Jaccard index-Based Assessing the Similarity of Research Fields in Dimensions

Serhiy Shtovba[0000-0003-1302-4899], Mykola Petrychko[0000-0001-6836-7843

Vinnytsia National Technical University, Khmelnytske Shose, 95, Vinnytsia, 21021, Ukraine
shtovba@vntu.edu.ua
petrychko.myckola@gmail.com

Abstract. We calculated the similarity of 10878 pairs of the research fields for 2010-2019 and for 2000-2009 on base of Dimensions data. The similarity of the research fields is assessed by Jaccard index. The most tied research fields for 2010-2019 are Specialist Studies in Education and Curricular and Pedagogy, whereas this pair holds rank #6 for 2000-2009. Dynamics of Jaccard index distributions does not confirm a hypothesis about increasing the share of interdisciplinary research over time. It is detected that the most collaborated research field for the both time intervals is Mechanical Engineering. This research field has maximum sum of Jaccard indexes with all the other research fields (so called stickiness index). It is detected that the most tied triad of research fields for the both time intervals are Commercial Services, Marketing, and Tourism. Distribution of similarity of the research fields, distribution of stickiness of single research field, and distribution of stickiness of triad of the research fields look like to Zipf law with 3 stable zones.

Keywords: scientometrics, research field, similarity, interdisciplinary, Dimensions, Jaccard index, stickiness index, distribution, categorization, Zipf law.

1 Introduction

The question of how strong the various research fields are interconnected has a long history. The relevance of the quantitative assessment of a similarity of the research fields has intensified recently, when we are increasingly hearing about interdisciplinary research, about PhD theses on boarding line of specialties, about multi-skill research team etc. Traditionally, a similarity of the research fields is assessed by experts. The assessment is based on subject matters under investigation and can be associated with an essentialist view. In [1] information-theoretic measure of linguistic similarity to investigate the organization and evolution of scientific fields is proposed. An analysis of almost 20M papers from the past three decades reveals that the linguistic similarity is related but different from experts and citation-based classifications, leading to an improved view on the organization of science.

Recently, citations-based approach to similarity assessment of the research fields is most widely used. This approach is based on Porter’s metrics of paper interdisciplinary [2]. The metrics is used as source data a number of citations from current paper

2019 DCSMart Workshop.

with certain research field to papers with other research fields and vice versa. A relation of the paper to some research fields is inherited by their journal status. 2D-distribution by research fields in axes "References to outside the research field – Citation from outside the research field" is shown in [3]. This 2D-distribution allows identifying most interdisciplinary (high stickiness) research fields.

Due to the different content of various papers of the same journal the journal-based classification to research field produces some errors. Moreover the journal-based classification would not allow categorize grants, patents, books etc. It is the reason to create in [4] a new approach to paper categorization approach using machine learning techniques. The title and abstract of the paper are source data for this categorization. That machine learning approach for paper categorization is implemented in Dimensions – the newest platform of indexing the research papers. Now, Dimensions is worldwide biggest one with good searching services, hence it is reasonable to try evaluating the similarities of the researches fields using Dimensions data and compare new detected dynamics of interdisciplinary with results of previous study in [5].

2 System of research fields in Dimensions

Dimensions uses a short version of the Australian and New Zealand Standard Research Classification. There are 154 fields of research and 22 domains. For example, domain 01 – Mathematical Sciences contains the following research fields: 0101 – Pure Mathematics; 0102 – Applied Mathematics; 0103 – Numerical and Computational Mathematics; 0104 – Statistics; 0105 – Mathematical Physics.

At the moment, Dimensions indexes 106M research papers. Each paper is assigned to one or several research fields. Most filled research field is Clinical Sciences with 9.3M papers. Distributions of papers by the research fields for 2010-2019 and for 2000-2010 are shown in Figure 1. Both distributions look pretty much as Zipf law in almost whole range.

