# 科技情报数据中有关作者名、作者机构、关键词的具体清洗方法

冯靖雯

2020年4月28日

# 内容提要

- 数据清洗、范畴与方法
- 常见的数据清洗算法（属性清洗算法）
- 作者名清洗
- 机构清洗
- 关键词清洗
- 实际应用

# 数据清洗

数据清洗目的：是为了提高数据质量而剔除数据中错误记录的一种技术手段，在实际应用中通常与数据挖掘技术、数据仓库技术、数据整合技术结合应用。

数据清洗技术的基本原理为：在分析数据源特点的基础上，找出数据质量问题原因，确定清洗要求，建立起清洗模型，应用清洗算法、清洗策略和清洗方案对应到数据识别与处理中，最终清洗出满足质量要求的数据。

数据清洗是数据分析、数据挖掘的前提，也是数据预处理的关键环节，可保证数据质量和数据分析的准确性。

# 数据清洗

数据清洗定义：

在原始数据库中找出可信度较低的字段，利用字段间的函数依赖关系清洗字段和数据，并对数据进行修复。

《面向知识服务的知识组织理论与方法》书中提到：

数据清洗（data cleaning）的目的是检测数据中存在的错误和不一致，提出或者改正它们，以提高数据的质量，透视信息的内涵，提升信息的价值，从而保障知识库的知识价值。

# 数据清洗的范畴与方法

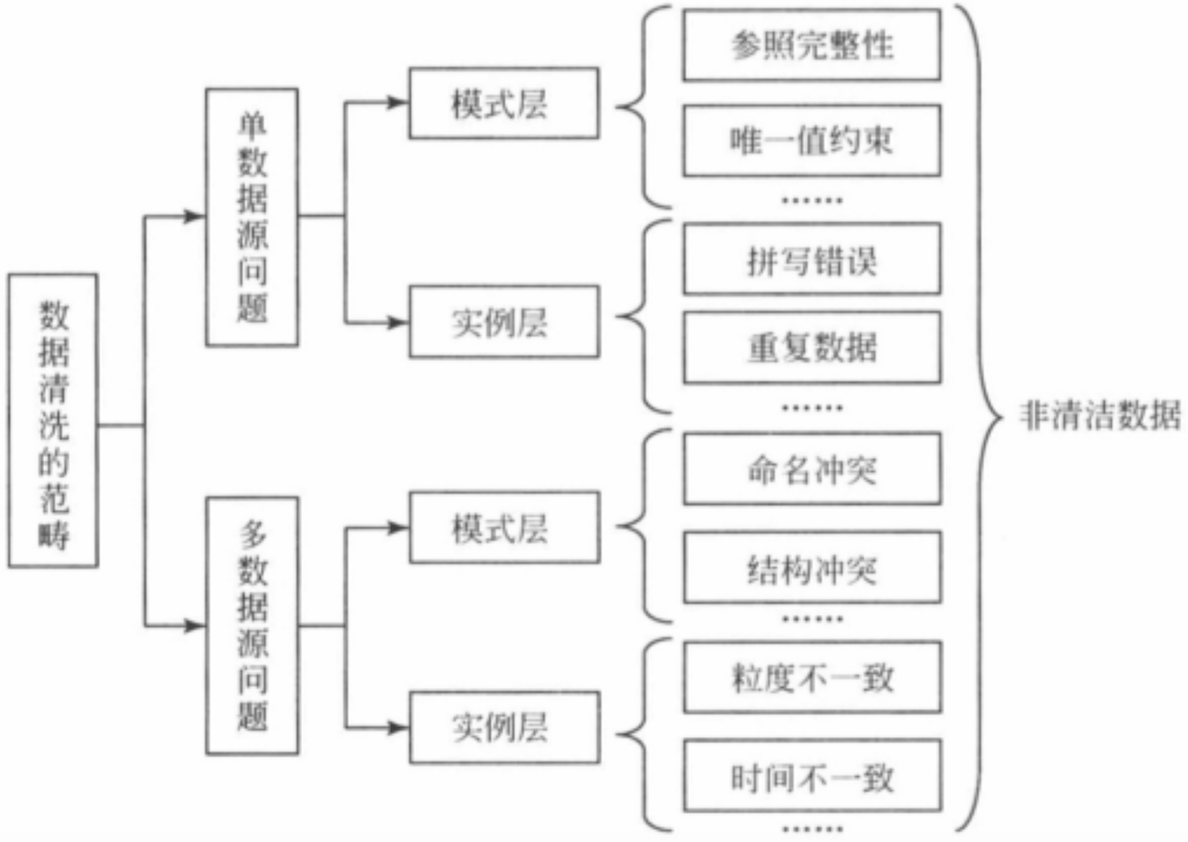



图 13-9　数据清洗的范畴

针对知识库中的非清洁数据（dirty data），对模式层的非清洁数据与实例层的非清洁数据分别提出了一些有针对性的方法，概括起来如表13-2所示。模式层针对知识库涉及过程的数据结构的缺陷，如缺乏完整性约束的设计；实例层关注数据内容的问题，如数据缺失、数据重复等。

# 数据清洗的范畴与方法

模式层针对知识库涉及过程的数据结构的缺陷，如缺乏完整性约束的设计；实例层关注数据内容的问题，如数据缺失、数据重复等。

**表 13-2　非清洁数据的分类及清洗方法**

| 层 | 对象 | 清洗方法 | 主要途径 | 解决范围 | 备注 |
|---|---|---|---|---|---|
| 模式层 | 数据结构设计不合理 | 结构冲突清洗方法 | 手工干预法（准确性高、效率低） | 类型冲突、关键字冲突 | 程序自身无法识别数据类型的冲突和关键字的冲突 |
| | | | 函数依赖法（存在必须满足依赖关系的条件的局限） | 依赖冲突 | 通过属性间的函数依赖关系，查找违反函数依赖关系的值 |
| | 属性约束不够 | 噪声数据清洗方法 | 分箱法 | 数字类型的数据 | 考察属性值周围的值来平滑属性的值 |
| | | | 人机结合法 | 通用性 | 对于计算机检测遗漏的可疑数据很难修正 |
| | | | 规则库 | 规则性强的数据 | 建立规则库，通过规则约束数据 |
| 实例层 | 属性值的非清洁数据 | 以统计和总结规律的方式计算错误值，并修正 | 统计法、聚类法、关联规则 | 属性值的错误值 | |
| | | 人工填写空缺值 | 手工检测 | 属性值为空值 | 可以使用属性的平均值、最大值、最小值或更为复杂的概率统计函数值填充空缺值 |
| | 重复数据 | 基于字段的清洗 | 编辑距离算法、Cosine 相似度函数算法 | 缩写字符串、字段重复 | 编辑距离算法最常用算法，易于实现 |
| | | 基于记录的清洗 | 排序邻居算法、优先队列算法 | 依赖关键字 | 对于阈值的选择敏感 |
| | 非关联数据（孤立数据） | 基于距离、深度、密度、聚类的方法 | 通过元数据找到知识的关联性 | 无相关相似的数据 | |

# 数据清洗的范畴与方法

表 13-3　非清洁数据的片段示例

| ID | 超市名称 | 归属城市 | 商品编码 | 联系电话 | 联系人 |
|---|---|---|---|---|---|
| 1 | Wal-Mart | Nanjing | J201301 | 025-83786382 | Lisa |
| 2 | Carrefour | Shanghai | N-201302 | 021-64458973 | Jerry |
| 3 | Wal-Mart | NJ | J201301 | 025-83786382 | Lisa |
| 4 | Walmart | Shanghai | N201303 | 62573489 | Lisa |
| 5 | Carrefour | Nanjing | J-201301 | 65737841 | Jerry |
| 6 | Mal-Mart | Nanjing | J201301 | 83786382 | Lisa |

以表13-3所示的一个非清洁数据片段为例，不难判断：
不完整的数据有：（ID：3、字段归属城市）“NJ”；
错误数据有：（ID：6、字段：超市名称）“Mal-Mart”；
重复数据：（ID：1、3、6）标识同一个实体。

# 数据清洗的范畴与方法

如“Wal-Mart”在同一个实体 (ID为1、3和6) 3条数据片段中出现了2次, 则清洁度为2/3, 如表4所示。

表4 清洁度表示示例

| ID | 超市名称 | 归属城市 | 商品编码 | 联系电话 | 联系人 |
|---|---|---|---|---|---|
| 1 | Wal-Mart(2/3)<br>Mal-Mart(1/3) | Nanjing(2/3)<br>NJ(1/3) | J201301(1.0) | 025-83786382(2/3)<br>83786382(1/3) | Lisa(1.0) |

# 数据清洗的范畴与方法

表 14-1　脏数据类型与出现原因

| 数据源 | 脏数据层次 | 类型 | 脏数据实例 | 出现原因 |
|---|---|---|---|---|
| 单数据源 | 模式层 | 字段约束 | 2012-13-01 | 不在约束范围内 |
| | | 属性依赖冲突 | AGE=32<br>BITHDAY=85-5-15 | 两个维度值不一致 |
| | | 唯一性冲突（如主键等） | 顾客 1（customerid = 1213，idcard = 12442）<br>顾客 2（customerid = 1214，idcard = 12442） | 两个不同的记录依附的主维度完全相同 |
| | | 参照完整性冲突 | Max institutionid=25<br>institutionid=26 | 超出设定的值范围，没有相应的对象数据 |
| | 实例层 | 单属性嵌入值过多 | 姓名=李爱国，江苏人，1977-5-16 | 单个属性值含有过多的信息 |
| | | 拼写错误 | Institution 输成了 institation | 数据输入错误、数据传输过程中发生的错误 |
| | | 空白值 | 姓名：李爱国<br>性别：男<br>年龄：_____（选填） | 用户不愿意泄露的信息、很多可选填的输入表单设计 |
| | | 噪音数据 | 4　5　8　9　25 | 真实的数据、人为故意的数据输入错误、数据采集设备有问题、数据传输过程中发生错误 |
| | | 数据重复 | 记录 1：学校名称 = Nanjing Unverisity，邮编=210093；<br>记录 2：学校名称 = Nanjin Unver，邮编=210093 | 指同一个现实实体在数据集合中用多条不完全相同的记录来表示，由于它们在格式、拼写上的差异，导致数据库管理系统不能正确识别 |
| | | 数据退化 | 025-3593597 | 电话号码的升位，数据变成无效数据 |

