Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2014) 000–000

Procedia
Computer Science

www.elsevier.com/locate/procedia

CRIS 2014

Common Map of Academia: augmenting bibliography research information data

Jakub Jurkiewicz[a*], Piotr Wendykier[a], Krzysztof Wojciechowski[a], Mateusz Fedoryszak[a], Piotr Jan Dendek[a]

[a]*Center of Open Science(CeON), Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, ul. Prosta 69 piętro 2, 00-838 Warszawa, Poland*

Abstract

Common Map of Academia (COMAC) is a system that keeps information about scientists and publications. It uses bibliographical information as a source of knowledge. In the article we show briefly this system. We present how we used that system for parts of Polish national CRIS: Polish national bibliography and newly created PolIndex – polish citation index. We present algorithms used for enhancing CRIS data: document deduplication, citations matching and authors identification. Finally we show performance statistics.

Keywords:, CRIS data enrichment, data mining, data deduplication, citation matching, author identification

1. Introduction

Bibliography is used to create key indicators for science evaluation. It also gives information about current research status. It is therefore very important input for CRIS. Obtaining complete and clean bibliographical data is very hard. In many scientific institutions, scientists have to manually fill bibliographic records about their works. It takes time and does not solve the problem of finding literature citations. Additionally, scientists often make mistakes or write incomplete data into CRIS. This problem could be solved by adding bibliographic data from existing bibliographic databases. The problem is lack of identifiers that would allow easy matching of author with articles. Additionally we need to resolve citations between data originating from different databases. Usually all automatic tools that could be used for this purpose are very hard to connect with existing systems and data sources.

[*] Corresponding author. Tel.: +48 22 8749425; fax: +48 22 8749401.
E-mail address: J.Jurkiewicz@icm.edu.pl

The team at CeON[1] created a system called the Common Map of Academia (COMAC)[2]. The system aims to collect and organize publicly accessible bibliographic metadata and research information and release the results under an open license. It also allows to process additional data on demand separately or together with open data available in system. This feature allows easy integration with existing CRIS.

To create the map, COMAC utilizes COntent ANalysis SYStem (CoAnSys)[3] – an Apache Hadoop[4]– based framework for mining scientific publications. CoAnSys is built from modules designed for different kinds of data augmenting. This article describes modules for author disambiguation, article deduplication and citation matching. For combining different modules into a single workflow we use Apache Oozie, while the input and output data are stored in the HBase[5]. A processing format used by CoAnSys is based on Protocol Buffers[6].

Common Map of Academia contains solely publicly available metadata coming from different sources. We use web pages harvested via CommonCrawl[7] and available on Amazon S3. To reduce pages that are not describing scholar publication we use datamining methods [1]. From CommonCrawl we have found nearly 5M scholar publications. We also used PubMed Central[8] open data, and we are ready to import Medline[9] data. From PubMed Central we have successfully imported around 620k of publications. We harvest publicly available repositories using the OAI-PMH protocol. OAI stays for Open Archive Initiative. Protocol OAI-PMH they created is widely used. It allows easy harvesting content of the most of open repositories. From open repositories we have harvested around 18M of documents. In our system we use parts of Polish Virtual Library[10]. COMAC also works on non open set of data for augmenting them i.e. enhance data provided by CRIS. Additionally CRIS could provide data to be source of metadata, but this data could be available under special agreements with publishers.

2. Usage scenarios

ICM is involved in POL-on[11]. This is a project for creation of Polish national CRIS. In the project ICM role is to take care of bibliographical parts. We are creating Polish National Bibliography (PBN)[12] and Polish Citation Index (PolIndex). PBN is a system tracking all publication written by polish scientists. We have used COMAC for augmenting data from this two systems.

Demanding and scenarios of using COMAC for the both systems were different. PBN demands two functionalities: correcting and enhancing of metadata coming from institutional CRIS and finding hints for scientists when they input data to the system. Therefore PBN usage scenario includes document deduplication, author identification and metadata enhancement. COMAC searched took metadata of scholar papers and searched through available documents for this metadata from another source. It were made by document deduplication. It allows enhancement of metadata. This functionality combined with author identification allows nice hints for adding papers to system by scientists. For PolIndex COMAC were simply used for citation matching in PolIndex provided data.

