Field-normalized scores based on Web of Science and Microsoft Academic data

A case study in computer sciences

Thomas Scheidsteger
Robin Haunschild
Sven Hug
Lutz Bornmann

SCIENTIFIC FACILITY
INFORMATION SERVICE CPT
MAX PLANCK INSTITUTE FOR SOLID STATE RESEARCH

September 14, 2018 – STI 2018, Leiden, The Netherlands

Outline

Motivation

Data Set for Case Study

Normalized Citation Counts & Statistical Measures

Summary & Outlook

Motivation to investigate Microsoft Academic (MA)

Promising new data source for evaluative bibliometrics

- size: currently more than 200 million documents
- functionality
  - free access to Web-GUI
  - inexpensive access to API
  - inexpensive access to Data Dump
  - search in several metadata
- *citation counts comparable* to Scopus, between WoS and Google Scholar
- only *one small study* using *normalized* data (Hug & Brandle, 2017), pointing out difficulties with field attributes
  - dynamic
  - fine-grained
  - incoherent hierarchy

Motivation to investigate Microsoft Academic (MA)

Promising new data source for evaluative bibliometrics

- size: currently more than 200 million documents
- functionality
  - free access to Web-GUI
  - inexpensive access to API
  - inexpensive access to Data Dump
  - search in several metadata
- *citation counts comparable* to Scopus, between WoS and Google Scholar
- only *one small study* using *normalized* data (Hug & Brandle, 2017), pointing out difficulties with field attributes
  - dynamic
  - fine-grained
  - incoherent hierarchy

Motivation to investigate Microsoft Academic (MA)

Promising new data source for evaluative bibliometrics

- size: currently more than 200 million documents
- functionality
  - free access to Web-GUI
  - inexpensive access to API
  - inexpensive access to Data Dump
  - search in several metadata
- *citation counts comparable* to Scopus, between WoS and Google Scholar
- only *one small study* using *normalized* data (Hug & Brandle, 2017), pointing out difficulties with field attributes
  - dynamic
  - fine-grained
  - incoherent hierarchy

Motivation to investigate Microsoft Academic (MA)

Promising new data source for evaluative bibliometrics

- size: currently more than 200 million documents
- functionality
  - free access to Web-GUI
  - inexpensive access to API
  - inexpensive access to Data Dump
  - search in several metadata
- *citation counts comparable* to Scopus, between WoS and Google Scholar
- only *one small study* using *normalized* data (Hug & Brandle, 2017), pointing out difficulties with field attributes
  - dynamic
  - fine-grained
  - incoherent hierarchy

Research Question

Research Question

Is it possible to calculate

- *field-normalized* citation scores in MA
- in *good agreement* with those
- from *established databases* as WoS?

Outline

Motivation

Data Set for Case Study

Normalized Citation Counts & Statistical Measures

Summary & Outlook

Choice of Data Set for Case Study

German Computer Science Institute

- comprehensive publication list on the web page
  - **2157** papers between *2005 and 2010*
- supposedly better coverage in MA than in WoS
- only restricted number of research fields

Search in WoS

Source: WoS in-house database

- maintained by the Max Planck Digital Library, Munich
- derived from SCI-E, SSCI, and AHCI (Clarivate Analytics)
- *address information for German research institutes and universities disambiguated and unified* by Competence Centre for Bibliometrics (CCB)

Data Set in WoS

- 1141 papers (52.9%) from the institute found in the CCB data alone.
- 51 further papers found by additional address search
- All **1192** papers **(55.3%)** have *at least one WoS subject category* – attached to the resp. *journals* and used for *field-normalization*.

Search in WoS

Source: WoS in-house database

- maintained by the Max Planck Digital Library, Munich
- derived from SCI-E, SSCI, and AHCI (Clarivate Analytics)
- *address information for German research institutes and universities disambiguated and unified* by Competence Centre for Bibliometrics (CCB)

Data Set in WoS

- 1141 papers (52.9%) from the institute found in the CCB data alone.
- 51 further papers found by additional address search
- All **1192** papers **(55.3%)** have *at least one WoS subject category* – attached to the resp. *journals* and used for *field-normalization*.