# 数据清洗的范畴与方法

| | | | | |
|---|---|---|---|---|
| 多数据源 | 模式层 | 命名冲突 | 数据源 1：南京大学<br>数据源 2：南大 | 同一实体不同的名称 |
| | | 结构冲突 | 数据源 1：客户存款=2300（字符型）<br>数据源 2：客户存款=2300（数值型） | 计量单位不一致、不同代码的不一致使用、一个代码有不一致的含义、相同的意义不同的代码，格式不同 |
| | 实例层 | 时间不一致 | 数据源 1：销售总量：2500（季度）<br>数据源 2：销售总量：15000（年度） | 不同时间层次上的数据在同一层次进行比较与计算 |
| | | 粒度不一致 | 数据源 1：销售总量：2500（市级）<br>数据源 2：销售总量：15000（省级） | 不同粒度大小的数据在同一环境下进行比较与计算 |
| | | 数据重复 | 数据源 1：销售总量：2500（南京市）<br>数据源 2：销售总量：2500（南京市） | 相同的数据在合并后的数据库中出现两次及以上 |

# 数据清洗

数据清洗层根据根据检测分析层提供的资源质量报告，选择适当的数据清洗策略对数据进行清洗。主要任务包括筛选数据、合并重复数据、补充完善缺失数据

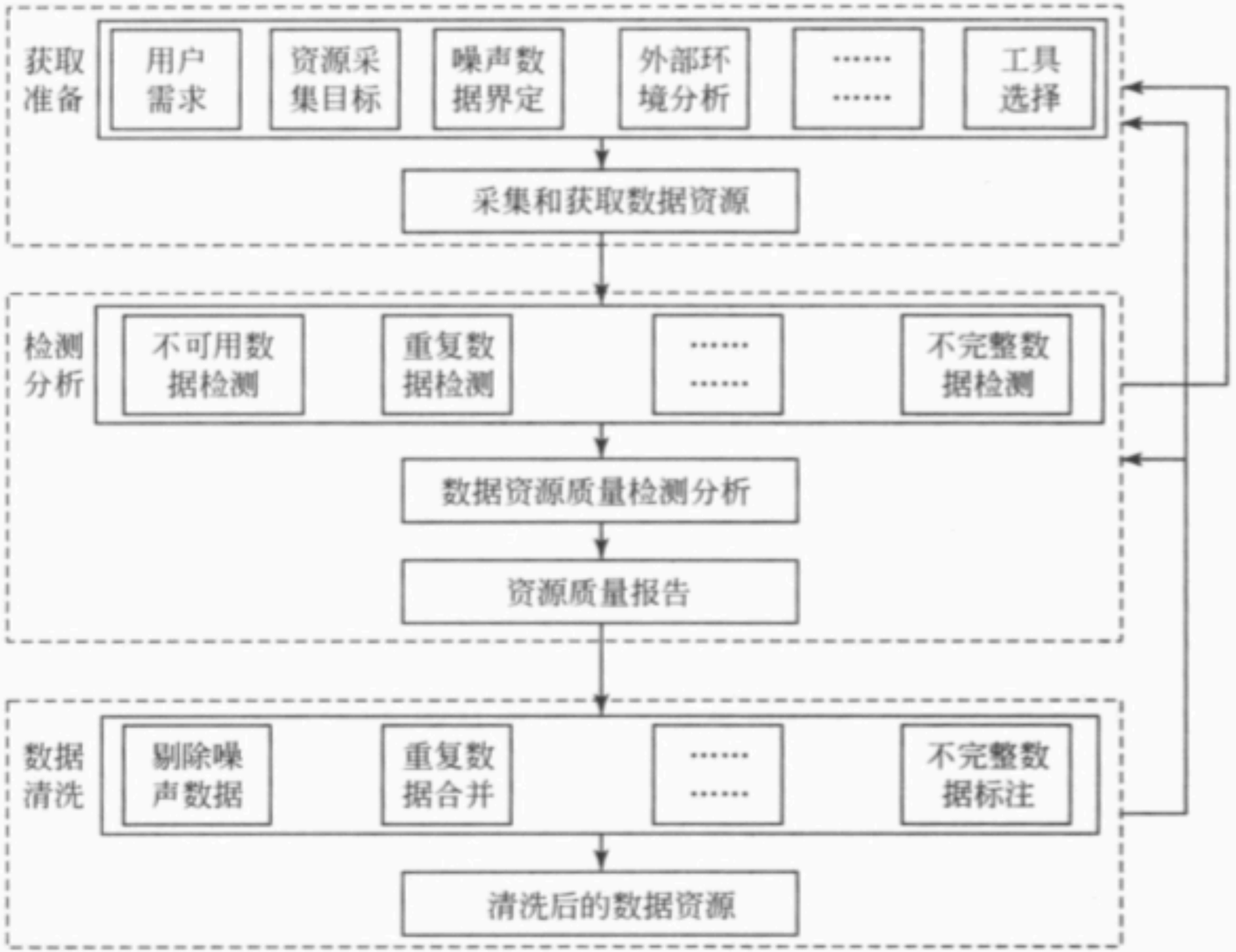
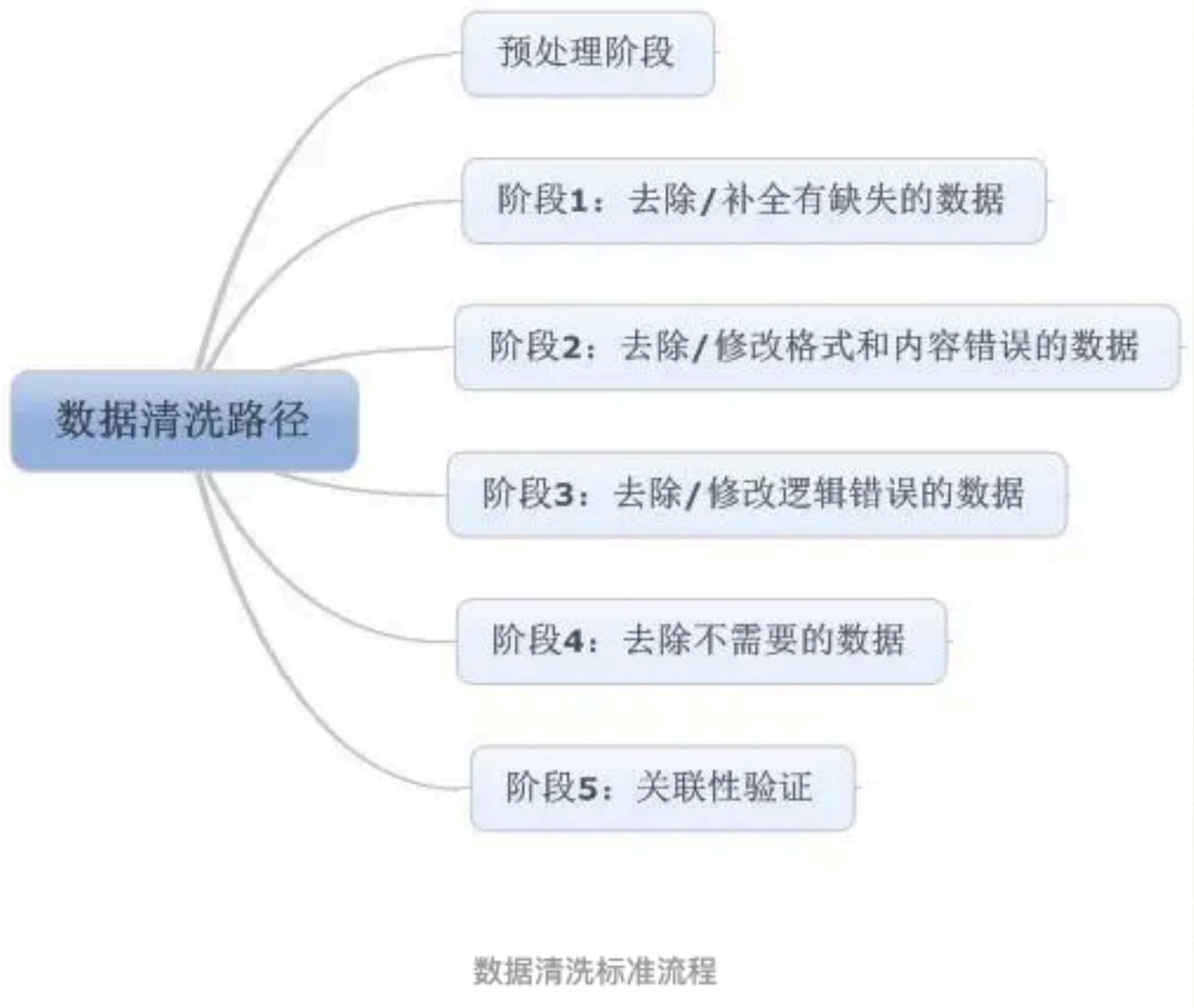



