A CiteSeer$^X$-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$^X$, 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$^X$-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


