A CiteSeer$^{\mathrm{X}}$-based dataset for automatic document metadata extraction

Zalán Bodó

Faculty of Mathematics and Computer Science, Babeş–Bolyai University

zbodo@cs.ubbcluj.ro

Metadata extraction [3] constitutes an important problem for search engines, digital libraries and scientific paper management systems like CiteSeer$^{\mathrm{X}}$[1], Mendeley[2], ResearchGate[3], Google Scholar[4], etc. It is usually considered to be a supervised learning task [4, 5], for which a large amount of labeled training data is needed.

In [1] we described a hybrid metadata extraction system that combines clustering and classification without the need of a conventional labeled dataset. Our initial CiteSeer$^{\mathrm{X}}$-based dataset was made up of 4217 metadata records, assembled automatically without applying any type of data cleaning. In this work we experiment with different record matching approaches [2] in order to clean the metadata and hence improve upon the performance of such extraction systems.

References

[1] Z. Bodó and L. Csató. A hybrid approach for scholarly information extraction. *Studia Universitatis Babeş–Bolyai Informatica*, 62(2):5–16, 2017.

[2] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. *IEEE Transactions on knowledge and data engineering*, 19(1):1–16, 2007.

[3] M. Granitzer, M. Hristakeva, R. Knight, K. Jack, and R. Kern. A comparison of layout based bibliographic metadata extraction techniques. In *WIMS*, pages 19:1–19:8. ACM, 2012.

[4] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic document metadata extraction using support vector machines. In *JCDL*, pages 37–48. IEEE Computer Society, 2003.

[5] F. Peng and A. McCallum. Accurate information extraction from research papers using conditional random fields. In *HLT-NAACL*, pages 329–336, 2004.

[1] citeseerx.ist.psu.edu

[2] www.mendeley.com

[3] www.researchgate.net

[4] scholar.google.com