Bibliographic units identification problem part 3

Vladimir Batagelj
UP FAMNIT Koper in IMFM Ljubljana

1339. sredin seminar
Ljubljana, 15. november 2023

Outline

Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si
Current version of slides (November 15, 2023 at 20:54): PDF
https://github.com/bavla/biblio/

Scopus

My grandson Mark is a handball player. He is also a student of Informatics at the Faculty of Social Sciences. To combine his interests we collected from WoS and Scopus all papers on handball.

He presented his analyses based on WoS data at Sunbelt 2022 and EuSN 2023. We intended to combine the WoS data with the Scopus data by first converting the Scopus descriptions into WoS descriptions and using WoS2Pajek afterward to produce the networks

Scopus2WoS

but the task is not as simple as we expected.

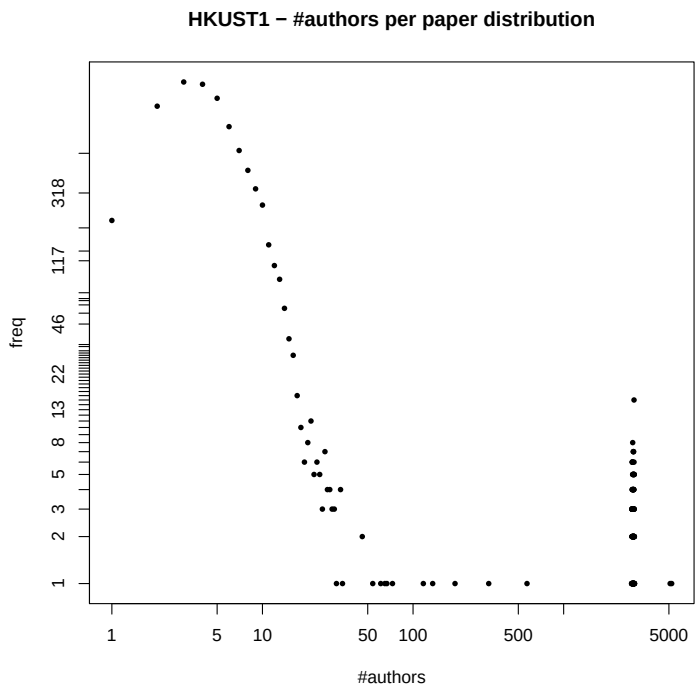
Scopus

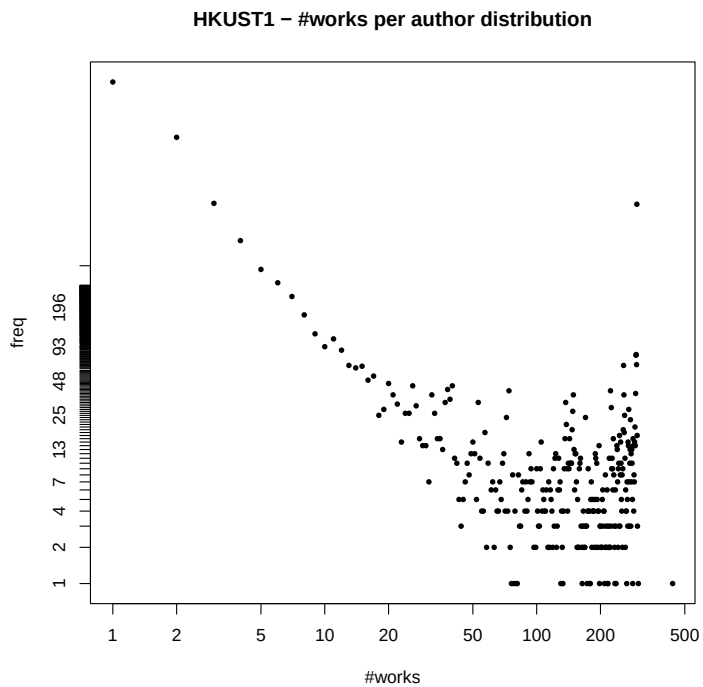
Some time ago, Nataliya Matveeva sent me data on the publications of young universities obtained from Scopus. I was pleasantly surprised to notice that several types of bibliographic units in Scopus have identifiers associated with them. This makes Scopus an interesting resource for building high-quality bibliographic networks.
vladowiki/scopus

It turned out that our decision to base Scopus2WoS on exported RIS files (similar to WoS format) was the worst possible because RIS files don't contain the IDs and references.