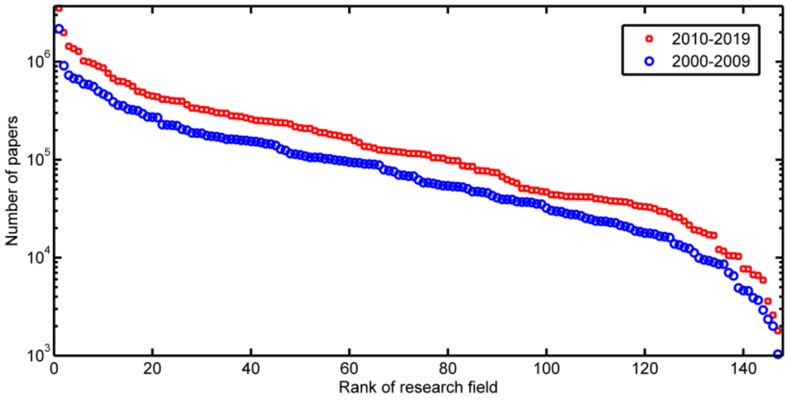


Fig. 1. Distribution of number of the papers in Dimensions by research fields in semilog format (research fields with less than 1000 publications for any of time intervals are excluded)

3 Jaccard index for two research fields

Dimensions provides a number of papers assigned with some research field. Additionally Dimensions provides number of above mentioned papers, which assigned with each other of the research fields. It allows obtain data for calculation a similarity of any pair of research fields using just 2 queries. For processing all the research fields it is necessary one query per one research field.

We propose to assess a similarity of two research fields *A* and *B* by Jaccard index:

$$S(A,B)=\frac{N_{A\cap B}}{N_A+N_B-N_{A\cap B}},$$

where N_A denotes a number of papers in research field *A*, N_B denotes a number of papers in research field *B*, $N_{A\cap B}$ denotes a number of papers assigned to research field *A* and to research field *B* simultaneously.

Consider the following example with the real data. There are 74519 papers assigned to Marketing in 2010 – 2019, 11843 of them also belong to Tourism. Research field Tourism has 36274 papers, hence the similarity of Marketing and Tourism equals to 11843 / (74519+36274-11843)=0.12.

Distributions of Jaccard index for two research fields are shown in Figure 2. For reliability reason tiny research fields with less than 1000 papers for any of the decades are out of consideration. In total 147 research fields are considered, hence the similarity of 10878 pairs of the research fields for 2010-2019 and for 2000-2010 are calculated. Most of the research fields have zero similarity. Both distributions look like as three-zone Zipf law: the first zone for high similar research fields; the second zone for low similar research fields; the third zone for tiny similar research fields. A strong decreasing the similarity over ranks takes place for the first and the third zones.

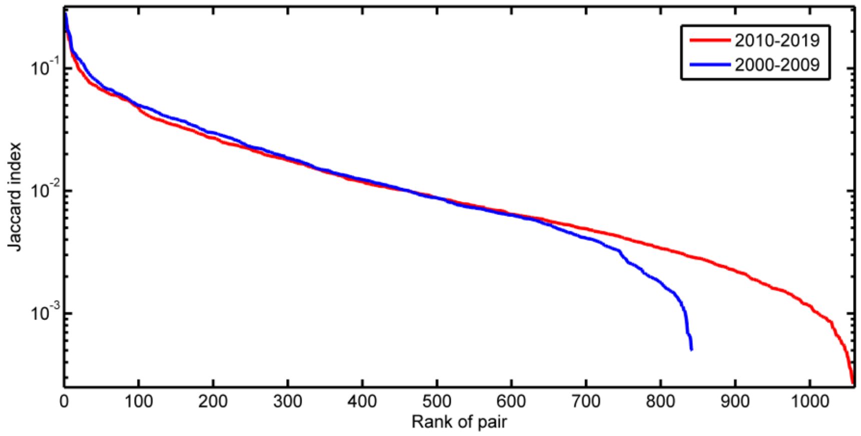


Fig. 2. Distribution of the similarity of pair of the research fields (semilog format)

Figure 2 shows that the leaders for 2000-2010 have higher Jaccard indexes than leaders for 2010-2019. But, distribution for 2010-2019 has more heavy tail. It allows to inference by eye that there is no significant difference in change of interdisciplinary research share over time. This conclusion concordances with a result in [5] about science is becoming more interdisciplinary during 1975 – 2005, but in small steps. Moreover, sum of Jaccard indexes for all the pairs of the research fields shows a slight decreasing of interdisciplinary research share over time. This sum equals to 19.6 for 2000-2010, and equals to 18.7 for 2010-2019.

