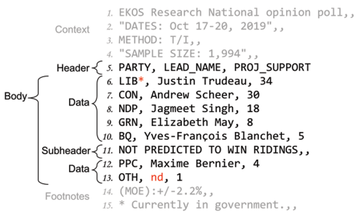

Figure 1: Example of a real Open CSV file, with a data table which does not follow expected CSV conventions. Without correctly automatically identifying table borders the resulting extracted data will be useless and can impact downstream applications relying on correct table extraction.

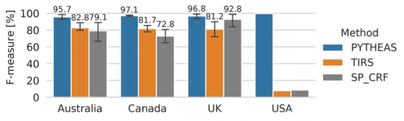

Figure 2a: Table discovery (Body and Header) with 10-fold cross-validation, averaged over all portals in each country in an International data set of 2500 files. Models are trained on 2000 Canadian Open Data files.

Figure 2b: Table discovery when growing the training set using active learning (uncertainty sampling) vs random sampling. Vertical lines show the training set size at which a method first reaches its final precision (precision when using the entire set of files). Pytheas's confidence measure is more effective at growing the training set, outperforming random sampling and reaching the plateau faster.

PROJECT SUMMARY

CSV is a popular Open Data format widely used in various domains for its simplicity and effectiveness in storing and disseminating data. Data published in this format has limited embedded metadata and often does not conform to strict specifications, making automated data extraction from CSV files a painful task.

We propose Pytheas: a principled method for automatically classifying lines in a CSV file and discovering tables within it based on the intuition that tables maintain a coherency of values in each column. We evaluate our methods over two manually annotated data sets: 2000 files sampled from four Canadian Open Data portals and 2500 additional files sampled from Canadian, US, UK, and Australian portals. Our comparison to state-of-the-art approaches shows that Pytheas is able to successfully discover tables with precision and recall of over 95.9% and 95.7%, respectively, compared to 89.6% precision and 81.3% recall. Furthermore, Pytheas's accuracy for correctly classifying all lines per CSV file is 95.6%, versus a maximum of 86.9% for compared approaches. Pytheas generalizes well to new data, with a table discovery Fmeasure above 95% even when trained on Canadian data and applied to data from different countries. Finally, we introduce a confidence measure for table discovery and demonstrate its value for efficiently building a labeled dataset.

REFERENCES

Christodoulakis, C., Munson, E. B., Gabel, M., Brown, A. D., & Miller, R. J. (2020). Pytheas: pattern-based table discovery in CSV files. Proceedings of the VLDB Endowment, 13(12), 2075-2089.

U OF T SYSTEMS AND NETWORKS GROUP (SYSNET)

Computer Science
UNIVERSITY OF TORONTO