图 5-4　资源获取与清洗总体框架示意图

# 数据清洗标准流程




**预处理阶段主要做两件事情：**

一是将数据导入处理工具。通常来说，建议使用数据库。如果数据量大（千万级以上），可以使用文本文件存储+Python操作的方式。

二是看数据。这里包含两个部分：一是看元数据，包括字段解释、数据来源、代码表等等一切描述数据的信息；二是抽取一部分数据，使用人工查看方式，对数据本身有一个直观的了解，并且初步发现一些问题，为之后的处理做准备。

# 常见的数据清洗算法（属性清洗算法）

## 1、空缺值的清洗

对于空缺值的清洗可以采取忽略元组，人工填写空缺值，使用一个全局变量填充空缺值，使用属性的平均值、中间值、最大值、最小值或更为复杂的概率统计函数值来填充空缺值。

## 2、不一致数据的清洗

对于有些事务，所记录的数据可能存在不一致。有些数据不一致，可以使用其他材料人工加以更正。例如，数据输入时的错误可以使用纸上的记录加以更正。知识工程工具也可以用来检测违反限制的数据。例如，知道属性间的函数依赖，可以查找违反函数依赖的值。此外，数据集成也可能产生数据不一致。

## 3、重复数据的清洗

目前消除重复记录的基本思想是“排序和合并”，先将数据库中的记录排序，然后通过比较邻近记录是否相似来检测记录是否重复。

## 4、异常值的清洗

异常值，即在数据集中存在不合理的值，又称离群点。对属性值进行一个描述性的统计，从而查看哪些值是不合理的。一般可以删除含有异常值的记录、将异常值视为缺失值，交给缺失值处理方法来处理，或者用平均值来修正。

# 作者名清洗

重名问题、全名简写问题
中文中同音不同字

连接到 ORCID

## Zhong, Nanshan

查看可能的匹配作者

作者 ID: 7102137996

归属机构:

1994-2019 Guangzhou Medical University, Guangzhou, China
1991-2019 Guangzhou Institute of Repiratory Disease, Guangzhou, China
2014-2019 National Clinical Research Center for Respiratory Disease, Guangzhou, China

收起 查看全部

其他姓名格式: Zhong, N. S. | Nan-Shan, Zhong | Shan Zhong, Nan | Zhong, Nan Shan | Zhong, Nan shan | Zhong, Na Shan | Zhong, Nan S. | Zhong, N. | Nanshan, Zhong

学科类别: Medicine | Biochemistry, Genetics and Molecular Biology | Immunology and Microbiology | Pharmacology, Toxicology and Pharmaceutics | Agricultural and Biological Sciences | Multidisciplinary | Environmental Science | 查看全部

# 作者名清洗

1、比较典型的就是头、尾、中间的空格，也有可能出现姓名中存在数字符号等问题，这种情况下，需要以半自动校验半人工方式来找出可能存在的问题，并去除不需要的字符

例如：张三/张 三

2、名字缩写

例如：Lee S./Lee Steven

| | A | B |
|---|---|---|
| 1 | 作者 | 作者 ID |
| 2 | Kankanamge N., Yigitcanlar T., Goonetilleke A. | 57207298965;6505536041;6604084914; |
| 3 | Rapp A., Tirabeni L. | 23398572100;57191836654; |
| 4 | Laspidou C.S., Mellios N.K., Spyropoulou A.E., Kofinas | 6602289466;56991051200;57197751724;56 |
| 5 | Smutny Z., Vehovar V. | 56028877100;55991627400; |
| 6 | Obamiro K., West S., Lee S. | 57190161875;57215198877;57215191910; |
| 7 | Feller D.J., Lor M., Zucker J., Yin M.T., Olender S., Ferris | 56326233200;55641704600;57055161800;8 |
| 8 | Pee L.G., Pan S.L., Li M., Jia S. | 23668683800;57203894889;57191224752;1 |
| 9 | Ilie V., Turel O. | 18133890800;12760214400; |
| 10 | Ong T.C., Duca L.M., Kahn M.G., Crume T.L. | 55901212200;26647313000;56775542400;8 |
| 11 | Fathi S., Sajadzadeh H., Mohammadi Sheshkal F., Aran | 57216292303;57215214632;57216293303;5 |
| 12 | Barros J.M., Duggan J., Rebholz-Schuhmann D. | 57195106401;7102588251;6507852707; |
| 13 | Kemp J., Zhang T., Inglis F., Wiljer D., Sockalingam S., ( | 57215644487;57210371009;57210373094;6 |
| 14 | Reuter C., Stieglitz S., Imran M. | 36857282000;8982225800;55598006500; |
| 15 | Kaufhold M.-A., Rupp N., Reuter C., Habdank M. | 57015128700;56735975400;36857282000;5 |
| 16 | Tugwell P., Knottnerus A. | 35374403700;57209196039; |
| 17 | Ebrahimpour Boroojeny A., Shrestha A., Sharifi-Zarchi | 57204360858;57201796639;55670807400;5 |
| 18 | Andrews S.M., Raspa M., Edwards A., Moultrie R., Turn | 57214871538;24554591700;56323788300;5 |
| 19 | Ebrahimpour Boroojeny A., Shrestha A., Sharifi-Zarchi | 57204360858;57201796639;55670807400;5 |
| 20 | Wachter T., Del Frari B., Edlinger M., Morandi E.M., Ma | 56437511000;57204584880;23102703300;5 |
| 21 | Onoda M. | 7202766183; |
| 22 | | |
| 23 | | |

# 作者名清洗

一篇文章可能有多个作者，统一作者也会发表多篇文章。
在传统的引文索引数据库中，仅存储了发文作者的姓名、地区和所在机构这几项简单的信息，如果能将作者的出生日期、性别、研究方向等详细的背景资料也在数据库中进行组织和存放，用户通过引文索引就可以获得更深层面的知识服务。
例如，可以查看某领域年轻学者的研究方向，可以分析和对比不同年龄段学者的研究特点，可以比较哪些学科更适合女性学者的发展，甚至可以对学科的发展前景作出预测。

# 机构清洗

传统引文索引中，对于机构信息的标注往往存在一些不足和误区，具体分析如下：

（1）作者的区分问题。一般来说，发文作者都会标注机构信息，但因工作或学习需要，同一作者常常会标注多个机构。另外，由于机构名称标注的不规范，有人用全称有人用简称，以及机构合并或更名等因素，相同的机构会出现多个不同的名称。

（2）跨机构合作问题。所谓跨机构是指处于不同机构的不同学者之间的合作，同一个作者标注多个机构的情况不属于跨机构合作的范畴。

（3）不同类型机构的表示问题。如何对机构进行分类并有有效地表示是提供深层次知识服务的前提。

# 机构清洗

检索策略及方法：在WOS数据库的检索页面中的地址栏输入“Nanjing normal univ”，在日期范围内输入“2015—2015”。选择数据SCI-EXPANDED、SSCI、A&HCI、CPCI-S、CPCI-SSH，检索时间2016年1月10日，共计检索出982条记录。

WOS数据库的地址字段就比较规范，下面是其中一条典型的地址字段记录：
[Lu, Si-Yuan;Zhou, Xing-Xing;Zhang, GuangShuai]Nanjing Normal Univ, Sch Comp Sci&Technol, Nanjing 210023, Jiangsu, Peoples R China;
[Wei, Ling]Shanghai Jiao Tong Univ, Sch Elect Informat&Elect Engn, Shanghai 200030, Peoples R China。

# 机构清洗

WOS的地址字段可以归纳为下面的模型:[authors (N, 1) ]address (N, 1) ;…[authors (N, x) ]address (N, x) ;…[authors (N, Y) ]address (N, Y) ;其中authors (N, x) 是第N条记录的第x个作者群，address (N, x) 是第N条记录的第x个作者群的共同署名机构。

# 机构清洗

将得到的数据记录中的C1字段复制到文本处理软件Emeditor中，利用正则表达式，查找“;\[”, 替换为“;/t[”。
处理后的文本记录可以表达为:[authors（N，1）]address（N，1）;…[authors（N，x）]address (N, x) ;…[authors (N, Y) ]address (N, Y) ;
再将处理后的数据复制到excel中，可以获得第N条记录中第x个作者群和机构信息为C1 (N, x) =[authors (N, x) ]address (N, x) ;
再对C1（N，x）分析，利用正则表达式， 查找“]”， 并替换为“]/t”，进而可以得到authors (N, x) 以及相应的address (N, x) 。
通过上述的文本处理和excel处理，就获得了所有记录的所有排序的署名作者群及相应的署名作者机构信息。

# 机构清洗

EmEditor是日本的江村软件公司（Emurasoft）所开发的一款在Windows平台上运行的文字编辑程序。EmEditor以运作轻巧、敏捷而又功能强大、丰富著称，得到许多用户的好评。 简单好用的文本编辑器

**正则表达式**，又称规则表达式。（英语：Regular Expression，在代码中常简写为regex、regexp或RE），计算机科学的一个概念。
正则表达式通常被用来检索、替换那些符合某个模式（规则）的文本。

# 机构清洗

在excel中对address（N，x）字段分析，依次遍历x=1…Y，判断address（N，x）是否包含“nanjing normal univ”，假设address（N，k）是第N条记录中第一个包含“nanjing normal univ”的地址信息，记录jg (N) =k, taget (N) =address (N, k) , authors (N) =authors (N, k) 。【在进行上述数据分析时发现有一条记录的整个C1字段中不包含“nanjing normal univ”，经研究发现这条记录之所以被检出，是因为在通讯作者字段（RP）中出现了“nanjing normal univ”。】

