Bibliometric-Enhanced arXiv

A Data Set for Paper-Based and Citation-Based Tasks

Tarek Saier and Michael Färber

UNI FREIBURG

Background

- Evaluation and real life applicability of paper based and citation based approaches heavily depend on the data set used.
- Existing data sets lack in
  - Size
  - Cleanliness
  - Cross-domain coverage
  - Global identifiers of cited documents
  - Data set interlinkage

Data Sources

- **arXiv.org**
  - Operates since 1991
  - 1.5 M papers
  - Monstly Physics, Mathematics, Computer Science
  - LaTeX sources of papers are available
- **Microsoft Academic Graph** [1]
  - Automatically generated
  - Metadata on 213 million publications
  - Data on papers, authors, journals, venues, ...

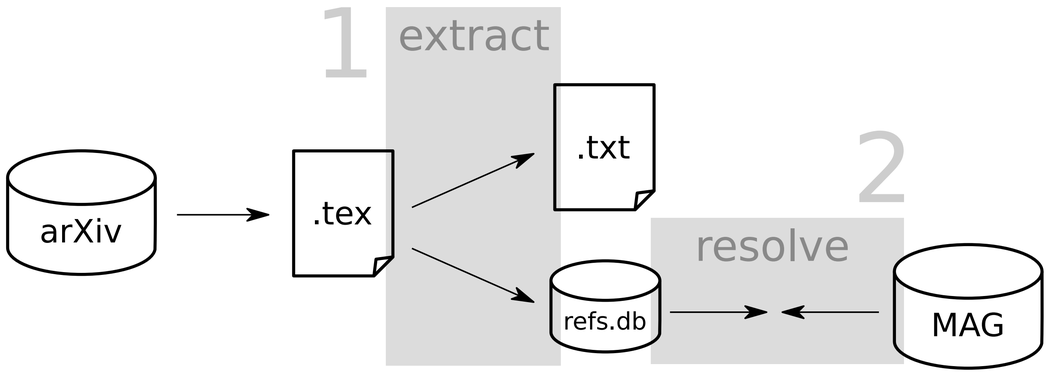
Creation Process

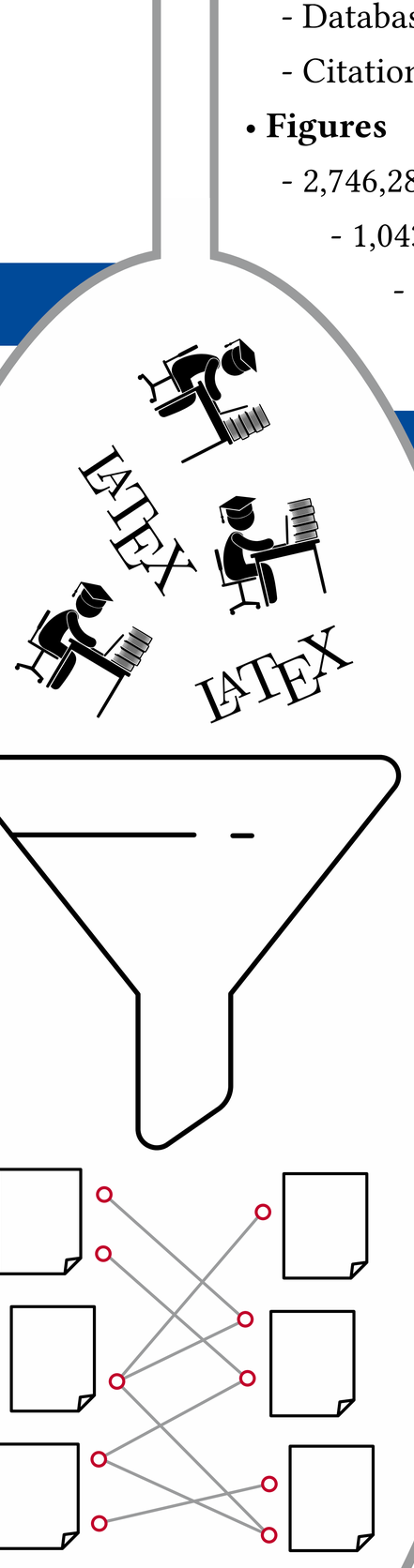
We use all arXiv data until December 2018 and build our data set in two steps:

- **Extraction**
  - Flatten source using *latexpand*
  - Convert to XML using *Tralics*
  - Save text and references, annotate citations within the text

  (success rate: 93% [all] 96% [2018]) (67 hours)
- **Matching**
  - Match reference strings to MAG paper records
  - Identify title (by arXiv ID, DOI, Neural ParsCit [2])
  - Match with MAG (by title→author name→#citations)

  (success rate: 43% [all] 59% [2018]) (10 × 119 hours)
- **Evaluation**
  - We check a random sample of 300 matched reference strings
  - Obtain 3 errors





Result

- **Format**
  - Citing papers' full plain text with citation annotations
  - Database with mapping from reference strings to MAG IDs
  - Citation context export CSV (+export script)
- **Figures**
  - 2,746,288 cited papers
    - 1,043,126 citing papers
      - 15,954,664 references
        - 29,203,190 contexts

DATA doi.org/10.5281/zenodo.2553523

CODE github.com/IllDepence/unarXive

Excerpt

1412.3684.txt

"It has over 79 million images stored at the resolution of FORMULA . Each image is labeled with one of the 75,062 non-abstract nouns in English, as listed in the Wordnet{{cite: 9ad20b7d-87d1-47f5-aeed-10a1cf89a2e2}}{{cite: 298db7f5-9ebb-4e98-9ecf-0bdda28a42cb}} lexical database."

refs.db

uuid	*in_doc*	*mag_id*	*reference_string*
9ad...	1412.3684	2081580037	George A. Mill...
298...	1412.3684	2038721957	Christiane Fell...

MAG

paperid	*originaltitle*	*publisher*
2038721957	WordNet : an electronic lexical database	MIT Press
2081580037	WordNet: a lexical data-base for English	ACM

Conclusion

- New data set of comparably large size
- Clean and accurate data by using LaTeX sources instead of PDFs
- Spanning multiple research domains enabling comparative analysis
- Flexible data format that is applicable to paper based and citation based tasks

References

[1] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. P. Hsu, and K. Wang, "An Overview of Microsoft Academic Service (MAS) and Applications," in Proceedings of the 24th International Conference on World Wide Web, WWW '15 Companion, (New York, NY,USA), pp. 243–246, ACM, 2015.

[2] A. Prasad, M. Kaur, and M.-Y. Kan, "Neural ParsCit: A Deep Learning Based ReferenceString Parser," International Journal on Digital Libraries, vol. 19, pp. 323–337, 2018.

Department of Computer Science, University of Freiburg, Germany | tarek.saier@uranus.uni-freiburg.de | michael.faerber@cs.uni-freiburg.de