Search in MA

Source: MA Data Dump of 165 million documents from August 2017

- imported and processed in locally maintained database
- about two thirds of them have a *Field of Study – algorithmically assigned on a per paper basis*

Data Set in MA

- refined address search with 14 different truncated address variants of the institute (13 false positive papers manually removed)
- total set of **2131** papers **(98.8%)** from the institute

Search in MA

Source: MA Data Dump of 165 million documents from August 2017

- imported and processed in locally maintained database
- about two thirds of them have a *Field of Study* – *algorithmically assigned on a per paper basis*

Data Set in MA

- refined address search with 14 different truncated address variants of the institute (13 false positive papers manually removed)
- total set of **2131** papers **(98.8%)** from the institute

Fields of Study in MA

Hierarchy of four levels (meanwhile two more)

- Level 0 (L0): 19
- Level 1 (L1): 290
- Level 2 (L2): 1490
- Level 3 (L3): 49531

Choosing L1

- compromise: granularity of the FoS
  vs. #publications per (FoS, PY).
- **290 L1 FoS vs. 262 WoS subject categories.**
- **1714 papers (80.4%)** of the institute **with at least one L1 FoS.**

Fields of Study in MA

Hierarchy of four levels (meanwhile two more)

- Level 0 (L0): 19
- Level 1 (L1): 290
- Level 2 (L2): 1490
- Level 3 (L3): 49531

Choosing L1

- compromise: granularity of the FoS
  vs. #publications per (FoS, PY).
- **290 L1 FoS vs. 262 WoS subject categories.**
- **1714 papers (80.4%)** of the institute **with at least one L1 FoS**.

Consolidated dataset used in this study

Match of institute’s papers via DOI

- 1379 papers (64.7%) with DOI in MA
- 622 (28.8%) with DOI in WoS
- **442 papers (20.5%)** could be matched
- **all** matched papers have **at least one L1 FoS**,

Affiliation check by random samples of 10%

- *none* of the matched papers incorrectly affiliated
- only 1% of the unmatched papers incorrectly affiliated

Consolidated dataset used in this study

Match of institute's papers via DOI

- 1379 papers (64.7%) with DOI in MA
- 622 (28.8%) with DOI in WoS
- **442 papers (20.5%)** could be matched
- **all** matched papers have **at least one L1 FoS**,

Affiliation check by random samples of 10%

- *none* of the matched papers incorrectly affiliated
- only 1% of the unmatched papers incorrectly affiliated

Outline

Motivation

Data Set for Case Study

Normalized Citation Counts & Statistical Measures

Summary & Outlook

NCS

Normalized Citation Score

$$NCS = \frac{c_i}{e_i}$$

- c_i: citation count of a focal paper,
- e_i: corresponding average citation count in the scientific field and publication year
  - **MA: L1 FoS**
  - **WoS: subject category**
  - **citations counted until end of 2016**
- NCS_{MA}:= arithmetic average over MA FoS
- NCS_{WoS}:= arithmetic average over WoS subject categories

NCS

Normalized Citation Score

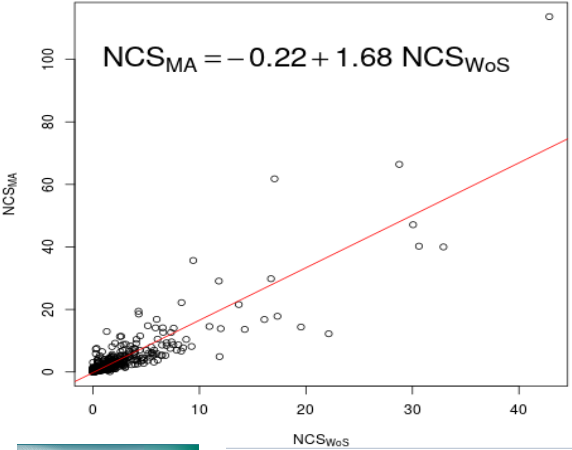
$$NCS = \frac{c_i}{e_i}$$

- c_i: citation count of a focal paper,
- e_i: corresponding average citation count in the scientific field and publication year
  - **MA: L1 FoS**
  - **WoS: subject category**
  - **citations counted until end of 2016**
- NCS_{MA}:= arithmetic average over MA FoS
- NCS_{WoS}:= arithmetic average over WoS subject categories

