Package ‘biblionetwork’

April 9, 2021

Title Create Different Types of Bibliometric Networks

Version 0.1.0

Date 2021-04-08

Maintainer Aurélien Goutsmedt <agoutsmedt@hotmail.fr>

Description Functions to find edges for bibliometric networks like bibliographic coupling network, co-citation network and co-authorship network. The weights of network edges can be calculated according to different methods, depending on the type of networks, the type of nodes, and what you want to analyse. These functions are optimized to be be used on large dataset. The package contains functions inspired by: Leydesdorff, Loet and Park, Han Woo (2017) <doi:10.1016/j.joi.2016.11.007>; Perianes-Rodriguez, Antonio, Ludo Waltman, and Nees Jan Van Eck (2016) <doi:10.1016/j.joi.2016.10.006>; Sen, Subir K. and Shymal K. Gan (1983) <http://nopr.niscair.res.in/handle/123456789/28008>; Shen, Si, Zhu, Danhao, Rousseau, Ronald, Su, Xinning and Wang, Dongbo (2019) <doi:10.1016/j.joi.2019.01.012>; Zhao, Dangzhi and Strotmann, Andreas (2008) <doi:10.1002/meet.2008.1450450292>.

URL https://github.com/agoutsmedt/biblionetwork, https://agoutsmedt.github.io/biblionetwork/

BugReports https://github.com/agoutsmedt/biblionetwork/issues

License MIT + file LICENSE

Encoding UTF-8

LazyData true

RoxygenNote 7.1.1

Imports data.table, Rdpack (>= 0.7)

Depends R (>= 2.10)

RdMacros Rdpack

Suggests knitr, rmarkdown

VignetteBuilder knitr

NeedsCompilation no

Author Aurélien Goutsmedt [cre, aut] (<https://orcid.org/0000-0002-3788-7237>),
François Claveau [aut] (<https://orcid.org/0000-0001-7129-7794>),
Alexandre Truc [aut] (<https://orcid.org/0000-0002-1328-7819>)

Repository CRAN

Date/Publication 2021-04-09 08:20:07 UTC

R topics documented:

Authors_stagflation 2
biblio_cocitation 3
biblio_coupling 4
coauth_network 6
coupling_entity 8
coupling_similarity 10
coupling_strength 12
Nodes_stagflation 13
Ref_stagflation 14

Index **15**

`Authors_stagflation` *List of Authors of the Articles and Books Explaining the 1970s US Stagflation.*

Description

A dataset associating the books and academic articles endeavouring to explain the US stagflation and their authors (`Nodes_stagflation` just takes the first author; here is the complete list of authors per document).

Usage

```
Authors_stagflation
```

Format

A data frame with 558 rows and 7 variables:

ItemID_Ref Identifier of the document published by the author

Author Author of the document

Order Use this as a label for nodes

Source

Goutsmedt A. (2020) “From Stagflation to the Great Inflation: Explaining the 1970s US Economic Situation”. Revue d’Economie Politique, Forthcoming 2021.

biblio_cocitation *Calculating the Coupling Angle Measure for Edges in a Co-citation Network*

Description

This function is basically the same as the `biblio_coupling()` function but it is explicitly framed for bibliographic co-citation network (and not for bibliographic coupling networks). It takes a data frame with direct citations, and calculates the number of times two references are citing together, and calculate a measure similar to the coupling angle value (Sen and Gan 1983): it divides the number of times two references are cited together by the square root of the product of the total number of citations (in the whole corpus) of each reference. The more two references are cited in general, the more they have to be cited together for their link to be important.

Usage

```
biblio_cocitation(
  dt,
  source,
  ref,
  normalized_weight_only = TRUE,
  weight_threshold = 1,
  output_in_character = TRUE
)
```

Arguments

`dt` The dataframe with citing and cited documents.

`source` The column name of the source identifiers, that is the documents that are citing.

`ref` The column name of the cited references identifiers. In co-citation network, these references are the nodes of the network.

`normalized_weight_only` If set to FALSE, the function returns the weights normalized by the cosine measure, but also simply the number of times two references are cited together.

`weight_threshold` Correspond to the value of the non-normalized weights of edges. The function just keeps the edges that have a non-normalized weight superior to the `weight_threshold`. In a large bibliographic co-citation network, you can consider for instance that being cited only once together is not sufficient/significant for two references to be linked together. This parameter could also be modified to avoid creating intractable networks with too many edges.

`output_in_character` If TRUE, the function ends by transforming the `from` and `to` columns in character, to make the creation of a tidygraph graph easier.

Details

This function uses data.table package and is thus very fast. It allows the user to compute the coupling angle on a very large data frame quickly.