Examples of bibliographic files exported from Scopus Truss.ris and Truss.csv on the topic of *k*-truss.

Scopus
Unusual distributions in Nataliya's data





The Hong Kong University of Science and Technology (HKUST) 2017–2019

5215 co-authors/participation in very large international projects.: The ATLAS collaboration, & The CMS collaboration (2019). Combinations of single-top-quark production cross-section measurements and $|f\ LV\ Vtb|$ determinations at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS experiments. Journal of High Energy Physics, 2019(5), Article 88. https://doi.org/10.1007/JHEP05(2019)088 / Springer

Multi-persons or hyperproductive labs?

Resources

A selection of some additional resources on the web that may come in handy in solving the problem of identifying bibliographic units.

DBLP computer science bibliography.
DBLP authors; How does dblp handle homonyms and synonyms?; dblp and ORCID 2020; Name disambiguation suffixes in dblp

Similar for mathematics **MathSciNet** and **zbMATH-Open**.
The Identification of Authors in the MR Database+2011; Author disambiguation.

VIAF Virtual International Authority File.
VIAF authors

Wikidata the free knowledge base.
Wikidata authors; How does Wikidata work?

schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the web.
Person; Book;

. . . Resources

Crossref makes research objects easy to find, cite, link, assess, and reuse.
crossref/rest-api; Crossref community

SCImago is both a Spanish consulting company (Scimago Lab) and a network of research groups.
Scimago journals; Assessing universities and other research-focused institutions

Citations from **Google** – use “
Google scholar; Google books; Citing

The multipersonality's effect on the results of bibliographic analyses

We would like to study the effect of multipersons in derived networks [2].

Let $\mathbf{M} = [m[u, v]]$ is a matrix on $U \times V$ and $\mathbf{C}_U = \{C_1, C_2, \ldots, C_k\}$ a partition of the set U, $\emptyset \subset C_i \subseteq U$ and $\bigcup_i C_i = U$. The set U is the (ground truth) set of real units (persons). The partition $\mathbf{C}_U$ corresponds to units (for example authors) identified by the network construction process. A cluster $C \in \mathbf{C}_U$ with $|C| > 1$ represents a multi-unit; and for $|C| = 1$ a correctly identified unit.

We introduce the *shrinking* transformation S of matrix $\mathbf{M}$ by partition $\mathbf{C}_U$ into $S(\mathbf{M}, \mathbf{C}_U) = \mathbf{S} = [s[C, v]]$ on $\mathbf{C}_U \times V$ determined by the rule

$$s[C, v] = \sum_{u \in C} m[u, v]$$

The shrinking transformation can be extended to a partition $\mathbf{C}_V$ of the set V by

$$s[u, C] = \sum_{v \in C} m[u, v]$$

Multiunits' effect

or

$$S(\mathbf{M}, \mathbf{C}_V) = S(\mathbf{M}^T, \mathbf{C}_V)^T$$

and to partitions $\mathbf{C}_U$ and $\mathbf{C}_V$ of both sets by

$$S(\mathbf{M}, (\mathbf{C}_U, \mathbf{C}_V)) = S(S(\mathbf{M}, \mathbf{C}_U), \mathbf{C}_V)$$

Consider now the case of two compatible matrices $\mathbf{M} = [m[u,t]]$ on $U \times T$ and $\mathbf{N} = [n[t,v]]$ on $T \times V$. For a partition $\mathbf{C}_U$ of the set U it holds

$$S(\mathbf{M} \cdot \mathbf{N}, \mathbf{C}_U) = S(\mathbf{M}, \mathbf{C}_U) \cdot \mathbf{N}$$