The following pairs of the research fields from top-halves of the distributions in Figure 2 changed their Jaccard index drastically during the last decade: Environmental Engineering and Manufacturing Engineering – +254.5%; Atmospheric Sciences and Aerospace Engineering – -67.1%.

Twenty most tied pairs of the research fields are ranked in Figure 3 and Figure 4. Similarity of all the research fields in Figure 3 and Figure 4 are reasonable. There is one significant change over years in the tops-20 – Specialist Studies in Education with Curricular and Pedagogy moved from #6 for 2000-2009 to #1 for 2010-2019.

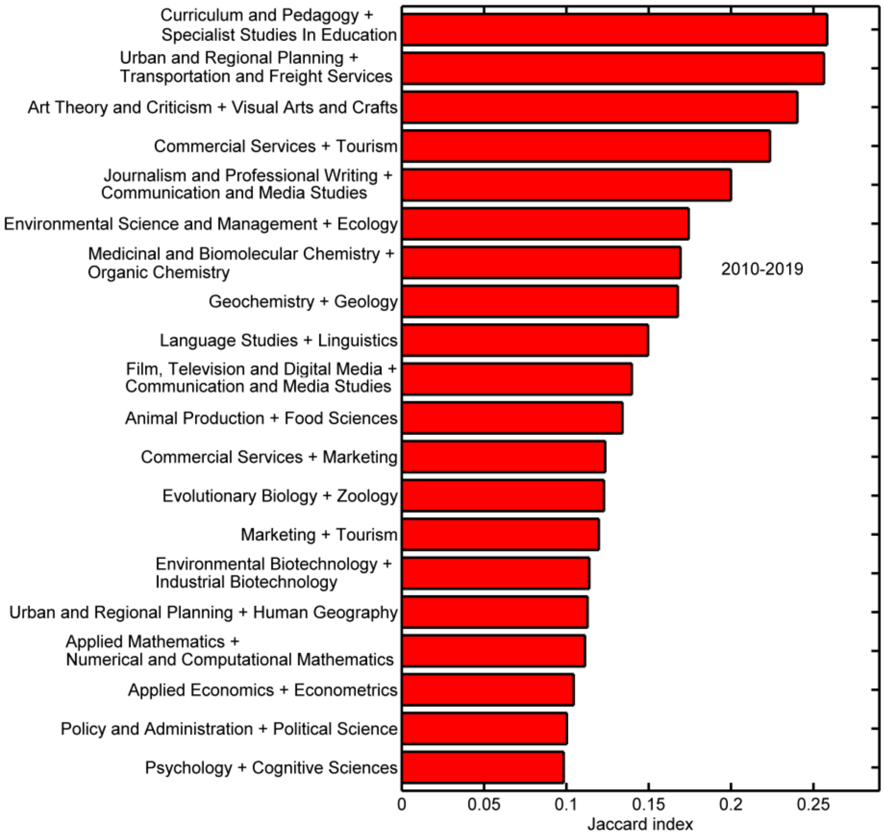


Fig. 3. Most tied pairs of the research fields for 2010-2019

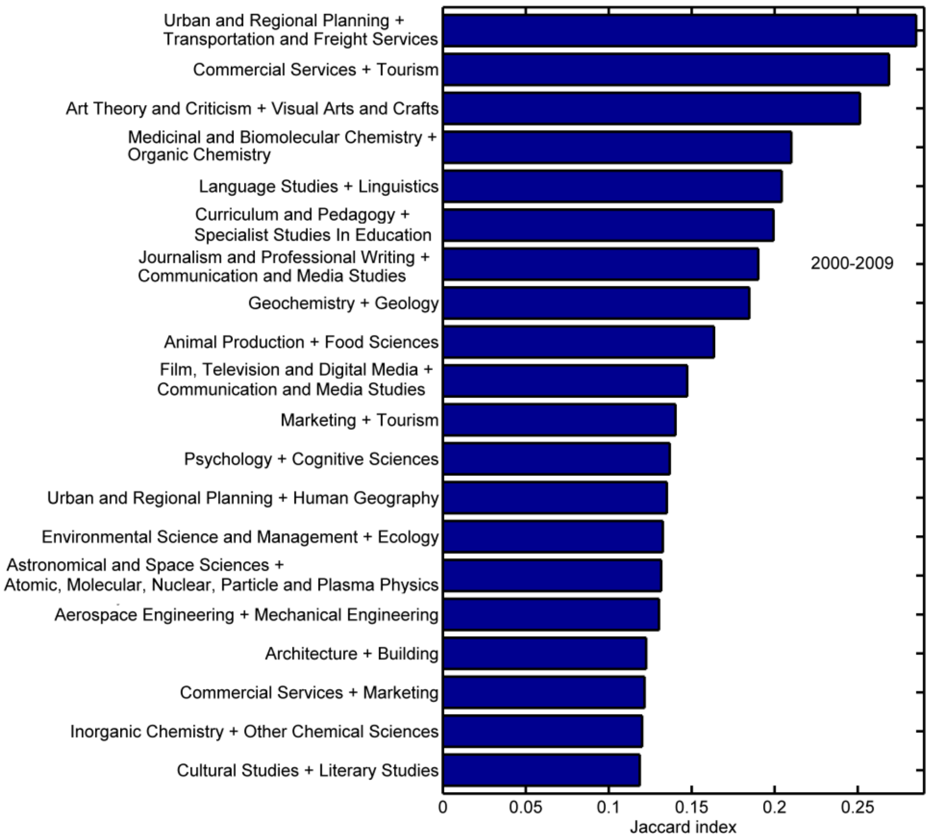


Fig. 4. Most tied pairs of the research fields for 2000-2009

As an example, Figure 5 shows similarity of research fields that belong to Information and Computer Sciences domain. There are 8 research fields (A), (B), …, (H) in the domain. They are