Value

A data.table with the articles (or authors) identifier in `from` and `to` columns, with one or two additional columns (the coupling angle measure and the number of shared references). It also keeps a copy of `from` and `to` in the `Source` and `Target` columns. This is useful is you are using the tidygraph package then, where `from` and `to` values are modified when creating a graph.

References

Sen SK, Gan SK (1983). “A Mathematical Extension of the Idea of Bibliographic Coupling and Its Applications.” *Annals of library science and documentation*, **30**(2). `http://nopr.niscair.res.in/bitstream/123456789/28008/1/ALIS%2030(2)%2078-82.pdf`.

Examples

```
library(biblionetwork)
biblio_cocitation(Ref_stagflation,
source = "Citing_ItemID_Ref",
ref = "ItemID_Ref")

# It is basically the same as:
biblio_coupling(Ref_stagflation,
source = "ItemID_Ref",
ref = "Citing_ItemID_Ref")
```

`biblio_coupling` *Calculating the Coupling Angle Measure for Edges*

Description

This function calculates the number of references that different articles share together, as well as the coupling angle value of edges in a bibliographic coupling network (Sen and Gan 1983), from a direct citation data frame. This is a standard way to build bibliographic coupling network using Salton’s cosine measure: it divides the number of references that two articles share by the square root of the product of both articles bibliography lengths. It avoids giving too much importance to articles with a large bibliography.

Usage

```
biblio_coupling(
  dt,
  source,
  ref,
  normalized_weight_only = TRUE,
  weight_threshold = 1,
  output_in_character = TRUE
)
```

Arguments

`dt` For bibliographic coupling (or co-citation), the dataframe with citing and cited documents. It could also be used

1. for title co-occurence network, with `source` being the articles, and `ref` being the list of words in articles titles;
2. for co-authorship network, with `source` being the authors, and `ref` the list of articles.

`source` The column name of the source identifiers, that is the documents that are citing. In a coupling network, these documents are the nodes of the network.

`ref` The column name of the cited references identifiers.

`normalized_weight_only`
If set to FALSE, the function returns the weights normalized by the cosine measure, but also the number of shared references.

`weight_threshold`
Corresponds to the value of the non-normalized weights of edges. The function just keeps the edges that have a non-normalized weight superior to the `weight_threshold`. In other words, if you set the parameter to 2, the function keeps only the edges between nodes that share at least two references in common in their bibliography. In a large bibliographic coupling network, you can consider for instance that sharing only one reference is not sufficient/significant for two articles to be linked together. This parameter could also be modified to avoid creating intractable networks with too many edges.

`output_in_character`
If TRUE, the function ends by transforming the `from` and `to` columns in character, to make the creation of a tidygraph network easier.

Details

This function implements the following weight measure:

$$\frac{R(A) \bullet R(B)}{\sqrt{L(A).L(B)}}$$

with $R(A)$ and $R(B)$ the references of document A and document B, $R(A) \bullet R(B)$ being the number of shared references by A and B, and $L(A)$ and $L(B)$ the length of the bibliographies of document A and document B.

This function uses data.table package and is thus very fast. It allows the user to compute the coupling angle on a very large data frame quickly.

This function is a relatively general function that can also be used

1. for co-citation networks (just by inversing the `source` and `ref` columns). If you want to avoid confusion, rather use the `biblio_cocitation()` function.

2. for title co-occurence networks (taking care of the length of the title thanks to the coupling angle measure);
3. for co-authorship networks (taking care of the number of co-authors an author has collaborated with on a period). For co-authorship, rather use the `coauth_network()` function.

Value

A data.table with the articles (or authors) identifiers in `from` and `to` columns, with one or two additional columns (the coupling angle measure and the number of shared references). It also keeps a copy of `from` and `to` in the `Source` and `Target` columns. This is useful is you are using the tidygraph package after, where `from` and `to` values are modified when creating a graph.

References

Sen SK, Gan SK (1983). "A Mathematical Extension of the Idea of Bibliographic Coupling and Its Applications." *Annals of library science and documentation*, **30**(2). http://nopr.niscair.res.in/bitstream/123456789/28008/1/ALIS%2030(2)%2078-82.pdf.

Examples

```
library(biblionetwork)
biblio_coupling(Ref_stagflation,
source = "Citing_ItemID_Ref",
ref = "ItemID_Ref",
weight_threshold = 3)
```

coauth_network *Creating Co-Authorship Network with Different Measures for Weights*

Description

This function creates an edge list for co-authorship networks from a data frame with a list of entities and their publications. The weight of edges can be calculated following different methods. The nodes of the network could be indifferently authors, institutions or countries.