To check this let's denote with $\mathbf{L}$ and $\mathbf{R}$ the left and right side of this expression. We have

$$l[C,v] = \sum_{u \in C} \mathbf{M} \cdot \mathbf{N}[u,v] = \sum_{u \in C} \sum_{t \in t} m[u,t] \cdot n[t,v]$$

and

$$r[C,v] = \sum_{t \in T} S(\mathbf{M}, \mathbf{C}_U)[C,t] \cdot n[t,v] = \sum_{t \in T} (\sum_{u \in C} m[u,t]) \cdot n[t,v] = l[C,v]$$

Multiunits' effect

For the partition $\mathbf{C}_V$ of the set V we get

$$S(\mathbf{M}\cdot\mathbf{N},\mathbf{C}_V) = S((\mathbf{M}\cdot\mathbf{N})^T,\mathbf{C}_V)^T = S(\mathbf{N}^T\cdot\mathbf{M}^T,\mathbf{C}_V)^T =$$

$$= (S(\mathbf{N}^T,\mathbf{C}_V)\cdot\mathbf{M}^T)^T = \mathbf{M}\cdot S(\mathbf{N}^T,\mathbf{C}_V)^T = \mathbf{M}\cdot S(\mathbf{N},\mathbf{C}_V)$$

Therefore

$$S(\mathbf{M}\cdot\mathbf{N},\mathbf{C}_V) = \mathbf{M}\cdot S(\mathbf{N},\mathbf{C}_V)$$

For partitions of both sets U and V we have

$$S(\mathbf{M}\cdot\mathbf{N},(\mathbf{C}_U,\mathbf{C}_V)) = S(S(\mathbf{M}\cdot\mathbf{N},\mathbf{C}_U),\mathbf{C}_V) =$$

$$= S(S(\mathbf{M},\mathbf{C}_U)\cdot\mathbf{N},\mathbf{C}_V) = S(\mathbf{M},\mathbf{C}_U)\cdot S(\mathbf{N},\mathbf{C}_V)$$

and finally

$$S(\mathbf{M}\cdot\mathbf{N},(\mathbf{C}_U,\mathbf{C}_V)) = S(\mathbf{M},\mathbf{C}_U)\cdot S(\mathbf{N},\mathbf{C}_V)$$

Multiunits' effect

For $C_u \in \mathbf{C}_U$ and $C_v \in \mathbf{C}_V$ we have

$$S(\mathbf{M} \cdot \mathbf{N}, (\mathbf{C}_U, \mathbf{C}_V))[C_u, C_v] = S(S(\mathbf{M} \cdot \mathbf{N}, \mathbf{C}_U), \mathbf{C}_V)[C_u, C_v] =$$

$$\sum_{z \in C_v} S(\mathbf{M} \cdot \mathbf{N}, \mathbf{C}_U)[C_u, z] = \sum_{z \in C_v} \sum_{w \in C_u} \mathbf{M} \cdot \mathbf{N}[w, z] = \sum_{w \in C_u} \sum_{z \in C_v} \mathbf{M} \cdot \mathbf{N}[w, z]$$

In a special case of singelton clusters $C_u = \{u\}$ and $C_v = \{v\}$ we get

$$S(\mathbf{M} \cdot \mathbf{N}, (\mathbf{C}_U, \mathbf{C}_V))[\{u\}, \{v\}] = \sum_{w \in \{u\}} \sum_{z \in \{v\}} \mathbf{M} \cdot \mathbf{N}[w, z] = \mathbf{M} \cdot \mathbf{N}[u, v]$$

We see that **the multi-units don't affect the values of relations between singletons in the derived networks**.

Conclusions

- importance of unique identifiers
- different publication cultures in different disciplines

Acknowledments

This work is supported in part by the Slovenian Research Agency (research program P1-0294, research program CogniCom (0013103) at the University of Primorska, and research projects J5-2557, J1-2481, and J5-4596), and prepared within the framework of the COST action CA21163 (HiTEc).

References I

 Anne-Wil Harzing: harzing.com

 Anne-Wil Harzing: Health warning: Might contain multiple personalities. The problem of homonyms in Thomson Reuters Essential Science Indicators. Scientometrics 105(3):2259-2270. paper