A Comparison of Metadata Extraction Techniques for Crowdsourced Bibliographic Metadata Management

Michael Granitzer
Knowledge Management Institute
Know-Center GmbH
Graz, Austria
mgranitzer@tugraz.at

Maya Hristakeva
Mendeley Ltd.
London, UK
maya.hristakeva@mendeley.com

Robert Knight
Mendeley Ltd.
London, UK
robert.knight@mendeley.com

Kris Jack
Mendeley Ltd.
London, UK
kris.jack@mendeley.com

ABSTRACT

Social research networks such as Mendeley and CiteULike offer various services for collaboratively managing bibliographic metadata and uploading textual artifacts. One core problem thereby is the extraction of bibliographic metadata from the textual artifacts. Our work investiages the use of Conditional Random Fields and Support Vector Machines, implemented in two state-of-the-art real-world systems, namely ParsCit and the Mendeley Desktop, for automatically extracting bibliographic metadata. We compare the systems' accuracy on two newly created real-world data sets gathered from Mendeley and Linked-Open-Data repositories. Our analysis shows that two-stage SVMs provide reasonable performance in solving the challenge of metadata extraction from user-provided textual artifacts.

Categories and Subject Descriptors

H.3.1 [**H.3.1 Content Analysis and Indexing**]: [Metadata Extraction]; I.2.7 [**Natural Language Processing**]: Text Analysis

Keywords

Metadata Extraction, Research Papers, Evaluation

1. INTRODUCTION

With the advancement of social research networks like Mendeley[1] and social bookmarking tools like CiteULike[2], metadata management is becoming more and more decentralized. Decentralized metadata management requires intelligent tools in order to reach a high metadata quality for creating a consistent bibliographic catalog out of user provided artifacts like PDF documents.

Support Vector Machines (SVM) [2, 3] and Conditional Random Fields (CRF) [1] have been successfully used in practice for extracting bibliographic metadata by large scale systems such as CiteSeer and Mendeley. However, evaluation conducted in previous work focuses on rather small data sets (see [4]) in very specific research fields. A detailed comparision of the available state-of-the-art systems on large, real world data available is yet missing.

Therefore, our work compares the accuracy of two bibliographic metadata extraction systems, namely ParsCit [1] and the Mendeley Desktop [3], on two noisy real-world data sets. As a result of our analysis, we consider the problem of metadata extraction for crowdsourced bibliographic metadata management as solved by a two-stage SVM approach combined with well engineered post-processing heuristics. We aim to extract authors and titles as metadata fields, since those are mostly sufficient to be used in lookup services for obtaining the full metadata record.

2. METADATA EXTRACTION SYSTEMS

ParsCit is one of today's metadata extraction forerunners and is based on CRFs that are tailored to the computer science domain. It employs the CRF++ implementation[4]. ParsCit already contains trained models and uses token identity, orthographic case, punctuation, numbers, locations and several dictionaries as features (see [1]). To have a fair comparison with previous experiments, we used ParsCit as it is and did not do any re-training on our data set. Since ParsCit does not have its own PDF to text converter, we used PDF Box[5] to generate corresponding test examples.

[1] `http://www.mendeley.com`
[2] `http://www.citeulike.org/`

[3] http://www.mendeley.com/download-mendeley-desktop/
[4] `http://crfpp.sourceforge.net/`
[5] `http://pdfbox.apache.org/`

SAC 2012 Mar 26-30 2012, Riva (Trento), Italy

The metadata extraction algorithm used by Mendeley Desktop relies on a two-stage SVM method as outlined in [2]. It treats metadata extraction from header text as a multi-class classification problem using SVMs. The idea is to first classify each line of the header text into title, author or other (e.g. multi-author) classes using text and formatting features. The line classification is then improved by using contextual information such as the predicted class labels of the neighboring lines assigned in the previous step. Finally, multi-author lines are segmented into a list of individual author names. This is done using a simple recursive descent parser which assumes that the line conforms to a simple punctuation-based grammar. In addition, the algorithm feature set is based on [2, 3] and uses: (i) character-level features; (ii) dictionary/word-list features (e.g. academic titles); (iii) layout features; (iv) independent line features (e.g. number of words on line); and (v) contextual line features (e.g. font size relative to previous and next line). The algorithm is implemented using the libsvm library [6] to train the SVM classifiers with an RBF kernel. In contrast to ParsCit, the Mendeley Desktop uses PDFNet [7] to extract text from imported pdf documents.

