Layout logical labelling and finding the semantic relationships between citing and cited paper content

Sergey Parinov*

Central Economics and Mathematics Institute of RAS,
Nakhimovsky pr. 47,
Moscow 117418, Russia
and
Russian Presidential Academy of
National Economy and Public Administration,
Prospect Vernadskogo 84/9,
Moscow 119571, Russia
Email: sparinov@gmail.com
*Corresponding author

Amir Bakarov

National Research University Higher School of Economics,
Myasnitskaya Ulitsa, 20,
Moscow 101000, Russia
Email: amirbakarov@gmail.com

Daniil Vodolazsky

Novosibirsk State University,
Ulitsa Pirogova, 1,
Novosibirsk, 630090, Russia
Email: daniil.vodolazsky@mail.ru

Abstract: Currently, large data sets of in-text citations and citation contexts are becoming available for research and developing tools. Using the "topic model" method to analyse these data, one can characterise thematic relationships between citation contexts from citing and the cited paper content. However, to build relevant topic models and to compare them accurately for papers linked by citation relationships we have to know the semantic labels of PDF papers' layout such as section titles, paragraph boundaries, etc. Recent achievements in papers' conversion from a PDF form into a rich attributed JSON format allow us to develop new approaches for the logical labelling of the papers' layout. This paper presents a re-usable method and open source software for the logical labelling of PDF papers, which gave good quality of a layout element's recognition for a set of research papers. Using these semantic labels we made a precise comparison of topic models built for citing and cited papers and we found some level of similarity between them.

Keywords: Cirtec project; in-text citation; citation contexts; research paper layout recognition; logical labelling; hierarchical topic models.

Reference to this paper should be made as follows: Parinov, S., Bakarov, A. and Vodolazsky, D. (2020) 'Layout logical labelling and finding the semantic relationships between citing and cited paper content', *Int. J. Metadata, Semantics and Ontologies*, Vol. 14, No. 1, pp.54–62.

Biographical notes: Sergey Parinov is Chief researcher of the CEMI RAS in Moscow. He has Doctor of Sciences degree (2001, Novosibirsk State University) in Computer Science. Currently he is a leader of the Cirtec software group to create a public source of the citation content/context data and to provide citation content analysis.

Amir Bakarov is first year PhD student at the National Research University Higher School of Economics. He got Master's degree in Linguistics and Bachelor's degree in Computer Science. His research interests are natural language processing, distributional semantics.

Daniil Vodolazsky is first year MSc student in Computer Science at the Novosibirsk State University. He got a BSc degree in Applied Mathematics and Informatics. His research interests include language models, machine translation and distributional semantics.

This paper is a revised and extended version of the paper entitled 'CRIS with in-text citations as interactive entities' presented at the '14th International Conference on Current Research Information Systems', Sweden, June 2018.

1 Introduction

The Cirtec project[1] funded by the Russian Presidential Academy of National Economy and Public Administration (RANEPA[2]) provides for the public use an open data set of in-text citations and citation contexts (Kogalovsky et al., 2019). It takes, as an input, the research papers' metadata from Research Information Systems (RIS) like RePEc (http://repec.org/), Socionet (https://socionet.ru/) and, directly, from paper provider organisations. Cirtec tools process the full text PDF of papers, available by links from their metadata and, as an output, they create citation data which is updated daily (Kogalovsky et al., 2019).

Some publishing systems, like the Public Library of Science (PLOS), provide papers in XML format (Bertin and Atanassova 2014, 2015, 2018). In such format all structural elements (paper layout) are marked up. In our case, papers in RePEc are available only in regular PDF files format, and the first pre-processing task is to recognise their layout, including section titles and paragraphs. Citation contexts provided by Cirtec have unique IDs, which allow researchers to link them accurately with the corresponding fragment of a paper in its hierarchical representation.

Integration of these citation data with RIS allows us to show the in-text citation as interactive elements (Parinov, 2019) in the user interface. Using the citation data, the Socionet RIS: a) produces computer-generated annotations to the content of PDF papers; b) links citation data with different information objects which are available in RIS, such as the profiles and contact data of papers' authors, profiles of affiliation organisations; c) allows RIS users to post reactions to the meaning of in-text citations.

In this way, a new type of interactivity appears in RIS. Users see annotations in PDFs located over the in-text citations. By clicking on the annotations, users have, on a screen, additional information about the related references and links to additional tools. Using the semantic layer of RIS, the additional tools get, for example, the contact data of the cited authors. It allows different ways of direct scholarly communication between the citing and cited authors. Socionet, in particular, allows the cited authors (and other users) some reactions to the meaning of citation contexts (Parinov, 2019).

If Socionet recognises a user as the author of the cited paper and this user clicks the in-text citation pointing at his cited paper, the system allows this user to express, publicly or privately, his reaction to how the citing author used their research output. According to a citation context, this reaction can be as simple as "agree/disagree", or can provide explanations by the cited author of what was wrong with using his outputs, or how it could be used properly. If the cited paper has several co-authors, the system allows them to express their "agree/disagree" with a reaction of one or more of them.