contoured by a blue rectangle in the Figure 5. Seven research fields that most tied with research fields (A), (B), …, (H) are also shown on Figure 5. Among those seven research fields Communications Technologies is most tied with Information and Computer Sciences domain.

Let's introduce stickiness index $G(A)$ of a research field A as a sum of all Jaccard indexes between A and other research fields as follows:

$$G(A) = \sum_{p:\ p \in \mathbf{F},\ p \neq A} S(A, p),$$

where **F** denotes a set of the research fields. Stickiness indexes allow to detect research fields with the highest attractive to collaboration, with the highest ability to interdisciplinary. The distributions of single research field stickiness indexes are shown in Figure 6. Both distributions have very heavy tails, hence difference in 5-10 positions is not important.

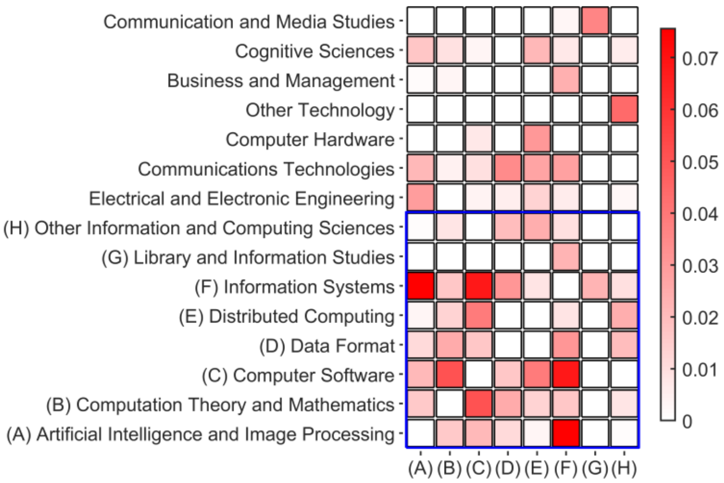


Fig. 5. Similarity of research fields that belong to Information and Computer Sciences domain (data for 2010-2019)

