Citation Content Analysis in the Cirtec project

Sergey Parinov

Russian Presidential Academy of National Economy and Public Administration (RANEPA),

Central Economics and Mathematics Institute of Russian Academy of Sciences

Cirtec project

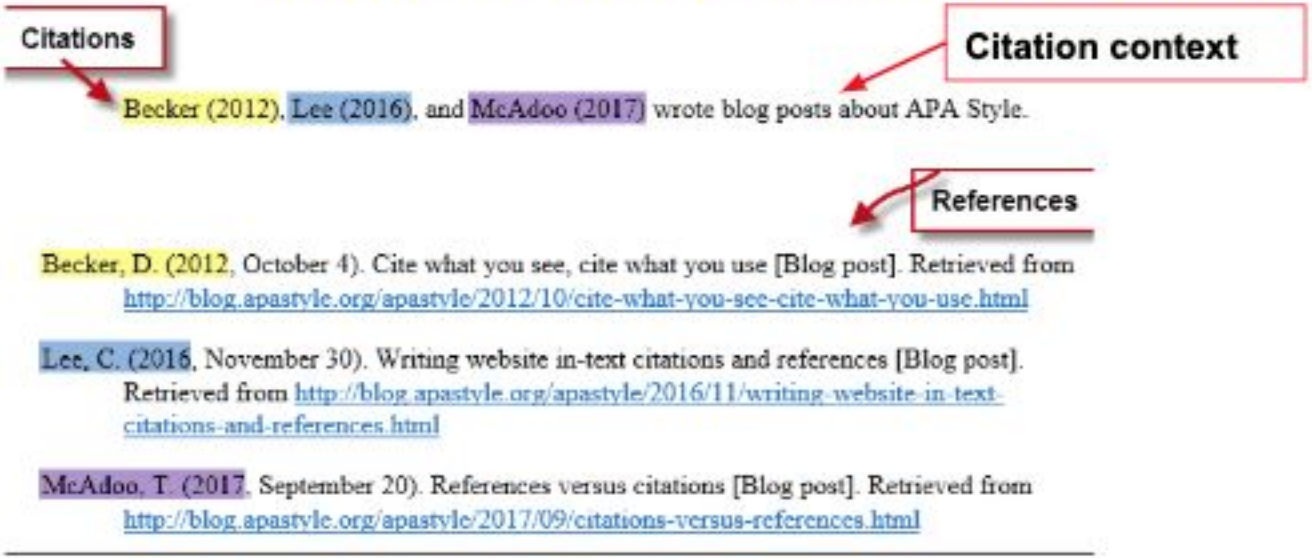
- Cirtec project started in 2017 has two main aims:
  - 1) to create a public service for processing available research papers full text in PDF in order to build and regularly update an open dataset of citation content data;
  - 2) to use the citation content data for developing methods of qualitative citation analysis and new indicators, which can improve visualization of scholarly cooperation
- This project is funded by the Russian Presidential Academy of National Economy and Public Administration (RANEPA, http://www.ranepa.ru/eng/)

The project mission

- The Cirtec project wants to communicate to the research community our lessons learnt and some resulting points for discussion, which can help with the understanding of:
  - who/what/why cites/is cited in research literature and why it's about research cooperation
  - How we can improve research assessment and evaluation
  - how we can improve global scholarly communication, which based on publications and create better scholarly cooperation at large

Citation content data: In-text Citation, References and Citation Contexts

- "Include an **in-text citation** when you refer to, summarize, paraphrase, or quote from another source. For every **in-text citation in** your paper, there must be a corresponding entry in your **reference** list", https://guides.libraries.psu.edu/apaquickguide/intext



https://blog.apastyle.org/apastyle/2017/09/references-versus-citations.html

Variations of the in-text citations

lead to a double dividend, according to Goulder [18]. This can be a strong argument in favour of an increasingly green tax system. After Bovenberg and de Mooij [5] who initially provided a refutation of the double dividend hypothesis, a large body of literature has deeply analyzed this issue. In particular, Goulder [18] and Ligthart [23] showed that the

quantity or quality of a certain resource (Freeman III, 1993; Cummings

regulation variables are defined as in Cole et al., 2005. The econometrics model is a panel with

and visualization techniques for scholarly literature to help a survey of research papers [MRC95] [Sma99]. Federico et al. [FHKM17] reported a large number of visual approaches to scholarly litera-