Correlation of NCS_{MA} and NCS_{WoS}

Correlation coefficients confirm linear relationship

- Pearson: $r_p = 0.87$ (Spearman: $r_s = 0.84$)



Concordance aka Reproducibility

Lin's concordance correlation coefficient

- for agreement on a continuous measure
- $\Rightarrow$ reproducibility of both scores

$r_{ccc} = 0.69[0.66, 0.72]$

- indicates a *strong* agreement (0.61-0.80)
  - according to Koch and Sporl (2007)
- both NCS show similar citation impact results

Concordance aka Reproducibility

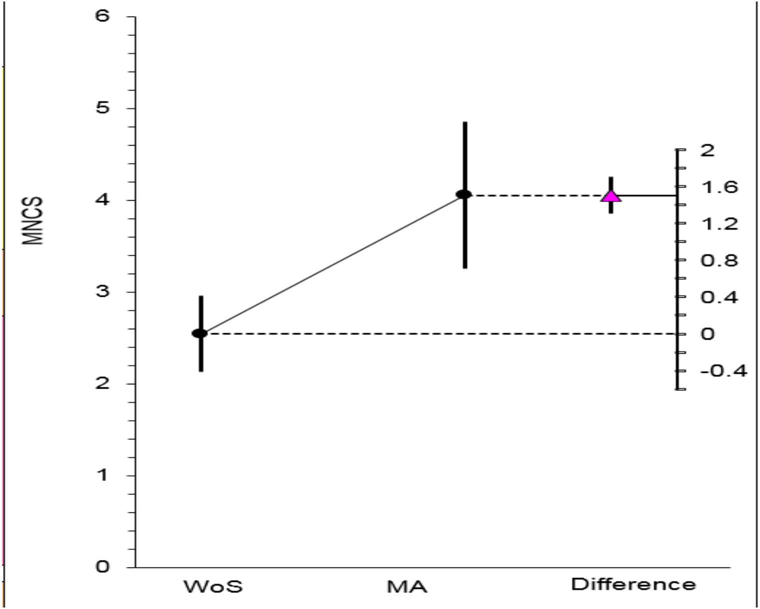
Lin’s concordance correlation coefficient

- for agreement on a continuous measure
- $\Rightarrow$ reproducibility of both scores

$r_{ccc} = 0.69[0.66, 0.72]$

- indicates a *strong* agreement (0.61-0.80)
  - according to Koch and Sporl (2007)
- both NCS show similar citation impact results

Mean of NCS (*paired design*, Cumming, 2012)



Mean of NCS - cont.

Difference between NCS_{MA} and NCS_{WoS}: 1.3 to 1.7

Proposed explanation:
field-specific citation rate e_i systematically lower for NCS_{MA} by inclusion of lesser cited document types and languages

Manually check random samples of 10%

	all DOI papers		DOI-matched papers	
Document Type	Publisher	MA	Publisher	MA
Conference Proc	**52%**	16%	**9%**	5%
Journal	44%	44%	91%	89%
Book	4%	-	-	-

English papers: only two thirds in our FoS

Mean of NCS - cont.

Difference between NCS_{MA} and NCS_{WoS}: 1.3 to 1.7

Proposed explanation:
field-specific citation rate e_i systematically lower for NCS_{MA}
by inclusion of lesser cited document types and languages

Manually check random samples of 10%

	all DOI papers		DOI-matched papers	
Document Type	Publisher	MA	Publisher	MA
Conference Proc	**52%**	16%	**9%**	5%
Journal	44%	44%	91%	89%
Book	4%	-	-	-

English papers: only two thirds in our FoS

Mean of NCS - cont.

