

TidyESL:

Make discipline statistics more easy in R

王敏杰

2020 年 6 月 15 日

四川师范大学

初衷

在做科研信息分析报告的时候，换位思考，对方肯定想获取一些信息，比如：

- 机构的科研产出多少？影响力多少？

现实需求

在做科研信息分析报告的时候，换位思考，对方肯定想获取一些信息，比如：

- 机构的科研产出多少？影响力多少？
- 与同行比较，孰高孰低？

现实需求

在做科研信息分析报告的时候，换位思考，对方肯定想获取一些信息，比如：

- 机构的科研产出多少？影响力多少？
- 与同行比较，孰高孰低？
- 是否进入 ESI 学科？哪些进入？同行都哪些进入了？

现实需求

在做科研信息分析报告的时候，换位思考，对方肯定想获取一些信息，比如：

- 机构的科研产出多少？影响力多少？
- 与同行比较，孰高孰低？
- 是否进入 ESI 学科？哪些进入？同行都哪些进入了？
- 现在阈值线多少？距离阈值线还差多少？

现实需求

在做科研信息分析报告的时候，换位思考，对方肯定想获取一些信息，比如：

- 机构的科研产出多少？影响力多少？
- 与同行比较，孰高孰低？
- 是否进入 ESI 学科？哪些进入？同行都哪些进入了？
- 现在阈值线多少？距离阈值线还差多少？
- 高被引论文多少？高被引贡献了多少？

现实需求

在做科研信息分析报告的时候，换位思考，对方肯定想获取一些信息，比如：

- 机构的科研产出多少？影响力多少？
- 与同行比较，孰高孰低？
- 是否进入 ESI 学科？哪些进入？同行都哪些进入了？
- 现在阈值线多少？距离阈值线还差多少？
- 高被引论文多少？高被引贡献了多少？
- 二级机构的贡献比是多少？

现实需求

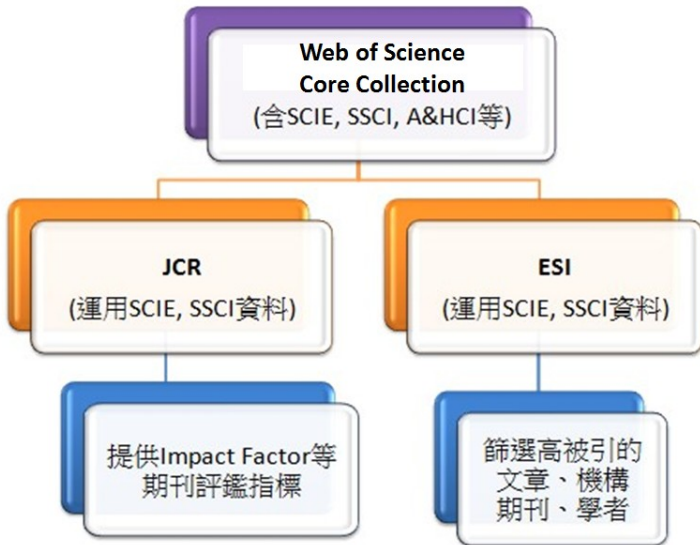
在做科研信息分析报告的时候，换位思考，对方肯定想获取一些信息，比如：

- 机构的科研产出多少？影响力多少？
- 与同行比较，孰高孰低？
- 是否进入 ESI 学科？哪些进入？同行都哪些进入了？
- 现在阈值线多少？距离阈值线还差多少？
- 高被引论文多少？高被引贡献了多少？
- 二级机构的贡献比是多少？
- 更多需求...

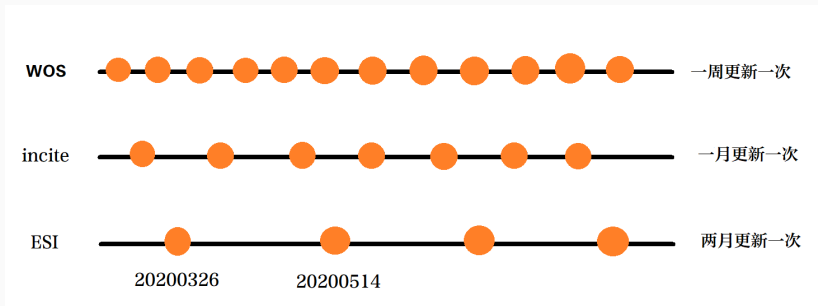
面临挑战

- 回答以上问题，往往需要
 - 多个数据源
 - 多个评价指标
 - 多个评估对象 (多个机构、多个学科)
 - 说明：评价机制越来越多元
- 带来的挑战
 - 多个数据集，如何整合？
 - 多个指标值，如何计量？
 - 思考：如何提高工作有效性、准确性