[1] http://www.ceon.pl
[2] http://comac.ceon.pl
[3] http://coansys.ceon.pl/
[4] https://hadoop.apache.org/
[5] https://hbase.apache.org/
[4] http://code.google.com/p/protobuf
[7] http://commoncrawl.org/
[8] http://www.ncbi.nlm.nih.gov/pmc/
[9] http://www.nlm.nih.gov/pubs/factsheets/medline.html
[10] http://bibliotekanauki.ceon.pl/
[11] http://polon.nauka.gov.pl/
[12] https://pbn.nauka.gov.pl/

2.1. Usage scenario for PBN

For PBN we use COMAC to improve data quality and ease data input. We used data provided by PBN, together with all other datasets. There are two usage scenarios: enhancement of data provided by journals and institutions (batch processing), and on-line prompting user while adding their articles. In both cases we use for deduplication, data already downloaded to COMAC system, as well as data provided by PBN. The workflow for PBN is presented on Fig. 1. COMAC is responsible for importing data into HBase (from different sources), converting them to text mining tools useable form, then running CoAnSys modules on this data, and finally indexing data into SOLR[13] index. COMAC uses Apache Oozie[14] for managing whole workflow.

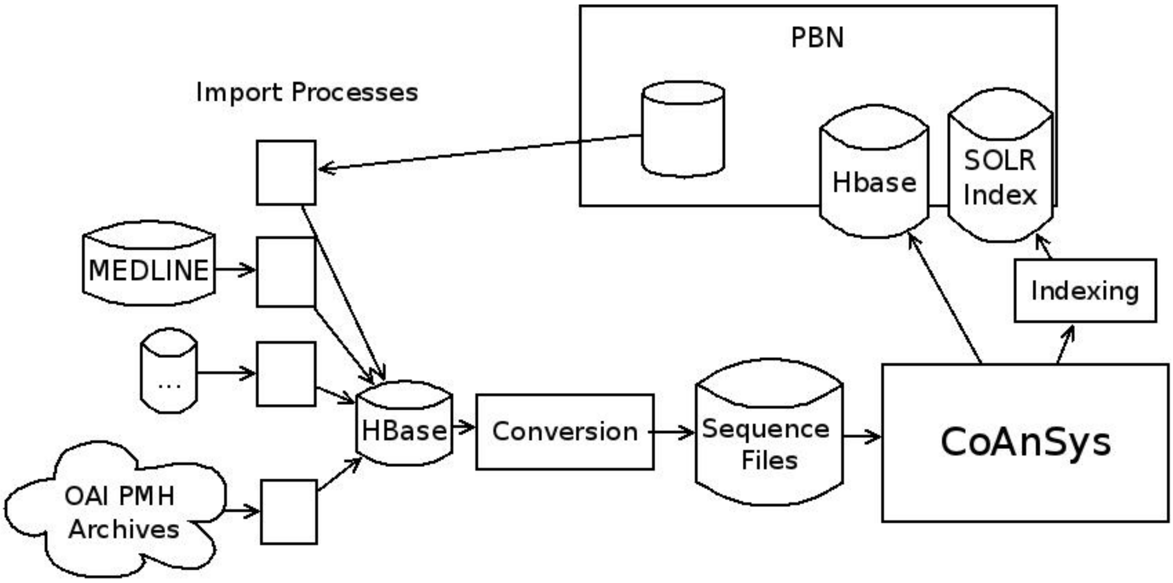

Fig. 1. Processing workflow for PBN

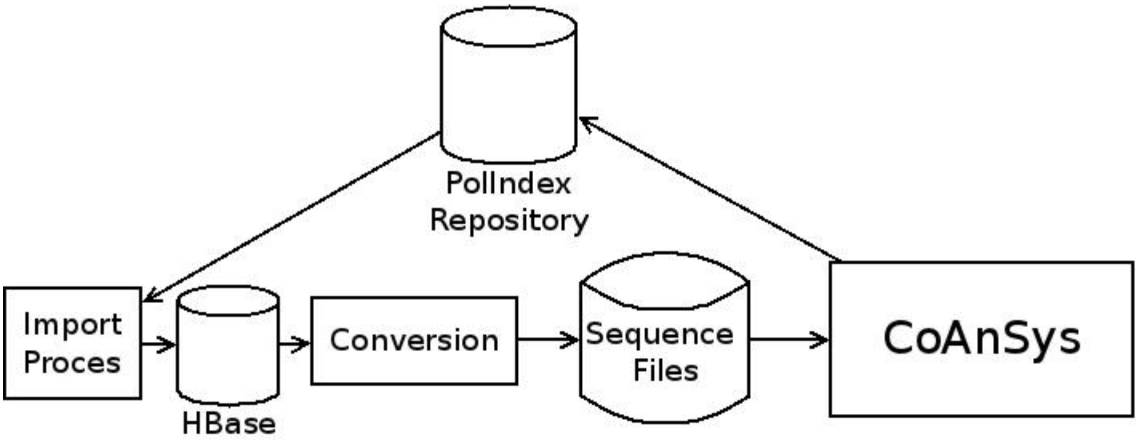

Fig. 2. Processing workflow for PolIndex

[13] https://lucene.apache.org/solr/
[14] https://oozie.apache.org/

2.2. Usage scenario for PolIndex

For PolIndex COMAC worked on closed set of articles provided by PolIndex. They were all provided by PolIndex. Workflow of COMAC for PolIndex is presented on Figure 1. CoAnSys is a part in which actual text mining is done. For PolIndex we actually do only citations matching. In this role COMAC were used as an input output system for CoAnSys. It contains importing data to the HBase storage and conversion to a text mining tool understandable format. Indexing is skipped for this solution. COMAC provides management for whole process (using Apache Oozie). However role of COMAC could seem limited, it is very important to have such system. There are numerous data recognition libraries, but without providing IO tools and workflow management, they are useless.

3. Workflow description

The usage of our system could be divided into four phases.

At the first phase the data are harvested to our system in an original format. This phase is made by import processes (Fig. 1, Fig. 2). During this phase a unique ID is assigned to each object. The data in the original format is stored as an attachment to the document. This phase is independent from others because it can be different for each data source. Also, the time needed to finish this step depends on the speed of a data provider. At the end of this phase all records are stored in the HBase.

In the second phase (Conversion on Fig. 1 and Fig. 2), bibliographical metadata are translated from different input formats to the internal format of CoAnSys. That format is based on Protocols Buffers which provides very efficient serialization and deserialization necessary for map-reduce jobs. Metadata records in this format are stored in Hadoop sequence files. Current implementation allows for translating metadata stored in Dublin Core, BWMETA and NLM format.

