Package: robustrao (via r-universe)

October 14, 2024

Type Package

Title An Extended Rao-Stirling Diversity Index to Handle Missing Data

Version 1.0-5

Date 2020-01-22

Author Maria del Carmen Calatrava Moreno [aut, cre], Thomas Auzinger [aut]

Maintainer Maria del Carmen Calatrava Moreno

<mc.calatrava.moreno@gmail.com>

Description A collection of functions to compute the Rao-Stirling diversity index (Porter and Rafols, 2009) <DOI:10.1007/s11192-008-2197-2> and its extension to acknowledge missing data (i.e., uncategorized references) by calculating its interval of uncertainty using mathematical optimization as proposed in Calatrava et al. (2016) <DOI:10.1007/s11192-016-1842-4>. The Rao-Stirling diversity index is a well-established bibliometric indicator to measure the interdisciplinarity of scientific publications. Apart from the obligatory dataset of publications with their respective references and a taxonomy of disciplines that categorizes references as well as a measure of similarity between the disciplines, the Rao-Stirling diversity index requires a complete categorization of all references of a publication into disciplines. Thus, it fails for a incomplete categorization; in this case, the robust extension has to be used, which encodes the uncertainty caused by missing bibliographic data as an uncertainty interval. Classification / ACM - 2012: Information systems ~ Similarity measures, Theory of computation ~ Quadratic programming, Applied computing ~ Digital libraries and archives.

Depends R (>= 3.2)

Imports doParallel (>= 1.0.10), gmp (>= 0.5-12), iterpc (>= 0.3.0), quadprog (>= 1.5-5), igraph (>= 1.0.1), foreach (>= 1.4.3)

License GPL-3

URL https://gitlab.com/mc.calatrava.moreno/robustrao.git

Encoding UTF-8

RoxygenNote 6.0.1

NeedsCompilation no

Date/Publication 2020-01-24 21:50:08 UTC

Repository https://mccalatravamoreno.r-universe.dev

RemoteUrl https://github.com/cran/robustrao

RemoteRef HEAD

RemoteSha df079e70f9902d119e85648fbd27d3c49829f06b

Contents

`LowerIndexBound` *Lower bound of the uncertainty interval of the Rao-Stirling diversity index.*

Description

This function computes the lower bound of the uncertainty interval of the Rao-Stirling diversity index, as explained in Calatrava et al. (2016). The computation involves the redistribution of uncategorized references to various disciplines. In order to avoid improbable redistributions of disciplines, a set of permissible disciplines for redistribution can be defined. Furthermore, the number of disciplines redistributed to uncategorized references can be limited.

Usage

```
LowerIndexBound(known.ref.counts, uncat.ref.count, similarity,
  permissible.disciplines = NULL, redistribution.limit = 4,
  max.batch.size = 131072)
```

Arguments

`known.ref.counts`
: A vector of positive integers. Each element represents the count of references to each discipline.

`uncat.ref.count`
: A positive integer denoting the number of references that are not categorized into disciplines.

`similarity`
: A positive semi-definite matrix that encodes the similarity between disciplines, as explained in Porter and Rafols (2009). The dimensions of this matrix are *n* x *n*, being *n* the total number of disciplines. The self-similarities (i.e. the diagonal elements) have to be 1.

`permissible.disciplines`
: A logical vector denoting to which disciplines uncategorized references can be distributed. Its length needs to be equal to the length of `known.ref.counts`. This argument is optional and leaving it unspecified or supplying NULL permits redistribution to all disciplines.

`redistribution.limit`
: A positive integer that limits the number of disciplines that each uncategorized reference can be redistributed to. This argument is optional and leaving it unspecified permits redistribution to all disciplines at once.

`max.batch.size`
: A positive integer that sets the size of the batch of candidates that is computed at once. This positive value determines the quantity of allocated memory and has to be reduced if corresponding errors arise. This argument is optional and leaving it unspecified sets it to a default value.

Value

The lower bound of the uncertainty interval of the Rao-Stirling diversity index.

Warning

This function solves a computationally intensive optimization problem. In order to reduce the search space it is recommended to provide the function with the vector of permissible disciplines and redistribution limit. When very dissimilar disciplines are referenced by the categorized references, a warning message is displayed to inform the user. Such cases require longer computation times. The dataset `pubdata2` contains an example of a publication that requires intensive computation in order to calculate the uncertainty interval of the Rao-Stirling diversity index.