WH2endogeneityJAN2562010EREMONTREAL.tex; 8/04/2010; 12:38; p.2

For the S&P500 and the DJE these extreme moves exceed two percent. This immediately suggests that macroeconomic news *does* move the markets.[5] The summary statistics confirm the usual rank ordering in

[5] This is consistent with Fair (2002), who finds most large moves in high-frequency S&P500 returns to be readily

3

et al., 1986; Mitchell and Carson, 1989). One of the main concerns

Variations in references

[1] W. Antweiller, B. R. Copeland and M.S. Taylor (2001), Is free Trade Good for the Environment, *American Economic Review*, **91**, pp. 877-908.

1- Acharya, R. C., & Coulombe, S. (2006). Research and Development Composition and Labour Productivity Growth in 16 OECD Countries, Industry Canada's working paper 2006-02, IC 60040

[BMS17] BERGER M., MCDONOUGH K., SEVERSKY L.: cite2vec: Citation-driven document exploration via word embeddings. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2017), 691–700. 2

Godfrey, L. G. 1978. Testing for multiplicative heteroskedasticity. *Journal of Econometrics* 8: 227–236.

—. 1988. *Misspecification tests in econometrics: The Lagrange multiplier principle and other approaches.* Cambridge: Cambridge University Press.

—. 1999. Instrument relevance in multivariate linear models. *Review of Economics & Statistics* 81(3): 550–552.