Potentially, RIS itself also can initiate some direct communications between citing and cited authors using the citation content meaning, such as: a) citation polarity; b) citation function; and/or c) topic models which characterise relationships between the citation contexts of the citing and cited papers.

Both the communications initiated by the citing/cited authors, within RIS, or by RIS services, are very promising for the research community, since such communication tools help it to build a modern digital representation for the scholarly cooperation networks, initially based on traditional publications and academic publishing systems (Parinov and Antonova, 2019). The development of such tools is much easier if we know the character of the relationships between the citing and cited papers.

It is well-known that researchers have different motivations to cite papers and, as a result, there are different types of citations in the research literature (Parinov and Antonova, 2019). If our primary focus is scholarly cooperation, the most important type is the "grateful" citation, which is an acknowledgement of the intellectual debt in relation to the cited publication. There are also "rhetorical" and "ritual" citations which are less related to the process of collective creation of scientific knowledge and scholarly cooperation (Parinov and Antonova, 2019).

The difference in these motivations is lost by the use of citation indexes to evaluate research performance. Measuring the success of scholars by the number of citations their publications receive, affects the essence of the citation process, because, in many cases, this indicator becomes the goal of scientific activity (Neylon, 2017).

Taking into account the situation with motivations and, since we use citation contexts for developing direct communication between the citing and cited authors, we should recognise the character of semantic relationships between the citing and cited papers (Swales, 1986; Ding et al., 2014). If we know this, we can organise useful communication for the related authors.

In this paper we use the "topic model" method to characterise the hierarchical relationships between the citation contexts of citing papers and cited papers using RePEc papers and Cirtec citation data. In our approach we consider a research paper as a hierarchical text structure: a citation context is a part of a paragraph of some paper's section, which is a part of a whole paper. We analyse hierarchical topic models of citing and cited papers, built for each of these nested text arrays: citation context -> paragraph -> section –> paper.

To build a hierarchical structure of topic models, we need a recognition of section titles and paragraph boundaries within the papers' content, which means making a semantic logical labelling of the papers' layout. As Tao et al. (2014b) wrote: "… the crucial determination of the semantic roles of the content, also known as logical labelling, remains an open problem". We created an open source tool for the semantic logical labelling of the papers' layout. It allowed us to build relevant topic models and to make their accurate comparison.

Section 2 presents a method to recognise research paper layout, including section titles and paragraph boundaries needed for the building of accurate topic models. In Section 3, we provide a background and present the used data for the topic modelling method. Section 4 presents the result of building the characteristics of the relationships between the citation contexts from the citing and cited papers. In the conclusion, we discuss results of using the topic modelling method for classifying the relationships between the citing and cited papers.

This paper is a revised and extended version of the paper "CRIS with in-text citations as interactive entities" (Parinov, 2019) published in the *Proceedings of the 14th International Conference on Current Research Information Systems*. Comparing with its older version, this paper provides new results on a specific aspect of semantic relationships between the citing and cited papers raised in the previous version.

2 Semantic logical labelling of the research papers' layout

We determine the semantic roles of research papers' content, also known as logical labelling, to specify elements of papers' layout, such as section titles, paragraphs, tables, figures, footnotes, formulas, headers, footers, etc. We need information about the text coordinates of these layout elements in a paper text in order to properly adjust borders of citation contexts, to exclude from the citation context inappropriate text and for more accurate analysis of citing and cited paper content.

Initially, document layout recognition is implemented on printed materials for browsing and retrieving the logical structure of documents (Paaß and Konya, 2011). Typically, it is a part of optical character recognition and used in computer vision to identify and categorise the regions of interest in the scanned image of a text document. The detection and labelling of the different zones (blocks) as text body, illustrations, math symbols, and tables embedded in a document is called a geometric layout analysis. To identify the logical roles of the selected zones inside the document one needs to make the semantic labels. This is called a logical layout analysis[3].

Another case are born-digital papers in PDF with more or less fixed layout. The logical labelling of such PDF strongly depends on the availability and completeness of the data about the PDF attributes characterised its layout. Only recently the open source solutions[4] able to produce such data became publicly available (see e.g. Kogalovsky et al., 2019). It explains why this area still has not plenty of available literature. See, e.g. the literature overview in Tao et al. (2014b) and Rahman and Finin (2017).

Tao et al. (2014b) used geometric, textual, typesetting and visual features for making the semantic labelling of the papers' layout. We use a similar approach based on common features of the research papers' layout: a) regular text lines have the same or close left/right indent, line spacing and font name/size; b) paragraphs have the same left indent for their first lines or bigger line spacing with neighbour paragraphs; c) section titles differ from regular text by font name/size and/or bigger line spacing with neighbour paragraphs.

Some authors propose the Conditional Random Fields model to learn the latent semantics of the PDF page content. Heuristic prior knowledge of PDF content and layout are interpreted to construct neighbourhood graphs and various pairwise clique templates for the modelling of multiple contexts (Tao et al., 2014a, 2014b).

In our case, we are implementing just a chain of logical rules. But it gave us satisfactory results of papers' layout elements recognition and labelling. To test this our approach we selected a set of PDF papers from the RePEc series National Bureau of Economic Research (NBER) Working Papers[5], provided by the NBER.