References

Calatrava Moreno, M. C., Auzinger, T. and Werthner, H. (2016) On the uncertainty of interdisciplinarity measurements due to incomplete bibliographic data. Scientometrics. DOI:10.1007/s11192-016-1842-4

Porter, A. and Rafols, I. (2009) Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. Scientometrics, Vol. 81, No. 3 (719-745). DOI:10.1007/s11192-008-2197-2

Examples

```
##EXAMPLE 1
#Load data
data(pubdata1)

#Get counts of citations of one of the publications in the dataset
counts <- pd1.count.matrix[,1]

#Get number of uncategorized references in the publication
uncat <- pd1.uncat.refs[1]

#Get vector of permissible disciplines.
logic.disciplines <- counts > 0
permissible <- PruneDisciplines(logic.disciplines, 0.233, pd1.similarity)

LowerIndexBound(counts, uncat, pd1.similarity, permissible)
```

`ParallelBoundIndices` *Parallel computation of the lower/upper bounds of the uncertainty interval of the Rao-Stirling diversity index.*

Description

This function allows the computation of the lower/upper bounds of the uncertainty interval of the Rao-Stirling index (Calatrava et al., 2016) in parallel threads. It includes the parallel computation of the permissible disciplines (i.e., function `PruneDisciplines`). The use of this function is recommended for an efficient computation of the lower and upper bounds of the uncertainty interval of the Rao-Stirling index. The computation of the lower bound is an NP-hard problem. Because the computation of the lower bound might require long computing times, this function creates a log file 'parallel-bounds-log.txt' in the user's workspace. The content of the log file is written during the execution of the function and indicates number of publications that have been processed.

Usage

```
ParallelBoundIndices(bound, count.matrix, uncat.refs, similarity,
  pruning = TRUE, tolerance = 1, redistribution.limit = 4, threads = 1,
  max.batch.size = 131072)
```

Arguments

`bound`	String that indicates which index to compute. Two values are valid: *upper* and *lower*.
`count.matrix`	Vector or matrix that contains the counts of references to different disciplines of a single publication (a vector) or of several publications (a matrix). If it is a vector its length is equal to the total number of disciplines. In case it is a matrix its dimensions are *n* x *m*, being n the total number of disciplines and m the number of publications for which the lower/upper bound will be computed.

`uncat.refs` Number of uncategorized references of a publication (a number) or several publications (a vector).

`similarity` A positive semi-definite matrix that encodes the similarity between disciplines, as explain in Porter and Rafols (2009). The dimensions of this matrix are n x n, being n the total number of disciplines. The self-similarities (i.e. values in the diagonal) have to be 1.

`pruning` Logical value that indicates whether the set of permissible disciplines will be calculated and used to avoid improbable redistributions of disciplines. This argument is optional and leaving it unspecified ignores the pruning of unlikely disciplines in the redistribution.

`tolerance` A real number in the interval [0,1]. This argument modulates the similarity between disciplines with which the strictness of the pruning of unlikely disciplines is controlled. A value of 0 allows all disciplines to participate in the redistribution process. A value of 1 permits no tolerance. This argument is optional and leaving it unspecified deactivates tolerances.

`redistribution.limit` A positive integer that limits the number of disciplines that each uncategorized reference can have redistributed. This argument is optional and leaving it unspecified will set the redistribution.limit to default.

`threads` A positive number that specifies the number of parallel threads that will be executed. This argument should be set according to the number of processor core in the CPU of the user. This argument is optional and leaving it unspecified will set the number of threads to default.

`max.batch.size` A positive integer that sets the size of the batch of candidates that is computed at once. This positive value determines the quantity of allocated memory and has to be reduced if corresponding errors arise. This argument is optional and leaving it unspecified sets it to a default value.

Value

The lower or the upper bound/s of the uncertainty interval of the Rao-Stirling index of one publication (an integer) or several publications (a vector).

Warning

This function solves a computationally intensive optimization problem. In order to reduce the search space it is recommended to provide the function with the vector of permissible disciplines and redistribution limit.