Usage

```
coauth_network(
  dt,
  authors,
  articles,
 method = c("full_counting", "fractional_counting", "fractional_counting_refined"),
  cosine_normalized = FALSE
)
```

Arguments

`dt` The data frame with authors (or institutions or countries) and the list of documents they have published.

`authors` The column name of the source identifiers, that is the authors (or institutions or countries).

`articles` The column name of the documents identifiers.

`method` Method for calculating the edges weights, to be chosen among "full_counting","fractional_counting" or "fractional_counting_refined".

`cosine_normalized`
Possibility to take into account the total number of articles written by two linked authors and to normalize the weight of their link using Salton's cosine.

Details

Weights can be calculated with:

1. the `"full_counting"` method: the linkds between authors correspond to their absolute number of collaborations.
2. the `"fractional_counting"` method which takes into account the number of authors in each article, following (Perianes-Rodriguez et al. 2016) equation:

$$\sum_{k=1}^{M} \frac{a_{ik}.a_{jk}}{n_k - 1}$$

   with M the total number of articles, $a_{ik}.a_{jk}$ which takes 1 if author i and j have co-written the article k, and n_k the number of authors for article k.
3. the `fractional_counting_refined` method, inspired by (Leydesdorff and Park 2017) which is similar to `fractional_counting` but which is formalised in a way that allows the sum of weights to equal the number of articles in the corpus:

$$\sum_{k=1}^{M} \frac{a_{ik}.a_{jk}.2}{n_k.(n_k - 1)}$$

   .

In addition, it is possible to take into account the total number of collaborations of two linked authors. If `cosine_normalized` is set to `True`, the weight calculated with one of the three methods above is divided by $\sqrt{C_i.C_j}$, with C_i being the number of articles co-written by author i.

Value

A data.table with the authors (or institutions or countries) identifier in `from` and `to` columns, with a `weight` column whose values depend on the method chosen. It also keeps a copy of `from` and `to` in the `Source` and `Target` columns. This is useful is you are using the tidygraph package then, where `from` and `to` values are modified when creating a graph.

References

Leydesdorff L, Park HW (2017). “Full and Fractional Counting in Bibliometric Networks.” *Journal of Informetrics*, **11**(1), 117–120. ISSN 17511577, https://linkinghub.elsevier.com/retrieve/pii/S1751157716303133.

Perianes-Rodriguez A, Waltman L, Van Eck NJ (2016). “Constructing Bibliometric Networks: A Comparison between Full and Fractional Counting.” *Journal of Informetrics*, **10**(4), 1178–1195. https://www.sciencedirect.com/science/article/pii/S1751157716302036?casa_token=AtzjmZ-1QmYAAAAA:2mlBPZsjGUleYi9mnybHODFw2RmMh3GHvRAuMYXygRm63cQOv07M4ixbAmJXuGq71tx2ug29baTp

Examples

```
library(biblionetwork)
coauth_network(Authors_stagflation,
authors = "Author",
articles = "ItemID_Ref",
method = "fractional_counting")
```

`coupling_entity` *Creating Coupling Networks at Entity Level*

Description

This function creates the edges of a network of entities from a direct citations data frame (i.e. documents citing references). Entities could be authors, affiliations, journals, *etc*. Consequently, coupling links are calculated using the coupling angle measure (like `biblio_coupling()`) or the coupling strength measure (like `coupling_strength()`. But it also takes into account the fact that an entity can cite several times a reference, and considers that citing 10 times a ref is more significant that citing it only once (see details).

Usage

```
coupling_entity(
  dt,
  source,
  ref,
  entity,
  weight_threshold = 1,
  output_in_character = FALSE,
  method = c("coupling_strength", "coupling_angle")
)
```

Arguments

`dt` The table with citing and cited documents.

`source` The column name of the source identifiers, that is the documents that are citing.

`ref` The column name of the cited references identifiers.

`entity` The column name of the entity (authors, journals, institutions) that are citing.

`weight_threshold`

Corresponds to the value of the non-normalized weights of edges. The function just keeps the edges that have a non-normalized weight superior to the `weight_threshold`. In other words, if you set the parameter to 2, the function keeps only the edges between nodes that share at least two references in common in their bibliography. In a large bibliographic coupling network, you can consider for instance that sharing only one reference is not sufficient/significant for two entities (above all when large entities like journals and institutions) to be linked together. This parameter could also be modified to avoid creating intractable networks with too many edges.

`output_in_character`

If TRUE, the function ends by transforming the `from` and `to` columns in character, to make the creation of a tidygraph network easier.