如果是第M条记录的CI字段中没有出现nanjing normal univ，则假定jg（M）=0，taget（M）=null。将address (N, x) , N=1…982, 进行遍历处理，数据清洗后获得三组数列jg (N) 、taget (N) 、authors (N) 。jg (N) 是南京师范大学在第N篇论文的机构排名，taget (N) 是南京师范大学的具体署名地址信息，authors (N) 是对应的作者群。

# 机构清洗

对jg (N) 分析可以了解982条记录中不同署名排序的发文情况。南京师范大学发表的WOS论文中第一署名机构发文573篇，占比58.4%，非第一作者机构发文409篇。

对非第一作者机构的发文部分， 再分析它们的address (N, 1) 即第一署名机构，可知南京师范大学与144家机构合作， 共发表论文419篇。其中发文1篇的97家， 发文2-3篇的22家。发文4篇以上的25家机构共计发文259篇， 占南京师范大学非第一机构合作论文数的61.8%，见表1。

表1 南京师范大学参与发文的第一作者机构分析

| 第一作者机构 | 篇数 | 第一作者机构 | 篇数 |
|---|---|---|---|
| Chinese Acad Sci | 37 | Ningbo Univ | 6 |
| Inst High Energy Phys | 32 | Beijing Normal Univ | 6 |
| Nanjing Univ | 30 | NE Normal Univ | 6 |
| Nanjing Med Univ | 16 | Univ Fed Rio de Janeiro | 5 |
| Hohai Univ | 14 | Jiangsu Acad Agr Sci | 5 |
| Southeast Univ | 12 | Zhejiang Sci Tech Univ | 5 |
| Nanjing Univ Sci & Technol | 11 | Peking Univ | 5 |
| Jiangnan Univ | 11 | Nanjing Forestry Univ | 5 |
| Nanjing Agr Univ | 8 | Nanjing Tech Univ | 5 |
| Jiangsu Univ | 7 | Zhejiang Univ | 4 |
| Nanjing Univ Aeronaut & Astronaut | 7 | Nanjing Univ Informat Sci & Technol | 4 |
| Nanjing Univ Posts & Telecommun | 7 | Nantong Univ | 4 |
| Anhui Univ | 7 | | |

# 机构清洗

如果是统计学院论文, 在地址栏仅输入学校的地址显然导致无关检索结果太多。
二级单位的名称缩写也可从已查得的论文的地址中找, 以确保其准确性。
例如, 中国科大化材学院下设4个系 (化学物理系、材料科学与工程系、化学系、高分子科学与工程系) , 拥有两个重点实验室 (中国科学院能量转换材料重点实验室、中国科学院软物质化学重点实验室) , 学院内部分教师双聘于合肥微尺度物质科学国家实验室, 那么, 在检索化材学院论文时地址栏的检索式可写成：Univ Sci&Technol China AND (Sch Chem&Mat Sci OR Dept Chem OR Dept Chem Phys OR Dept Mat Sci&Engn OR Dept Polymer Sci&Engn OR CAS Key Lab Soft Matter Chem OR CAS Key Lab Mat Energy Convers OR Hefei Natl Lab Phys Sci Microscale) 。

# 机构类型编码

发文机构数量庞大、类型众多，对机构编码的过程，实质上是对众多发文机构进行归类整序的过程，同时也是知识化的过程。

经过机构编码，引文索引可以根据用户不同的知识需求，对编码进行组配查询，已达到特殊要求的统计分析结果。

# 机构类型编码

从科学分析角度对科研机构进行分类，可以将所有机构划分：高等院校、科研院所、党政机关、文化团体、企业、军队系统、非高校教育单位和其他共八种类型，分别用数字1~7和9表示（在整个编码系统中，对于其他类别统一都用数字9表示）。对于每种类型再根据具体情况编制下级机构类型代码。

# 机构类型编码

对于高校，可以将其划分为“985”工程院校、“211”院校、教育部直属院校、中央其他部门所属院校、地方本科院校、高职（专科）院校六种类别，分别用数字1~6表示；再下一级体现高校的专业特征，分为：综合性院校、师范类院校、医学类院校、工科类院校、体育类院校、艺术类院校、军事类院校、财经类院校和其他专业院校九种类别，用数字1~9表示。当然，也可以根据实际情况增加院校类别。有些高校可能会对应多个编码，本编码规定一律高靠，比如，南京大学可能对应的机构类型编码有“111”“121”和“131”三种，本编码系统只为南京大学取“111”作为编码，可在进行统计分析时，通过一定算法来区分。在对某一类学校进行统计对比是，借助编码就可以方便地进行统计分析。例如，对全国财经类院校进行统计分析时，可以利用代码“1x8”抽取相关要素具体完成统计，其中“x”可以为任何数字，也可以是指定某几个数字

# 机构类型编码

对于科研院所也可以划分为三级进行编码。首先分为：国家所属、省部所属、专业学会所属、其他等。然后再将其细分为：自然科学类、工程科学类、医药学类、社会科学类、其他等。例如，中国社会科学院为国家所属的社会科学院科研机构。再如，江苏省中医药研究院为江苏省所属的医药学类研究所。

# 机构类型编码

其他类型机构都可以按纵横两个方面再划分两级编码。纵向是指行政上的上下级或所属关系，如，国家级、省市级等。横向是指机构的类别或属性，如高校按学科划分等。所有机构类型编码均由三位数字组成。在实际的机构编码过程中，可以根据资源设计的机构状况，根据实际需要来设置机构编码。

# 关键词清洗

为了更好地计算、分析和观察关键词所形成的网络，对所提取的关键词数据，进行了一定的数据清洗操作。

在Scopus数据库中分别提取图书情报领域和人工智能领域的文献，通过对表示学科分类号的字段“ASJC-CODE”进行提取，两个学科领域的ASJC-CODE分别为 3309 和 1702 。从提取的文献中，获得图书情报领域2008年所有文献的关键词共 111 444个，人工智能领域关键词共 128 732个（2007~2011年）。鉴于关键词本身分布的不规范性，为了充分挖掘关键词之间存在的语义关系，按照表8-1的清洗规则完成了对两个学科的关键词数据清洗，主要进行了名词单复数转换、不必要的符号的删除、恢复全称这三种类型的清洗。最终得到图书馆情报领域关键词91 225个，人工智能领域关键词92 928个。

# 关键词清洗

表 8-1　数据清洗规则与实例

| 序号 | 数据清理操作 | 原数据 | 清洗后数据 |
|---|---|---|---|
| 1 | 大小写转换 | Web mining | web mining |
| 2 | 去掉简称 | support vector machines（SVMs） | support vector machine |
| 3 | 删除括号“( )” | (almost) cyclostationary processes | almost cyclostationary processe |
| 4 | 删除逗号“,” | (k, n) threshold | k n threshold |
| 5 | 删除短横“-” | dempster-shafer theory | dempster shafer theory |

# 关键词清洗

| 序号 | 数据清理操作 | 原数据 | 清洗后数据 |
| --- | --- | --- | --- |
| 6 | 将序数词转换为数字 | (second) preimages | 2 preimage |
| 7 | 删除星号“*” | 3 * -connected | 3 connected |
| 8 | 复数转单数 | abstracts | abstract |
| 9 | 删除其他符号 | H<inf>â^ž</inf> decentralized | decentralized |
| 10 | 恢复全称 | SVM | support vector machine |
| 11 | 删去斜杠“/” | 3D/stereo scene analysis | 3d stereo scene analysis |

# 字段匹配问题

字段匹配是记录匹配的基础，可以分为字符型字段匹配和数值型字段匹配。对于数值型字段的匹配，用到的基本上都是精确匹配理论，一般直接进行简单的相等比较。但字符型匹配一般用到的是模糊匹配知识，其中很大程度和字符的类型有关。

且不同于语言下的字符型匹配算法有较大的的区别。例如英文中的缩写、简写、姓名标识方式、地址标识方式等，同汉语有很大差异。汉语和英文之间存在的诸多差异，使得字符型字段匹配算法，两者基本上不能通用。考虑到大部分的字段值都是字符型的，且重复记录之间的差异往往也是由字符型字段值引起的。