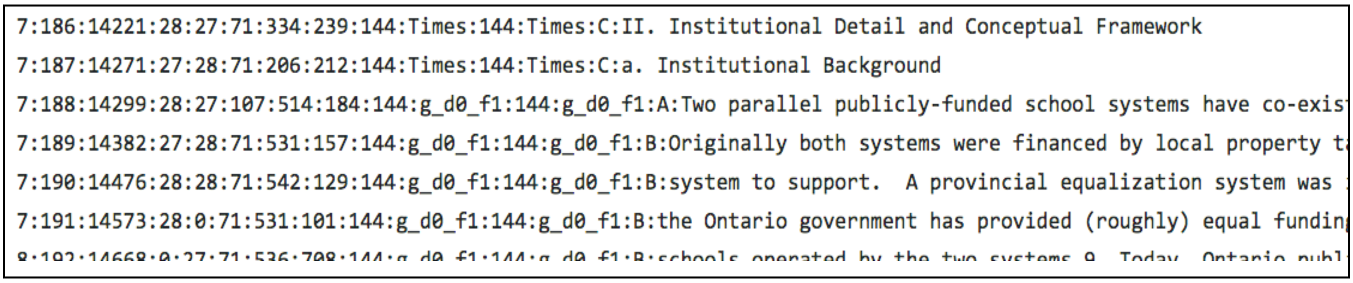
This set of research papers and results of their recognitions are available at our Github repository[6]. Each folder in this repository contains two PDF and two TXT files. The PDF files are: 1) an initial version of a paper (with the label "orig" in the file name); and 2) a version of this PDF with marked-up structural elements. The marked-up version of PDF files are used to visually check the quality of structural element recognition.

The TXT file with the label "orig" in its file name is a result of a two-step pre-processing of the paper PDF: 1) converting the initial PDF into JSON[7]; and 2) counting all the attributes for each text line of the paper[8]. This file is the input for making a recognition. The second TXT file is the recognition output. It differs from the first one by added symbols for each text line, which are the semantic labels of its structural element type.

Currently the recognition of structural elements was restricted by defining the section titles and the text paragraphs for papers with the one-column layout only. The open source of this software available at Github[9]. Tao et al. (2014b) wrote about a solution that recognises 16 semantic logical labels. But they did not provided their tools for open public non-commercial re-use.

A fragment of the output data file[10] used to make a semantic logical labelling of a paper is in Figure 1.

Figure 1 A fragment of data file with recognition results



```
7:186:14221:28:27:71:334:239:144:Times:144:Times:C:II. Institutional Detail and Conceptual Framework
7:187:14271:27:28:71:206:212:144:Times:144:Times:C:a. Institutional Background
7:188:14299:28:27:107:514:184:144:g_d0_f1:144:g_d0_f1:A:Two parallel publicly-funded school systems have co-exis
7:189:14382:27:28:71:531:157:144:g_d0_f1:144:g_d0_f1:B:Originally both systems were financed by local property t
7:190:14476:28:28:71:542:129:144:g_d0_f1:144:g_d0_f1:B:system to support.  A provincial equalization system was
7:191:14573:28:0:71:531:101:144:g_d0_f1:144:g_d0_f1:B:the Ontario government has provided (roughly) equal fundin
8:192:14668:0:27:71:536:708:144:g_d0_f1:144:g_d0_f1:B:schools operated by the two systems.9  Today, Ontario publ
```

The information in each line of this data file (see Figure 1), separate values of which are divided by the symbol ":", contains 14 attributes of each line. These attributes have the following meanings (from left to right):

1 page number of the text line;

2 serial number of this line in a paper;

3 serial number (or text coordinate) of the first symbol of a text line;

4 number of pixels of vertical space to a previous text line (line spacing with upper line);

5 number of pixels to a text line below (line spacing with a line below);

6 horizontal graphic coordinate (in pixels) of the first symbol of a text line (left text line intend);

7 horizontal graphic coordinate (in pixels) of the last symbol of a text line (right text line intend);

8 vertical graphic coordinate (in pixels) of a text line;

9 font size of the first symbol of a text line;

10 font-name of the first symbol of a text line;

11 font size of the last symbol of a text line;

12 font-name of the last symbol of a text line;

13 structural element label of a text line, which is added to this file as a result of recognition procedure described below;

14 text line itself.

There are the following labels for structural element: "A" the first line of a paragraph; "B" regular text line; "C" section title; "?" other.

Using data from the input TXT file, a recognition for a specific paper works by following two steps:

1 Counting of common characteristics: a) the most used font name (regular font name) and font size (regular font size); b) the most used left border and right border; c) the most used (regular) line spacing. Since values in "b" and "c" can have rounding errors, we consider as equal values having differences up to 5%.