The augmenting of the data is performed in the third phase (CoAnSys Fig.1 Fig.2). It covers document deduplication, citation matching and author disambiguation.

3.1. Document deduplication

Document deduplication executes a single map-reduce job to create sets of duplicates. In the map step, the whole set of metadata records is divided into clusters using the prefix of documents' titles as a blocking variable. In the reduce step, the sequence of comparisons is executed on each pair of documents within each cluster. These comparisons include computation of edit distance on authors' names, titles, and the names of a journal/conference. Other metrics are used to compare DOIs, years and issues. At the output of reduce step we obtain small clusters that should contain only duplicates. The representative of each cluster is chosen in the last step of the process. If any metadata information is missing for that record, it is copied from the other documents in the cluster.

3.2. Citation matching

Citation matching consists of two main parts. In the first part, heuristic search is used to find candidate documents that match a given citation sting. It is assumed that the first few tokens of citation string contains author names. Other pieces of metadata, that are easy to locate, are publication year and page numbers. Document store is queried for records with at least one matching author, year and page numbers varying by at most 1. Those are returned as candidates. If no candidates are found, the algorithm resorts to weaker heuristic that does not take into account pages information.

In the second part, pair wise similarities between a citation and candidate document metadata are computed and the candidate with highest score is selected. To obtain this similarity, various similarity metrics such as common tokens fraction, common q-grams fraction, edit distance and longest common character subsequence are defined on metadata extracted from bibliographic entries using CRF-based parser provided by CERMINE [2]. The overall result is obtained by combining them by means of linear SVM. This part is presented in details in [3] while the description of the first part is outdated.

3.3 Author disambiguation

Author disambiguation is conceptually build of 4 steps: extracting contributions data from each document, grouping contributions with the same value of a given hash function, calculating affinity between each pair of contributions in the same group and finally executing clustering of each group of contributors. An unique ID is assigned to objects in each resulting cluster. The affinity within a pair is calculated against a previously computed model, created with Linear Support Vector Machines on our training set. LSVM assigns weights to chosen features and finds the minimum value of similarity [4, 5].

- In the last phase, the augmented data is exported. They are inserted to the HBase and indexed in Apache Solr – a search server that provides an efficient querying. In addition, we provide export to RDF format, which will allow to perform scientometric analysis on that data.