# 实际应用

数据库：Web of Science

检索词：主题=library and information science

年份：2017年-2020年

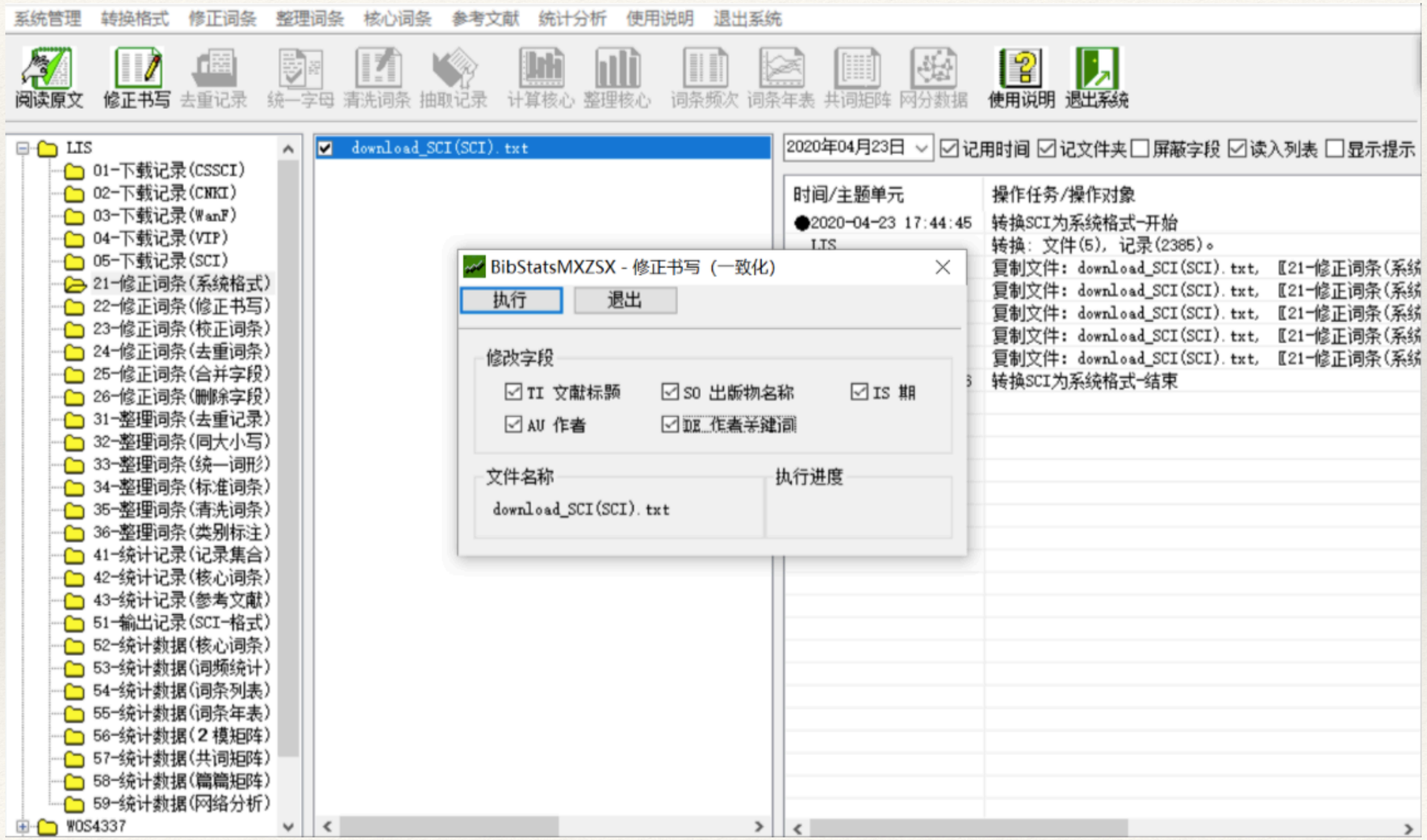
检索时间：2020年4月23日

数据结果：2835条

清洗软件：Bibstats

# 实际应用

1、创建主题单元、转换格式
2、修正书写



# 实际应用

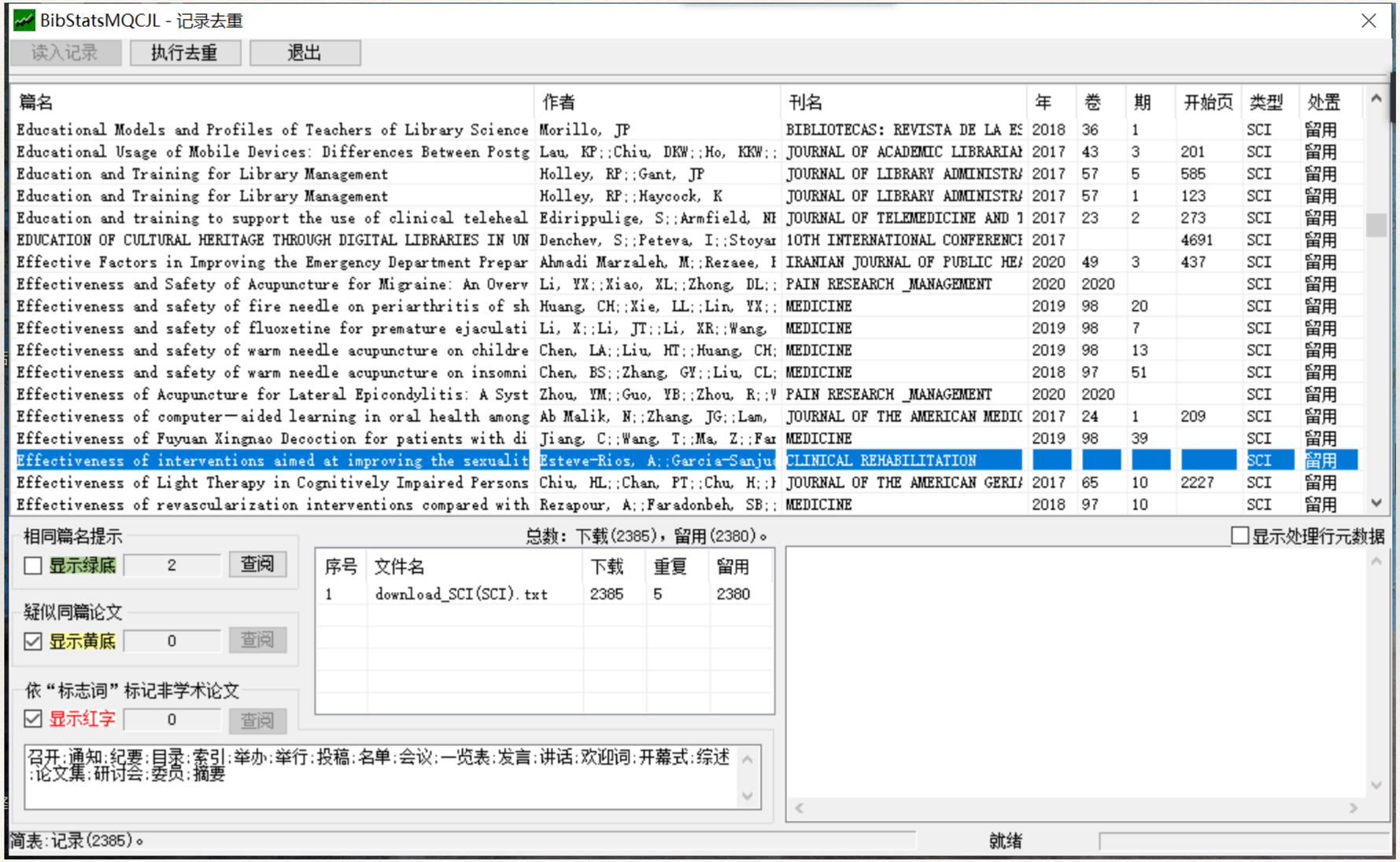
## 3、校正词条：提取处理内容-暂存处理内容

AF：作者全名
AU：作者
C1：作者或机构地址
C2：[C1->作者]
C3：[C1->机构（1级）和地址]
C4：[C1->机构（1级+2级）和地址]
DE：作者关键词

AF字段：词条(8145)
AU字段：词条(8150)
C1字段：词条(8163)
C2字段：词条(10426)
C3字段：词条(4753)
C4字段：词条(5271)
DE字段：词条(12876)

# 实际应用

4、去重词条：AF、AU、DE有重复，其他字段无重复




| 篇名 | 作者 | 刊名 | 年 | 卷 | 期 | 开始页 | 类型 | 处置 |
|---|---|---|---|---|---|---|---|---|
| Educational Models and Profiles of Teachers of Library Science | Morillo, JP | BIBLIOTECAS: REVISTA DE LA ES | 2018 | 36 | 1 | | SCI | 留用 |
| Educational Usage of Mobile Devices: Differences Between Postg | Lau, KP;;Chiu, DKW;;Ho, KKW;; | JOURNAL OF ACADEMIC LIBRARIAI | 2017 | 43 | 3 | 201 | SCI | 留用 |
| Education and Training for Library Management | Holley, RP;;Gant, JP | JOURNAL OF LIBRARY ADMINISTRA | 2017 | 57 | 5 | 585 | SCI | 留用 |
| Education and Training for Library Management | Holley, RP;;Haycock, K | JOURNAL OF LIBRARY ADMINISTRA | 2017 | 57 | 1 | 123 | SCI | 留用 |
| Education and training to support the use of clinical teleheal | Edirippulige, S;;Armfield, NI | JOURNAL OF TELEMEDICINE AND T | 2017 | 23 | 2 | 273 | SCI | 留用 |
| EDUCATION OF CULTURAL HERITAGE THROUGH DIGITAL LIBRARIES IN UN | Denchev, S;;Peteva, I;;Stoyar | 10TH INTERNATIONAL CONFERENCI | 2017 | | | 4691 | SCI | 留用 |
| Effective Factors in Improving the Emergency Department Prepar | Ahmadi Marzaleh, M;;Rezaee, I | IRANIAN JOURNAL OF PUBLIC HEA | 2020 | 49 | 3 | 437 | SCI | 留用 |
| Effectiveness and Safety of Acupuncture for Migraine: An Overv | Li, YX;;Xiao, XL;;Zhong, DL;; | PAIN RESEARCH _MANAGEMENT | 2020 | 2020 | | | SCI | 留用 |
| Effectiveness and safety of fire needle on periarthritis of sh | Huang, CH;;Xie, LL;;Lin, YX;; | MEDICINE | 2019 | 98 | 20 | | SCI | 留用 |
| Effectiveness and safety of fluoxetine for premature ejaculati | Li, X;;Li, JT;;Li, XR;;Wang, | MEDICINE | 2019 | 98 | 7 | | SCI | 留用 |
| Effectiveness and safety of warm needle acupuncture on childre | Chen, LA;;Liu, HT;;Huang, CH; | MEDICINE | 2019 | 98 | 13 | | SCI | 留用 |
| Effectiveness and safety of warm needle acupuncture on insomni | Chen, BS;;Zhang, GY;;Liu, CL; | MEDICINE | 2018 | 97 | 51 | | SCI | 留用 |
| Effectiveness of Acupuncture for Lateral Epicondylitis: A Syst | Zhou, YM;;Guo, YB;;Zhou, R;;Y | PAIN RESEARCH _MANAGEMENT | 2020 | 2020 | | | SCI | 留用 |
| Effectiveness of computer—aided learning in oral health among | Ab Malik, N;;Zhang, JG;;Lam, | JOURNAL OF THE AMERICAN MEDIC | 2017 | 24 | 1 | 209 | SCI | 留用 |
| Effectiveness of Fuyuan Xingnao Decoction for patients with di | Jiang, C;;Wang, T;;Ma, Z;;Far | MEDICINE | 2019 | 98 | 39 | | SCI | 留用 |
| Effectiveness of interventions aimed at improving the sexualit | Esteve-Rios, A;;Garcia-Sanjus | CLINICAL REHABILITATION | | | | | SCI | 留用 |
| Effectiveness of Light Therapy in Cognitively Impaired Persons | Chiu, HL;;Chan, PT;;Chu, H;;H | JOURNAL OF THE AMERICAN GERIA | 2017 | 65 | 10 | 2227 | SCI | 留用 |
| Effectiveness of revascularization interventions compared with | Rezapour, A;;Faradonbeh, SB;; | MEDICINE | 2018 | 97 | 10 | | SCI | 留用 |

