Drahomira Herrmannova (drahomira.herrmannova@open.ac.uk) & Petr Knoth (petr.knoth@open.ac.uk)

Semantometrics: Towards Fulltext-based Research Evaluation

Q: Would you rate the **quality** of a movie based **only** on the number of views?

Aim

To understand the properties and behaviour of the semantometric *contribution* measure, which uses semantic similarity of publications to estimate research contribution, in comparison with established research evaluation metrics.

Semantometrics are a new class of research evaluation metrics which build on the premise that full-text is needed to assess the value of a publication.

Semantometrics

Contribution to the discipline assessed by using the article manuscript.

Problems of citation-based measures

- Sentiment, semantics, context and motives [Nicolaisen, 2007]
- Popularity and size of research communities [Brumback, 2009; Seglen, 1997]
- Differences between types of research papers [Seglen, 1997]
- Require complete data (inconsistency across systems)
- …

Possibilities for semantometrics

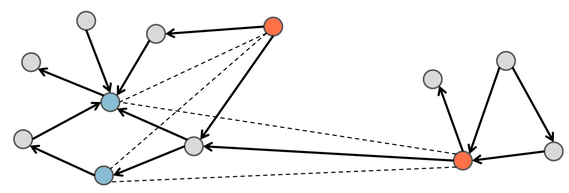
- Detecting good research practices were followed (sound methodology, research data/code shared …)
- Detecting paper type …
- Analysing citation contexts (tracking facts propagation) …
- Detecting the sentiment of citations …
- Normalising by size of community that is likely to read the research …
- …

Alternative metrics

- Alt-/Webo-metrics etc.
  - Impact still dependent on the number of interactions in a scholarly communication network (downloads, views, readers, tweets, etc.)

Semantometric 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*.

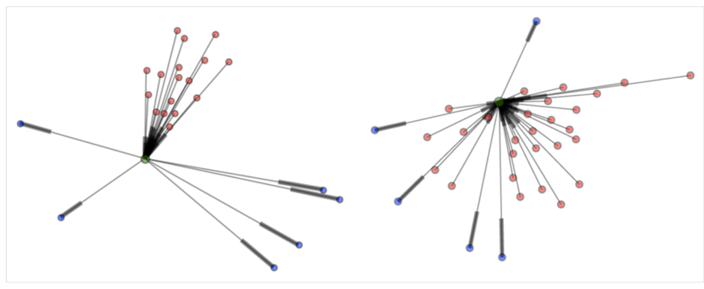


Semantometric contribution

- Based on semantic distance between citing and cited publications
  - Cited publications – state-of-the-art in the domain of the publication in question
  - Citing publications – areas of application

Practical example

- Below- and above-average publication in terms of *contribution* value



Experiment

- Evaluation of the contribution measure in comparison with established research evaluation metrics
  - Citation counts obtained from the Microsoft Academic Graph (MAG) (bibliometric data)
  - Usage data (readership) obtained from Mendeley (altmetric data)
  - Research articles aggregated by the Open Access Connecting Repositories (CORE) system (representative sample for the study)

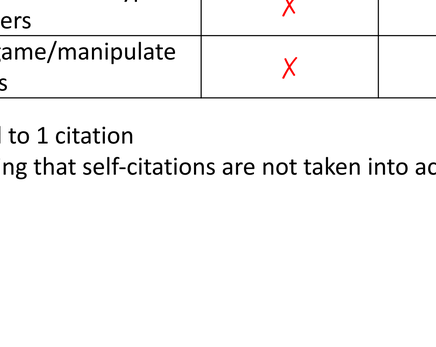
Dataset statistics

Articles from CORE matched with MAG	**1,655,835**
Average number of received citations	16.09
Standard deviation	66.30
Max number of received citations	13,979
Average readership	15.94
Standard deviation	42.17
Max readership	15,193
Average contribution value	0.89
Standard deviation	0.0810
Total number of publications	**12,075,238**

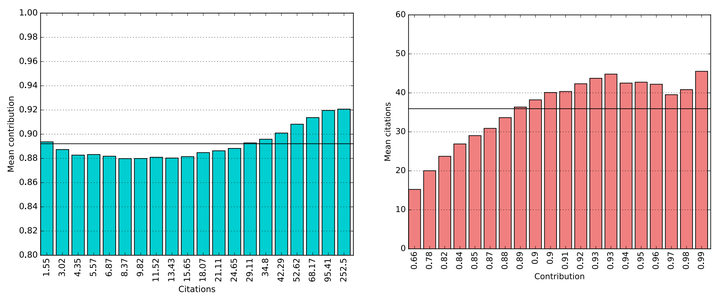
Experiment – results

- No direct correlation between contribution measure and citations/readership
- When working with mean citation, readership and contribution values a clear behavioral trend emerges

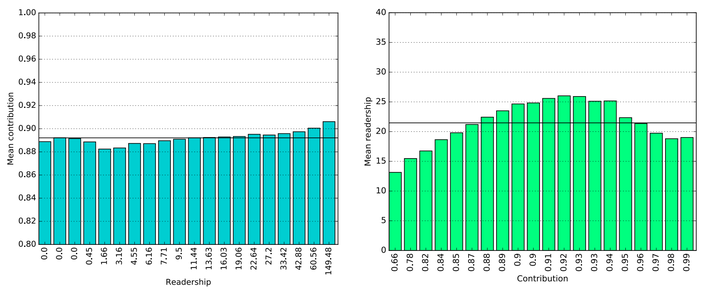
Relation between citations and readership



Relation between mean contribution and citations



Relation between mean contribution and readership



Current impact metrics vs semantometrics

Unaffected by	**Current impact metrics**	**Semantometrics**
Citation sentiment, semantics, context, motives	X	
Popularity & size of res. communities	X	
Time delay	X	X/ *
Skewness of the citation distribution	X	
Differences between types of res. papers	X	
Ability to game/manipulate the metrics	X	X/ **

* reduced to 1 citation

** assuming that self-citations are not taken into account

References

Xiaolin Shi, Jure Leskovec, and Daniel A Mcfarland (2010) *Citing for High Impact.*

M. E. J. Newman (2004) *Coauthorship networks and patterns of scientific collaboration.*

R. Lambiotte and P. Panzarasa (2009) *Communities, know- ledge creation, and information diffusion*.