2 Labelling of paper's layout element by implementing a chain of logical rules for each line of the input TXT file:

- *font size checks*
  - ○ if the font size is not the same at the beginning and the end of the line, then the label of this line is "other" (the symbol "?" for this line in an output file);
  - ○ otherwise, if the line's font size is bigger than the regular font size, the label of this line is the "section title";
  - ○ otherwise, if the line's font size is less than the regular font size, the label of this line is the "other";
- *line spacing checks*
  - ○ if the line spacing is less than the regular line spacing, then the label is "other";
  - ○ if the line spacing is bigger than the regular and if the font name is not regular or the line starts from a number (e.g. "2" or "II"), then the label is the "section title", etc.

To label paragraphs (the first and other lines of paragraphs) we also make comparing of their left indent with regular one, and use values of the right and left border of the lines. See the whole set of rules in the Python code, lines 178–273[11].

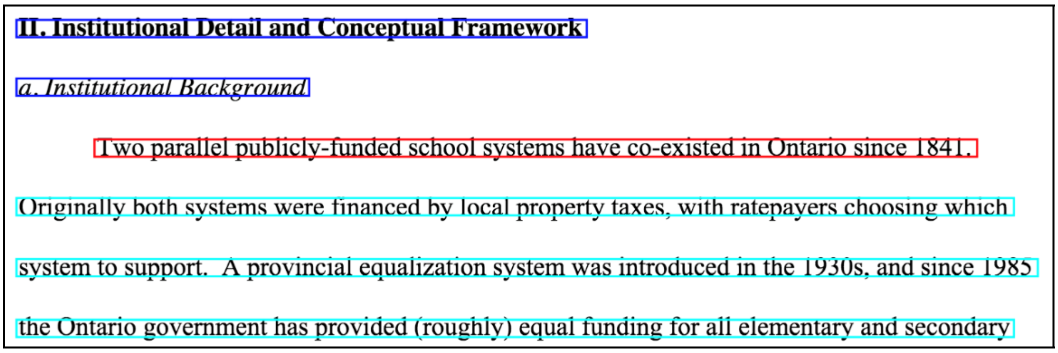
Using such logical rules we produce the output data files described above and illustrated by Figure 1. This output data is also used to visualise recognition results as a markup for an initial PDF paper.

Figure 2 shows a visualisation of data displayed at Figure 1 as a markup for a fragment of following paper:

Card, D., Dooley, M.D. and Payne, A.A. (2010) 'School competition and efficiency with publicly funded Catholic schools', *American Economic Journal: Applied Economics*, Vol. 2, No. 4, pp.150–76. http://www.nber.org/papers/w14176.pdf

For this paper our algorithm recognised section titles (with blue border), the first line of a paragraph (with red border) and regular text lines (with green border)[12].

Figure 2 A fragment of a marked up paper



II. Institutional Detail and Conceptual Framework

a. Institutional Background

Two parallel publicly-funded school systems have co-existed in Ontario since 1841.

Originally both systems were financed by local property taxes, with ratepayers choosing which

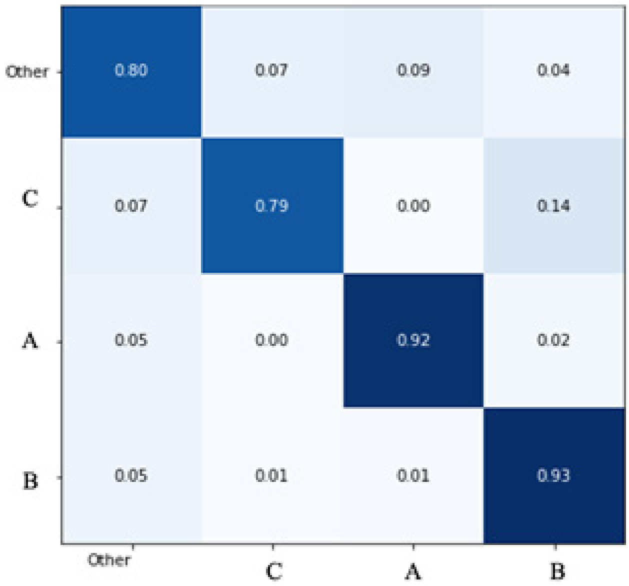
system to support. A provincial equalization system was introduced in the 1930s, and since 1985

the Ontario government has provided (roughly) equal funding for all elementary and secondary

To measure the recognition quality for our approach we selected some papers from the NBER Working Papers series, which are linked by the citation relationships, and can be used for the similarity analysis between the citing and cited papers (see the next section). From the "Introduction" section of these selected papers and to the "References" section, we manually counted the numbers of the section titles and paragraphs. They contain: 99 lines as "C" (section titles); 497 as "A" (the first line of a paragraph); 4851 as "B" (regular text line); and 826 as "?" (other) which are mainly footnotes, formulas, table content, captions.

Comparing these numbers with the results of our recognition we found (see Figure 3 with cross labels ratios of recognition quality) that our algorithm recognised correctly: 79% of lines with the section titles (label "C"); 92% of the paragraph first lines (label "A"); 93% of regular text lines (label "B"); and 80% of the label "other". Numbers outside the main diagonal of a matrix at Figure 3 show the ratios of wrong recognitions.