| 序号 | 文件名 | 下载 | 重复 | 留用 |
|---|---|---|---|---|
| 1 | download_SCI(SCI).txt | 2385 | 5 | 2380 |

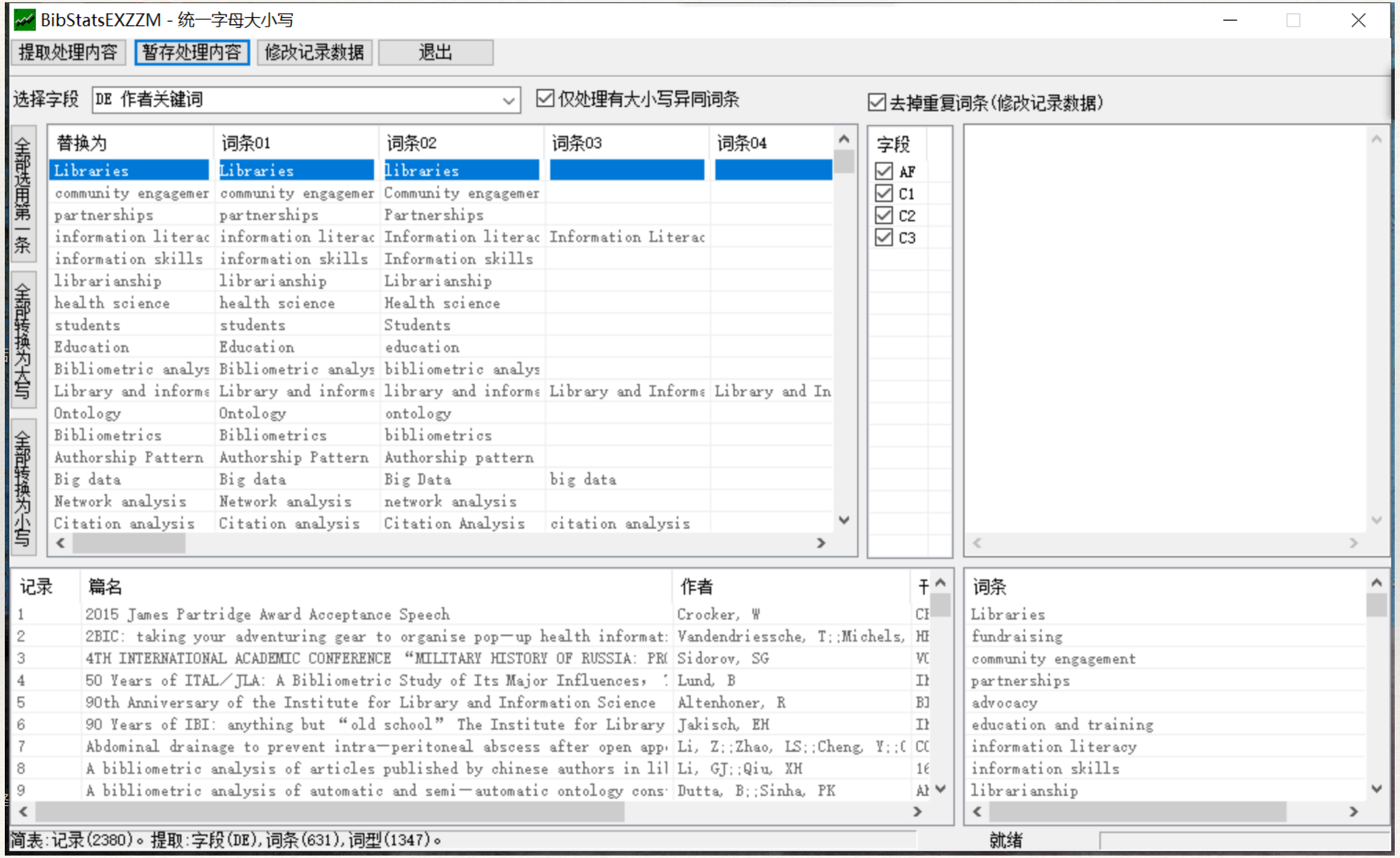
召开;通知;纪要;目录;索引;举办;举行;投稿;名单;会议;一览表;发言;讲话;欢迎词;开幕式;综述;论文集;研讨会;委员;摘要

AF：作者全名
AU：作者
C1：作者或机构地址
C2：[C1->作者]
C3：[C1->机构（1级）和地址]
C4：[C1->机构（1级+2级）和地址]
DE：作者关键词

清洗后记录：2380条

# 实际应用

## 5、统一字母大小写




AF：作者全名
AU：作者
C1：作者或机构地址
C2：[C1->作者]
C3：[C1->机构（1级）和地址]
C4：[C1->机构（1级+2级）和地址]
DE：作者关键词

DE字段：词条(631)、词型(1347)

# 实际应用

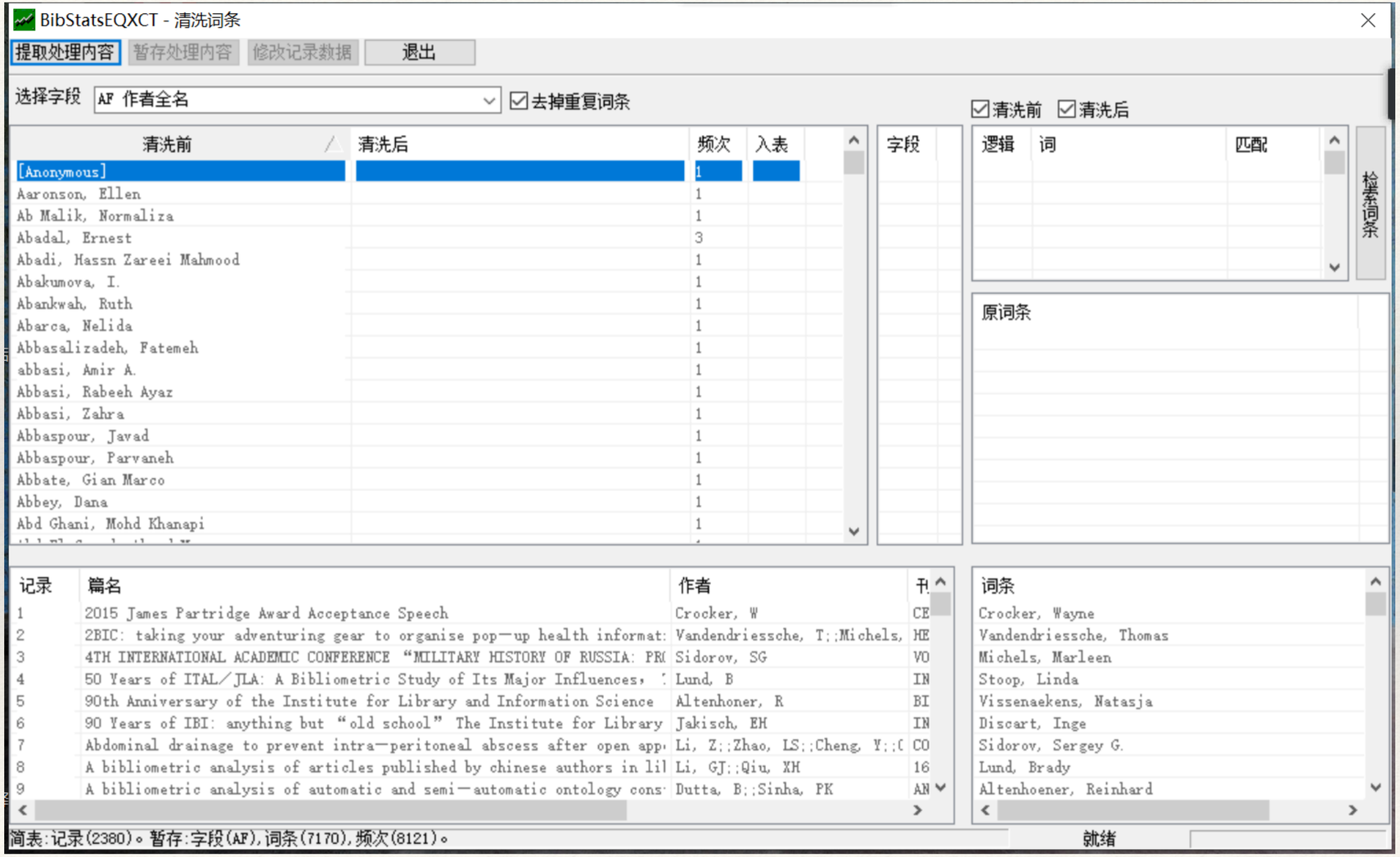
## 6、统一词形

AF：作者全名
AU：作者
C1：作者或机构地址
C2：[C1->作者]
C3：[C1->机构（1级）和地址]
C4：[C1->机构（1级+2级）和地址]
DE：作者关键词

