BRIDGING ANOTHER GAP BETWEEN RESEARCH ASSESSMENT AND INFORMATION RETRIEVAL

THE DELINEATION OF DOCUMENT ENVIRONMENTS

WOLFGANG GLÄNZEL & BART THIJS

Centre for R&D Monitoring and Dept MSI, KU Leuven, Belgium

ECOOM

Structure of presentation

1. INTRODUCTION
2. DOCUMENT-SPACE CLUSTERING VS. INDIVIDUAL ENVIRONMENTS
3. DATA SOURCE AND METHODOLOGY
4. RESULTS
5. CONCLUSIONS

The combination of cognitive mapping and citation analysis proved useful in several aspects.

Citation-based links are preferably used

- in the framework of subject classification.
- to support information retrieval
- to identify topic-related reference standards for citation analysis.

All three applications reflect *distinct tasks*, have *different features* and require their *specific methodologies*.

Introduction

In previous project we have dealt both with bibliographic-coupling based and hybrid clustering exercises for the cognitive clustering of large document spaces to assist the subject classification and to detect emerging research topics within given scientific disciplines, and to assist the retrieval of scientific documents.

The third task, which is related with, but not identical with subject classification, aims at the identification of proper reference standards for building and applying citation indicators.

- ☞ This is necessary because the scholarly communication patterns and practices may vary within the same subject and forcing a joint discipline standard based on cognitive assignment would not account for these peculiarities.

Introduction

Intuitively, two solutions offer themselves a possible approaches: The bottom-up solution as proposed by Schubert and Braun (1986) by using *related records* as reference standard, and the top-down solution suggested by Waltman and van Eck (2012) by clustering the document space "down" to the necessary micro-environments.

- ☞ Both methods are prone to biases and errors and tend to produce incomplete environments or insufficient document assignment.

In what follows, we will have a look at these effects and the prons and cons of the two approaches and then we will propose a new method to minimise the above-mentioned shortcomings.

Document-space clustering vs. individual environments

Top-down high-resolution clustering is very tempting. Fully automated processes could create hard or fuzzy but unlabelled micro environments, which is not problematic in the present context. However, there are serious drawbacks:

- Micro clusters form static environments and flexibly them is problematic.
- The large amount of singletons and minute clusters requires supplementary algorithms.

Document-space clustering vs. individual environments

Schubert and Braun (1993) paved the way to a new solution. They proceed from an experimental data set and applied bibliographic coupling as a local alternalive solution to the usual journal-based reference standards.

- About 15 years later, Zitt and Small (2008) and Moed (2010) proposed new impact measures for journals to eliminate the biases caused by the subject-specific citation behaviour.
- In 'source' or ´citing side' normalisation still the question of the citation window remains a critical issue.
- By contrast, BC-based records create flexible environments and do not need citations windows.

Document-space clustering vs. individual environments

Two issues remain:

1. Minute environments – with the effect that a document creates, so to speak, its own standard.
2. Largely overlapping environments that might require further micro-clustering.

In the present study we scope the properties of coupling-based environments in the context of citation indicators proceeding from Schubert's and Braun's idea and draft a solution for creating proper reference standards.

Document-space clustering vs. individual environments

Research question:

Can “related-record” based environments serve as a promising alternative to cognitive subject categories or direct citation links in providing reference standards for citation rates of individual documents and, if so, what are the limitations and possible solutions to overcome those?

Data source and methodology

The data sets:

1. All "citable" documents (about 1.4 million records) indexed in the 2013 volume of WoS.
2. 21,039 papers from 2013 with at least one Belgian address.
3. 11,514 documents (1998–2013) in the discipline of scientometrics.

Five networks were obtained:

- The first one is built on the complete 2013 document set.
- The other four networks can be considered as bi-partite networks with the primary nodes coming from the selected paper sets and the secondary nodes those that share at least one reference with a primary node (with same or different publication years of the secondary notes, respectively).

On the environments originated from these networks, we created expected citation rates (using a three year citation window).

1. For each node p:

$$ECRE_p = \frac{\sum_{t=1}^{s} bc_{pi} \cdot cit_i}{\sum_{t=1}^{s} bc_{pi}},$$

   bc_{pi} being the BC Salton-similarity based link strength between p and the secondary node i.

2.

$$ECRE = \sum_{j}^{p} ECRE_i \quad \text{and} \quad ENCR = \frac{\sum_{j}^{p} cit_j}{ECRE},$$

   where *ENCR* is the "Environment Normalized Citation Rate".

Using this approach, publications cannot contribute to their own reference score.