学科评价用到的数据源

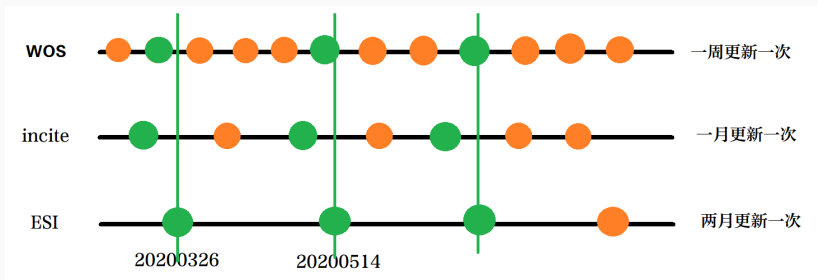


痛点 1



三个库更新时间不同步！如何匹配？

可接受的办法

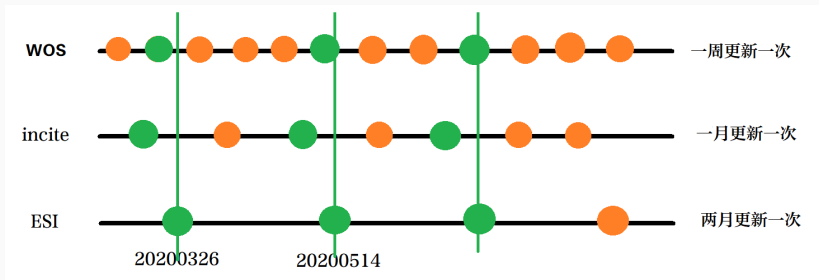


三个库：

- ESI 库
- incite 库
- WOS 库

一个妥协的办法：一旦 ESI 库更新，立马下载另两个库

痛点 2

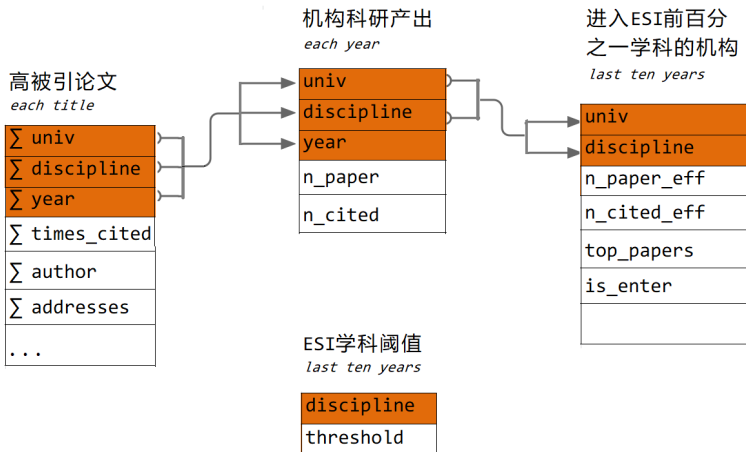


三个库**内容不同**:

- ESI 库: 包含 (ESI 学科阈值线, 高被引论文等)
- Incite 库: 包含某机构历年科研产出和影响力
- WOS 库: 包含某机构科研产出和影响力的题录清单

如何合并?

可接受的办法

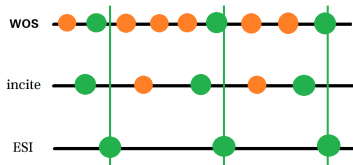


关联表，通过关键词合并，但统计口径呢...