Figure 3 Per-label quality of recognition, by ratio



Some section titles (label "C") are recognised as a paragraph's line (14%) or "other" (7%), mostly because of negligence in papers' formatting: the titles are written with a regular font or without proper indent. This case is hard for improving by a recognition algorithm.

Some first lines of paragraphs (label "A") are recognised as "other" (5%) or a regular line (2%), mostly because we should analyse more accurately the left indent and filtering out table text, list items, etc.

Some regular text lines of paragraphs (label "B") are recognised as "others" (5%), mostly because of improper formatting, e.g. they do not have a proper left indent.

In average, the recognition quality by F1-score[13] is: for the section titles – 0.58; for the paragraph's first lines – 0.85; for the paragraph's regular lines – 0.96; for the "others" – 0.75.

Tao et al. (2014b) wrote "… documents are generally designed in their own styles, and a single universal scheme is unlikely to exist". Typically, papers belonged to the same series, e.g. as 26642 papers of the "NBER Working Papers" series, have more or less similar layout that promise the same quality of recognition for them. Of course, it does not guarantee good quality of recognition for papers from other series with different layouts. It is hard to create universal solution, but we provide all created software as an open source and anybody can add into it more rules for recognition of other layouts.

3 Topic modelling method and used data

The "topic modelling" method[14] is one of the approaches to the analysis of semantic relationships between text documents. In our case, topic modelling allows us to discover text latent semantic structures and to build abstract "topics" that occur in a collection of citing and cited papers. In order to find "topical" relationships we compare the topics built for citation contexts from the citing papers and the ones built for the cited papers.

Building the topics means discovering particular words appearing in the papers' texts more or less frequently. As an approach for finding semantic similarities between texts, topic modelling has several limitations: for instance, a citing author can re-write a citing research output without any terms or keywords from the cited paper to which the output belonged. In our research we assume that analysed citation contexts and their related cited papers have an intersection between terms or keywords.

Methods of machine learning and natural language processing, to which the topic modelling method also belongs, are highly popular in the area of citation content analysis. They were implemented for various tasks such as the automated classification of citation polarity (Yousif et al., 2018), the automated building of citation ontologies (Bakhti et al., 2018) and the measurement of text similarities (Knoth and Herrmannova, 2014).

The topic modelling method is usually used beyond the area of citation content analysis. It is used for tasks such as author classification (Kongthon et al., 2008; Lu and Wolfram, 2012), an analysis of temporal thematic changes in scientific papers (Jensen et al., 2016) and an analysis of research article content (Leydesdorff and Nerghes, 2017). In citation content analysis, such a method was applied for analysing knowledge flows across countries, through publication and citation data (Hassan and Haddawy, 2015), for discovering topical relationships embedded in citations (Kim et al., 2016) and for the joint modelling of text and citations in the topic modelling framework (Nallapati et al., 2008).

We build topic models on the basis of a merge of two corpora: a collection of NBER Working Papers full texts (prepared as described in the previous section), and Cirtec citation data. As a method for building topic models, we used the Latent Dirichlet Allocation (LDA) method (Blei et al., 2003), one of the most ubiquitous topic modelling algorithms. We consider it to be a more appropriate one for our study than other similar algorithms (e.g. Probabilistic Latent Semantic Analysis[15], PLSA), because it allows us to build meaningful topics for short text fragments (citation contexts). In a nutshell, LDA builds a distribution of topics over a collection of text documents, and creates a distribution of words over each of these topics. To this end, one can describe any text document with a topic. The topic here is a set of words pretending to represent the semantic content of a text.

The research papers layout recognition method presented in the previous section allows us to treat papers as hierarchical text structures: a citation context is considered as a part of a paragraph of a paper's section, and that section is considered as a part of the whole paper. We make an extraction of the citation context boundaries more precise by excluding text lines not belonging to the "parent" paragraph (section titles, headers, footers, etc.).

With an LDA model, we built topics characterising citation contexts from the perspective of the paper's hierarchical structure: topics are extracted from the nested texts fragments of: a) a paragraph, b) a section, c) a paper as a whole. Such hierarchical structure of topics allows us to analyse relationships between citing and cited papers for each of these nested sequence pairs: citation context –> paragraph –> section –> paper.

LDA has adjustable parameters such as a number of topics per document and a number of words per topic. In our experiments we have tested different variations of these parameters. A suitability of combination of the parameters was estimated by a heuristic evaluation of adequacy of the obtained topics. In the reported results, LDA was configured to generate one topic (consisting of 5 words) per text. Such a combination of parameters allowed us to obtain the most representative topics. We have trained a separate topic model for each of the hierarchical text levels. Texts were previously tokenised and lemmatised with the help of UDPipe[16] (we used an EWT model, version 2.3). In this study we have used an LDA algorithm implemented in the Gensim[17] library. The model was trained with 1 corpus pass on the aforementioned corpus. Source code for model training and inferring is available at our repository[18].

4 Topical similarity between citing and cited papers

In order to represent the semantic difference between the compared texts we have quantitatively compared the corresponding topics. To make such a comparison we have counted the word intersection ratio in these topics: the number of words that exist in both topics to the number of words in a topic (the latter was equal to 5 all the time). This ratio is close to the Jaccard similarity coefficient[19], but more interpretable for our setting.

