A New Semantic Similarity Based Measure for Assessing Research Contribution

Petr Knoth & Drahomira Herrmannova

Knowledge Media institute, The Open University

Current impact metrics

- Pros: simplicity, availability for evaluation purposes
- Cons: insufficient evidence of quality and research contribution

Problems of current impact metrics

- Sentiment, semantics, context and motives [Nicolaisen, 2007]
- Popularity and size of research communities [Brumback, 2009; Seglen, 1997]
- Time delay [Priem and Hemminger, 2010]
- Skewness of the distribution [Seglen, 1992]
- Differences between types of research papers [Seglen, 1997]
- Ability to game/manipulate citations [Arnold and Fowler, 2010; Editors, 2006]

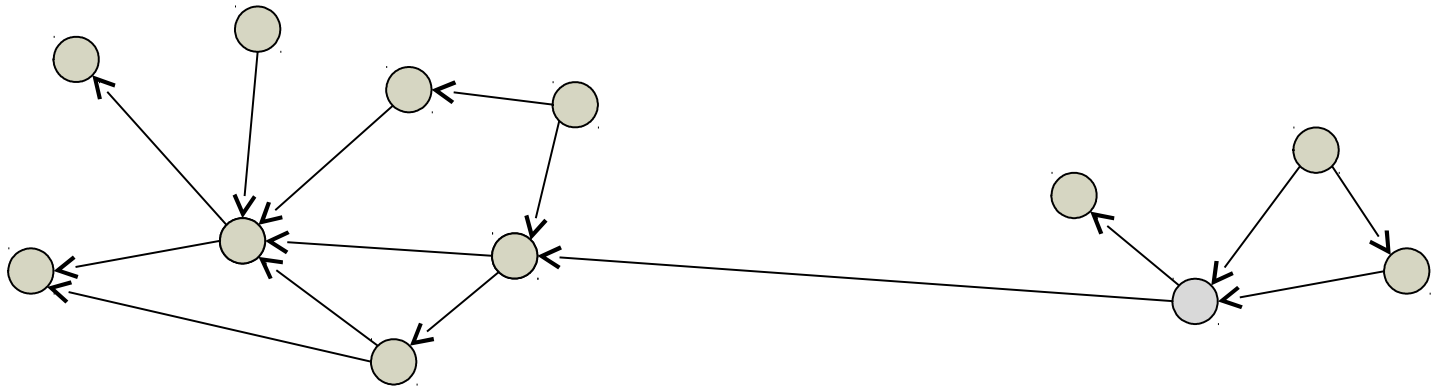
Alternative metrics

- Alt-/Webo-metrics etc.
  - Impact still dependent on the number of interactions in a scholarly communication network
- Full-text (**Semantometrics**)
  - Contribution to the discipline dependent on the content of the manuscript.

Approach

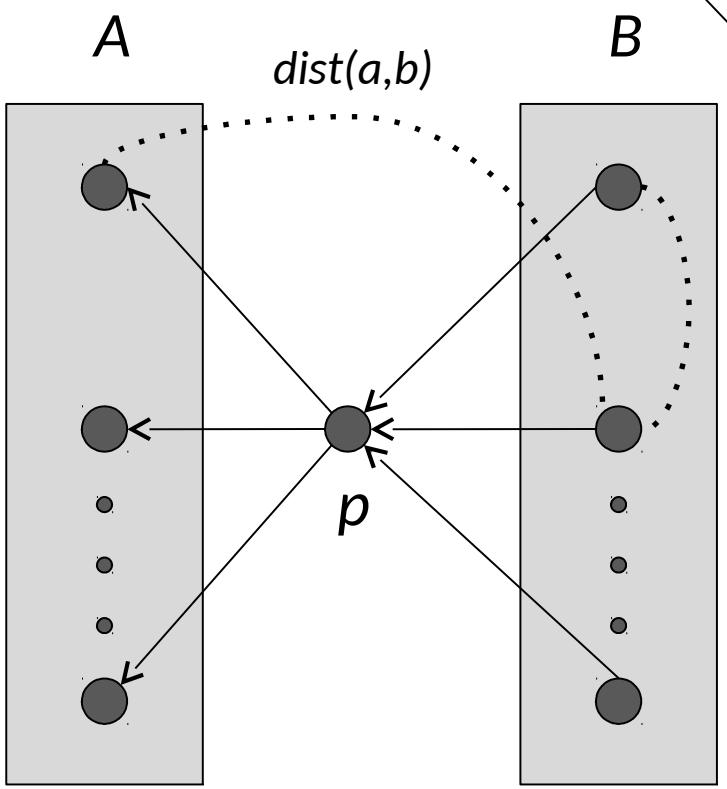
Premise: Full-text needed to assess publication's research contribution.