DE字段：词条(723)、词型(1729)

# 实际应用

## 7、清洗词条



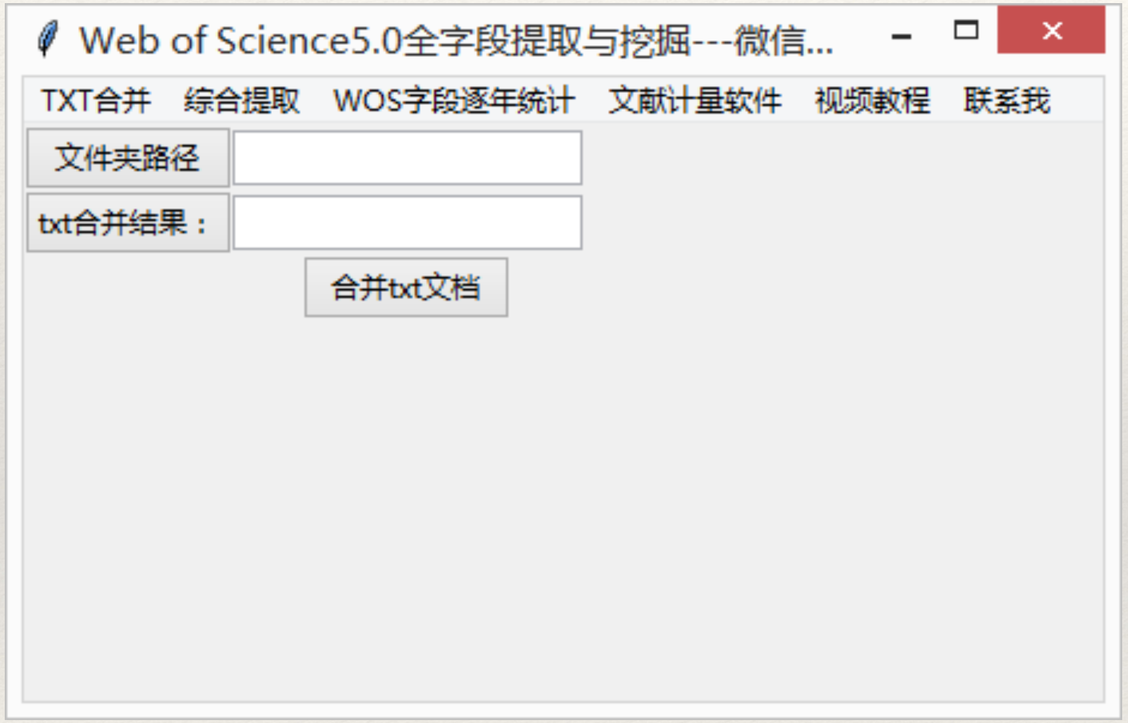

AF：作者全名
AU：作者
C1：作者或机构地址
C2：[C1->作者]
C3：[C1->机构（1级）和地址]
C4：[C1->机构（1级+2级）和地址]
DE：作者关键词

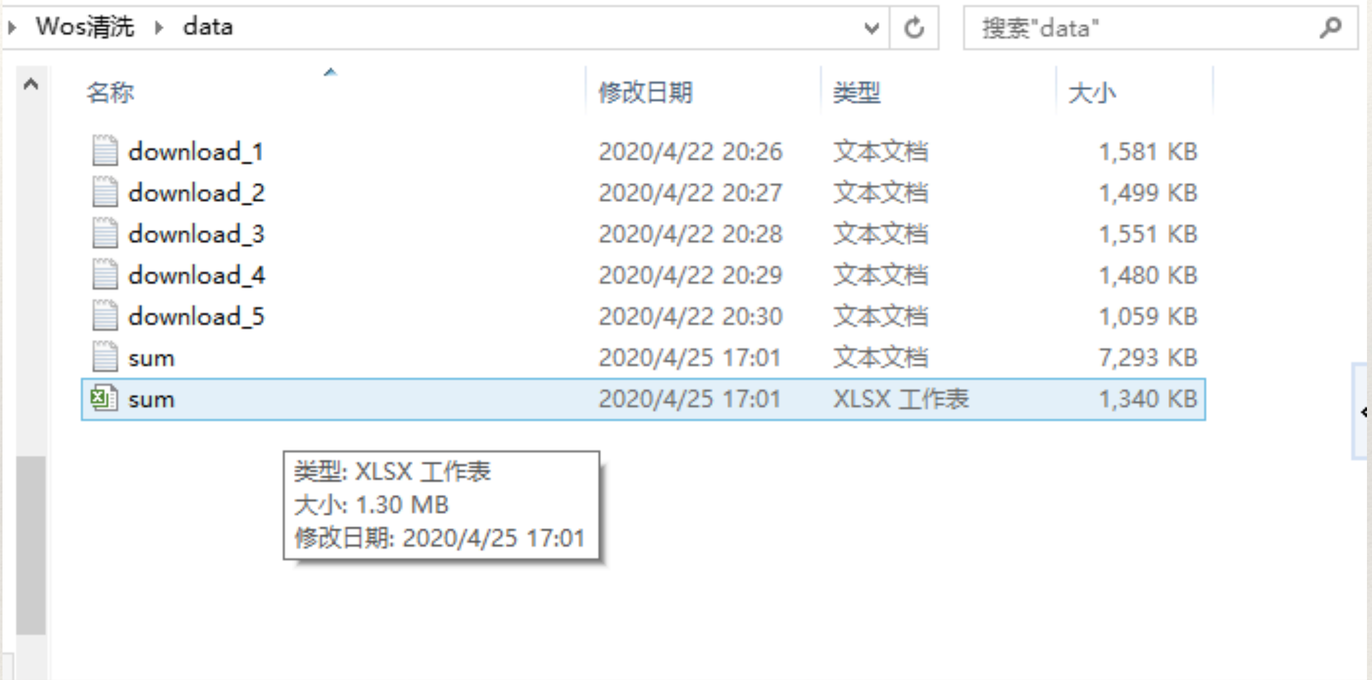
# 实际应用

| 清洗前 | 清洗后 |
| --- | --- |
| AF字段：词条(8145) | AF字段：词条(7170)、频次(8121) |
| AU字段：词条(8150) | AU字段：词条(6870)、频次(8120) |
| C1字段：词条(8163) | C1字段：词条(7779)、频次(8087) |
| C2字段：词条(10426) | C2字段：词条(9098)、频次(10330) |
| C3字段：词条(4753) | C3字段：词条(2522)、频次(4707) |
| C4字段：词条(5271) | C4字段：词条(3893)、频次(4771) |
| DE字段：词条(12876) | DE字段：词条(9406)、频次(10288) |

# 补充资料

一个小程序


Web of Science5.0全字段提取与挖掘---微信...
TXT合并
综合提取
WOS字段逐年统计
文献计量软件
视频教程
联系我
文件夹路径
txt合并结果：
合并txt文档


Wos清洗 ▸ data
搜索"data"
名称
修改日期
类型
大小
download_1 2020/4/22 20:26 文本文档 1,581 KB
download_2 2020/4/22 20:27 文本文档 1,499 KB
download_3 2020/4/22 20:28 文本文档 1,551 KB
download_4 2020/4/22 20:29 文本文档 1,480 KB
download_5 2020/4/22 20:30 文本文档 1,059 KB
sum 2020/4/25 17:01 文本文档 7,293 KB
sum 2020/4/25 17:01 XLSX 工作表 1,340 KB
类型: XLSX 工作表
大小: 1.30 MB
修改日期: 2020/4/25 17:01

