26_10_2020

Eligibility criteria - Finding Venues

SJR score and journals selection

In order to sort the selected journals (i.e. the journals that met the imposed threshold relative to the nr. of retrieved papers), the SJR indicator (averaged across the time period 2017-2021) was used.

The SJR2 indicator (Vicente et al., 2012) measures the prestige of a scientific journal over a journal citation network. In this network, nodes represent journals and edges the citation relationships between journals.

The score computation is divided into two phases.

Phase 1

At the start of the first phase, each journal in the network is assigned a starting prestige value of 1/N. The starting value is updated iteratively using the following formula:

$$PSJR2_i = \overbrace{\frac{(1-d-e)}{N}}^{1} + \overbrace{e \cdot \frac{Art_i}{\sum_{j=1}^{N} Art_j}}^{2} + \overbrace{\frac{d}{PSJR2D} \cdot \left[\sum_{j=1}^{N} Coef_{ji} \cdot PSJR2_j\right]}^{3}$$

Where:

- *d* is the constant 0.9
- *e* is the constant 0.0999
- *N* is the number of journals in the repository
- *Art_i* represents the number of citable primary documents in journal *i*

Additionally, the coefficient *Coef_j_i* is computed (before the beginning of each iteration) as follows:

$$Coef_{ji} = \frac{(Cos_{ji} \cdot C_{ji})}{\sum_{h=1}^{N}(Cos_{jh} \cdot C_{jh})}$$

Where:

- *C_j_i* represents the numbers of citations from journal j to journal i
- *Cos_j_i* is the cosine between cocitation profiles of journals j and I

Cocitation is defined as: "the frequency with which two documents are cited together" (Small, 1973).

The cosine of the cocitation profiles is expressed as follows:

$$Cos_{ij} = \frac{\sum_{h=1,h\neq i,h\neq j}^{N} Cocit_{ih} Cocit_{hj}}{\sqrt{\sum_{h=1,h\neq i,h\neq j}^{N}(Cocit_{ih})^2}\sqrt{\sum_{h=1,h\neq i,h\neq j}^{N}(Cocit_{jh})^2}}$$

Where:

- *Cocit_j_i* is the cocitation of journals j and i

Finally, the factor *PSJR2D* is expressed by the formula:

$$PSJR2D = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{(Cos_{ji} \cdot C_{ji})}{\sum_{h=1}^{N}(Cos_{jh} \cdot C_{jh})} \cdot PSJR2_j$$

This factor represents the total prestige distributed in the current iteration.

The individual contributions to the prestige indicator can therefore be identified in the parts 1,2 and 3 of the preceding formula.

These parts represent respectively:

1. A base prestige value derived from being part of the SJR repository
2. The prestige determined by the number of articles included in the journal
3. The prestige related to the number, "importance", and "closeness" of citations

received by the journal

Phase 2

Since the PSJR value computed in phase 1 is a size-dependent metric (i.e. larger journals will end up having higher prestige values), the final value by dividing the Phase 1 PSJR value by the ratio of citable documents each journal has with regards to the total:

$$SJR2_i = \frac{PSJR2_i}{\left(Art_i / \sum_{j=1}^{N} Art_j\right)} = \frac{PSJR2_i}{Art_i} \cdot \sum_{j=1}^{N} Art_j$$

Methodology for final venues selection

The final venues selection was performed by accounting for both the number of associated publications resulting from the issued query, and the prestige as defined by either the SJR score (average 2017-2021) for journals or the CORE and GGS ratings for conferences.

For journals, the cut-off value related to the number of retrieved publications was set to 30 documents (i.e. all journals with fewer than 30 articles retrieved from the issued query were discarded). This threshold allowed for an initial selection of 23 candidate journals.

On the other hand, the minimum number of publication for a given candidate conference was set to 20 articles for all conferences except the ones published in the ACL Anthology were the cut-off value was set to 10. As previously mentioned, this choice is justified by the limitation imposed by the ACL Anthology repository which limited the search process to the papers titles and abstracts (as opposed to the other repositories which also included other metadata such as the full document content, tags, etc.).

This allowed for the creation of a subset of 23 journals and 12 conferences. This initial subset was further refined by means of a second selection process based on the venues prestige established by the selected metrics.
In this context, the best scoring journals belonging to the upper quartile (Q3) with regards to the SJR score were selected.
Additionally, all conferences that had at least a “A” CORE rating and a “A-” GGS rating were selected from the pool of conferences that met the initial requirements.
This decision led to the selection of the following 6 journals:

- Pattern Recognition

- Journal of Informetrics
- Information Sciences
- Decision Support Systems
- Knowledge-Based Systems
- Expert Systems with Applications

And the following 10 conferences:

- International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)
- ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)
- Annual Meeting of the Association for Computational Linguistics (ACL)
- ACM Conference on Information and Knowledge Management (CIKM)
- Conference on Empirical Methods in Natural Language Processing (EMNLP)
- North American Chapter of the Association for Computational Linguistics (NAACL)
- Machine Learning and Knowledge Discovery in Databases (ECML PKDD)
- International Conference on Computational Linguistics (COLING)
- Conference of the European Chapter of the Association for Computational Linguistics (EACL)
- Advances in Information Retrieval (ECIR)

Information sources

Based on the list of selected venues, the repositories on which the literature search process is performed are:

- ACL Anthology (ACL, EMNLP, NAACL, COLING, EACL)
- ACM Digital Library (SIGIR, SIGKDD, CIKM)
- Springer (ECML PKDD, ECIR)
- ScienceDirect (Pattern Recognition, Journal of Informetrics, Information Sciences, Decision Support Systems, Knowledge-Based Systems, Expert Systems with Applications)

Search strategy

In order to remain consistent with the exploratory search process used to gather the set of venues utilised in this systematic review, it has been decided to gather the relevant work using the query:

- `"topic label*" OR ("topic model*" AND "label*")`

As previously stated, this query allows for the collection of

- All work containing the root term `topic label` (e.g. topic labelling, topic labels, etc…) and
- All work containing the root term `topic model` (e.g. topic modelling, topic models,

etc…) together with the root term `label` (e.g. labelling, labels, etc…)

Previous formulae in Latex

```
PSJR2_i =
\overbrace{\frac{(1-d-e)}{N}}^\text{1} +
\overbrace{e \cdot \frac{Art_i}{\sum_{j=1}^{N} Art_j}}^\text{2} +
\overbrace{\frac{d}{PSJR2D} \cdot\left [ \sum_{j=1}^{N} Coef_{ji} \cdot PSJR2_j
\right ]}^\text{3}

Coef_{ji} = \frac{(Cos_{ji} \cdot  C_{ji})}{\sum_{h=1}^{N} (Cos_{jh} \cdot
C_{jh})}

Cos_{ji} = \frac
{\sum_{h=1,h \neq i, h \neq j}^{N} Cocit_{ih} Cocit_{hj}}
{\sqrt{\sum_{h=1,h \neq i, h \neq j}^{N}(Cocit_{ih})^{2}
\sqrt{\sum_{h=1,h \neq i, h \neq j}^{N}(Cocit_{jh})^{2}}}}

PSJR2D = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{(Cos_{ji} \cdot C_{ji})}
{\sum_{h=1}^{N}(Cos_{jh} \cdot C_{jh})} \cdot PSJR2_j

SJR2_i = \frac{PSJR2_i}{(Art_i / \sum_{j=1}^{N} Art_j)} = \frac{PSJR2_i}{Art_i}
\cdot \sum_{j=1}^{N} Art_j
```

#Thesis/Temporary notes#