科技情报数据清洗与数据整理

汇报人：法慧

01

基本概念

数据清洗

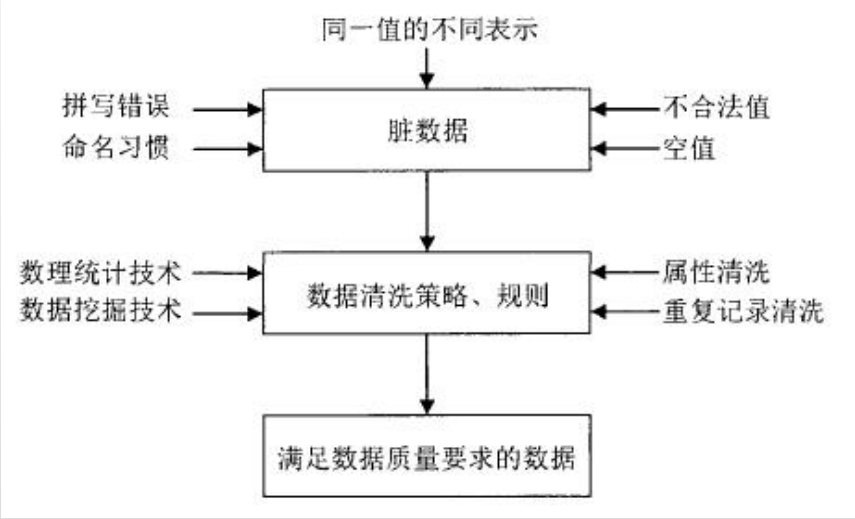
数据清洗(Data cleaning)– 对数据进行重新审查和校验的过程，目的在于删除重复信息、纠正存在的错误，并提供数据一致性。

数据清洗从名字上也看的出就是把“脏”的“洗掉”包括检查数据一致性，处理无效值和缺失值等。按照一定的规则把“脏数据”“洗掉”，这就是数据清洗。

清洗类型

残缺/空值数据 错误数据 重复数据 不一致性数据 离群/异常数据

清洗原理



02

数据清洗实现方式

Summary of Report Work Summary Business Report Work Plan Mid-year Work Summary

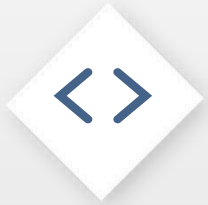
王曰芬等在《数据清洗研究综述》和杨辅祥等《数据清理综述》中指出数据清洗方式一般分为全人工清洗、全机器清洗、特定应用领域数据清洗、无关数据清洗四种清洗方式

全人工清洗

手工实现 ,通过人工检查 ,只要投入足够的人力物力财力 , 也能发现所有错误 ,但效率低下 。 在大数据量的情况下 ,几乎是不可能的 。

全机器清洗

通过专门编写的应用程序 , 这种方法能解决某个特定的问题 , 但不够灵活 , 特别是在清理过程需要反复进行(一般来说 , 数据清理一遍就达到要求的很少) 时 ,导致程序复杂 ,清理过程变化时,工作量大 。 而且这种方法也没有充分利用目前数据库提供的强大数据处 理能力。



特定应用领域清洗

解决某类特定应用域的问题 ,如根据概率统计 学原理查找数值异常的记录 ,对姓名 、 地址 、 邮政编码等 进行清理 , 这是目前研究得较多的领域 , 也是应用最成功的一类 。

领域无关数据清洗

与特定应用领域无关的数据清洗 ,这一部分的研究主要集中在清理重复的记录上 。

这 4 种实现方法 , 由于后两种具有某种通用性 , 较 大的实用性 ,引起了越来越多的注意 。 但是不管哪种方法 ,大致都由三个阶段组成 :①数据分析 、 定义错误类 型 ;②搜索 、 识别错误记录 ;③修正错误 。