Difference between NCS_{MA} and NCS_{WoS}: 1.3 to 1.7

Proposed explanation:
field-specific citation rate e_i systematically lower for NCS_{MA}
by inclusion of lesser cited document types and languages

Manually check random samples of 10%

	all DOI papers		DOI-matched papers	
Document Type	Publisher	MA	Publisher	MA
Conference Proc	**52%**	16%	**9%**	5%
Journal	44%	44%	91%	89%
Book	4%	-	-	-

English papers: only two thirds in our FoS

Agreement between NCS_{MA} and NCS_{WoS}

Characteristic Scores and Scales (CSS) by Glanzel et al. (2016)

- 4x4-Contingency Table

		NCS_{MA}			
		poorly cited	fairly cited	remarkably cited	outstandingly cited
NCS_{WoS}	poorly cited	**291**	23	**1**	0
	fairly cited	32	**50**	8	0
	remarkably cited	0	13	**7**	2
	outstandingly cited	0	0	4	**7**

- Agreement (= share of diagonal entries): **81%**
- only 1 paper (0.2%) more than one class apart

Outline

Motivation

Data Set for Case Study

Normalized Citation Counts & Statistical Measures

Summary & Outlook

Summary & Conclusion

Summary

- Focusing on *journal papers* only, we compared *field-normalized scores* based on WoS resp. MA for an anonymous computer science institute.
- ⇒ substantial correlation of both scores ($r_p, r_s > 0.8$)
- ⇒ substantial Lin's concordance $r_{ccc} \sim 0.7$
- ⇒ significantly higher impact of paper set in MA, probably due to inclusion of lesser cited document types
- ⇒ CSS show high level of agreement in all four classes

Conclusion

It **is possible and reasonable** to calculate **field-normalized citations scores from FoS (L1) in MA** in good agreement with the resp. scores based on WoS subject categories.

Summary & Conclusion

Summary

- Focusing on *journal papers* only, we compared *field-normalized scores* based on WoS resp. MA for an anonymous computer science institute.
- ⇒ substantial correlation of both scores ($r_p, r_s > 0.8$)
- ⇒ substantial Lin's concordance $r_{ccc} \sim 0.7$
- ⇒ significantly higher impact of paper set in MA, probably due to inclusion of lesser cited document types
- ⇒ CSS show high level of agreement in all four classes

Conclusion

It **is possible and reasonable** to calculate **field-normalized citations scores from FoS (L1) in MA** in good agreement with the resp. scores based on WoS subject categories.

Limitations & Outlook

Limitations

- Computer Science only
- papers with DOI only
- no distinction of document types

Outlook

- apply more comprehensive *paper matching* procedures
- compare also with *Scopus*
- evaluate separately according to *document type* - as far as available in MA - currently and in the future
- for a fairer comparison with WoS focus on *other subject fields*
- . . .

Limitations & Outlook

Limitations

- Computer Science only
- papers with DOI only
- no distinction of document types

Outlook

- apply more comprehensive *paper matching* procedures
- compare also with *Scopus*
- evaluate separately according to *document type* - as far as available in MA - currently and in the future
- for a fairer comparison with WoS focus on *other subject fields*
- ...

References

- CCB: http://www.bibliometrie.info.
- Cumming, G. (2012). *Understanding the new statistics: effect sizes, confidence intervals, and meta-analysis.* London, UK: Routledge.
- Glanzel, W., Debackere, K., & Thijs, B. (2016). *Citation classes: a novel indicator base to classify scientific output.*
- Hug, S.E, Brandle, M.P. (2017). *The coverage of Microsoft Academic: analyzing the publication output of a university*, Scientometrics 113:1551-1571, doi: 10.1007/s11192-017-2535-3
- Lin, L. I. (1989). *A concordance correlation-coefficient to evaluate reproducibility.* Biometrics, 45(1), 255-268. doi: 10.2307/2532051.
- Lin, L. I. (2000). *A Note on the Concordance Correlation Coefficient.* Biometrics, 56(1), 324-325. doi: 10.1111/j.0006-341X.2000.00324.x.
- Liu, J., Tang, W., Chen, G., Lu, Y., Feng, C., Tu, X.M. (2016). *Correlation and agreement: overview and clarification of competing concepts and measures.* Shanghai Archives of Psychiatry, 28(2), 6. doi: 10.11919/j.issn.1002-0829.216045.
- Koch, R., & Sporl, E. (2007). *Statistical methods for comparison of two measuring procedures and for calibration: Analysis of concordance, correlation and regression in the case of measuring intraocular pressure.* Klinische Monatsblatter Fur Augenheilkunde, 224(1), 52-57. doi: 10.1055/s-2006-927278.
- Microsoft Academic Graph; download from https://aminer.org/open-academic-graph on August 15, 2017

53 23 16
IVS