Variations of the citation context

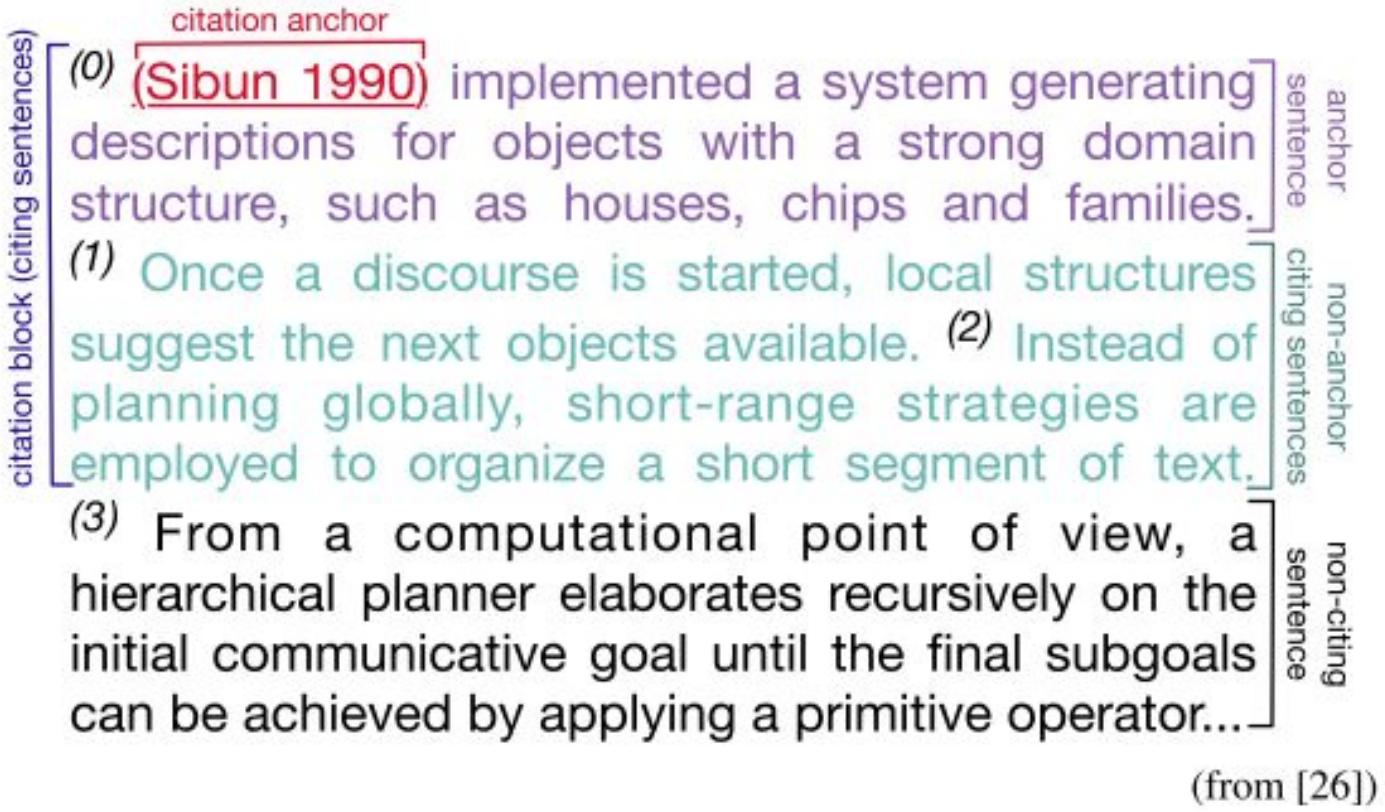


Fig. 1 An example multi-sentence citation block with following non-citing sentence.

Kaplan, D., Tokunaga, T., & Teufel, S. (2016). Citation block determination using textual coherence. Journal of Information Processing, 24(3), 540-553.

Lessons and needed improvements

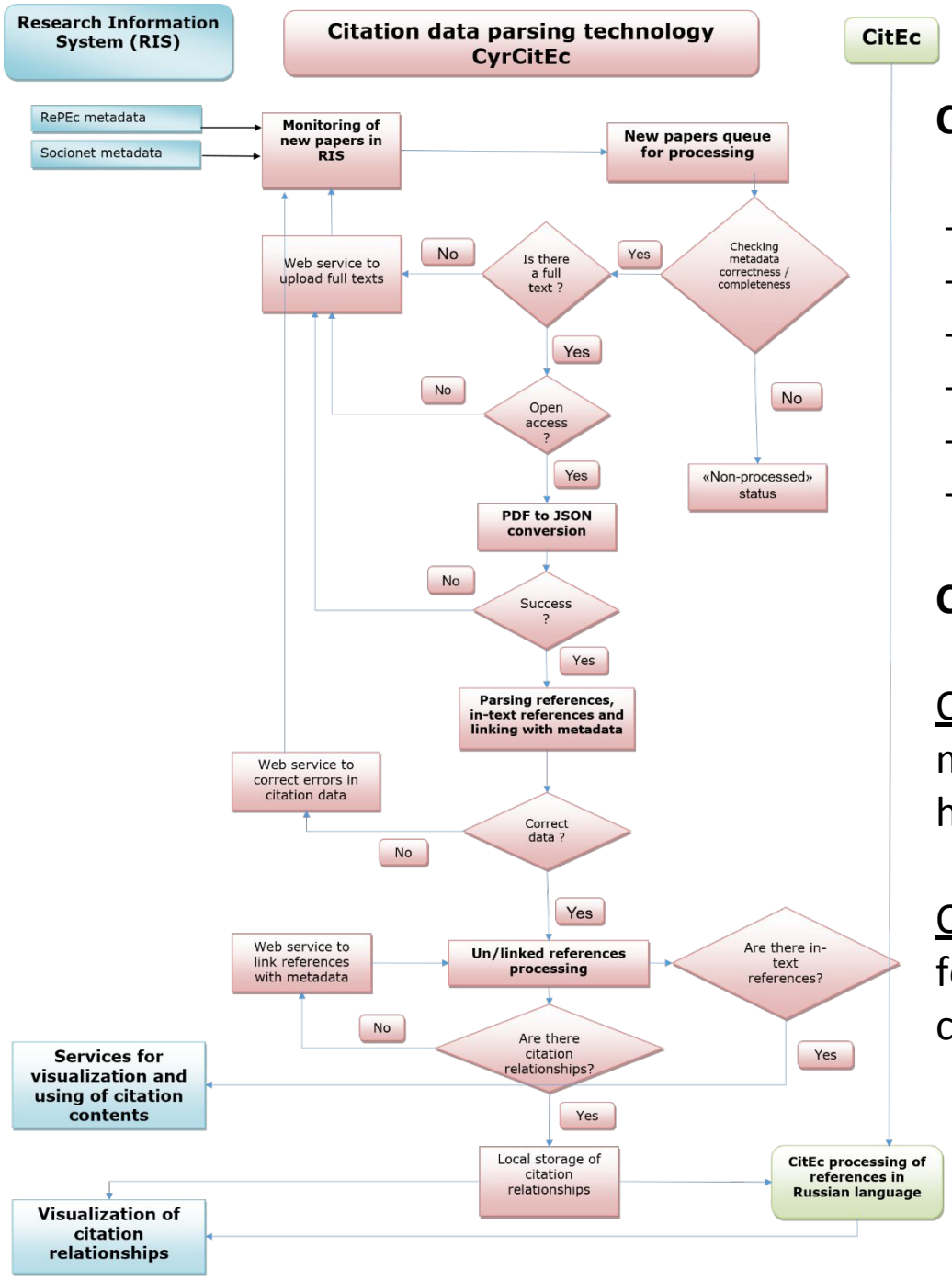
- "… an author's reasons for citing in a particular way at a particular time are controlled by an internalized set of norms" (Cronin 1984)
- Lessons learnt and discussion points:
  - Scholars tend to use different ways of citing. Citations do not always match references.
  - Scholars should use one of standardized ways for citing
- Needed improvements
  - While writing a paper, an author should be able to make digital links between: a) a reference and metadata of the same paper (e.g. by DOI); b) an in-text citation and a reference that means changing a way of making papers