`method` Choose the method you want to use for calculating the edges weights: either `"coupling_strength"` like in the `coupling_strength()` function, or `"coupling_angle"` like in the `biblio_coupling()` function.

Details

Coupling links are calculated depending of the number of references two authors (or any entity) share, taking into account the minimum number of times two authors are citing each references. For instance, if two entities share a reference in common, the first one citing it twice (in other words, citing it in two different articles), the second one three times, the function takes two as the minimum value. In addition to the features of the coupling strength measure (see `coupling_strength()`) or the coupling angle measure (see `biblio_coupling()`), it means that, if two entities share two reference in common, if the first reference is cited at least four times by the two entities, whereas the second reference is cited at least only once, the first reference contributes more to the edge weight than the second reference. This use of minimum shared reference for entities coupling comes from Zhao and Strotmann (2008). It looks like this for the coupling strength:

$$\frac{1}{L(A)} \cdot \frac{1}{L(A)} \sum_j Min(C_{Aj}, C_{Bj}).(log(\frac{N}{freq(R_j)}))$$

with C_{Aj} and C_{Bj} the number of time documents A and B cite the reference j.

Value

A data.table with the entity identifiers in `from` and `to` columns, with the coupling strength or coupling angle measures in another column, as well as the method used. It also keeps a copy of `from` and `to` in the `Source` and `Target` columns. This is useful is you are using the tidygraph package then, where `from` and `to` values are modified when creating a graph.

References

Zhao D, Strotmann A (2008). "Author Bibliographic Coupling: Another Approach to Citation-Based Author Knowledge Network Analysis." *Proceedings of the American Society for Informa-*

tion Science and Technology, **45**(1), 1–10. https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/meet.2008.1450450292.

Examples

```
library(biblionetwork)
Ref_stagflation$Citing_ItemID_Ref <- as.character(Ref_stagflation$Citing_ItemID_Ref)
# merging the references data with the citing author information in Nodes_stagflation
entity_citations <- merge(Ref_stagflation,
                          Nodes_stagflation,
                          by.x = "Citing_ItemID_Ref",
                          by.y = "ItemID_Ref")

coupling_entity(entity_citations,
                source = "Citing_ItemID_Ref",
                ref = "ItemID_Ref",
                entity = "Author.y",
                method = "coupling_angle")
```

coupling_similarity *Calculating the Coupling Similarity Measure for Edges*

Description

This function calculates a refined similarity measure of coupling links, from a direct citation data frame. It is sinpired by (Shen et al. 2019). To a certain extent, it mixes the `coupling_strength()` function with the cosine measure of the `biblio_coupling()` function.

Usage

```
coupling_similarity(
  dt,
  source,
  ref,
  weight_threshold = 1,
  output_in_character = TRUE
)
```

Arguments

dt	The table with citing and cited documents.
source	The column name of the source identifiers, that is the documents that are citing. In bibliographic coupling, these documents are the nodes of the network.
ref	The column name of the references that are cited.

`weight_threshold`
: Corresponds to the value of the non-normalized weights of edges. The function just keeps the edges that have a non-normalized weight superior to the `weight_threshold`. In other words, if you set the parameter to 2, the function keeps only the edges between nodes that share at least two references in common in their bibliography. In a large bibliographic coupling network, you can consider for instance that sharing only one reference is not sufficient/significant for two articles to be linked together. This parameter could also be modified to avoid creating intractable networks with too many edges.

`output_in_character`
: If TRUE, the function ends by transforming the `from` and `to` columns in character, to make the creation of a tidygraph network easier.

Details

The function use the following formalisation:

$$\frac{R_S(A) \bullet R_S(B)}{\sqrt{R_S(A).R_S(B)}}$$

1. with

   $$R_S(A) \bullet R_S(B) = \sum_j \sqrt{log(\frac{N}{freq(R_j)})}$$

   that is a measure similar to the coupling strength measure;

2. and

   $$R_S(A).R_S(B) = \sum_j \sqrt{log(\frac{N}{freq(R_j(A))})}.\sum_j \sqrt{log(\frac{N}{freq(R_j(B))})}$$

   which is the separated sum for each article of the normalized value of a citation. It is the cosine measure of documents A and B but adapted to the spirit of the coupling strength.

Value

A data.table with the articles identifiers in `from` and `to` columns, with the similarity measure in another column. It also keeps a copy of `from` and `to` in the `Source` and `Target` columns. This is useful is you are using the tidygraph package then, where `from` and `to` values are modified when creating a graph.

References

Shen S, Zhu D, Rousseau R, Su X, Wang D (2019). “A Refined Method for Computing Bibliographic Coupling Strengths.” *Journal of Informetrics*, **13**(2), 605–615. `https://linkinghub.elsevier.com/retrieve/pii/S1751157716300244`.

