CERMINE — automatic extraction of metadata and references from scientific literature

Dominika Tkaczyk, Pawel Szostek, Piotr Jan Dendek, Mateusz Fedoryszak and Lukasz Bolikowski

Interdisciplinary Centre for Mathematical and Computational Modelling
University of Warsaw

11th IAPR International Workshop on Document Analysis Systems
7-10 April 2014
The goal

CERMINE — automatic extraction of metadata and references from scientific literature

Dominika Tkaczyk, Paweł Szostek, Piotr Jan Dendeck, Mateusz Fedoryszak and Łukasz Bolikowski
Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, ul. Presta 69, 00-838 Warszawa, Poland
Email: {d.tkaczyk, p.dendeck, m.fedoryszak, l.bolikowski}@icm.edu.pl, pawel.szostek@gmail.com

Abstract — CERMINE is a comprehensive open source system for extracting metadata and parsed bibliographic references from scientific articles in born-digital form. The system consists of a modular workflow, whose architecture and modules, including training and evaluation, enables effortless modifications and replacements of individual components and simplifies further architecture expanding. The implementations of most steps are based on supervised and unsupervised machine-learning techniques, which simplifies the process of adjusting the system to new document layouts. The paper describes the overall workflow architecture, provides details about individual implementations and reports evaluation methodology and results. CERMINE service is available at http://cermine.econ.pl

Keywords — document analysis, metadata extraction, bibliographic references extraction, PDF processing, zone classification

I. Introduction

The amount of literature stored in digital libraries nowadays is huge and constantly growing. A fully functional, modern digital library system in order to provide high quality services

The first version of the system was presented in [1]. Since then we introduced the following improvements:

Workflow architecture was reorganized. The new version contains two parallel paths: metadata extraction path and parsed references extraction path.

• New reading order resolving step was added. In this step we compute the order in which the elements of the document should be read.

• The implementations of many workflow steps were improved or replaced, including zone classification, references extraction and parsing.

• We introduced new classification models based on documents from PubMed [2].

• We performed the evaluation of key workflow steps and the whole metadata extraction path using a large dataset composed of documents from PubMed [2].

CERMINE web service, as well as the source code, can be now accessed online at http://cermine.econ.pl.
The goal

• performing the evaluation of the whole references extraction path using the PubMed-based dataset,
• the evaluation of other similar systems using the same dataset and comparing the extraction results.

ACKNOWLEDGMENTS

The work is supported by the National Centre for Research and Development (NCBiR) under Grant No. SP/I/1/77065/10 by the Strategic scientific research and experimental development program: “Interdisciplinary System for Interactive Scientific and Scientific-Technical Information”.

REFERENCES


The motivation

- There are documents without metadata.
- Metadata information may be incomplete or incorrect.
Requirements

The **metadata extraction system** should be:

- comprehensive,
- automatic,
- modular,
- open and widely available,
- easily applicable,
- flexible and able to **adapt to new layouts**, and
- well tested.
The process

PDF

Basic structure extraction

Metadata extraction

References extraction

<XML>
<title>Syste...
<author>M.K...
<author>J.I...
<journal>J...
<date>2009...
</XML>

<XML>
<ref>
<author>M.K.
<title> Sys..
<journal> J...
... 
</ref>
<ref> ...
</XML>

<JATS>
<front>
<meta><title>
</front>
<back>
<ref>1. <aut
<ref>2. <aut
</back>

D.Tkaczyk et al. (ICM UW) CERMINE DAS 7-10 April 2014 6 / 21
The process

Basic structure extraction

PDF
BT
/F13 10 Tf
250 720 Td
(PDF) Tj
ET

Metadata extraction

References extraction

<XML>
<title>Syste...
<author>M.K...
<author>J.I...
<journal>J...
<date>2009...
</XML>

<JATS>
<front>
<meta><title
</front>
<back>
<ref>1. <aut
<ref>2. <aut
</back>

Basic structure extraction

D.Tkaczyk et al. (ICM UW)
CERMINE
DAS 7-10 April 2014
Basic structure extraction

- Character extraction — **iText** library
- Page segmentation — **Docstrum**
- Reading order resolving — bottom-up **heuristic-based**
- Initial zone classification — **SVM** (*metadata, references, body and other*)
The output

**TrueViz** XML format:

- **hierarchical structure** containing: pages, zones, lines, words, characters
- all elements have **bounding boxes**
- **reading order** is given
- zones have **labels**

```xml
<Page>
  <PageID Value="0"/>
  <Zone>
    <ZoneID Value="0"/>
    <ZoneCorners>
      <Vertex x="55.320"y="34.295"/>
      <Vertex x="235.704"y="58.295"/>
    </ZoneCorners>
    <ZoneNext Value="1"/>
    <Category Value="TITLE"/>
  </Zone>
  <Line>
    <Word>
      <Character>
```