思考与 tidyESI

tidyESI

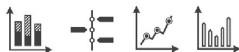
tidyESI 是一个 R 语言宏包，主要用于规整 ESI 和 incites 等数据格式 (下载办法参考[附录](#))，想法来源**bibliometrix**宏包



⚠ 检索的时间范围要一致，
比如都是近十年

⚠ 用什么关键词检索，
就用什么文件名保存

tidyESI



perlatex / tidyESI

Unwatch 1 Star 1 Fork 0

<> Code Issues 0 Pull requests 0 Actions Projects 0 Wiki Security 0 Insights Settings

A package to tidy ESI data Edit

Manage topics

12 commits 1 branch 0 packages 0 releases 1 contributor CC0-1.0 1.15 MB

Branch: master New pull request Create new file Upload files Find file Clone or download

perlatex type		Latest commit 6f38f94 6 days ago
R	adjust add_high_cited	6 days ago
data-raw	add_high_cited	7 days ago
data	add_high_cited	7 days ago
man	type	6 days ago
tests	init	13 days ago
.Rbuildignore	update	13 days ago
.gitignore	init	13 days ago

- <https://github.com/perlatex/tidyESI>
- `devtools::install_github("perlatex/tidyESI")`

函数

功能 1: 读取 incites 数据

```
library(tidyESI)
library(tidyverse)
# 一个机构
df <- read_incites(
  "./data_incites/Sichuan Normal University.csv")
```

univ	discipline	year	n_paper	n_cited
Sichuan Normal University	Physics	2000	6	104
Sichuan Normal University	Physics	2001	6	100
Sichuan Normal University	Physics	2002	22	303
Sichuan Normal University	Physics	2003	30	179
Sichuan Normal University	Physics	2004	26	242

功能 1: 读取 incites 数据

多个机构

```
dt <- here::here("data_incites_normal_univ") %>%  
  fs::dir_ls(regex = "\\*.csv$") %>%  
  read_incites()
```

univ	n_paper	n_cited
Anhui Normal University	5309	79687
Beijing Normal University	33283	511749
Capital Normal University	8167	92554
Central China Normal University	12420	242724
Chongqing Normal University	2942	28601

功能 2: 读取高被引论文数据

单个机构

```
sicnu <- read_highcited(  
  "./data_highcited/Sichuan Normal University.xlsx")
```

year	authors
2014	QIAO, JS;KONG, XH;HU, ZX;YANG, F;JI, W
2010	SUN, CL;LI, H;YU, DG;YU, MA;ZHOU, XA;LU, XY;HUANG, K;Z
2015	ZHANG, XC;ZHAO, GP;FANGOHR, H;LIU, JP;XIA, WX;XIA, J;M
2015	LIANG, S;WANG, T;HU, X;LUO, J;LI, W;WU, X;DUAN, Y;JIN, F
2018	HAN, JR;LIU, ZC;MA, YJ;CUI, GW;XIE, FY;WANG, FX;WU, YP
2016	WEI, GW

功能 2: 读取高被引论文数据

多个机构

```
hc <- here::here("data_highcited_normal_univ") %>%  
  fs::dir_ls(regex = "\\\\.xlsx$") %>%  
  tidyESI::read_highcited()
```

univ	n_paper_high	n_cited_high
Anhui Normal University	41	5465
Beijing Normal University	359	53527
Capital Normal University	45	7949
Central China Normal University	208	43351
Chongqing Normal University	40	2516

功能 3: 读取进入 ESI 前百分之一学科机构清单

多个机构

```
top <- here::here("data_isEnter_normal_univ") %>%  
  fs::dir_ls(regex = "\\\\.xlsx$") %>%  
  tidyESI::read_top_one_percent()
```

univ	discipline	top_papers	is_enter_top
Anhui Normal University	Chemistry	20	TRUE
Beijing Normal University	Environment/Ecology	53	TRUE
Beijing Normal University	Geosciences	72	TRUE
Beijing Normal University	Chemistry	35	TRUE

演示数据

为了演示方便，我这里只选取部分

```
df_sub <- df %>% dplyr::filter(  
  year %in% c("2018", "2019"),  
  discipline %in% c("Physics", "Chemistry")  
)  
df_sub
```

univ	discipline	year	n_paper	n_cited
Sichuan Normal University	Physics	2018	62	396
Sichuan Normal University	Physics	2019	82	90
Sichuan Normal University	Chemistry	2018	91	1656
Sichuan Normal University	Chemistry	2019	102	471

