lirec: An R package for generating literature recommendations based on a set of given articles

Payman Tajalli-Seifi[1], Tobias Schmidt[2]

[1,2]Department of Statistical Bioinformatics, Institute of Functional Genomics, University of Regensburg, 93040, Regensburg, Germany

[1]tajalli-seifi.payman@stud.uni-regensburg.de, [2]tobias2.schmidt@ukr.de

INTRODUCTION

- A review of prior, relevant literature is an essential feature of any academic project[1].
- The most common way of doing literature research is searching by keywords, i.e., you enter some keywords into a search engine such as google scholar and you end up with a list of articles matching your keywords.
- Although, Using an effective keyword search will produce some initial insight into the domain one wishes to study, because of its several limitations, some additional steps should be used to advance the literature search, such as **backward references search**, even more comprehensive **multi-level of backward references search[2]**.
- Backward references search refers to reviewing the references of the interested articles[2].
- A multi-level backward references search can be done by pulling the "**references of the references**" [2].
- To the best of our knowledge, there is not yet an easy-to-use, open-source software to address this topic.

APPROACH

- We present here an R package, called **lirec** (**li**terature **rec**ommendations), which is designed for generating and visualizing literature recommendations based on a set of given articles. The recommended articles can be either printed as table or visualized as graph.
- This package receives a list of interested article's (DOI) as a character and the metadata of each article (e.g., year, first author, URL, abstract) will be retrieved from **crossref.org** through using **rcrossref** package[3]. Finally, the output can be either
  - a data table of references of interested articles or references of the references depending on the level of backward references search.
  - a plot including interested articles, the references and their connections.

INSTALLATION AND FUNCTIONS

```
# From Github (development version)
devtools::install_github("toscm/lirec")
# From CRAN (stable version)
install.packages("lirec") # not yet available

# Important note:
The Github repository is private at this moment,
for access permissions,please contact
tajalli-Seifi.Payman@stud.uni-regensburg.de
```

Workflow	Function	Parameters	Description
Retrieving data	get_refs	search_depth	The level of backward references search
		doi_na_drop	Removes the row if the reference has no doi, in ref_table
Constructing citation table	articles_citation_freq	sorted	Sorts articles in citation table regarding citation count
		top_n	Selects top n most cited articles from citation table
Visualization	plot_as_graph	min_cit	Minimum citation count
		circles	Search-depth circles
		links	Links between to articles if one cited another one

WORKFLOW

Retrieveing data

- **get_refs (list_of_dois, search_depth = 1, doi_na_drop = FALSE)**
- If any input doi was wrong, the function will indicate that and will ask the user whether carry on without that wrong doi (Error handling).
- The output is a data table with several columns.
- First few rows with "DOI = NA" are the input articles.
- The URLs of the input articles are accessible.

Constructing citation table

- **articles_citation_freq (ref_data, sorted = TRUE, top_n = 0)**
- The input "ref_data" should be initiated using "get_refs" function.
- If "top_n = 0", the function will return only DOIs and Cit.Freq, otherwise it returns DOIs, Cit.Freq, SearchDepth, author and year.

Visualization

- **plot_as_graph (ref_data, min_cit = 2, links=F, circles=T)**
- Each (dashed) circle presents the articles with individual search depth.
- This function depending on the number of "min_cit" can indicate which search depths are involved in the plot.
- The links presents the connection between two articles even in different search depths.

EXAMPLE

ERROR HANDLING

```
> doi_1 <- '10.1007/s11306-018-1449-2'
> doi_2 <- '10.jH'
> doi_3 <- '10.1111/ijfs.14794'
>
> dois_error_test <- c(doi_1,doi_2,doi_3)
>
> get_refs(dois_error_test,search_depth = 2)
<simpleWarning: 404 (client error): /works/10.jH - Resource not found.>
Continiou without that Doi?
Yes: 1
No: 2
```

Input data

```
doi_1 <- '10.1007/s11306-018-1449-2'
doi_2 <- '10.1038/s41592-021-01197-1'
doi_3 <- '10.1111/ijfs.14794'
doi_4 <- '10.1039/D1AY00173F'
doi_5 <- '10.1007/978-3-030-51652-9_4'
doi_6 <- '10.1007/s11306-019-1493-6'
doi_7 <- '10.1002/0471142727.mb3004s114'
doi_8 <- '10.4155/bio-2019-0014'
doi_9 <- '10.1002/mas.20108'
doi_10 <- '10.1007/s11306-010-0254-3'

dois <- c(doi_1,doi_2,doi_3,doi_4,doi_5,
          doi_6,doi_7,doi_8,doi_9,doi_10)

ref_data <- get_refs(dois, search_depth=2)
```

Retrieving data (First step)

```
> ref_data|> as_tibble() |>  print(n=23)
# A tibble: 65,836 × 9
```