Available citation content data sources

- Some publishers (e.g. PLOS, PubMed, Elsevier, etc.) and aggregators (e.g. Microsoft Academic) provide full text papers as XML files (using the Journal Article Tag Suite (JATS) schema) with marked up text elements, including in-text citations and references
- There are special data sets, e.g. “InTeReC: In-text Reference Corpus - Single References Dataset” [1] provided about 300K sentences
- Our Cirtec project provides Open Citation Content Data [2] with about 2.4 ml citation content records parsed from about 270K papers

[1] https://zenodo.org/record/1203737

[2] Kogalovsky et al. (2018). Open Citation Content Data. In *Research Conference on Metadata and Semantics Research* (pp. 355-364). Springer, Cham.



Cirtec Technology:

- Takes papers from RePEc and Socionet
- Returns citation data to RePEc/Socionet
- Integrated by data with CitEc/RePEc
- Uses PDF.js to convert PDF to JSON
- Stores citation data as XML files
- Provides open access to produced data

Cirtec Outputs:

Open source software to parse papers' metadata and full text PDFs available at https://github.com/citeccyr

Open service to process papers' PDFs for extracting citation data including citation contexts

See more in (Kogalovsky et al. 2018)

Examples of citation content data

and cross-hedging effects. The problem from an econometric perspective is that, without any additional restrictions or modeling assumptions, the contemporaneous coefficients in $\beta_{h0}^{h'}$ are not identified.