功能 4: 添加高被引论文信息

添加该机构各学科高被引逐年信息

```
df_sub %>%
```

```
  add_high_cited(discipline, source = hc,  
                 scope = "each year") %>%
```

```
  select(-univ)
```

discipline	year	n_paper	n_cited	n_paper_high	n_cited_high
Physics	2018	62	396	2	151
Physics	2019	82	90	1	18
Chemistry	2018	91	1656	11	908
Chemistry	2019	102	471	8	193

功能 4: 添加高被引论文信息

添加该机构各学科高被引汇总信息 (取决于检索周期)

```
df_sub %>%
```

```
  add_high_cited(discipline, source = hc,  
                 scope = "all year") %>%
```

```
  select(-univ)
```

discipline	year	n_paper	n_cited	n_paper_high	n_cited_high
Physics	2018	62	396	6	2303
Physics	2019	82	90	6	2303
Chemistry	2018	91	1656	21	1656
Chemistry	2019	102	471	21	1656

功能 5: 添加阈值

tidyESI 宏包会定期更新最新阈值

添加最新的阈值线

```
df_sub %>%
```

```
  add_esi_threshold(discipline, date = "last") %>%
```

```
  select(-univ)
```

discipline	year	n_paper	n_cited	20200514
Physics	2018	62	396	21050
Physics	2019	82	90	21050
Chemistry	2018	91	1656	8188
Chemistry	2019	102	471	8188

功能 5: 添加阈值

添加全部的阈值线

```
df_sub %>%
```

```
  add_esi_threshold(discipline, date = "all") %>%  
  select(-univ)
```

discipline	year	n_paper	n_cited	20200326	20200514
Physics	2018	62	396	21325	21050
Physics	2019	82	90	21325	21050
Chemistry	2018	91	1656	8502	8188
Chemistry	2019	102	471	8502	8188

功能 5: 添加阈值

添加指定日期的阈值线

```
df_sub %>%  
  add_esi_threshold(discipline,  
                    date = c("20200326", "20200514")) %>%  
  select(-univ)
```

discipline	year	n_paper	n_cited	20200326	20200514
Physics	2018	62	396	21325	21050
Physics	2019	82	90	21325	21050
Chemistry	2018	91	1656	8502	8188
Chemistry	2019	102	471	8502	8188

功能 6: 添加学科信息

添加是否进入 *ESI* 前百分之一学科

df_sub %>%

add_is_enter_top(univ, discipline, source = top)

#> Please make sure the source information is up to date.

univ	discipline	year	n_paper	n_cited	is_enter_top
Sichuan Normal University	Physics	2018	62	396	NA
Sichuan Normal University	Physics	2019	82	90	NA
Sichuan Normal University	Chemistry	2018	91	1656	NA
Sichuan Normal University	Chemistry	2019	102	471	NA

功能 7: 汉化学校名

univ	discipline	year
Sichuan Normal University	Physics	2018
Sichuan Normal University	Physics	2019
Sichuan Normal University	Chemistry	2018
Sichuan Normal University	Chemistry	2019

```
df_sub %>% select(-starts_with("n_")) %>%  
  tidyESI::add_univ_name_cn(univ)
```

univ	discipline	year	univ_cn
Sichuan Normal University	Physics	2018	四川师范大学
Sichuan Normal University	Physics	2019	四川师范大学
Sichuan Normal University	Chemistry	2018	四川师范大学
Sichuan Normal University	Chemistry	2019	四川师范大学

功能 8: 汉化学科名

univ	discipline	year	n_paper	n_cited
Sichuan Normal University	Physics	2018	62	396
Sichuan Normal University	Physics	2019	82	90
Sichuan Normal University	Chemistry	2018	91	1656
Sichuan Normal University	Chemistry	2019	102	471

```
df_sub %>% add_discipline_cn(discipline)
```

univ	discipline	year	n_paper	n_cited	discipline
Sichuan Normal University	Physics	2018	62	396	物理学
Sichuan Normal University	Physics	2019	82	90	物理学
Sichuan Normal University	Chemistry	2018	91	1656	化学
Sichuan Normal University	Chemistry	2019	102	471	化学