We analysed the cross-topics relationships for the case when a paper cited two other papers. The first case: the citing paper has two mentions of the cited paper and two citation contents to analyse. The second case: the cited paper has 5 citation contexts in the citing paper. These papers are:

Citing paper: Figlio, D., Karbownik, K. and Salvanes, K.G. (2016) 'Education research and administrative data', *Handbook of the Economics of Education*, Elsevier, Vol. 5, pp.75–138. Available online at: https://www.nber.org/papers/w21592.pdf

which has two citation contexts for the first cited paper –

Cited paper: Einav, L. and Levin, J. (2013) *The data revolution and economic analysis.* NBER Working Paper No. 19035. Available online at: https://www.nber.org/papers/w19035.pdf

and five citation contexts for the second cited paper-

Cited paper: Card, D., Dooley, M.D. and Payne, A.A. (2010) 'School competition and efficiency with publicly funded Catholic schools', *American Economic Journal: Applied Economics*, Vol. 2, No. 4, pp.150–76. Available online at: https://www.nber.org/papers/w14176.pdf

Table 1 presents data to analyse case 1: the relationships between the citing paper and the first cited paper mentioned as the in-text citation "(Einav and Levin, 2013)". Table 2 presents data for case 2: topics for the citing paper and the second cited paper.

For case 1, both citation contexts for this cited paper belong to the same paragraph. Thus, they have common topics built for the upper hierarchical text fragments, as a paragraph and a section. Where possible, in Table 1, we highlighted the words from the topics into the related text fragments. Comparing the topics built for the citation contexts as well as for the whole text of the cited paper we have computed the word intersection ratio for each citation case.

As one can see, the intersection ratio is non-zero in all the cases. Owing the fact that the topics pretend to be representative of the texts, we consider that the non-zero word intersection ratio reports the existence of semantic relationship between the citation contexts and the cited paper content.

Table 1 Related text fragments, citation contexts and topic models for case 1

Citing paper: Figlio, D., Karbownik, K. and Salvanes, K.G. (2016) 'Education research and administrative data', *Handbook of the Economics of Education*, Elsevier, Vol. 5, pp.75–138. **Abstract**: Thanks to extraordinary and exponential improvements in data storage and computing capacities, it is now possible to collect, manage, and analyse data in magnitudes and in manners that would have been inconceivable just a short time ago. As the world has developed this remarkable capacity to store and analyse data, so have the world's governments developed large-scale, comprehensive data files on tax programs, workforce information, benefit programs, health, and education. While these data are collected for purely administrative purposes, they represent remarkable new opportunities for expanding our knowledge. This chapter describes some of the benefits and challenges associated with the use of administrative data in education research. We also offer specific case studies of data that have been developed in both the Nordic countries and the USA, and offer an (incomplete) inventory of data sets used by social scientists to study education questions on every inhabited continent on earth.	*Built topics*: data, teacher, school, student, use

Table 1 Related text fragments, citation contexts and topic models for case 1 (continued)

Section title: II. The benefits of using administrative data in education research	data, administrative, study, set, possible Word intersection ratio = 0.2
Built topic for the paragraph: study, experiment, data, natural, make	
Citation contexts and in-text citation	*Built topics*
Prefix: Administrative data sets also provide novel types of variables typically not found in non-administrative data *In-text citation*: (Einav and Levin, 2013). *Suffix:* They can offer new opportunities, for instance, to look at measures of delinquency, of changing geographical location, of social networks, and of health instances that are nearly impossible to study in any other manner.	*For citation context*: data, instance, delinquency, manner, find Word intersection ratio = 0.2
Prefix: The real-time nature of administrative data also provides new opportunities to study the effects of educational policies and practices that are very recent; and offers the chance for researchers to make their scholarship much more relevant to the specific policy decisions that policymakers must make right away than are studies that make use of retrospective information *In-text citation*: (Einav and Levin, 2013) *Suffix*: And of course, natural experiments need not be rare events to be better-studied using administrative data sets: Because natural experiments are unannounced, and often occur via chance or quirks, it is very difficult to set up a prospective study that will permit the evaluation of a natural experiment; with administrative data that cover a population…	*For citation context*: data, administrative, study, experiment, make Word intersection ratio = 0.2
Cited paper: Einav, L. and Levin, J. (2013) *The data revolution and economic analysis*, NBER Working Paper No. 19035. *Abstract*: Many believe that "big data" will transform business, government and other aspects of the economy. In this article we discuss how new data may impact economic policy and economic research. Large-scale administrative datasets and proprietary private sector data can greatly improve the way we measure, track and describe economic activity. They also can enable novel research designs that allow researchers to trace the consequences of different events or policies. We outline some of the challenges in accessing and making use of these data. We also consider whether the big data predictive modelling tools that have emerged in statistics and computer science may prove useful in economics.	*Built topics for the paper*: data, use, economic, large, information Word intersection ratio = 0.4