References

Calatrava Moreno, M. C., Auzinger, T. and Werthner, H. (2016) On the uncertainty of interdisciplinarity measurements due to incomplete bibliographic data. Scientometrics. DOI:10.1007/s11192-016-1842-4

Porter, A. and Rafols, I. (2009) Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. Scientometrics, Vol. 81, No. 3 (719-745). DOI:10.1007/s11192-008-2197-2

Examples

```
#Load data
data(pubdata1)

#Get upper bound indices of the uncertainty interval of the Rao-Stirling diversity index.
ParallelBoundIndices("upper", pd1.count.matrix, pd1.uncat.refs, pd1.similarity, TRUE, 0.233, 4, 2)

#Get lower bound indices of the uncertainty interval of the Rao-Stirling diversity index.
ParallelBoundIndices("lower", pd1.count.matrix, pd1.uncat.refs, pd1.similarity, TRUE, 0.233, 4, 2)

#When many references of a publication are uncategorized, a warning message is displayed
#to inform the user. Such cases require longer computation times.
```

`PruneDisciplines` *Set of permissible disciplines for redistribution.*

Description

Computes the set of disciplines to which uncategorized references can be redistributed. This set is computed taking into account the mutual similarities of the already referenced disciplines, as explained in Calatrava et al. (2016). This function allows to set a tolerance of similarity that only permits similar disciplines to participate in the redistribution process. Therefore, it avoids redistributions that include very dissimilar and improbable disciplines.

Usage

```
PruneDisciplines(r, tolerance = 1, similarity)
```

Arguments

`r`	A logical vector indicating which disciplines are referenced by the current document. Its length is equal to the total number of disciplines.
`tolerance`	A real number in the interval [0,1]. This argument modulates the similarity between disciplines with which the strictness of the pruning of unlikely disciplines is controlled. A value of 0 allows all disciplines to participate in the redistribution process. A value of 1 permits no tolerance. This argument is optional and leaving it unspecified deactivates tolerances.
`similarity`	A positive semi-definite matrix that encodes the similarity between disciplines, as explained in Porter and Rafols (2009). The dimensions of this matrix are *n* x *n*, being *n* the total number of disciplines. The number of rows and the number of columns of this matrix needs to be equal to the length of `r`. The self-similarities (i.e. the diagonal elements) have to be 1.

Value

A logical vector indicating to which disciplines a reference redistribution is permissible.

References

Calatrava Moreno, M. C., Auzinger, T. and Werthner, H. (2016) On the uncertainty of interdisciplinarity measurements due to incomplete bibliographic data. Scientometrics. DOI:10.1007/s11192-016-1842-4

Porter, A. and Rafols, I. (2009) Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. Scientometrics, Vol. 81, No. 3 (719-745). DOI:10.1007/s11192-008-2197-2

Examples

```
#Load data
data(pubdata1)

#Get counts of citations of one of the publications in the dataset
counts <- pd1.count.matrix[,1]

#Get logical vector indicating which disciplines are referenced by the publication
logic.disciplines <- counts > 0

PruneDisciplines(logic.disciplines, 0.233, pd1.similarity)
```

---

pubdata1 *pubdata1*

Description

Small example dataset with 5 publications that have most of their references categorized into disciplines. The dataset contains the following information: A matrix of counts of referenced disciplines for each publication, a vector of counts of uncategorized references in each publication, and a matrix that contains a measure of similarity between disciplines.

Usage

```
data("pubdata1")
```

Format

`pd1.count.matrix` the count of referenced disciplines for each publication

`pd1.uncat.refs` the count of referenced disciplines for each publication

`pd1.similarity` between disciplines as given in Porter and Rafols, 2009.

References

Porter, A. and Rafols, I. (2009) Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. Scientometrics, Vol. 81, No. 3 (719-745). DOI:10.1007/s11192-008-2197-2

pubdata2 *pubdata2*

Description

Small example dataset with 2 publications. The first publication references rather dissimilar disciplines and has uncategorized references. Therefore, the computation of the interval of uncertainty of the Rao-Stirling index requires longer computation time. The dataset contains the following information: A matrix of counts of referenced disciplines for each publication, a vector of counts of uncategorized references in each publication, and a matrix that contains a measure of similarity between disciplines.