功能 9: issn 映射到 ESI 学科

id	issn
1	2053-1583
2	2329-7662

```
dt %>% mutate(  
  discipline = map_issn_to_esi_discipline(issn)  
)
```

id	issn	discipline
1	2053-1583	Materials Science
2	2329-7662	Engineering

功能 10: 二级机构映射

(目前仅对四川师范大学)

coll

Sichuan Normal Univ, Coll Geog & Resource Sci,
Sichuan Normal Univ, Coll Geog & Resources Sci,
Sichuan Normal Univ, Coll Geog & Resources,

```
tb %>% mutate(  
  coll_name = map_addr_to_college_name(coll)  
)
```

coll	coll_name
Sichuan Normal Univ, Coll Geog & Resource Sci,	地理与资源科学学院
Sichuan Normal Univ, Coll Geog & Resources Sci,	地理与资源科学学院
Sichuan Normal Univ, Coll Geog & Resources,	地理与资源科学学院

下一步

未来打算

- 支持更多文件格式
 - `read_highcited(*.xlsx, *.csv)`
- 解析 WOS 题录
 - `parse_title_highcited()`
- 解析高被引题录
 - `parse_title_wos()`
- 用户可添加二级机构的映射关系表
 - `import_college_name()`

感谢聆听, 欢迎批评

38552109@qq.com

附录

ESI 学科阈值线

ESI 学科阈值

近十年，各 ESI 学科被引次数排在前 1% 的作者或机构的最低被引频次

tidyESI

tidyESI 宏包会自动更新最新阈值

incites 数据

某个机构历年科研产出和影响力等

获取办法


- 点击进入 web of science(SCI)，页面最上面一排 incites
- 点击菜单第一个 Analysis 下拉选项中的 Organizations
- 设置
 - 选择 Time period 2000 - 2020
 - 在 By_Research_Output 栏下的 Research Area 选择第二个 Essential Science Indicators
 - 选择 Document Type = Review + Article
 - 更新 Update results
- 检索
 - 输入 sichuan normal university
 - 点击 sichuan normal university 下出现的小框框中选择 Research Areas 后，点击 Refocus
- 下载
 - 点击 Benchmarks 下载（注意要勾选 Trend Data，不勾就是汇总表）
 - 重命名 Sichuan Normal University.csv (注意与检索词一致，很重要)
 - 每个学校一张表

Web of Science **InCites** Journal Citation Reports Essential Science Indicators EndNote Publons 38552109@qq.com ▾ Help English ▾

InCites


Clarival Analytics

Analysis ▾ Reports ▾ My Folders My Organization


 New in InCites

Explore InCites Data


Create dynamic tables and graphs based on your needs.




Researchers




Organization




Locations



Research Areas



Journals, Books,
Conference
Proceedings



Funding Agencies

InCites System Reports

Learn about research performance, collaboration and most cited documents.

--	--	--	--	--

The screenshot shows the Incites Web of Science Documents interface. The left sidebar contains various filters and settings, while the main area displays a table of search results for 'Sichuan Normal University'.

Annotations:

- 1:** Points to the 'Entity Type' dropdown menu, which is currently set to 'Organizations'.
- 2:** Points to the 'Document Type' dropdown menu, which is currently set to 'Article'.
- 3:** Points to the 'Research Area' dropdown menu, which is currently set to 'Essential Science Ind...'.
- 4:** Points to the 'Research Areas' dropdown menu in the main results area, which is currently set to 'Research Areas'.
- 5:** Points to the 'Refocus' button in the main results area.
- 6:** Points to the 'Benchmarks' button in the top right corner of the main results area.

Table Data:

Name	Rank	Web of Science Documents	Category Normalized Citation Impact	Times Cited	% Docs Cited
Sichuan Normal University	1	3,991	1.12	34,076	70.28%