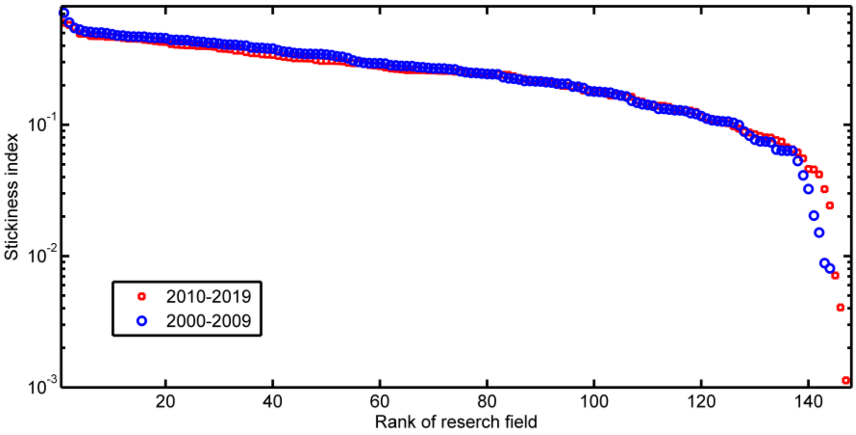


Fig. 6. Distribution of the research field stickiness indexes (semilog format)

The highest stickiness index has Mechanical Engineering for both time intervals (Figure 7 and Figure 8). Comparing the top-rank lists, it is found that Mechanical Engineering has lost their gap. Some significant differences in Figure 6 and Figure 8 are as follows: Environmental Science and Management jumped from #17 to #2, Condensed Matter Physics descended from #2 to #20+. The following research fields from the top halves of the distributions changed their stickiness indexes significantly during the last decade:

Ecological Applications –	+46.7%;
Aerospace Engineering –	-37.7%;
Atmospheric Sciences –	-28.2%;
Optical Physics –	+27.6%;
Language Studies –	-27.4%;
Astronomical and Space Sciences –	-25.9%;
Genetics –	+25.0%;
Environmental Science and Management –	+24.6%;
Condensed Matter Physics –	-24.6%;
Soil Sciences –	+23.9%;
Curriculum and Pedagogy –	+22.7%;
Civil Engineering –	-20.5%.