To overcome this problem, we follow the approach of a recent series of papers by Rigobon (2003), Rigobon and Sack (2003a,b, 2004), and Sentana and Fiorentini (2001), who use the conditional heteroskedasticity in the high-frequency data to identify the contemporaneous response coefficients. The idea is straightforward. Assuming that the innovations in (4.1) are conditionally uncorrelated but

```
<reference num="35" start="47738" end="47848" author="Rigobon" title="2003 Identification through heteroskedasticity" year="
  <from_pdf>
    Rigobon, R., 2003. Identification through heteroskedasticity. Review of Economics and Statistics 85, 777792.
  </from_pdf>
 </reference>
<reference num="36" start="47850" end="47985" author="Rigobon Sack" title="2003a Measuring the reaction of monetary policy t
market" year="2003a">
  <from_pdf>
    Rigobon, R., Sack, B., 2003a. Measuring the reaction of monetary policy to the stock market. Quarterly Journal of Economi
    639-669.
  </from_pdf>
 </reference>
<reference num="37" start="47987" end="48102" author="Rigobon Sack" title="2003b Spillovers across" year="2003b">
  <from_pdf>
    Rigobon, R., Sack, B., 2003b. Spillovers acros
  </from_pdf>
 </reference>
<reference num="38" start="48103" end="48222" aut
year="2004">
  <from_pdf>
    Rigobon, R., Sack, B., 2004. The impact of mo
  </from_pdf>
```

```
<intextref>
 <Prefix>
    The problem from an econometric perspective is that, without any additional
    restrictions or modeling assumptions, the contemporaneous coefficients in are not
    identified. To overcome this problem, we follow the approach of a recent series of
    papers by
  </Prefix>
 <Suffix>
    Rigobon and Sack (2003a,b, 2004), and Sentana and Fiorentini (2001), who use the
    conditional heteroskedasticity in the high-frequency data to identify the
    contemporaneous response coefficients. The idea is straightforward.
  </Suffix>
  <Start>33571</Start>
  <End>33586</End>
  <Exact>Rigobon (2003)</Exact>
  <Reference>35</Reference>
</intextref>
```

http://cirtec.ranepa.ru/data/RePEc/aah/create/2007-20/summary.xml

What can we measure based on citation content data?

- "… the words and sentences around citations are analyzed to get to know information about characteristics of the cited work, reasons to cite, and decision rules of the citing authors" (Tahamtan and Bornmann 2018)
- ".. citations are measurable indicators, logically linked to interesting theoretical variables (e.g. scientific productivity, communication units or whatever), but the correct functional form of this linkage is unknown." (Porter, 1977)
- We can count, basically:
  - in-text citations at all and for each reference,
  - co-cited authors within common citation context,
  - spatial distribution of in-text citations over paper's sections
- We can analyze:
  - co-occurrence of words, e.g. common phrases
  - appearance of special words, e.g. polarity and function terms, topic models, etc.
  - similarity between groups of words, etc.

Method: Common phrases within citation contexts

- Method: **n**-gram is a contiguous sequence of **n** words from a given sample of text
- It is used to recognize lexical clichés, cue (hint) words or common phrases within citation contexts that would allow an automatic classification of the citation contexts
- It helps to find a "relation of frequent n-gram patterns in citation contexts with the rhetorical structure of scientific articles" (Bertin et al. 2016)
- It is also used for "automatic annotation of citation contexts; to identify sentences that can be potentially annotated with citation acts" (Bertin et al. 2016)