03

数据清洗一般过程

根据不同的任务要求与环境特点数据清洗执行过程也不同，根据对一般清洗工具的总结，数据清洗的一般过程可分为四个环节

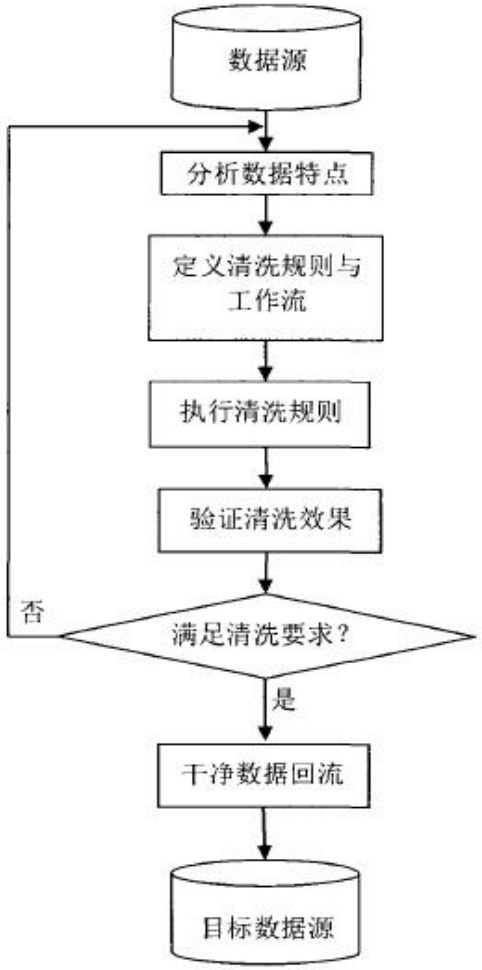


图 2.4 数据清洗一般过程

序号	清洗方法	描述
1	格式化	根据数据定义的标准格式，对于一些格式不一致或不标准的数据，进行格式化处理
2	合并/删除	对于重复记录，根据业务规则进行合并，并删除重复的数据
3	替换	将不符合规范的值替换为符合规范的值
4	分割	可以将单个属性分割成多个属性，或者将多个属性合并成一个属性，主要是为了消除模式冲突

