CitEc to CitEcCyr. A stab at distributed citation systems

Thomas Krichel
RANEPA & Open Library Society

Köln 2017-03-30

acknowledgement

- Work described here has been funded by the Russian Academy for the National Economy and State Service more precisely known as «Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации»
- I have benefited from some email exchange with Min-Yen Kan.

talk structure

- general background.
- a look at Russian references.

- [The longer I talk, the less I know what I am talking about.]

CitEc

- A system to do autonomous citation indexing for documents in the RePEc digital library.
- Founded and operated by José Manuel Barrueco Cruz since 2001.
- It does
  - parsing of references from full-text working papers
  - parsing of references strings or structured reference data provided by publishers
  - citation data for RePEc.
- Public data is exported to other RePEc services.

input sources

- Bibliographic data from RePEc, may contain reference strings
- Some full-text papers that are not freely available for scanning of references.
- Stored back copies of RePEc working paper full-text.
- Importance of full blown autonomous reference parsing is in steady decline.

tools

- MySQL database
- poppler PDF text extractor
- ParsCit
- in-house Perl scripts

sustainability

- Server sponsored by INOMICS GmBH.
- No other external funding.
- Mainly work contributed by José Manuel Barrueco Cruz (JMBC).
- Not likely to follow the like “op cit” and CiteBase.

CitEcCyr

- A funded project that combine the resources of two unfunded project
  - CitEc
  - SocioNet [*next slide*]
- Basic objective is to use CitEc technology stack to build a citation database for Russian publications.
- The name is still confusing.

two-part system

- SocioNet, despite its name, has evolved into a cross-disciplinary aggregator with heavy Russian presence.
- CitEcCyr will handle the Russian papers and papers in Russian from RePEc that CitEc is handling very badly at this time.
- We are at the verge of a decentralized system.

aim for openess

- The aim is for software and data to be reusable.
- Outputs should reproducible.
- We also need some form of coordination between nodes.
- For that we already have some simple protocols.

Fraga

- A simple metadata format that specifies the results of automated citation analysis.
- http://fraga.openlib.org
- Fully implemented for CitEc.
- Documents not processed by one node can be picked up by another node.

Lafka

- A protocol to gather full-text files and store them in WARCS.
- Partly implemented software is in operation on RePEc and some SocioNet contents.
- The CitEc public full-text file will be merged into the RePEc Lafka collection to form a main component of ArchEc ... at same stage.
- http://lafka.openlib.org

DiCit

- A simple replacement to save us the pains of OAI-PMH.
- Relies on XML, RelaxNG and rsync.
- The idea is that rsync is used on files in XML format, the format of which is specified in RelaxNG files.
- A node can be specified in a single XML file.
- Fully implemented for a part of the SocioNet data.
- http://dicit.openlib.org

CitEcCyr

- Other aspects of CitEcCyr include the usage of citation data to feed annotation services for documents.
- JMBC's and my job is to build a Russian citation index.
- For me that means building an index of Russian language-reference strings citing Russian document.
- We may do other languages that use Cyrillic letters and that use a similar citation style.

state of play (maybe)

- Scopus and WoS don't do any indexing of document described in Cyrillic. References to these documents have to be transliterated.
- There is a tendency to translate and/or transliterate the references.
- There is also Российский индекс научного цитирования, but it is not an open project either.

state of work: ParsCit

- CitEc uses ParsCit basically out of the box, it just works.
- This is pretty amazing since it essentially uses a built-in binary crf++ model.
- That is possible because ParsCit was built with computer science in mind. The citation style is similar to the one in economics.
- Written in Perl but understandable.

ParsCit specificity

- I have studied the section parsing part.
- Regular expressions that need changing can't be done without changing the source code. I wrote a library.
- Complicated structure of dictionaries make them hard to extend.
- Parts of the code a literally duplicated rather than placed into libraries.
- Parts of the code deal with Omnipage, and that's something we don't have.

data source

- НЭИКОН have a repository.
- The metadata has about 900k reference strings.
- However, that data also contains references to non-Russian language documents, mostly written in the language of that document.
- I suspect that most cases of transliteration or translation occurs in non-Russian documents written by Russian authors when they refer to a Russian paper.

limitations

- We need to parse references for authors, titles and years.
- If a reference does not have all of these, we discard it.
- Thus we don't look at
  - laws and other government documents
  - patents

rombas

- I exclude reference strings that don't contain a Cyrillic char.
- In order to deal with translation, transliterations and to exclude non-Russian references, I create the “roman bags”, aka rombas.
- I partition the references by the number of Roman chars they contains.

romba stats

0000.txt: 272171
0001.txt: 12676
0002.txt: 10940
0003.txt: 6683
0004.txt: 4085
0005.txt: 2479
0006.txt: 2124
0007.txt: 1632
0008.txt: 1132
0009.txt: 984

example 1

<a>Кон Е. Л., Фрейман В. И., Южаков А. А.
<t>Проблема оценки качества обучения в вузах с системой подготовки «бакалавр-магистр» (на примере технических направлений)</t>
// Открытое образование.
<y>2013.</y> № 1. С. 23–31.

example 2

<a>Дмитриченко, Э. И. <t>Разработка и исследование процесса магнитно-электрического шлифования деталей машин :</t> автореф. дис. ... канд. техн. наук: 05.02.08 / Э. И. Дмитриченко; Моск. автомобилестроит. ин-т. – М., <y>1991.</y> – 17 с.

example 3

<a>Ласкин М.Б., Русаков О.В., Джаксумбаева О.И., Ивакина А.А. <t>Особенности рыночной стоимости на рынке недвижимости при логарифмически нормальном распределении.</t> «Имущественные отношения в Российской Федерации», <y>2016</y> г., в печати

example 4

<a>Марченков Ю.В., Рябчиков М.М., Шульгин М.А.<t>Сравнительная характеристика различных видов послеоперационной анальгезии у больных с онкологическими заболеваниями легких.</t>Общая реаниматология.<y>2011;</y> 7 (3): 32—37.

authors

- The fact that no first names are used makes it easier to track authors.
- I got a started stock of Russian surnames from a web site.
- Then scanned the references for further surnames as co-authors. I have a risk of overfitting.
- I will also use the data from the bibliography.

years

- Easy to spot as a year number at the end of the string.
- ParsCit uses a “location” indicator that goes from 1 to 12.
- I suspect that these are just categories and that crf++ does not actually use numbers.

titles

- Titles appear almost invariably to appear after the list of author.
- There is no distinct punctuation mark.
- Thus the end of the title is difficult to track.

tailicity

- The // or when represent, will give a good indication of the end of the title.
- This allows us to parse all references for term after the //. Look at all references that contain //, look at all the token, and evaluate how often they appear after the
- This is “tailicity” of the token.

thus potential features

- surnames
- initials
- // and /
- tailicity
- location
- year potential

token border

- The approach of ParsCit takes is to tokenize at spaces.
- This will not be able to work when fields are not separated by blanks.
- I have seen this in a few cases.

Спасибо за внимане!

Томас Крихель

http://openlib.org/home/krichel