Method: Topic models of citation contexts

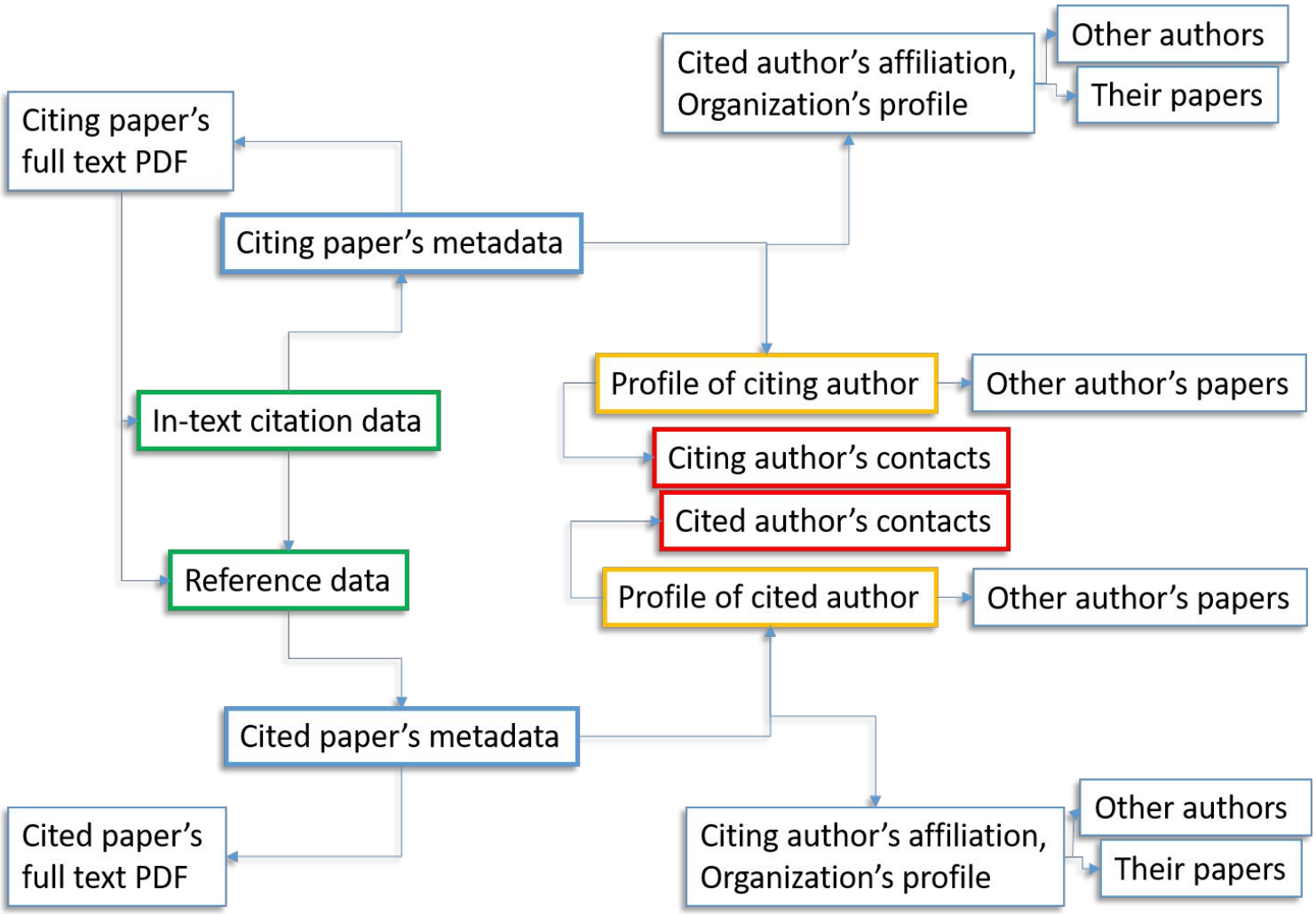
- Method: a technique for finding a collection of words i.e. **topic** from a group of documents that represents the information in the group
- Latent Dirichlet Allocation (LDA) was proposed in 2000. Now it is the most popular topic modelling technique which assumes documents are produced from a mixture of topics and each topic as a mixture of words
- LDA tries to figure out what topics would create those documents in the first place based on a rule that similar words occur in similar contexts
- After a number of iterations, a steady state is achieved where the document topic and topic term distributions are fairly good

Method: Semantic similarity of citation contexts

- Method: Word embedding. It maps words or phrases from a large corpus of text to vectors of real numbers. One of implementations is Word2vec created in 2013. It produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space [1]
- Word2Vec parameters to control:
  - Dimensions of the vector space (about 300). Quality of word embedding increases with higher dimensionality.
  - The size of the context window (about 5). It determines how many words before and after a given word would be included as context words of the given word.

[1] https://en.wikipedia.org/wiki/Word_embedding

CRIS Semantic Layer helps with grouping citation data



Grouping citation data for references, authors, publishers/journals, etc.

- We can group citation data:
  - For references cited by different authors (in-text citations and citation contexts)
    - We can characterize how a cited paper was used
  - For authors, citation data linked with their papers (references, in-text citations and citation contexts)
    - We can characterize citation relationships between author's papers and papers of other authors, e.g. authors' collusions
  - For publishers/journals, citation data linked with their papers (references, in-text citations and citation contexts)
    - It will characterize citation relationships between publisher's papers/authors and others, e.g. publishers' collusions