3. DATA SETS

Previous work on metadata extraction from academic research papers focused on the Computer Science domain and used rather small data sets with no layout information [2, 4]. In order to take advantage of layout information and to consider noise resulting from PDF to text conversion, we created two real-world test data sets: the e-prints; and the Mendeley data sets. Both reflect real-world data obtained from existing archives and social research networks. Fuzzy string matching based on the Levensthein distance has been used to generate the ground truth. To keep the data set as clean as possible, we excluded artifacts with ambiguous annotations (e.g. overlapping fields) or incomplete metadata.

Our first data set, the *e-prints Data Set*, has been created by crawling the e-prints RDF Repository provided by the RKB-Explorer project[8] and downloading all available pdfs. From this data set, we took all journals and conferences that have more than 10 assigned PDFs resulting in 2,452 PDFs plus metadata. For our experiments, the data set has been narrowed down to three groups of presumably similar publication styles by using regular expression patterns on the journal names. Three groups have been created: Physical Reviews (215 publications); the British Medical Journal (138 publications); and all publications belonging to IEEE (344 publications). The data set, including preprocessed, layout and sequence annotated data as well as an detailed analysis of the results is available for downloaded[9]. Manual inspection of the data revealed particularly challenging aspects like (i) multiple articles contained in one PDF; (ii) front matters from institutional repositories making metadata occurrences more frequent but less consistent in style; and (iii) mismatching Metadata due to incorrect or abbreviated metadata fields (e.g. abbreviated title and forenames).

[6] http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[7] http://www.pdftron.com/pdfnet/
[8] http://eprints.rkbexplorer.com/
[9] http://team-project.tugraz.at/2011/10/17/metadata-extraction-e-prints-data-set/

Experiment	$Prec_a$	Rec_a	$Prec_t$	Rec_t
ParsCit				
BMJ	0.30	0.21	0.29	0.21
IEEE	0.60	0.48	0.82	0.82
Physical Reviews	0.87	0.48	0.84	0.81
Mendeley				
BMJ	0.27	0.20	0.29	0.26
IEEE	0.53	0.27	0.73	0.67
Physical Reviews	0.78	0.41	0.81	0.54

Table 1: Results on the e-prints Data Set.

The second dataset, the Mendeley Data Set, consists of 20,672 publications sampled from the 20 million pdfs available in the Mendeley research network. The sample was chosen to reflect different sources and research fields, namely Nature (1,399 publications, 7 journals) as general science literature, ACM (266 publications, 11 journals) and IEEE (1,481 publications, 48 journals) as computer science literature, BMC (7,572 publications, 9 journals) as bio-medical literature, Physical Reviews (8,815 publications, 5 journals) as Physics literature as well as arXiv (1,006 publications, 1 journal) as an Open Access publisher with mixed literature. In total the data set consists of 81 different journals and proceedings with each journal having more than 20 artifacts. Manual inspection of the training data showed that journals included in the groups for Physical Reviews and BMC are most consistent in style, while Nature, ACM and IEEE are least consistent in style. Over all groups multi-column and single column journals are mixed as well as pre-prints, together with published versions. Compared to the e-prints data set the Mendeley data set contains less noise.

4. RESULTS

In evaluating the results, we estimated precision ($Prec$) and recall (Rec) for author (subscript a) and title (subscript t) fields. An extracted metadata is considered as correctly identified if it has a Levenshtein similarity higher than 0.7 to the original metadata. We chose 0.7 since PDF extraction errors in titles hinder string equality (e.g. titles including chemical compounds). We used 3-fold cross-validation on the Mendeley data set to re-train the Mendeley Desktop in order to compare the pre-given models with trained models (runs are depicted as $Mendeley_{trained}$). The original Mendeley Desktop comes with a model trained on around 1,000 publications from very different domains including physics and biology; the ParsCit model has been trained on the Computer Science domain [1]. The obtained precision/recall figures are shown in tables 1 and 2.

Results on the e-prints data set show that title extraction seems to be less challenging than author extraction. However, the pre-trained models from Mendeley and ParsCit performed particularly poorly on the medical domain. Hence, without re-training neither systems are able to adapt to different domains (i.e. research fields). We also conducted experiments using layout features, which have been omitted due to space reasons. Layout features play a role, but cannot outperform a combination of layout and semantic features with well-known heuristics. Overall, extraction performance is acceptable for titles in computer science and physics papers, but not so strong for author recall in all groups.