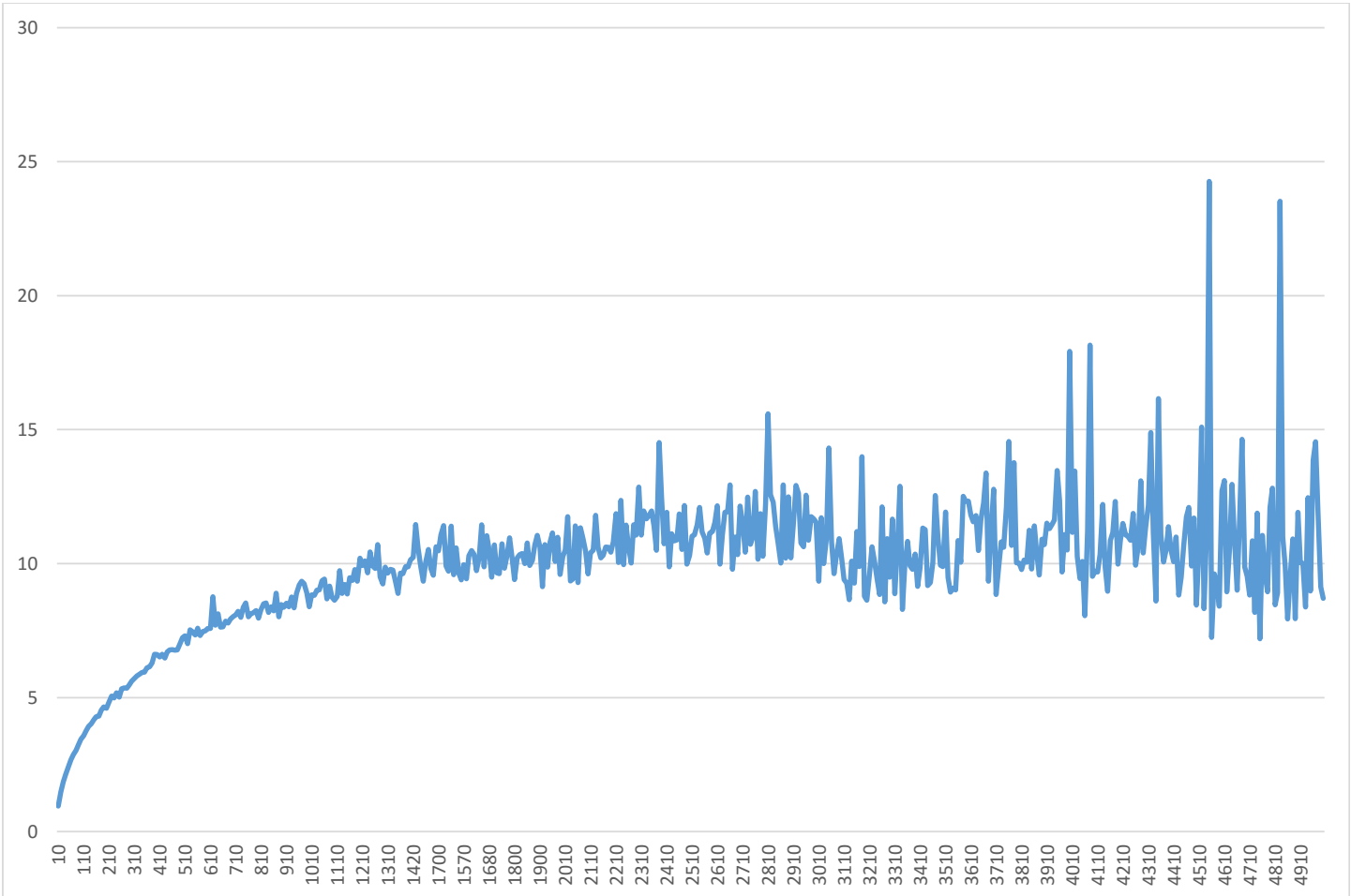
Number of documents in each of the five networks and share of publications with no or a small environment (<20)

Paper set	Papers	Share of singletons	Share of small environment papers
Papers in 2013	1,391,192	2.35%	10.74%
Belgium (links in 2013)	21,039	1.83%	
Belgium (all links)		1.02%	2.40%
SCIM (same year links)	11,514	4.75%	
SCIM (all links)		2.91%	6.05%

Source: WoS Core Collection

Results

Conditional Average Citations by size of environment



Source: WoS Core Collection

Results

Citation score, expected and normalized citation rates in the five networks

Paper set	Publications	Citations	NMCR	ECRE	ENCR
Papers in 2013	1,358,483	7,665,501	1.04	9,452,757.41	0.81
Belgium (links in 2013)	20,654	172,959	1.56	141,904.54	1.22
Belgium (all links)	20,824	173,029	1.55	117,844.99	1.47
SCIM (same year links)	10,967	45,319	1.23	50,481.65	0.90
SCIM (all links)	11,179	45,507	1.22	43,327.63	1.05

Source: WoS Core Collection

Results

Comparison of CSS-classes across networks and with the population standard

Paper set and reference		Class 1	Class 2	Class 3	Class 4
Papers in 2013	ECOOM	70.0%	21.4%	6.2%	2.4%
	2013 Links	73.8%	19.5%	4.4%	2.2%
Belgium	ECOOM	58.3%	27.1%	9.7%	4.9%
	2013 links	64.8%	25.0%	7.1%	3.1%
	all links	57.9%	29.6%	8.7%	3.7%
SCIM	ECOOM	59.8%	25.6%	9.7%	5.0%
	same year	68.9%	22.1%	6.3%	2.7%
	all links	65.9%	25.3%	6.5%	2.3%

Source: WoS Core Collection

Conclusions

Using related records helps bridge the gap between cognitive classification and "natural citation" environments.

Advantages:

- Very good "natural" standard for citation analysis, which uses cognitive links without the limitations of direct citations.
- No static solution , can be used for any publication period (one year or more) and allows for dynamic approaches.
- Fuzzy "classification" since environments may share documents. This is close to the traditional cognitive base expected from environments.

Conclusions

Nonetheless, biases encountered in previous approaches persist as they inherent in citation processes.

Limitations:

- A reasonable amount of documents has no or tiny environments, which are not suited to serve as reference standard.
- Large size-variety of environments: This might affect statistical reliability of expected citation rates.
- A number of documents is part of large or largely overlapping environments (*core documents*).

Conclusions

Possible improvements:

- Combination with text-based methods in applying "hybrid" environments (cf. Glänzel & This, *Scientometrics*, 2011; 2012).
- Additional weighting of citations by number of environments to which a publication contributes

Expected effects:

- Reduction of the number of tiny environments and of influence of core documents
- Possibly increase of discriminative power, while keeping the positive properties.

Thank you very much for your attention!