图 2.5 清洗规则执行顺序

04

数据清洗的基本处理方法

缺失、空值数据——补充

1.删除包含空值的记录

2.自动补全

手工补全缺失值

不一致数据——统一

在分析不一致产生原因的基础上，利用各种变换函数、格式化函数、汇总分解函数去实现清洗

噪声、错误数据——删除或修正

1.分箱

2.回归

3.计算机检查和人工检查相结合

4.聚类

重复数据——合并

在分析不一致产生原因的基础上，利用各种变换函数、格式化函数、汇总分解函数去实现清洗

05

常用清洗方法优缺点

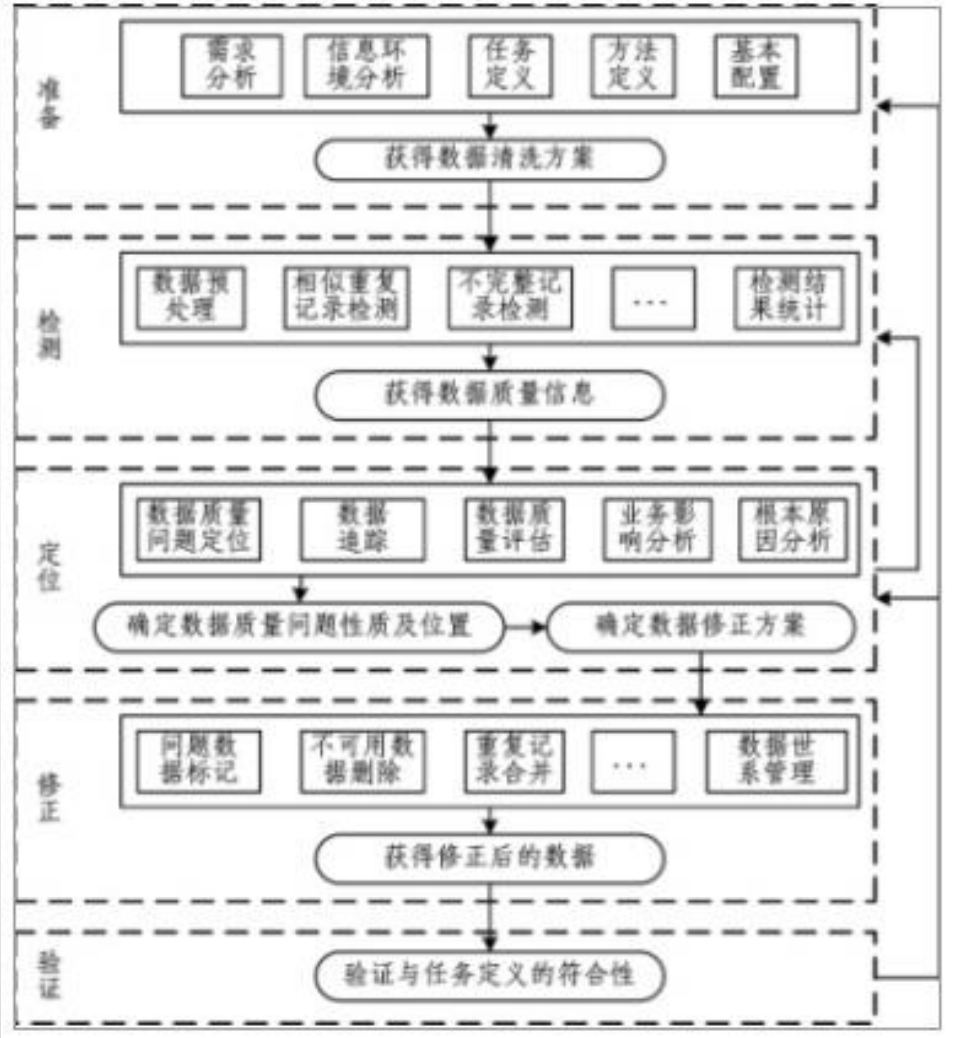
缺失值处理	适用条件	优缺点
删除法	样本量大，缺失值所占样本比例较少	优点：简单、易行 缺点：1. 损失样本量，造成资源浪费，容易丢弃隐藏信息 2. 削弱统计功效 3. 样本量较小时，数据的客观性和结果的正确性会受到严重影响 4. 缺失数据所占比例较大，且缺失值非随机分布时，可能导致数据发生偏离，得出错误结论
均值插补	缺失值为数值型：平均值来填充 缺失值为非数值型：众数填充	优点：简便、快速 缺点：1. 这种方法会产生有偏估计 2. 均值补差法是建立在完全随机缺失的假设上，会造成变量的方差和标准差变小
回归插补	处理变量之间的相关关系	优点：利用了尽可能多的信息，较前几种方法得到的缺失值更有说服力 缺点：1. 容易忽视随机误差，低估标准差和其他未知性质的测量值，且随缺失信息的增多而严重 2. 研究者必须假设存在缺失值所在的变量与其他变量存在线性关系，很多时候这种关系是不存在的
极大似然估计	适用于任何总体	优点：估计量具有一致性和有效性 缺点：并非所有缺失值都能求得似然估计量，解似然方程时，可能难以求解或根本写不出有限形式的解

噪声过滤处理	适用条件	优缺点
回归	建立在稳定数据变量基础上	优点：分析多因素模型时，更加简单和方便，去噪效果好 缺点：1. 直接采用非平稳时间序列建立回归模型，很容易产生“伪回归”问题 2. 存在着因果关系的变量间建立的回归预测模型的预测效果较差
均值平滑	有序列特征的变量	优点：简单、计算速度快 缺点：次方法去噪导致信号的细节和边缘模糊
离群点分析	1. 数据和检验类型要充分 2. 预先知道样本空间中数据集的分布特征	优点：建立在标准的统计学技术之上，当数据和检验的类型十分充分时，检验有效 缺点：1. 绝大多数是针对单个属性的，而数据挖掘要求多维空间挖掘离群点 2. 数据分布可能是未知的，统计学方法在数据不充分的情况下，不能确保所有的离群点被发现
小波法	对图像、信号去噪	优点：1. 低熵性 2. 多分率，能非常好的刻画信号的非平稳特征，如：边缘、尖峰、断点等 3. 去相关性，噪声在变换后有自化趋势，小波域比时域更利于去噪 4. 可以得到信号的最优估计

06

数据清洗一般过程

提出一个数据清洗的一般性系统框架，该框架由准备、检测、定位、修正、验证 5 部分组成，简称为 P D L M V



曹建军等《数据清洗及其一般性系统框架》

07

举个栗子

数据清洗

一般来说，清洗可以分为两个基本的任务：错误检测，即发现数据中潜在的错误、重复或缺失等；数据修复，即针对发现的错误，对数据进行修复。下面结合一个具体的实例分别进行介绍。错误检测任务旨在发现影响数据质量的错误因素。一般将错误因素划分为4类，下面通过一个例子进行说明。

	姓氏	名字	年龄/岁	工作单位	所在城市
t1	张	三	40	中国人民大学	上海
t2	李	四	5	上海交通大学	上海
t3	王	五	35	<缺失>	北京
t4	三	张	40	人大	北京