We have created simple web interface for fast evaluation of our results, but real strength of our system is shown in cooperation with POLINDEX and PBN. In POLINDEX our system resolves citations which is crucial for their goal. In PBN our system helps person that inputs data, with recommendation based on documents found in other data sources.

4. COMAC efficiency and quality

The big advantage of our system is efficiency and relatively short time for response. All phases, except harvesting, took usually less then one day on our cluster. Our results are also very promising. Our testbed was Hjadoop cluster containing 4 nodes and master node. Each worker node has four AMD Opteron 6174 processors (48 cores in total), 192 GB of RAM, four 600 GB disks connected in RAID 5 array with an access to 7TB LUN of NetApp disk storage over Fiber Channel. The master has 8-core CPU, 32 GB of RAM and 64 GB storage. For the most phases we have tested set of 30M of documents (open and provided by PolIndex and PBN).

Data conversion phase for 30M of documents takes around 2 hours.

Citation matching modules has been tested on subset of data containing almost 749k documents (over 2.8M citations). For about 140 of them manually curated citation links information was available. They contained over 1200 citations of which about 130 were resolvable (i.e. metadata records of referenced documents were present in the set). To ensure that test is close to real situation we run citation matching process on the whole set of documents, but calculated metrics only on the subset. Results showed 71% precision and 69% recall. The whole matching process took less than 4 hours.

Document deduplication run on full set of 30 millions of documents from different sources found 5 millions of duplicates. Document deduplication takes around 3 hours on full set of 30 millions of documents.

Author disambiguation algorithm has been combined into our system recently and would be tested in the next few weeks. The linear SVM model has been trained on sample of 100 thousands of pairs of contributors, where decision parameter has been the answer to question if a pair of contributors represent the same person. 5 features were used, which are number of intersecting: co-author surnames, classification codes, ISSN, keywords, elements in bag of keywords. In vitro tests give error of 20.54% on 3-fold cross-validation [6].

In final phase indexing of 30 millions documents in SOLR takes about 4 hours.

Our system showed reasonable quality and efficiency. As expected the most of time is spent on actual work – running CoAnSys and not on conversion and indexing. COMAC proves that we can apply CoAnSys library in CRIS system. It is now used for PBN and PolIndex, and waits for feedback from users.

5. Conclusion

In COMAC we successfully joined scientifically created bibliographical analysis library with large set of data and created a useful system. The system were applied for two independent parts of real-life national CRIS. This bridges the gap between scientific development and real life systems. Our system showed high efficiency and ease to use. It provides acceptable quality. However we still need to improve this quality. Our system is currently working with scientific articles. We need to add other forms of scientific communication – books, datasets and so on…

COMAC currently provides only basic, test, web user interface for browsing output. Additionally we are ready to expose data in RDF form – as an RDF triple files. In future we have to improve our web interface and expose our data world wide.

References

1. J. Jurkiewicz and A. Nowiński. Towards Finding Scholarly Articles in Internet Using Hadoop MapReduce with Oozie Workflow In: *Challenges of Modern Technology; 2013; Vol 4; Is. 4; p. 3-6*
2. D. Tkaczyk, Ł. Bolikowski, A. Czeczko, and K. Rusek. A modular metadata extraction system for born-digital articles. In: *10th IAPR International Workshop on Document Analysis Systems*; 2012. p.11–16.
3. M. Fedoryszak, D. Tkaczyk and Ł. Bolikowski. Large Scale Citation Matching Using Apache Hadoop. In: *Research and Advanced Technology for Digital Libraries*. Springer Berlin Heidelberg; 2013; 8092. p. 362-365
4. P. J. Dendek, Ł. Bolikowski and M. Łukasik. Evaluation of Features for Author Name Disambiguation Using Linear Support Vector Machines. In: *Proceedings of the 10th IAPR International Workshop on Document Analysis Systems*; 2012; p. 440–444
5. P. J. Dendek, M. Wojewódzki and Ł. Bolikowski. Author disambiguation in the YADDA2 software platform. In: *Intelligent Tools for Building a Scientific Information Platform*. Springer; 2013; p. 131–143
6. P. J. Dendek, Ł. Bolikowski, M. Łukasik. Evaluation of Features for Author Name Disambiguation Using Linear Support Vector Machines. In: *Proceedings of the 10th IAPR International Workshop on Document Analysis Systems*. 2012; p. 440–444