D.Tkaczyk et al. (ICM UW) CERMINE DAS 7-10 April 2014 9 / 21
The process

PDF

Basic structure extraction

Metadata extraction

References extraction

<JATS>

<front>
  <meta><title>
</meta>
  <back>
    <ref>1. <aut>
    <ref>2. <aut>
  </back>
</front>

<XML>
  <title>Syste...
  <author>M.K...
  <author>J.I...
  <journal>J...
  <date>2009...
</XML>

<XML>
  <ref>
    <author>M.K.
    <title>Sys..
    <journal>J...
    ...
  </ref>
  <ref>...
</XML>
Metadata extraction

- Metadata zone classification — **SVM** (abstract, bib_info, type, title, affiliation, author, keywords, correspondence, dates and editor)
- Metadata extraction — simple **rule-based**
Zone classification

- classifiers are based on LibSVM library
- a zone is represented by 78 features: geometrical, lexical, sequential, formatting, heuristics
- the best SVM parameters were found by:
  - a grid-search over 3-dimensional space of kernel function types and C (penalty parameter) and γ coefficients
  - at every grid point a 10-fold cross-validation was performed
  - we chose the parameters that gave the best mean accuracy
- initial classifier was trained on 964 documents with 155,144 zones in total
- metadata classifier was trained on 1,934 documents and 45,035 metadata zones in total
The process

PDF

Basic structure extraction

Metadata extraction

References extraction

<XML>
  <title>Syste...
  <author>M.K...
  <author>J.I...
  <journal>J...
  <date>2009...
</XML>

<JATS>
  <front>
    <meta><title
</front>
  <back>
    <ref>1. <aut
    <ref>2. <aut
</back>
  <XML>
    <ref>
      <author>M.K.
      <title>Sys..
      <journal>J...
      ...
    </ref>
    <ref>
      ...
    </ref>
</XML>

D.Tkaczyk et al. (ICM UW) CERMINE DAS 7-10 April 2014 13 / 21
Parsed reference extraction

- Reference strings extraction — **K-means clustering**
- Reference parsing — **CRF**
Reference strings extraction

clustering text lines into two sets: first lines and the rest

unsupervised **K-means algorithm** with **Euclidean distance**

**5 features** (based on length, indentation, space between lines and the text)
Reference parsing


- **Conditional Random Fields** token classifier based on **GRMM** and **MALLET** packages
- **42 constant features** + the most popular **words** + features of **two preceding** and **two following** tokens
- the classifier was trained on **1000 citations** from **Cora-ref + PubMed**
GROTOAP2 dataset

- **GROund Truth for Open Access Publications**
- built automatically from **PubMed Central Open Access Subset**
- ~ **60k ground truth files** in TrueViz format with corresponding PDF files
## Results

<table>
<thead>
<tr>
<th></th>
<th>avg. precision</th>
<th>avg. recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>initial zone classifier</td>
<td>91.74%</td>
<td>87.31%</td>
</tr>
<tr>
<td>metadata zone classifier</td>
<td>92.49%</td>
<td>93.83%</td>
</tr>
<tr>
<td>reference parsing</td>
<td>90.18%</td>
<td>89.51%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
<th>avg. adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td>journal title</td>
<td>68.68%</td>
<td>49.23%</td>
<td></td>
</tr>
<tr>
<td>volume</td>
<td>97.57%</td>
<td>78.57%</td>
<td></td>
</tr>
<tr>
<td>issue</td>
<td>52.50%</td>
<td>56.64%</td>
<td></td>
</tr>
<tr>
<td>pages</td>
<td>51.37%</td>
<td>34.71%</td>
<td></td>
</tr>
<tr>
<td>year</td>
<td>98.79%</td>
<td>89.18%</td>
<td></td>
</tr>
<tr>
<td>DOI</td>
<td>93.60%</td>
<td>57.46%</td>
<td></td>
</tr>
<tr>
<td>ISSN</td>
<td>44.29%</td>
<td>3.01%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>avg. precision</td>
<td>avg. recall</td>
<td></td>
</tr>
<tr>
<td>authors</td>
<td>87.19%</td>
<td>82.07%</td>
<td></td>
</tr>
<tr>
<td>affiliations</td>
<td>70.13%</td>
<td>59.44%</td>
<td></td>
</tr>
<tr>
<td>keywords</td>
<td>61.11%</td>
<td>68.37%</td>
<td></td>
</tr>
</tbody>
</table>
Future work

- a new extraction path for extracting structured full text
- the evaluation of the entire references extraction path
- comparing the results to other similar systems
**CERMINE web service**: http://cermine.ceon.pl

**CERMINE source code**: https://github.com/CeON/CERMINE

**GROTOAP2**: http://cermine.ceon.pl/grotoap2/
Thank you!

Questions?

Dominika Tkaczyk
d.tkaczyk@icm.edu.pl