Experiment/Group	$Prec_a$	Rec_a	$Prec_t$	Rec_t
Parscit	**0.77**	**0.50**	**0.75**	**0.69**
ACM	0.77	0.53	0.78	0.78
arXiv	0.89	0.58	0.90	0.61
BMC	0.74	0.32	0.26	0.22
IEEE	0.75	0.55	0.87	0.81
Nature	0.86	0.33	0.45	0.40
Physical Reviews	0.89	0.42	0.88	0.53
Mendeley	**0.79**	**0.54**	**0.86**	**0.81**
ACM	0.86	0.61	0.90	0.85
arXiv	0.83	0.50	0.91	0.68
BMC	0.94	0.86	0.76	0.74
IEEE	0.73	0.47	0.86	0.81
Nature	0.91	0.60	0.91	0.84
Physical Reviews	0.84	0.43	0.97	0.90
$Mendeley_{trained}$	**0.81**	**0.62**	**0.94**	**0.91**
ACM	0.84	0.69	0.88	0.84
arXiv	0.84	0.63	0.94	0.92
BMC	0.94	0.91	0.96	0.95
IEEE	0.75	0.54	0.94	0.91
Nature	0.91	0.67	0.94	0.92
Physical Reviews	0.85	0.59	0.99	0.98
all groups	0.81	0.61	0.93	0.91

Table 2: Extraction results on the Mendeley Data Set for the different systems. Bold rows show the average performance of the system, while the other rows show the performance on a particular groups.

Similar to the e-prints corpus, the title extraction problem can be solved with higher accuracy than author extraction on the Mendeley data set. While title extraction can be considered as fairly good, this is not true for author extraction. Recall figures are especially low. Training Mendeley improves author extraction recall and reduces recall-variance among journals, but has negligible effects on author extraction precision. The post-processing heuristic may be responsible for this result. Both ParsCit and Mendeley apply post-processing which has a high impact on the author-extraction precision.

Re-training Mendeley's two-stage SVM shows an impressive precision and recall improvement for title extraction and a good recall improvement for author extraction. This is especially true for BMC and arXiv. With an average precision of 0.94 and an average recall of 0.91 Mendeley provides satisfactory accuracy. Also, training Mendeley on all journals independent of the group to which they belong does not lead to poorer performance (see table 2 last line "all groups"). Hence, the SVMs provide enough model complexity to scale in terms of data set complexity. Also variances among journal groups are lower than for ParsCit for title extraction and equal for author extraction. Hence, Mendeley's two stage SVM is reliable across groups. One reason for the more reliable SVM results may lie in the bad splitting of sequences extracted from PDFs. Since every line break marks the beginning of a new sequence, sequences belonging to the same metadata field but ranging over several lines are broken apart. This happens especially for longer titles or formats with one-column titles like Nature. While the SVM context model can recover from such bad splits, CRFs cannot. CRFs do not consider labeling information from previous sequences and hence may more easily fail in finding bad splits.

In regards to title extraction on different domains (e.g. different journals), performance varies considerably across domains. ParsCit's domain focus on Computer Science becomes clear. Precision and recall are very high for ACM and IEEE, but low for all other groups (see table 2). Although Mendeley's standard model has been trained on only 1,000 PDFs from varying research fields, it shows surprisingly good performance and stability across groups. A performance that can be improved by training Mendeley on this data set. Nevertheless, comparing ParsCit to Mendeley seems to support the hypothesis in favor of generating more heterogeneous than homogeneous data sets for bootstrapping metadata extraction.

5. CONCLUSION

In our work, we compared two different systems for extracting bibliographic metadata from real-world PDF artifacts. Together with strong de-duplication techniques, we consider the problem of metadata extraction for crowdsourced bibliographic metadata management as solved by the two-stage SVM approach combined with well-engineered heuristics.

Acknowledgement

This work has been funded by the European Commission as part of the TEAM IAPP project (grant no. 251514) within the FP7 People Programme (Marie Curie) and by the EUROSTARS project 4811 MAKIN'IT. The Know-Center is funded within the Austrian COMET Program under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency (FFG).

6. REFERENCES

[1] *ParsCit: An open-source CRF Reference String Parsing Package*. European Language Resources Association, 2008.

[2] 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'03*, pages 37–48, 2003.

[3] H. Han, E. Manavoglu, H. Zha, K. Tsioutsiouliklis, C. L. Giles, and X. Zhang. Rule-based word clustering for document metadata extraction. In *Proceedings of the 2005 ACM symposium on Applied computing - SAC '05*, page 1049, New York, New York, USA, 2005. ACM Press.

[4] K. Seymore, A. McCallum, and R. Rosenfeld. Learning hidden Markov model structure for information extraction. In *Proceedings of AAAI 99 Workshop on Machine Learning for Information Extraction*, pages 37–42, 1999.