Examples

```
library(biblionetwork)
coupling_similarity(Ref_stagflation,
source = "Citing_ItemID_Ref",
ref = "ItemID_Ref")
```

coupling_strength *Calculating the Coupling Strength Measure for Edges*

Description

This function calculates the coupling strength measure (following Vladutz and Cook 1984 and Shen et al. 2019) from a direct citation data frame. It is a refinement of `biblio_coupling()`: it takes into account the frequency with which a reference shared by two articles has been cited in the whole corpus. In other words, the most cited references are less important in the links between two articles, than references that have been rarely cited. To a certain extent, it is similar to the tf-idf measure.

Usage

```
coupling_strength(
  dt,
  source,
  ref,
  weight_threshold = 1,
  output_in_character = TRUE
)
```

Arguments

`dt` The data frame with citing and cited documents.

`source` the column name of the source identifiers, that is the documents that are citing.

`ref` the column name of the references that are cited.

`weight_threshold` Corresponds to the value of the non-normalized weights of edges. The function just keeps the edges that have a non-normalized weight superior to the `weight_threshold`. In other words, if you set the parameter to 2, the function keeps only the edges between nodes that share at least two references in common in their bibliography. In a large bibliographic coupling network, you can consider for instance that sharing only one reference is not sufficient/significant for two articles to be linked together. This parameter could also be modified to avoid creating intractable networks with too many edges.

`output_in_character` If TRUE, the function ends by transforming the `from` and `to` columns in character, to make the creation of a tidygraph graph easier.

Value

A data.table with the articles identifiers in `from` and `to` columns, with the coupling strength measure in another column. It also keeps a copy of `from` and `to` in the `Source` and `Target` columns. This is useful is you are using the tidygraph package then, where `from` and `to` values are modified when creating a graph.

References

Shen S, Zhu D, Rousseau R, Su X, Wang D (2019). “A Refined Method for Computing Bibliographic Coupling Strengths.” *Journal of Informetrics*, **13**(2), 605–615. https://linkinghub.elsevier.com/retrieve/pii/S1751157716300244.

Vladutz G, Cook J (1984). “Bibliographic Coupling and Subject Relatedness.” *Proceedings of the American Society for Information Science*, **21**, 204–207.

Examples

```
library(biblionetwork)
coupling_strength(Ref_stagflation,
source = "Citing_ItemID_Ref",
ref = "ItemID_Ref")
```

`Nodes_stagflation` *Articles and Books Explaining the 1970s US Stagflation.*

Description

A dataset containing the books and academic articles endeavouring to explain what happened in the US economy in the 1970s, as well as all the articles and books cited at least twice by the first set of articles and books (on the stagflation).

Usage

```
Nodes_stagflation
```

Format

A data frame with 558 rows and 7 variables:

ItemID_Ref Identifier of the document

Author Author of the document

Author_date Use this as a label for nodes

Year Year of publication of the document

Title Title of the document

Journal Journal of publication of the document (if an article)

Type If "Stagflation", the document is listed as an explanation of the US stagflation. If "Non-Stagflation", the document is cited by a document explaining the stagflation

Source

Goutsmedt A. (2020) “From Stagflation to the Great Inflation: Explaining the 1970s US Economic Situation”. Revue d’Economie Politique, Forthcoming 2021.

`Ref_stagflation` *Articles and Books Explaining the 1970s US Stagflation.*

Description

A dataset containing all the articles and books cited by the books and academic articles endeavouring to explain what happened in the US economy in the 1970s.

Usage

```
Ref_stagflation
```

Format

A data frame with 4416 rows and 6 variables:

Citing_ItemID_Ref Identifier of the citing document

ItemID_Ref Identifier of the cited document

Author Author of the cited document

Year Year of publication of the cited document

Title Title of the cited document

Journal Journal of publication of the cited document (if an article)

Source

Goutsmedt A. (2020) “From Stagflation to the Great Inflation: Explaining the 1970s US Economic Situation”. Revue d’Economie Politique, Forthcoming 2021.

Index

∗ **datasets**
    Authors_stagflation, 2
    Nodes_stagflation, 13
    Ref_stagflation, 14

Authors_stagflation, 2

biblio_cocitation, 3
biblio_cocitation(), *5*
biblio_coupling, 4
biblio_coupling(), *3*, *8–10*, *12*

coauth_network, 6
coauth_network(), *6*
coupling_entity, 8
coupling_similarity, 10
coupling_strength, 12
coupling_strength(), *8–10*

Nodes_stagflation, 13

Ref_stagflation, 14