Common Map of Academia: augmenting bibliography research information data

Jakub Jurkiewicz, Piotr Wendykier, Krzysztof Wojciechowski, Mateusz Fedoryszak, Piotr Jan Dendek

ICM, University of Warsaw

Rome, 13th of May 2014
Short introduction: Need for easy application of scientific methods

- Really nice methods for author identification, documents metadata deduplication
- A lots of possible data sources
- Very hard joining data, methods and practical implementation
- Need for COMAC - something that joins data, algorithms and results to public
Short introduction: Need for easy application of scientific methods

- Really nice methods for author identification, documents metadata deduplication
- A lots of possible data sources
- Very hard joining data, methods and practical implementation
- Need for COMAC - something that joins data, algorithms and results to public
Short introduction: Need for easy application of scientific methods

- Really nice methods for author identification, documents metadata deduplication
- A lots of possible data sources
- Very hard joining data, methods and practical implementation
- Need for COMAC - something that joins data, algorithms and results to public
Short introduction: Need for easy application of scientific methods

- Really nice methods for author identification, documents metadata deduplication
- A lots of possible data sources
- Very hard joining data, methods and practical implementation
- Need for COMAC - something that joins data, algorithms and results to public
Short introduction: Need for easy application of scientific methods

- Really nice methods for author identification, documents metadata deduplication
- A lots of possible data sources
- Very hard joining data, methods and practical implementation
- Need for COMAC - something that joins data, algorithms and results to public
Table of contents

1 Motivation

2 COMAC - common map of academia

3 Workflows in COMAC
Table of contents

1 Motivation

2 COMAC - common map of academia

3 Workflows in COMAC
1. Motivation

2. COMAC - common map of academia

3. Workflows in COMAC
Need of bibliographic data

- Bibliographic data we need them for:
  - citations search
  - evaluation of citations
  - cooperation analysis

- Supplying data manually to CRIS system is
  - time consuming
  - erroneous
Need of bibliographic data

- Bibliographic data we need them for:
  - citations search
  - evaluation of citations
  - cooperation analysis

- Supplying data manually to CRIS system is
  - time consuming
  - erroneous
Need of bibliographic data

- Bibliographic data we need them for:
  - citations search
  - evaluation of citations
  - cooperation analysis
- Supplying data manually to CRIS system is
  - time consuming
  - erroneous
Multiply bibliographic datasources

- OAI-PMH - and open archives - 18M documents
- Common Crawl - 5M scholar documents after filtering
- PubMed Central - 600 k documents
- Open and closed bibliographical databases - PubMed, publishers databases - depending on licence
Methods for automatic bibliographic data processing

- methods described in scientific articles
- libraries - but hard to use
- ...
- COMAC
What is COANSYS

- COANSYS - COntent ANalizys System - library containing
  - author identification
  - article metadata deduplication
  - citation matching

- Only library with API
What is COANSYS

- COANSYS - COntent ANalizys System - library containing
  - author identification
  - article metadata deduplication
  - citation matching
- Only library with API
• Map - system for creating a graph of articles and authors - and in future institutions
• COANSYS - applied to data - so with input and output
• Workflows
• Data
  • Common Crawl
  • OAIPMH
  • PMC
  • other sources
Map - system for creating a graph of articles and authors - and in future institutions

COANSYS - applied to data - so with input and output

Workflows

Data

- Common Crawl
- OAIPMH
- PMC
- other sources
Output

- to interested systems
- extremely simplified, test interface to processed data
- processed data in form of RDF triples
Efficiency and quality

Fast Hadoop cluster.

- Data conversion for 30M of documents takes around 2 hours.
- Indexing of 30 M documents in SOLR takes about 4 hours

Only test results

- citations matching - 749 K documents, 2.8M citations
  - Time: 4 hours
  - 140 documents with manually recognized citations - 1200 citations in total, 130 resolved inside set, 71 % precision , 69 % recall
- Document deduplication
  - 30 M of documents - 3 hours
  - found 5M duplicates
- Author identification
  - only test results: 100 thousands of pairs of contributors
  - error of 20.54% on 3-fold cross-validation
Efficiency and quality

Fast Hadoop cluster.

- Data conversion for 30M of documents takes around 2 hours.
- Indexing of 30 M documents in SOLR takes about 4 hours

Only test results

- citations matching - 749 K documents, 2.8M citations
  - Time: 4 hours
  - 140 documents with manually recognized citations - 1200 citations in total, 130 resolved inside set, 71 % precision, 69 % recall
- Document deduplication
  - 30 M of documents - 3 hours
  - found 5M duplicates
- Author identification
  - only test results: 100 thousands of pairs of contributors
  - error of 20.54% on 3-fold cross-validation
Efficiency and quality

Fast Hadoop cluster.
- Data conversion for 30M of documents takes around 2 hours.
- Indexing of 30 M documents in SOLR takes about 4 hours

Only test results
- Citations matching - 749 K documents, 2.8M citations
  - Time: 4 hours
  - 140 documents with manually recognized citations - 1200 citations in total, 130 resolved inside set, 71% precision, 69% recall
- Document deduplication
  - 30 M of documents - 3 hours
  - Found 5M duplicates
- Author identification
  - Only test results: 100 thousands of pairs of contributors
  - Error of 20.54% on 3-fold cross-validation
Efficiency and quality

Fast Hadoop cluster.
- Data conversion for 30M of documents takes around 2 hours.
- Indexing of 30 M documents in SOLR takes about 4 hours

Only test results
- citations matching - 749 K documents, 2.8M citations
  - Time: 4 hours
  - 140 documents with manually recognized citations - 1200 citations in total, 130 resolved inside set, 71 % precision, 69 % recall

Document deduplication
- 30 M of documents - 3 hours
- found 5M duplicates

Author identification
- only test results: 100 thousands of pairs of contributors
- error of 20.54% on 3-fold cross-validation
Efficiency and quality

Fast Hadoop cluster.
- Data conversion for 30M of documents takes around 2 hours.
- Indexing of 30 M documents in SOLR takes about 4 hours

Only test results
- citations matching - 749 K docuemnts, 2.8M citations
  - Time: 4 hours
  - 140 documents with manually recognized citations -1200 citations in total, 130 resolved inside set, 71 % precision , 69 % recall

Document deduplication
- 30 M of documents - 3 hours
- found 5M duplicates

Author identification
- only test results: 100 thousands of pairs of contributors
  - error of 20.54% on 3-fold cross-validation
Efficiency and quality

Fast Hadoop cluster.
- Data conversion for 30M of documents takes around 2 hours.
- Indexing of 30 M documents in SOLR takes about 4 hours

Only test results
- citations matching - 749 K documents, 2.8M citations
  - Time: 4 hours
  - 140 documents with manually recognized citations -1200 citations in total, 130 resolved inside set, 71 % precision , 69 % recall

Document deduplication
- 30 M of documents - 3 hours
- found 5M duplicates

Author identification
- only test results: 100 thousands of pairs of contributors
- error of 20.54% on 3-fold cross-validation
Remarks and Conclusion

- Our software proved acceptable efficiency
- Our software has good quality of test set
- We have applied system to real life CRIS!!!
- Re-computation is easy so it's easy to improve system after improving system parts
Remarks and Conclusion

- Our software proved acceptable efficiency
- Our software has good quality of test set
- We have applied system to real life CRIS!!!
- Re-computation is easy so it's easy to improve system after improving system parts
Future work

- creating nice web user interface to the system itself
- publishing and improving (more predicates) in RDF data
- improving data mining methods and introducing new algorithms
Questions?  http://comac.ceon.pl