Table 2 Topics for case 2

Text	*Topics*	*Word intersection ratio*
Citing paper	data, teacher, school, student, use	–
Citation context 1	data, government, maintain, country, individual	0.2
Citation context 2	study, data, event, rare, ability	0.2
Citation context 3	data, likely, problem, less, administrative	0.2
Citation context 4	data, research, administrative, conduct, develop	0.2
Citation context 5	administrative, data, experiment, find, purpose	0.2
Cited paper	school, student, test, enrolment, grade	0.4

Data in Table 2 illustrates that the non-zero word intersection ratio also exists for the second case. Each citation topic in Table 2 has at least one common word with the cited paper topic. Moreover, certain words in the topics have close meanings in different forms: e.g. "study" in the topic of Citation context 2 is semantically related to the words "school" and "student" in the cited paper topic. Therefore, we can assume that a semantic relationship between the content of compared fragments of the citing and cited papers exists.

To ensure that the topic model has, indeed, captured meaningful semantic information (and is not a statistical error), we have tried to sample sets of five random words from the compared papers' fragments. In most of the experiments we have obtained sets of words with zero intersection rate. So, according to this evaluation procedure, topic models exceeded our baseline (random word sampling), and we think that topic modelling is applicable for finding the semantic relationships between the considered texts.

5 Conclusions and outlook

In this paper, we have presented a study (made for the Cirtec project) aiming to prepare a methodology for the long term project task: the creation of direct communications between citing and cited authors. Such a task, technically, should be based on RIS tools and Cirtec citation data. We analysed the thematic similarities between the citation contexts of the citing and cited papers from these datasets, and compared the papers' content by using the "topic modelling" method. We also created open source tools for the semantic logical labelling of the papers' layout. Such tools allowed us to build, for this study, the hierarchical structure of topic models (and, therefore, to make a more accurate comparison of topics). Lastly, we found similarities of topics for the compared fragments of the papers. In general, our experiments confirmed the hypothesis that the thematic relationships between the analysed citation contexts of the citing papers and the content of the cited papers exist and could be measured through the topic modelling method.

These results can be easily reproduced and developed, since all the created and used tools are open source. At the same time, we came to the conclusion that "topic models" is a bit crude for such purposes. If one has to know the characteristics of the thematic similarity between the related fragments of the citing and cited papers, we hypothesise that it could be better to employ the n-gram method[20], which provides the most frequently repeated professional terms and lexical cliché. We intend to use this method for our future research and development.

Acknowledgements

A part of this study, which is a development of the approach for extracting citation content data, with the focus on the supercomputer simulation of interactions among the agents and the research community environment, is funded by a RSF grant (project No. 19-18-00240).

References

Bakhti, K., Niu, Z., Yousif, A. and Nyamawe, A.S. (2018) 'Citation function classification based on ontologies and convolutional neural networks', *International Workshop on Learning Technology for Education in Cloud*, Springer, Cham, pp.105–115.

Bertin, M. and Atanassova, I. (2014) 'A study of lexical distribution in citation contexts through the IMRaD standard', *Proceedings of the 1st Workshop on Bibliometric-enhanced Information Retrieval co-located with 36th European Conference on Information Retrieval (ECIR'14)*, pp.5–12.

Bertin, M. and Atanassova, I. (2015) 'Factorial correspondence analysis applied to citation contexts', *BIR@ ECIR*, pp.22–29.

Bertin, M. and Atanassova, I. (2018) 'InTeReC: in-text reference corpus for applying natural language processing to bibliometrics', *Proceedings of the 7th Workshop on Bibliometric-enhanced Information Retrieval (BIR)*, Grenoble, France, pp.54–62.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) 'Latent dirichlet allocation', *Journal of Machine Learning Research*, pp.993–1022.

Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X. and Zhai, C. (2014) 'Content-based citation analysis: the next generation of citation analysis', *Journal of the Association for Information Science and Technology*, Vol. 65, No. 9, pp.1820–1833.

Hassan, S.U. and Haddawy, P. (2015) 'Analyzing knowledge flows of scientific literature through semantic links: a case study in the field of energy', *Scientometrics*, Vol. 103, No. 1, pp.33–46.

Jensen, S., Liu, X., Yu, Y. and Milojevic, S. (2016) 'Generation of topic evolution trees from heterogeneous bibliographic networks', *Journal of Informetrics*, Vol. 10, No. 2, pp.606–621.

Kim, H.J., An, J., Jeong, Y.K. and Song, M. (2016) 'Exploring the leading authors and journals in major topics by citation sentences and topic modeling', *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, pp.42–50.

Knoth, P. and Herrmannova, D. (2014) 'Towards semantometrics: a new semantic similarity based measure for assessing a research publication's contribution', *D-Lib Magazine*, Vol. 20, No. 11, p.8.

Kogalovsky, M., Krichel, T., Lyapunov, V., Medvedeva, O., Parinov, S. and Sergeeva, V. (2019) 'Open citation content data', in Garoufallou, E., Sartori, F., Siatri, R. and Zervas, M. (eds): *Communications in Computer and Information Science Metadata and Semantic Research*, Springer, Cham, pp.355–364.