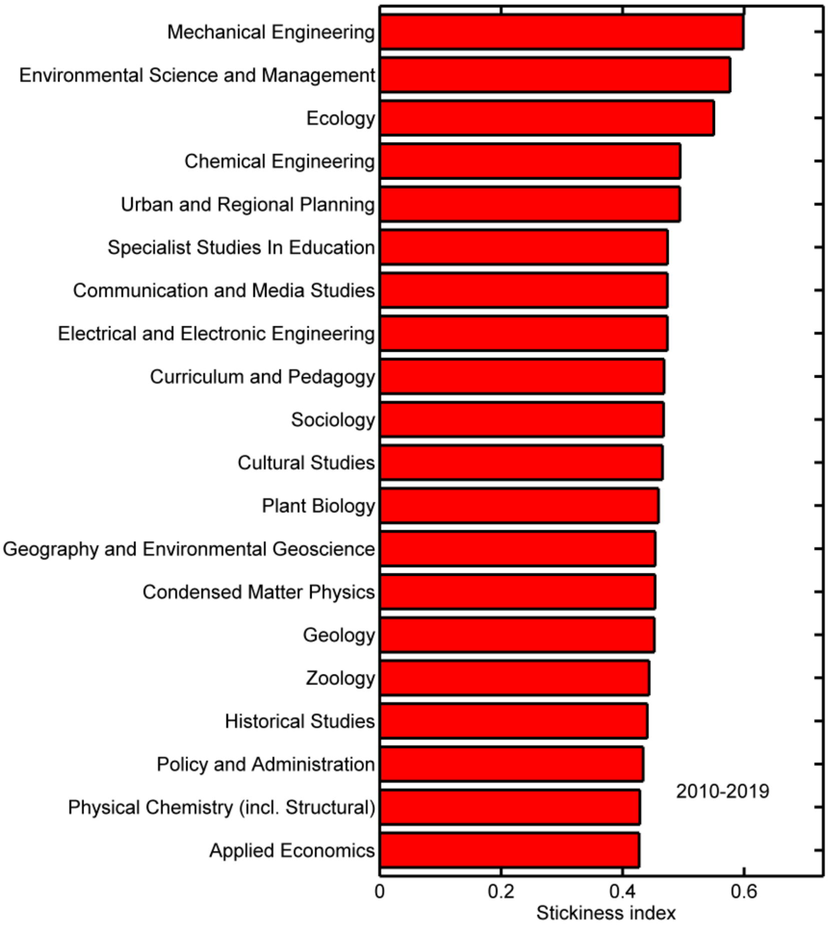


Fig. 7. Most collaborative single research fields for 2010-2019

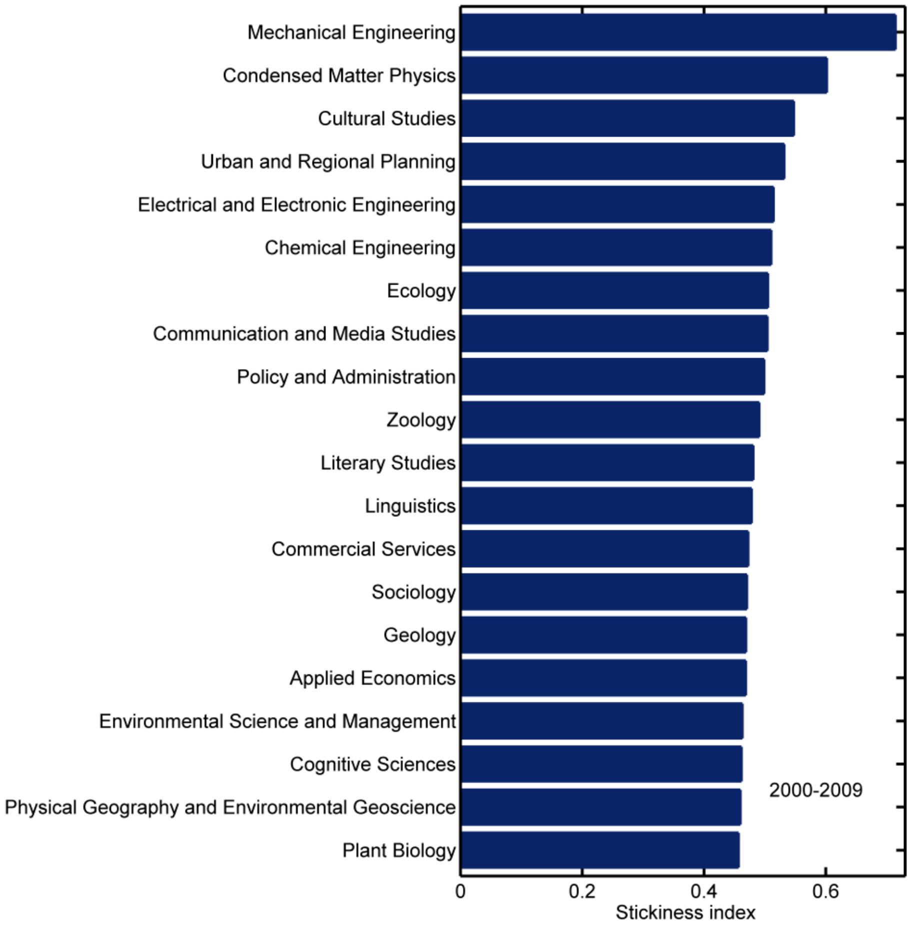


Fig. 8. Most collaborative single research fields for 2000-2009

4 Similarity of triad of research fields

A similarity of triad of research fields (*A*, *B*, *C*) is proposed assess by so called triad stickiness index $G(A, B, C)$ as follows:

$$G(A, B, C) = S(A, B) + S(A, C) + S(B, C)$$

We found 110451 triads for 2000 – 2009 and 135503 triads for 2010 – 2019 with non-zero stickiness index. Distributions of the triad stickiness indexes (Figure 9) and distributions of Jaccard indexes of pairs of research fields (Figure 2) have the same shape.

Dozens of the most tied triads are shown in Table 1 and Table 2. All the triads are reasonable. Leader for the both time intervals is Commercial Services, Marketing and

Tourism. For this triad we constructed iteratively a map of science with nearest research fields (Figure 10). On each iteration we choose a research field that maximizes new net stickiness index. The iteration traces are present at right part of Figure 10. It starts with Business and management, and finishes on Historical Studies. Figure 10 shows that research fields with the same domains are stronger tied each other than with research fields from different domains. It may considerate as an indirect support for proposed stickiness index.