Grouping citation data for references

Unique references	pubs	cits	distribution	common phrases	topic models
Beck, T., Demirgüç-Kunt, A., & Levine, R (2003) Law and finance why does legal origin matter Journal of Comparative Economics 31 pp	82	166	16/44/70/24/12	estimation technique (54/88)	variable, estimation, legal, technique, instrumental
Mlachila, M., Tapsoba, R., & Tapsoba, S. J. A (2014) A Quality of Growth Index for	36	162	15/39/36/17/55	inclusive development (54/218)	index, et, al, mlachila, growth
Pesaran, M. H., Shin, Y., & Smith, R. J (2001) Bounds testing approaches to the analysis of level relationships	60	159	6/11/63/62/17	critical values (73/259)	cointegration, test, value, variable, fstatistic

Grouping citation data for authors

Three groups of papers for each author:

(1) own papers of a specific author;

(2) papers cited by this author;

(3) citing papers for this author.

Citation contexts based indicators for groups of papers related with an author

Groups of Pubs	pubs	cits	with	distribution	common phrases	topic models
Author's papers	109	2204	2300	668/613/358/256/309	exchange rate (314/988)	firm, uncertainty, model, use, effect
Cited papers	142	4052	4402	1225/944/731/577/575	exchange rate (378/988)	patent, use, asset, firm, model
Citing papers	69	1360	1408	407/404/229/159/161	exchange rate (296/988)	rate, exchange, firm, uncertainty, use

Source page: http://cirtec.ranepa.ru/groups/authors/Christopher-Baum/

Towards a research cooperation visualization

- There are three types of cooperative scholars: “suppliers”, “producers” and “consumers” of research outputs
- Using citation data we can analyze relationships
  - “producer” ⟷ “supplier”
  - “producer” ⟷ “consumer”

Analysis of citation relationships “author–suppliers”

- We make more precise analysis of citation data related with papers of an author “A” and papers/authors cited by “A”:
  - a distribution of cited authors by numbers of their in-text citations, by common phrases and topic models
  - a share of citations produced by “A” in total statistics for cited authors
  - what authors/papers co-cited by ”A”, with what common phrases and topic models, with what frequency
  - a distribution of cited author by location of their in-text citation, common phrases and topic models for subgroups authors cited in the “introduction/discussion” and in the “method/results” sections
  - frequency, locations, common phrases and topic models for self-citations of “A”

Towards an analysis of citation relationships “author – consumers”

- We are going to analyze the citation data related with papers/authors that citing an author “A” and papers of “A”:
  - a distribution of citings of “A” by numbers of in-text citations, by common phrases and topic models
  - a share of citations produced by citing authors in total statistics for the author ”A”
  - with what authors/papers the author ”A” is co-cited, with what common phrases and topic models, with what frequency
  - a distribution of locations of in-text citation for “A”, common phrases and topic models for citings “A” in the “introduction/discussion” and in the “method/results” sections

Lessons and needed improvements

- Using available citation content data we can characterize three types of cooperative scholars “suppliers”, “producers” and “consumers” and relationships between them
- This data allow us to do something useful for huge number of scholars who usually cooperate by writing, reading and citing research papers of each other

Further development

- Now we have a lot of data about researcher's place in scholarly cooperation network
- Why don't use this to improve research assessment and evaluation?
- Evaluating research performance by this way we create more "healthy" motivations for researchers

Use new data to make better research papers

- Because we can:
  - Give authors tools to make digital links between paper's metadata and a reference, between a reference and in-text citations
  - Allow authors a previewing of their new outputs within existing research cooperation network
  - Organize some kind of a completion among suppliers to be cited by an author, etc