	InputDoi	DOI	SearchDepth	author	year	journ…¹	artic…²	url	abstr…³
1	10.1007/s11306-018-1449-2	NA	0	Beale	2018	Metabo…	Review…	http…	NA
2	10.1038/s41592-021-01197-1	NA	0	Alseekh	2021	Nat Me…	Mass s…	http…	NA
3	10.1111/ijfs.14794	NA	0	Adebo	2021	Int. J…	Applic…	http…	NA
4	10.1039/D1AY00173F	NA	0	Misra	NA	Anal. …	Advanc…	http…	<p>Gas…
5	10.1007/978-3-030-51652-9_4	NA	0	Prodhan	2021	NA	Compre…	http…	NA
6	10.1007/s11306-019-1493-6	NA	0	Diez-Simon	2019	Metabo…	Mass s…	http…	NA
7	10.1002/0471142727.mb3004s114	NA	0	Fiehn	2016	CP Mol…	Metabo…	http…	NA
8	10.4155/bio-2019-0014	NA	0	Segers	2019	Bioana…	Analyt…	http…	<jats:…
9	10.1002/mas.20108	NA	0	Dettmer	2007	Mass S…	Mass s…	http…	NA
10	10.1007/s11306-010-0254-3	NA	0	Kopka	2011	Metabo…	Quanti…	http…	NA
11	10.1007/s11306-018-1449-2	10.1007/s11306-015-0839-y	1	H Abbiss	2015	Metabo…	NA	NA	NA
12	10.1007/s11306-018-1449-2	10.1007/978-1-4939-6747-6_23	1	R Adusumil…	2017	Method…	NA	NA	NA
13	10.1007/s11306-018-1449-2	NA	1	NA	NA	NA	NA	NA	NA
14	10.1007/s11306-018-1449-2	NA	1	JW Allwood	2008	Physio…	NA	NA	NA
15	10.1007/s11306-018-1449-2	10.1016/j.trac.2009.12.004	1	B Alvarez-…	2010	TrAC T…	NA	NA	NA
16	10.1007/s11306-018-1449-2	10.1002/elps.201500352	1	EG Armitage	2015	Electr…	NA	NA	NA
17	10.1007/s11306-018-1449-2	10.1016/j.envres.2012.12.001	1	JP Arrebola	2013	Enviro…	NA	NA	NA
18	10.1007/s11306-018-1449-2	10.1002/rcm.7505	1	S Baldwin	2016	Rapid …	NA	NA	NA
19	10.1007/s11306-018-1449-2	10.1021/ac501710y	1	MP Barrow	2014	Analyt…	NA	NA	NA
20	10.1007/s11306-018-1449-2	10.5936/csbj.201301009	1	J Bartel	2013	Comput…	NA	NA	NA
21	10.1007/s11306-018-1449-2	10.3390/ijms18010024	1	D Beale	2017	Intern…	NA	NA	NA
22	10.1007/s11306-018-1449-2	10.1016/j.scitotenv.2018.03.106	1	DJ Beale	2018	Scienc…	NA	NA	NA
23	10.1007/s11306-018-1449-2	10.1016/j.cresci.2011.10.026	1	DJ Beale	2012	Corros…	NA	NA	NA

```
# … with 65,813 more rows, and abbreviated variable names ¹journal.title, ²article.title, ³abstract
```

Inputs information

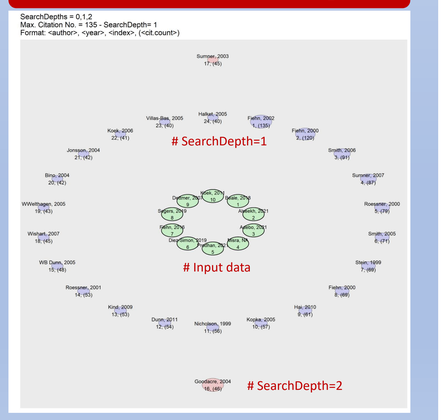
References information

Constructing citation table (Second step)

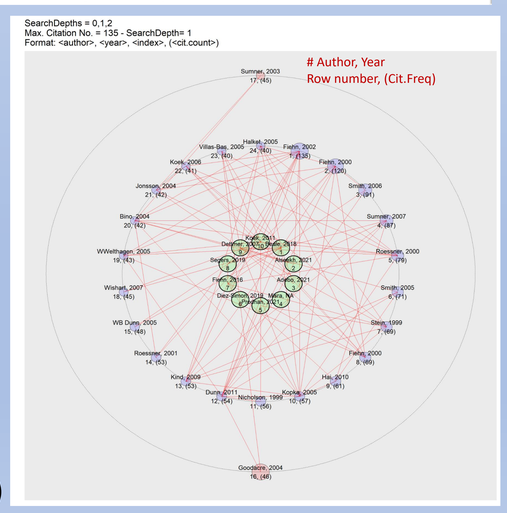
```
> articles_citation_freq(ref_data, top_n = 8) |> as_tibble()
# A tibble: 8 × 5
```

	DOIs	Cit.Freq	SearchDepth	author	year
	<chr>	<db1>	<db1>	<chr>	<db1>
1	10.1023/a:1013713905833	135	1	Fiehn	2002
2	10.1038/81137	120	1	Fiehn	2000
3	10.1021/ac051437y	91	1	Smith	2006
4	10.1007/s11306-007-0082-2	87	1	Sumner	2007
5	10.1046/j.1365-313x.2000.00774.x	79	1	Roessner	2000
6	10.1097/01.ftd.0000179845.53213.39	71	1	Smith	2005
7	10.1016/s1044-0305(99)00047-1	69	1	Stein	1999
8	10.1021/ac991142i	69	1	Fiehn	2000

Visualization (Third step)



plot_as_graph (ref_data, min_cit=45)



FUTURE DIRECTIONS

- This package is still experimental and maturing, we are continuous improving its function and collecting user feedback. Your comments are very valuable
- Add caching mechanism
- Applying other APIs
- Add new plotting functions

REFERENCES

1- Jane Webster and Richard T. Watson (2002), https://www.jstor.org/stable/4132319

2- Yair Levy and Timothy J. Ellis (2006), https://doi.org/10.28945/479

3- https://github.com/ropensci/rcrossref

SOURCE CODE AND LINKS

Package's repo: https://github.com/toscm/lirec (private)