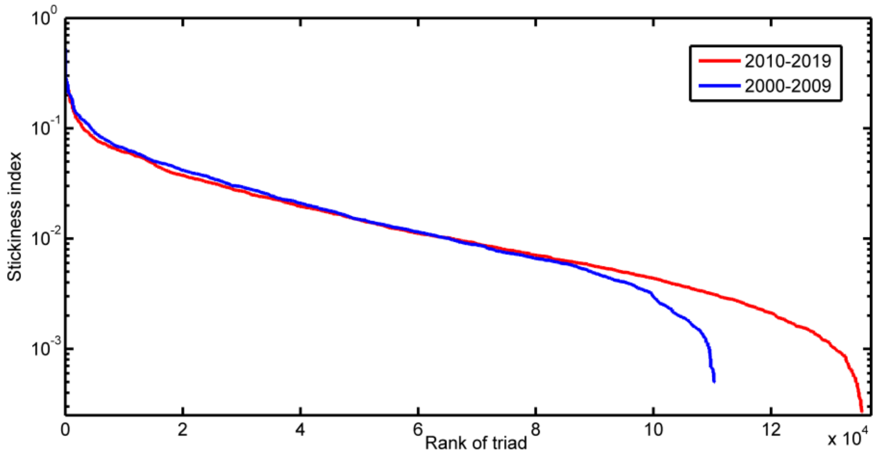
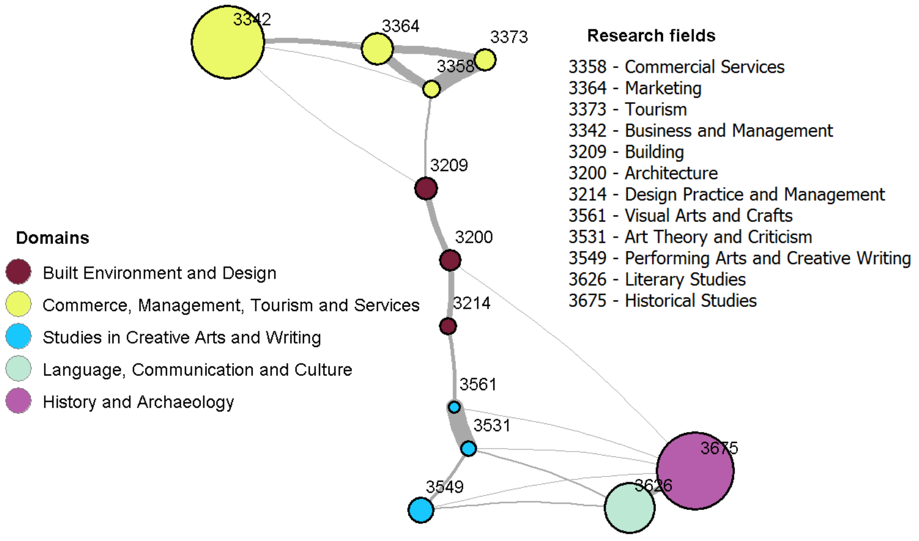


Fig. 9. Distribution of triad stickiness indexes (semilog format)



3342
3364
3373
3358
3209
3200
3214
3561
3531
3549
3626
3675
Research fields
3358 - Commercial Services
3364 - Marketing
3373 - Tourism
3342 - Business and Management
3209 - Building
3200 - Architecture
3214 - Design Practice and Management
3561 - Visual Arts and Crafts
3531 - Art Theory and Criticism
3549 - Performing Arts and Creative Writing
3626 - Literary Studies
3675 - Historical Studies
Domains
● Built Environment and Design
● Commerce, Management, Tourism and Services
● Studies in Creative Arts and Writing
● Language, Communication and Culture
● History and Archaeology

Fig. 10. Map of science for the most tied triad of research fields
(edge thickness equals to Jaccard index, vertex area equals to a number of papers)