Usage

```
data("pubdata2")
```

Format

`pd2.count.matrix` the count of referenced disciplines for each publication

`pd2.uncat.refs` the count of referenced disciplines for each publication

`pd2.similarity` between disciplines as given in Porter and Rafols, 2009.

References

Porter, A. and Rafols, I. (2009) Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. Scientometrics, Vol. 81, No. 3 (719-745). DOI:10.1007/s11192-008-2197-2

RaoStirling *Rao-Stirling diversity index based on the counts of cited disciplines.*

Description

This function calculates the Rao-Stirling diversity index of one or several publications, based on the count of citations of the publication(s) to different disciplines.

Usage

```
RaoStirling(count.matrix, similarity)
```

Arguments

count.matrix	Vector or matrix that contains the counts of references to different disciplines of a single publication (vector) or of several publications (matrix). If count.matrix is a vector its length is equal to the total number of disciplines. In case it is a matrix its dimensions are *n* x *m*, being n the total number of disciplines and m the number of publications for which the lower/upper bound will be computed.
similarity	A positive semi-definite matrix that encodes the similarity between disciplines, as explain in Porter and Rafols (2009). The dimensions of this matrix are *n* x *n*, being *n* the total number of disciplines. The number of rows and the number columns of this matrix need to be equal to the number of rows of count.matrix. The self-similarities (i.e. the diagonal elements) have to be 1.

Value

The Rao-Stirling diversity index of one or several publications.

References

Porter, A. and Rafols, I. (2009) Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. Scientometrics, Vol. 81, No. 3 (719-745). DOI:10.1007/s11192-008-2197-2

Examples

```
#Load data
data(pubdata1)

#Get Rao-Stirling diversity index of all publications in the dataset
RaoStirling(pd1.count.matrix, pd1.similarity)

#Get Rao-Stirling diversity index of one publication of the dataset
RaoStirling(pd1.count.matrix[,2], pd1.similarity)
```

UpperIndexBound — *Upper bound of the uncertainty interval of the Rao-Stirling diversity index.*

Description

This function computes the upper bound of the uncertainty interval of the Rao-Stirling diversity index, as explained in Calatrava et al. (2016). The computation involves the redistribution of uncategorized references to various disciplines. In order to avoid improbable redistributions of disciplines, a set of permissible disciplines for redistribution can be defined. Furthermore, the number of disciplines redistributed to uncategorized references can be limited.

Usage

```
UpperIndexBound(known.ref.counts, uncat.ref.count, similarity,
  permissible.disciplines = NULL, redistribution.limit = 4)
```

Arguments

known.ref.counts	A vector of positive integers. Each element represents the count of references to each discipline.
uncat.ref.count	A positive integer denoting the number of references that are not categorized into disciplines.
similarity	A positive semi-definite matrix that encodes the similarity between disciplines, as explained in Porter and Rafols (2009). The dimensions of this matrix are *n* x *n*, being *n* the total number of disciplines. The self-similarities (i.e. the diagonal elements) have to be 1.
permissible.disciplines	A logical vector denoting to which disciplines uncategorized references can be distributed. Its length needs to be equal to the length of known.ref.counts. This argument is optional and leaving it unspecified or supplying NULL permits redistribution to all disciplines.
redistribution.limit	A positive integer that limits the number of disciplines that each uncategorized reference can have redistributed. This argument is optional and leaving it unspecified will set the redistribution.limit to default.

Value

The upper bound of the uncertainty interval of the Rao-Stirling diversity index.

References

Calatrava Moreno, M.C., Auzinger, T. and Werthner, H. (2016) On the uncertainty of interdisciplinarity measurements due to incomplete bibliographic data. Scientometrics. DOI:10.1007/s11192-016-1842-4

Porter, A. and Rafols, I. (2009) Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. Scientometrics, Vol. 81, No. 3 (719-745). DOI:10.1007/s11192-008-2197-2

Examples

```
#Load data
data(pubdata1)

#Get counts of citations of one of the publications in the dataset
counts <- pd1.count.matrix[,1]

#Get number of uncategorized references in the publication
```

```
uncat <- pd1.uncat.refs[1]

#Get vector of permissible disciplines.
logic.disciplines <- counts > 0
permissible <- PruneDisciplines(logic.disciplines, 0.233, pd1.similarity)

UpperIndexBound(counts, uncat, pd1.similarity, permissible)
```

Index