Kongthon, A., Haruechaiyasak, C. and Thaiprayoon, S. (2008) 'Enhancing the literature review using author-topic profiling', *International Conference on Asian Digital Libraries*, Springer, Berlin, Heidelberg, pp.335–338.

Leydesdorff, L. and Nerghes, A. (2017) 'Co-word maps and topic modeling: a comparison using small and medium-sized corpora (N < 1,000)', *Journal of the Association for Information Science and Technology*, Vol. 68, No. 4, pp.1024–1035.

Lu, K. and Wolfram, D. (2012) 'Measuring author research relatedness: a comparison of word-based, topic-based, and author cocitation approaches', *Journal of the American Society for Information Science and Technology*, Vol. 63, No. 10, pp.1973–1986.

Nallapati, R.M., Ahmed, A., Xing, E.P. and Cohen, W.W. (2008) 'Joint latent topic models for text and citations', *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp.542–550.

Neylon, C. (2017) *Citation metrics are making headlines, but what does citation really mean? JISC blog.* Available online at: https://jisc.ac.uk/blog/citation-metrics-are-making-headlines-but-what-does-citation-really-mean-08-dec-2017

Paaß, G. and Konya, I. (2011) 'Machine learning for document structure recognition', *Modeling, Learning, and Processing of Text Technological Data Structures*, Springer, Berlin, Heidelberg, pp.221–247.

Parinov, S. (2017a) 'Semantic attributes for citation relationships: creation and visualization', in Garoufallou, E., Virkus, S., Siatri, R. and Koutsomiha, D. (eds): *Communications in Computer and Information Science Metadata and Semantic Research*, Springer, Cham, Vol. 755, pp.286–299.

Parinov, S. (2017b) 'Extraction and visualisation of citation relationships and its attributes for papers in PDF', *International Journal of Metadata, Semantics and Ontologies (IJMSO)*, Vol. 12, No. 4, PP.195–203. Doi: 10.1504/IJMSO.2017.093626.

Parinov, S. (2019) 'CRIS with in-text citations as interactive entities', *Procedia Computer Science*, Vol. 146, pp.20–28.

Parinov, S. and Antonova, V. (2019) 'Global scholarly collaboration: in Connecting the knowledge commons – from projects to sustainable infrastructure', *Proceedings of the 22nd International Conference on Electronic Publishing–Revised Selected Papers*, OpenEdition Press, pp.57–75.

Rahman, M.M. and Finin, T. (2017) 'Understanding the logical and semantic structure of large documents', *arXiv preprint arXiv:1709.00770*, pp.1–10.

Swales, J. (1986) 'Citation analysis and discourse analysis', *Applied Linguistics*, Vol. 7, No. 1, pp.39–56.

Tao, X., Tang, Z. and Xu, C. (2014b) 'Contextual modeling for logical labeling of PDF documents', *Computers and Electrical Engineering*, Vol. 40, No. 4, pp.1363–1375.

Tao, X., Tang, Z., Xu, C. and Wang, Y. (2014a) 'Logical labeling of fixed layout PDF documents using multiple contexts', *Proceedings of the 11th IAPR International Workshop on Document Analysis Systems (DAS)*, Tours, France, pp.360–364. Doi: 10.1109/DAS.2014.54.

Yousif, A., Niu, Z., Nyamawe, A.S. and Hu, Y. (2018) 'Improving citation sentiment and purpose classification using hybrid deep neural network model', *International Conference on Advanced Intelligent Systems and Informatics*, Springer, Cham, pp.327–336.

Notes

[1] http://cirtec.ranepa.ru/

[2] https://www.ranepa.ru/eng/

[3] https://en.wikipedia.org/wiki/Document_layout_analysis

[4] there are also commercial solutions, e.g. Docparser; most of them work as online services

[5] https://sociorepec.org/collection.xml?h=repec:nbr:nberwo&l=en

[6] https://github.com/citations-ai/linelabeler/tree/master/experiment1

[7] see details about this in (Parinov 2017a, 2017b)

[8] Victor Lyapunov made this software

[9] https://github.com/citations-ai/linelabeler

[10] See the full data file at https://github.com/citations-ai/linelabeler/blob/master/experiment1/repec:nbr:nberwo:14176/repec:nbr:nberwo:14176:l.txt

[11] https://github.com/citations-ai/linelabeler/blob/5c3815b2b61d60da2cabb2a65f1805984a938404/dataline.py#L178

[12] See the full marked up paper at https://github.com/citations-ai/linelabeler/blob/master/experiment1/repec:nbr:nberwo:14176/repec:nbr:nberwo:14176:l.pdf

[13] https://en.wikipedia.org/wiki/F1_score

[14] https://en.wikipedia.org/wiki/Topic_model

[15] https://en.wikipedia.org/wiki/Probabilistic_latent_semantic_analysis

[16] https://github.com/ufal/udpipe

[17] https://radimrehurek.com/gensim/

[18] https://github.com/bakarov/cirtec/blob/master/topic_modeling/experiment_1/topic_modeller.py

[19] https://en.wikipedia.org/wiki/Jaccard_index

[20] https://en.wikipedia.org/wiki/N-gram