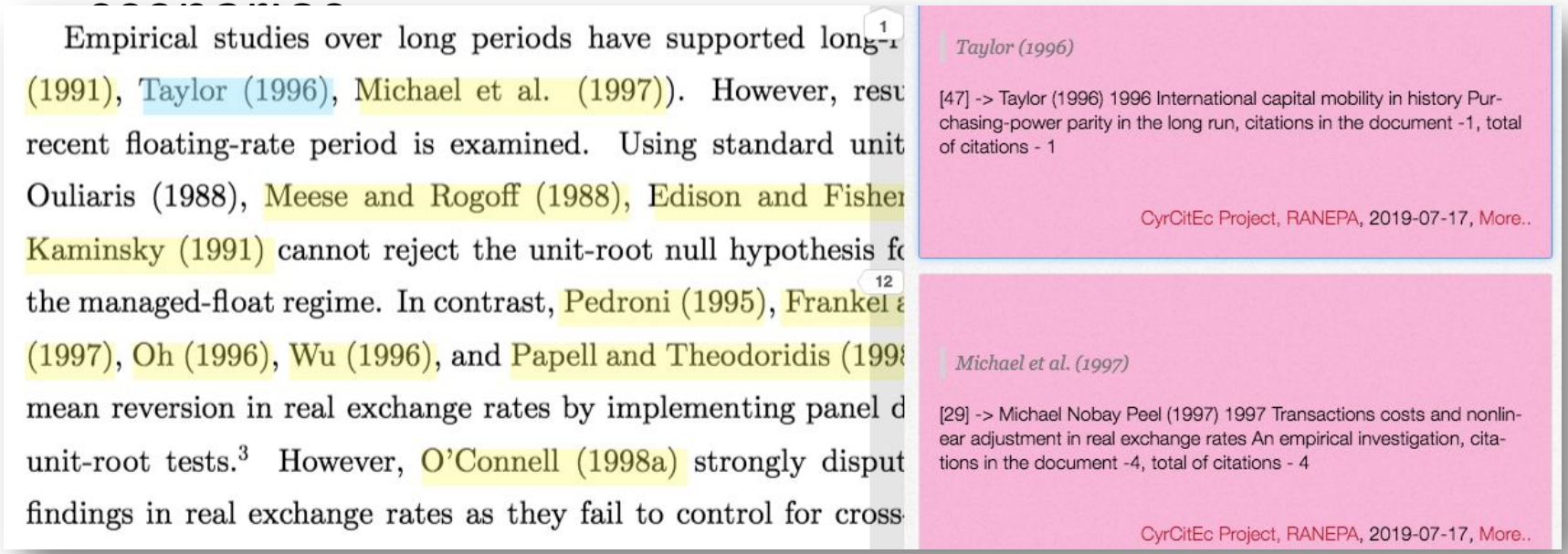
Use cooperation to make better research

- Today, citations are the one-way passive communication between “supplier” and “consumer” of research outputs
- Why not consider these relationships to be interactive? We can think of citations as “p2p” communication channel inside scholars’ collective work [1].

[1] Parinov, S. (2019). CRIS with in-text citations as interactive entities. *Procedia computer science*, *146*, 20-28.

A pilot tool for making in-text citations as interactive elements

- PDF full text has in-text citations as computer-built annotations (we used Hypothes.is software)
- Click on such annotation can run different scenarios



https://socionet.ru/pdfviewer.xml?h=repec:boc:bocoec:404&file=http://fmwww.bc.edu/EC-P/wp404.pdf&l=en

Possible use cases

- If a system recognizes a user as the author of cited or citing paper, it allows him to express publicly or privately his reaction:
  - the author of the cited paper can express his reaction on meaning of the citation
  - the co-author of the cited paper can express the “agree” or “disagree” with a comment made by another co-author;
  - the author of the citing paper can express nothing for the citation itself, but the “agree” or “disagree” for comments to his/her citations.
- According to a citation context, this reaction can provide explanations of the cited author what was wrong with using his outputs, or how it could be used properly, etc.
- A reader can specify only “agree” or “disagree”;
- The system itself also can initiate some direct communications between citing and cited authors using the citation content meaning, such as citation polarity or citation function

Conclusion: there is an ambitious goal behind a scene

- Implementing this approach we can make one more step towards solving an ambitious task:
  - converting a global latent scholarly cooperation network, currently visible only as citations in research papers, into an effective mechanism of global direct scholarly communication between three types of cooperative scholars: “suppliers”, “producers” and “consumers” of research outputs
- It will benefit each researcher, because everybody has potential collaborators among readers of their papers
- CRIS is in a center of this because it is an interface for such interactions

If you share this vision, contact us

Oxana Medvedeva, RANEPA, Cirtec project head,
email - oxana.medvedeva.1984@gmail.com

Sergey Parinov, CEMI RAS, RANEPA, Cirtec project development group leader,
email - sparinov@gmail.com