Hypothesis: Added value of publication *p* can be estimated based on the semantic distance from the publications cited by *p* to publications citing *p*.



Contribution measure

Average distance of the set members



$$Contribution(p) = \frac{\overline{B}}{\overline{A}} \times \frac{1}{|B| \times |A|} \times \sum_{a \in A, b \in B, a \neq b} dist(a,b)$$

$$\overline{X} = \begin{cases} 1 & |A|=1 \vee |B|=1 \\ \frac{1}{|X|(|X|-1)} \times \sum_{x_1 \in X, x_2 \in X, x_1 \neq x_2} dist(x_1, x_2) & |A|>1 \wedge |B|>1 \end{cases}$$

dist(b1,b2)

$$dist(a,b) = 1 - sim(a,b)$$

Datasets

- Requirements
  - Availability of full-text
  - Density
  - Multidisciplinarity

Datasets (present as table)

- Examined datasets
  - CORE
  - Open Citation Corpus
  - ACM Dataset
  - DBLP+Citation
  - KDD Cup Dataset
  - iSearch Collection
- However…

- TABLE

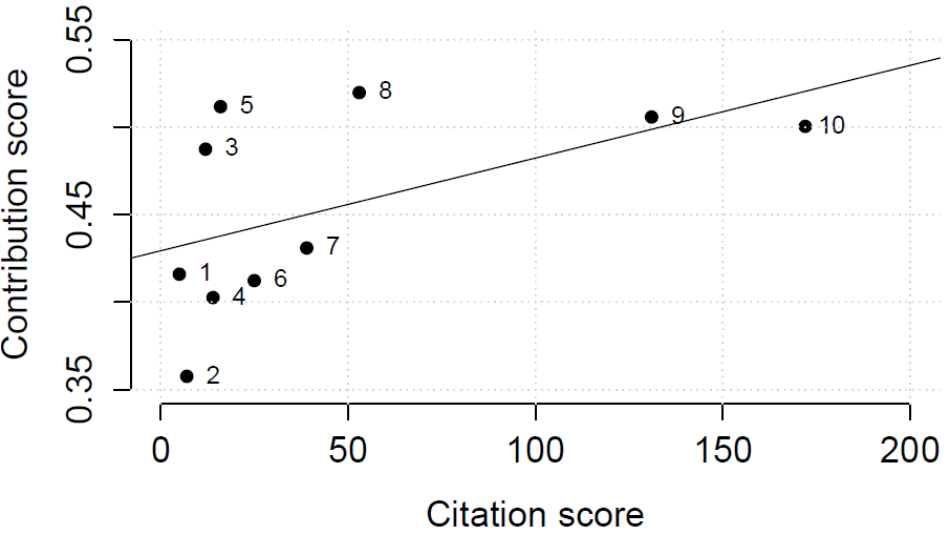
Our dataset

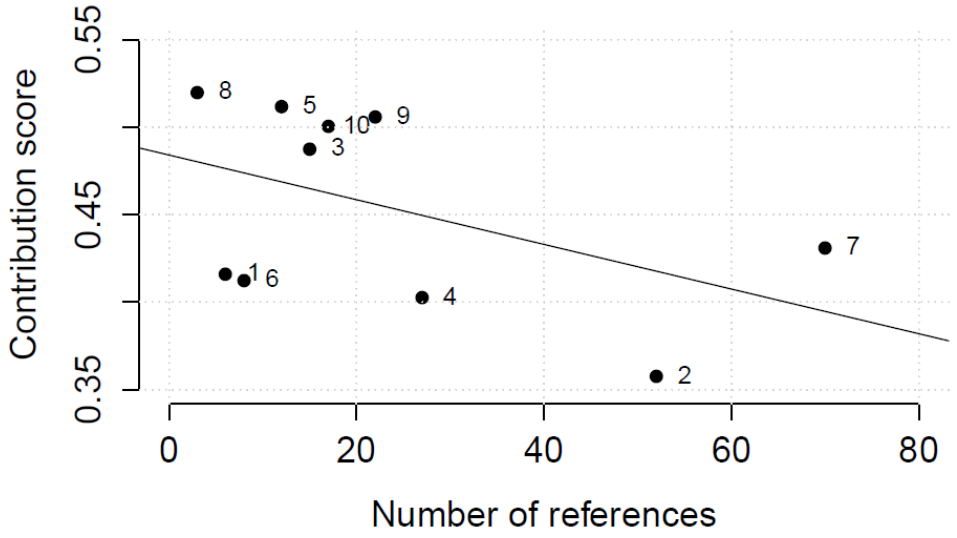
- 10 seed publications from CORE with varying level of citations
- missing citing and cited publications downloaded manually
- only freely accessible English documents were downloaded
- in total 716 documents (~50% of the complete network)
- 2 days to gather the data

Results

Publication no.	|B| (Citation score)	|A| (No. of references)	Contribution
1	5 (9)	6 (8)	0.4160
2	7 (11)	52 (93)	0.3576
3	12 (20)	15 (31)	0.4874
4	14 (27)	27 (72)	0.4026
5	16 (30)	12 (21)	0.5117
6	25 (41)	8 (13)	0.4123
7	39 (71)	70 (128)	0.4309
8	53 (131)	3 (10)	0.5197
9	131 (258)	22 (32)	0.5058
10	172 (360)	17 (20)	0.5004
	474 (958)	232 (428)	

Results


Contribution score
0.35
0.45