Table 1. Most tied triads for 2010–2019

Rank	Research fields in the triad	Stickiness index
1	Commercial Services Marketing Tourism	0.4669
2	Film, Television and Digital Media Journalism and Professional Writing Communication and Media Studies	0.3894
3	Urban and Regional Planning Transportation and Freight Services Human Geography	0.3835
4	Education Systems Curriculum and Pedagogy Specialist Studies In Education	0.3469
5	Curriculum and Pedagogy Specialist Studies In Education Linguistics	0.3179
6	Curriculum and Pedagogy Specialist Studies In Education Sociology	0.3150
7	Ecological Applications Environmental Science and Management Ecology	0.3074
8	Geochemistry Geology Geophysics	0.2992
9	Design Practice and Management Art Theory and Criticism Visual Arts and Crafts	0.2949
10	Civil Engineering Urban and Regional Planning Transportation and Freight Services	0.2942
11	Art Theory and Criticism Performing Arts and Creative Writing Visual Arts and Crafts	0.2897
12	Curriculum and Pedagogy Specialist Studies In Education Psychology	0.2834

Table 2. Most tied triads for 2000–2009

Rank	Research fields in the triad	Stickiness index
1	Commercial Services Marketing Tourism	0.5307
2	Urban and Regional Planning Transportation and Freight Services Human Geography	0.4203
3	Film, Television and Digital Media Journalism and Professional Writing Communication and Media Studies	0.3936
4	Civil Engineering Urban and Regional Planning Transportation and Freight Services	0.3593
5	Language Studies Linguistics Literary Studies	0.3493
6	Geochemistry Geology Geophysics	0.3435
7	Design Practice and Management Art Theory and Criticism Visual Arts and Craft	0.3293
8	Building Commercial Services Tourism	0.3216
9	Geochemistry Geology Physical Geography and Environmental Geoscience	0.3105
10	Architecture Urban and Regional Planning Transportation and Freight Services	0.3095
11	Art Theory and Criticism Visual Arts and Crafts Curatorial and Related Studies	0.3077
12	Environmental Science and Management Urban and Regional Planning Transportation and Freight Services	0.3028

5 Conclusion

For the first time similarity of the research fields in form of Jaccard index are assessed by Dimensions data. Source information for Jaccard index calculation is the statistics of categorized papers in Dimensions. It is calculated the similarity of 10878 pairs of the research fields for 2010-2019 and for 2000-2009. The most tied pair of the research fields for 2010-2019 is Specialist Studies in Education and Curricular and Pedagogy, whereas they hold rank #6 for 2000-2009. The following pairs of the research fields changed their Jaccard index drastically during the last decade: Environmental Engineering and Manufacturing Engineering jumped on 254.5%, Atmospheric Sciences and Aerospace Engineering dropped on 67.1%.

Stickiness index of a research field is introduced as an assessment of its attractiveness to be a member in an interdisciplinarity group. Stickiness index is a sum of all Jaccard indexes related to current research field. The highest stickiness index has Mechanical Engineering for both time intervals. The following research fields changed their stickiness indexes drastically during the last decade: Ecological Applications jumped on 47%, Aerospace Engineering dropped on 38%.

The most tied triads of research fields are detected on base of the proposed stickiness index. All the triads are reasonable. Leader among the triads is Commercial Services, Marketing and Tourism for the both time intervals.

Possible applications of proposals similarity index is categorization of researchers, universities, journals, which act in various research topics. Similarity index allows taking into account tails in the research profile for more correct categorization.

References

1. Dias, L., Gerlach, M., Scharloth, J., Altmann, E.G.: Using text analysis to quantify the similarity and evolution of scientific disciplines. Royal Society Open Science. **5**, (2018). doi: 10.1098/rsos.171545.
2. Porter, A.L., Cohen, A.S., Roessner, J.D., Perreault, M.: Measuring researcher interdisciplinarity. Scientometrics. **72**, 117–147 (2007). doi: 10.1007/s11192-007-1700-5.
3. Noorden, V.R.: Interdisciplinary research by the numbers. Nature. **525**, 306-307 (2015). doi: 10.1038/525306a.
4. Herzog, C., Kierkegaard, B.L.: Response to the letter “Field classification of publications in Dimensions: a first case study testing its reliability and validity”. Scientometrics. **117**. 641-645 (2018). doi: 10.1007/s11192-018-2854-z.
5. Porter, A., Rafols, I.: Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. Scientometrics. **81**, 719-745 (2009). doi: 10.1007/s11192-008-2197-2.