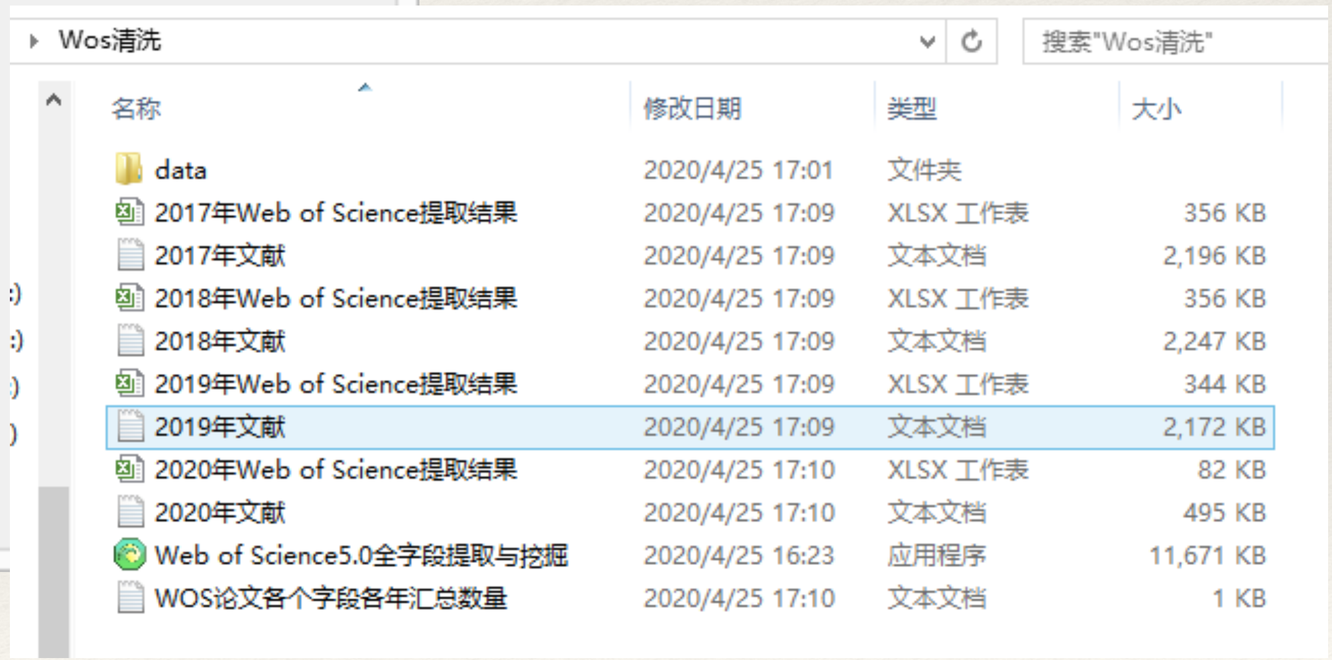
# 补充资料

可进行初步遍历、去重

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 2362 | 2017 | DEVELOP | GEOGRAPHICALLY | | ANNALS ( | EISMAN, E | EISMAN, E | FLORIDA I | USA | 5 |
| 2363 | 2017 | INTERVENTIONS TO | | SMEAR-P | COCHRAN | MHIMBIRA | MHIMBIRA | UNIV LIVE | ENGLAND | 5 |
| 2364 | 2017 | HEALTH IN | ADOLESC | ADHEREN | COGNITIV | KESELMA | KESELMA | US NATL I | USA | 0 |
| 2365 | 2017 | INTERDISC | INTERDISCIPLINARIT | | EM QUES | NETO, JAI | DOS SANT | UNIV EST/ | BRAZIL | 2 |
| 2366 | 2017 | MODELS ( | INFORMATION PRAC | | EM QUES | ROCHA, J | PEREIRA | UNIV FED | BRAZIL | 3 |
| 2367 | 2017 | ANALYSIS | RESEARCH PRODUC | | EM QUES | BUFREM, | BUFREM, | UNIV FED | BRAZIL | 2 |
| 2368 | 2017 | INTERNAT | SCIELO CITATION IN | | EM QUES | ALENCAR | DE MENE | UNIV FED | BRAZIL | 1 |
| 2369 | 2017 | SUBJECT | SUBJECT | ACCESS; | EM QUES | DE MEDE | DE MEDE | UNIV FED | BRAZIL | 0 |
| 2370 | 2017 | A PRIMER | EMOTION | POSITIVE | EMOTION | MATTESO | MATTESO | KENT STA | USA | 0 |
| 2371 | 2017 | CONSTRU | LIBRARIAI | LIBRARIAI | EMOTION | WALLACH | WALLACH | UNIV SOU | USA | 0 |
| 2372 | 2017 | INFORMATION SYSTEMS IN CL | | | FARMACI | GABARRC | MARTINEZ | UNIV MIGU | SPAIN | 4 |
| 2373 | 2017 | PRESERVATION OF INDIGENO | | | HANDBOC | HIKWA, L; | HIKWA, L | NATL UNIV | ZIMBABW | 1 |
| 2374 | 2017 | THE ATTR | STONECUTTING ART | | IZVESTIYA | BUDRINA, | BUDRINA, | YEKATER | ITALY; RU | 1 |
| 2375 | 2017 | A METHOI | ECONOMI | WILLINGN | PROCEED | PUTRI, RA | PUTRI, RIZ | INDONESI | INDONESI | 0 |
| 2376 | 2017 | SOCIAL M | SOCIAL MEDIA; SOC | | PROCEED | AJEGBOM | AJEGBOM | FED UNIV | NIGERIA | 0 |
| 2377 | 2017 | PREVALE | URBAN HE | PHYSICAL | REVISTA I | DE SA, TH | DE SA, TH | UNIV SAO | BRAZIL; U | 1 |
| 2378 | 2017 | AGAIN AB | ARCHIVAL SCIENCE | | SKHIDNOI | BEZDRAB | BEZDRAB | UKRAINIA | UKRAINE | 0 |
| 2379 | 2017 | ARISING T | IOFFE INSTITUTE; N/ | | SOCIOLO | D'YAKOV, | D'YAKOV, | RUSSIAN | RUSSIA | 0 |
| 2380 | 2017 | THE CONTRIBUTION OF INDIVI | | | SPORTS M | ABELL, B; | ABELL, BI | BOND UNI | AUSTRALI | 18 |
| 2381 | 2017 | N OF 1 TR | N OF 1; AI | RANDOMI | SYSTEMA | DEMEYIN, | DEMEYIN, | UNIV EXE | ENGLAND | 8 |
| 2382 | 2017 | DIABETES | DIABETES | HEALTH IN | SYSTEMA | KUSKE, S | KUSKE, S | HEINRICH | GERMANY | 7 |
| 2383 | 2017 | RISK FAC | TUBEROU | NEUROLC | SYSTEMA | MITCHELL | MITCHELL | ROYAL CH | AUSTRALI | 4 |
| 2384 | 2017 | ADVERSE | OPIOID US | MAINTENA | SYSTEMA | MOURAVS | MOURAVS | ST JOSEP | CANADA; | 1 |
| 2385 | 2017 | ASSOCIAT | DIABETES | RISK-FAC | SYSTEMA | TEGEGNE | TEGEGNE | UNIV GRO | NETHERL | 4 |
| 2386 | | | | | | | | | | |

# 补充资料

## Wos字段逐年统计




| 名称 | 修改日期 | 类型 | 大小 |
| --- | --- | --- | --- |
| data | 2020/4/25 17:01 | 文件夹 | |
| 2017年Web of Science提取结果 | 2020/4/25 17:09 | XLSX 工作表 | 356 KB |
| 2017年文献 | 2020/4/25 17:09 | 文本文档 | 2,196 KB |
| 2018年Web of Science提取结果 | 2020/4/25 17:09 | XLSX 工作表 | 356 KB |
| 2018年文献 | 2020/4/25 17:09 | 文本文档 | 2,247 KB |
| 2019年Web of Science提取结果 | 2020/4/25 17:09 | XLSX 工作表 | 344 KB |
| 2019年文献 | 2020/4/25 17:09 | 文本文档 | 2,172 KB |
| 2020年Web of Science提取结果 | 2020/4/25 17:10 | XLSX 工作表 | 82 KB |
| 2020年文献 | 2020/4/25 17:10 | 文本文档 | 495 KB |
| Web of Science5.0全字段提取与挖掘 | 2020/4/25 16:23 | 应用程序 | 11,671 KB |
| WOS论文各个字段各年汇总数量 | 2020/4/25 17:10 | 文本文档 | 1 KB |

# 补充资料

## Wos论文各个字段各年汇总数量

WOS论文各个字段各年汇总数量 - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

| 出版年 | 标题总数 | 关键词总数 | 期刊总数 | 作者数(简称)总数 | 作者数(全称)总数 | 机构总数 | 国家总数 |
|---|---|---|---|---|---|---|---|
| 2017 | 745 | 2276 | 336 | 2120 | 2157 | 985 | 89 |
| 2018 | 724 | 2508 | 354 | 2282 | 2313 | 986 | 85 |
| 2019 | 697 | 2315 | 310 | 2357 | 2414 | 961 | 78 |
| 2020 | 155 | 554 | 93 | 572 | 579 | 325 | 53 |

# 总结

1、当前数据清洗的研究重点主要放在了重复记录的清洗上，对属性清洗的研究不是很多。属性清洗主要集中在单个字段属性的清洗上，多个字段之间的属性清洗还很不成熟。对字符型属性的自动识别与清洗有待进一步的研究。

2、完全针对中文数据清洗的研究比较少。中文数据清洗一直没有受到足够关注，以前的算法是针对英文而提出的，不适用于中文数据的清洗。

3、在数据集动态增长的情况下，要多关注如何利用已经完成的数据清洗的结果作为辅助信息，避免重复清洗，来提高数据清洗算法效率。

# 参考文献

[1] 苏新宁.面向知识服务的知识组织理论与方法[M].北京:科学出版社, 2014.
[2]数据挖掘中常用的数据清洗方法[EB/OL]. [2020-4-19]. https://www.jianshu.com/p/c2b15fcf7100
[3]卢星.大数据的数据清洗技术及运用[J].电子技术与软件工程,2019(09):157.
[4]毕强.面向知识服务的知识组织研究的重要成果——简评《面向知识服务的知识组织理论与方法》[J].图书情报工作,2017,61(19):146-149.
[5]何春建.从WOS地址字段提取二级机构数据的半自动数据清洗方法[J].新世纪图书馆,2017(08):56-58+70.
[6]蒋勋,徐绪堪.面向知识服务的知识库逻辑结构模型[J].图书与情报,2013(06):23-31.
[7]刘贤玉,周小东.基于Web of Science快速统计学校(学院)论文的技巧[J].图书情报工作,2013,57(S2):210-212+207.
[8]周奕辛. 数据清洗算法的研究与应用[D].青岛大学,2005.
[9]数据挖掘中常用的数据清洗方法[EB/OL]. [2020-4-20].
https://www.cnblogs.com/tomcattd/p/3372341.html

谢谢！