高被引论文

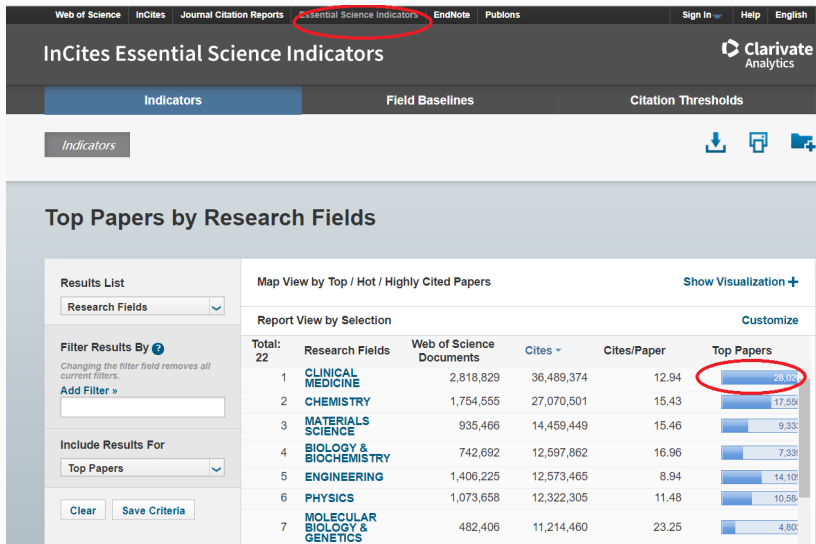
高被引论文

同一年同一个 ESI 学科发表论文的被引次数按由高到低进行排序，排在前 1% 的论文

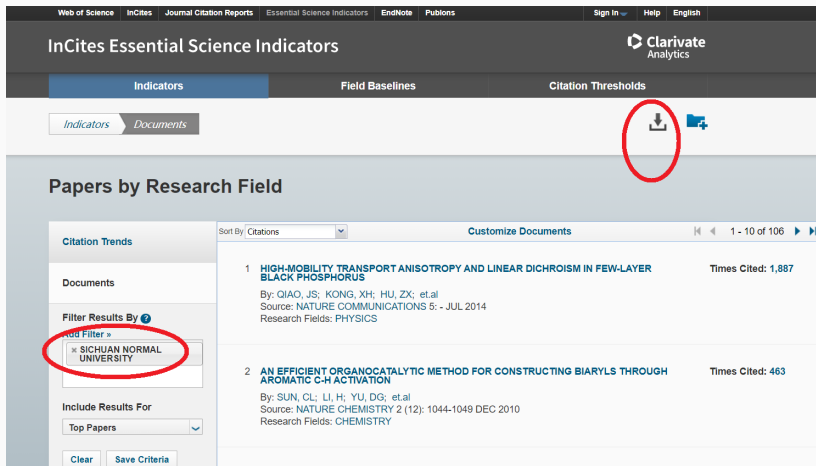
获取方法

- 打开 web of science, 点击 Essential Science Indicators
- 点击页面右边 “Top Papers” 下的进度条，进入高被引论文查询页面
- 点击页面左边 “Add Filter”，选择 “institution”，进入机构高被引论文查询页面
- 检索框中输入 Beijing Normal University，查询机构高被引论文详情
- 下载 XLS 格式
- 重命名 Beijing Normal University.csv (与检索词一致，很重要)

高被引论文



高被引论文



Web of Science InCites Journal Citation Reports Essential Science Indicators EndNote Publons Sign In Help English

InCites Essential Science Indicators

Clarivate Analytics

Indicators Field Baselines Citation Thresholds

Indicators Documents

Papers by Research Field

Citation Trends

Sort By: Citations Customize Documents 1 - 10 of 106

Documents

Filter Results By

Add Filter

x SICHUAN NORMAL UNIVERSITY

Include Results For

Top Papers

Clear Save Criteria

- HIGH-MOBILITY TRANSPORT ANISOTROPY AND LINEAR DICHROISM IN FEW-LAYER BLACK PHOSPHORUS** Times Cited: 1,887
By: QIAO, JS; KONG, XH; HU, ZX; et.al
Source: NATURE COMMUNICATIONS 5: - JUL 2014
Research Fields: PHYSICS
- AN EFFICIENT ORGANOCATALYTIC METHOD FOR CONSTRUCTING BIARYLS THROUGH AROMATIC C-H ACTIVATION** Times Cited: 463
By: SUN, CL; LI, H; YU, DG; et.al
Source: NATURE CHEMISTRY 2 (12): 1044-1049 DEC 2010
Research Fields: CHEMISTRY

- 打开 <https://www.webofknowledge.com/>，进入核心合集
- 输入学校全名：比如 Sichuan Normal University
- 选择“机构扩展”检索
- 选择时间范围：“2010-2020 年”
- 选择“SCI/SSCI/A&HCI”
- 点击检索
- 文档类型精炼：“Article + Review”
- 一次显示最多 50 条，一次下载最多 500 条
- 选择“其他类型下载”+“全记录”+“win UTF”
- 下载保存