0.55
0
50
100
150
200
Citation score
1
2
3
4
5
6
7
8
9
10


Contribution score
0.35
0.45
0.55
0
20
40
60
80
Number of references
1
2
3
4
5
6
7
8
9
10

Current impact metrics vs Semantometrics

Unaffected by, CROSS (red), TICK (green) ✔

- Sentiment, semantics, context and motives ✔
- Popularity and size of research communities ✔
- Time delay [Reduced to 1 citation] ✔
- Skewness of the distribution ✔
- Differences between types of research papers
- Ability to game/manipulate citations [solved providing that self-citations not allowed] ✔

TABLE

Conclusions

- Full-text necessary
- Semantometrics are a new class of methods.
- We showed one method to assess the research contribution

References

- Jeppe Nicolaisen. 2007. Citation Analysis. Annual Review of Information Science and Technology, 41(1):609-641.
- Douglas N Arnold and Kristine K Fowler. 2010. Nefarious numbers. Notices of the American Mathematical Society, 58(3):434-437.
- Roger A Brumback. 2009. Impact factor wars: Episode V -- The Empire Strikes Back. Journal of child neurology, 24(3):260-2, March.
- The PLoS Medicine Editors. 2006. The impact factor game. PLoS medicine, 3(6), June.

References

- Jason Priem and Bradely M. Hemminger. 2010. Scientometrics 2.0: Toward new metrics of scholarly impact on the social Web. First Monday, 15(7), July.
- Per Ottar Seglen. 1992. The Skewness of Science. Journal of the American Society for Information Science, 43(9):628-638, October.
- Per Ottar Seglen. 1997. Why the impact factor of journals should not be used for evaluating research. BMJ: British Medical Journal, 314(February):498-502.