Web data management : bad academic practices in business study

Priyanka Aravindan and Simon Coumes

Overview

Data sources

Articles and citations :

- Arxiv
- Google scholar
- Semantic scholar

Affiliations :

- DBLP
- Google scholar

Due to issues (mostly linked with speed), we were unable to get researchers affiliations.

The **Semantic scholar** database gave us articles and citation.

Measures

We want to look at :

- Citation rings
- Parasites adding their name to too many papers
- Excessive self citation

The semantic scholar dataset

Size and limitations :

- about 150 GB of metadata
- 10 GB left after filtering only business articles
- fast queries $\rightarrow$ we can download the whole dataset

Structure

- List of metadata per article
- All outbound and inbound citations
- Field
- Author ID

Filter the data

A first step was to filter the data and restrain ourselves to articles in the field of business. Total new size : 10 GB.
It is no longer true that all citations point to an article in the database.

Preprocessing

We compute a few useful lookup tables in advance.

Notably, precompute the matrix of citations from A to B for every pair (A, B) of researchers.

Citation rings

We define a citation ring as a set of 3 authors such that, for each of these authors, most of their citations comme from the others.
Consider four time periods :

1. before 1990
2. 1990 - 2000
3. 2000 - 2010
4. after 2010

Citation rings : algorithm

- For each author A consider all pairs of other authors that together form more than half the citations received by A.
- Check if they form a citation ring.

citation rings : results

Time period	number of rings
before 1990	13
1990 - 2000	48
2000 - 2010	123
after 2010	286

No point in jumping to conclusions. We didn't investigate this.

Self citation

We look at the ration of self citation over total received citations per author. This is easy thanks to the previous preprocessing.

time period	more than half	mean self citation score	median
before 1990	1.6%	3.7%	0
1990 - 2000	0.9%	3%	0
2000 - 2010	0.5%	2.2%	0
after 2010	0.2%	1.2%	0

Parasites

Hello

Conclusion

- Data issues : no institution affiliations
- very heterogenous dataset
- Some suspect results