数据清洗中错误检测的示例

07

举个栗子

数据检测

（1）异常值

异常值是指明显不符合属性语义的取值。现有的代表性解决方案主要是基于统计和距离的方法解决。

（2）结构性错误

结构性错误是指数据不符合特定领域语义要求的完整性约束。

（3）数据重复

记录重复在真实数据中十分普遍，其原因是多方面的，比如数据可能由不同的机构提供，或者数据整合自组织的内外部渠道。

（4）数据缺失

数据缺失是指数据的部分属性不存在于数据库中后续的分析过程。针对数据缺失，现有的方法是采用缺失值插补技术进行修复，更为有效的办法是采用最大可能性的数据值并进行推理。

数据检测

数据修复任务是指根据检测出的错误对数据进行更新，以达到纠正错误的目的。与错误检测相比，数据修复的挑战性更大，因为通常缺乏对修复进行指导的信号。为了应对这一挑战，现有的方法往往采用外部知识或一些定量的统计指标。

08

已有研究实例展示

潘玮,牟冬梅,等.关键词共现方法识别领域研究热点过程中的数据清洗方法[J].图书情报工作,2017

□ 1 生物医学专利分析领域关键词数据清洗

数据清洗解决方案具体实施过程如下:

①选择 Pubmed 数据库作为数据来源。通过专家咨询，制定检索式为(patent * [Title] AND analy* [Title]) OR (“patent analysis” [Title / Abs - tract])，截至 2016 年 4 月共得到 271 条记录。

②用字符“DT”对数据记录中的关键词进行统一标识。

③利用 Citespace Ⅲ的 Utility Functions [28] 功能去除原始数据中的重复或相似重复记录。

④组织领域专家对缺失的关键词进行标注。

⑤通过人工排查的方法去除错误包含错误关键词的记录。

⑥对被识别为研究热点的关键词进行人工整合。

08

已有研究实例展示

潘玮,牟冬梅,等.关键词共现方法识别领域研究热点过程中的数据清洗方法[J].图书情报工作,2017

2 研究热点识别结果对比分析

运用 Citespace Ⅲ分别对原始数据及清洗后数据的研究热点进行识别，结果见下图。图中各区块为使用聚类算法得到的类团，字体的大小与关键词的词频成正比，这些类团为数据挖掘出的研究热点。

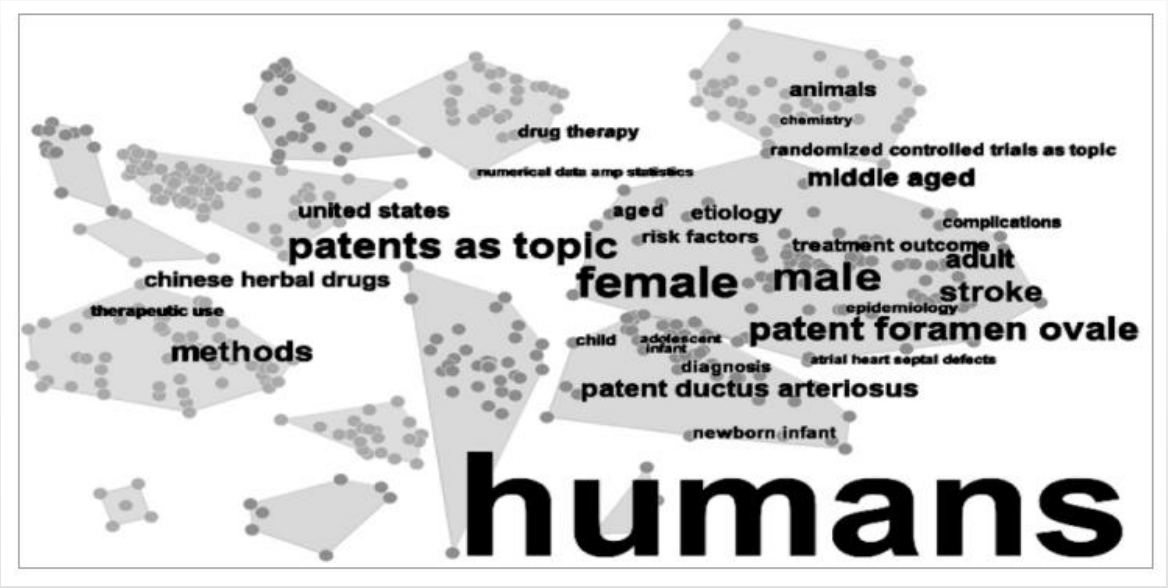


图 2 未经数据清洗识别出的研究热点

图 3 经数据清洗后识别出的研究热点

08

已有研究实例展示

潘玮,牟冬梅,等.关键词共现方法识别领域研究热点过程中的数据清洗方法[J].图书情报工作,2017

3 数据清洗研究讨论

造成图 2 所得结果准确性较差的原因在于:

①未经清洗的数据中包含大量错误的关键词，其中有 39 篇文献包含关键词patent ductus arteriosus，53 篇文献包含关键词 patent fo-ramen ovale等。这些错误关键词被作为研究热点识别出来，从而降低了结果的准确性。其产生的原因在于在制定检索式时没有考虑 patent 在医学领域中有“开放的、未闭的”含义，以及 Chinese patent medicines 的固定搭配，但是这种情况除了该领域内的专家，其他人在检索之前是很难预见的，因此若不经数据清洗这种错误很难避免。

②未经清洗的数据存在较多的关键词缺失情况，使得一些与医药专利分析相关的关键词无法作为研究。

09

基本概念

数据整理

数据整理主要是指对原始数据进行加工处理，使之系统化、条理化，以符合统计分析的需要，同时用图表形式将数据展示出来，以便简化数据，使之更容易理解和分析。

方法

(1)归纳法: 可应用直方图、分组法、层别法及统计解析法。
(2)演绎法: 可应用要因分析图、散布图及相关回归分析。
(3)预防法: 通称管制图法，包括Pn管制图、P管制图、C管制图、U管制图、管制图、X-Rs管制图。

基本步骤

1.根据研究目的设计整理方案。
2.统计数据的审核与检查。
3.数据分组和汇总，并计算各项指标。
4.通过统计表或统计图，显示整理结果。
5.统计资料的积累、保管和公布。

背景介绍

数据整理是为了使数据更好地服务于数据分析而对数据进行的审查和转换的过程，它是整个数据分析流程中最占用精力的过程。从技术上讲，数据整理包含前期数据解析与结构化处理、数据质量评估与数据清洗、数据集成和提纯等过程。由于问题的复杂性，数据整理过程通常不是完全自动化的，而是需要用户介入的反复迭代和交 互的过程。数据可视化、用户反馈与交互在整个过程中都发挥了重要作用。数据整理是由数据可视化领域的Jeffery Heer教授（华盛顿大学）和数据库领域的Joseph M. Hellerstein教授（加州大学伯克利分校）等人较早提出来并持续开展系列研究的。他们还将研究成果进行了产业化，成功创立了以数据整理为主业的Trifacta公司。

数据整理的核心技术

数据的结构化处理

数据结构化处理首先要对原始数据进行 解析，提取出需要的信息，再进一步将其转换成结构化数据。

数据质量评估与数据清洗

处理后的数据还要进行质量评估，如果发现数据中存在问题，则采取进一步的数据清洗措施。这个过程称作数据质量评估。伴随着数据质量问题的发现，用户可以定义一些数据清洗规则，批量化地处理数据中存在的质量问题，提高数据清洗的效率。

数据规范化

数据融合与摘取

数据融合是数据集整合的过程，有些分析任务未必需要全部整合后的数据，可能仅需要一部分数据支撑分析任务。在这种情况下，需要从数据集中提取部分数据（如一些样本或者数据片段），降低数据量，供数据分析模型实现分析操作。这一过程称作数据摘取，它需要根据任务的特点摘取相关数据。

公开共享

问题与展望

(1)识别主要集中在数值型、字符串型字段 。识别数值型字段之间的关系异常很不成熟与实用。数据挖掘算法在数据清理中的应用需要加强 。
(2)尽管识别重复记录受到最多的关注,采取了许多措施 ,但识别效率与识别精度问题,并不令人满意 。特别是在记录数据非常多时 ,耗时太多 ,有待于更好的算法 。
(3)以前数据清理主要集中在结构化的数据上 , 半 结构化的数据(如 XML 数据[3])已受到越来越多的重视。特别是由于 XML 自身所具有的特点(通用性 、自描述 性),在数据清理中应受到重视 。
(4)用户友好性 ,很多系统都提供了描述性语言 , 但基本上都是经过某种已有语言(如 SQL ,XML)根据自己 需要经过扩展实现 ,不能很好地满足

08

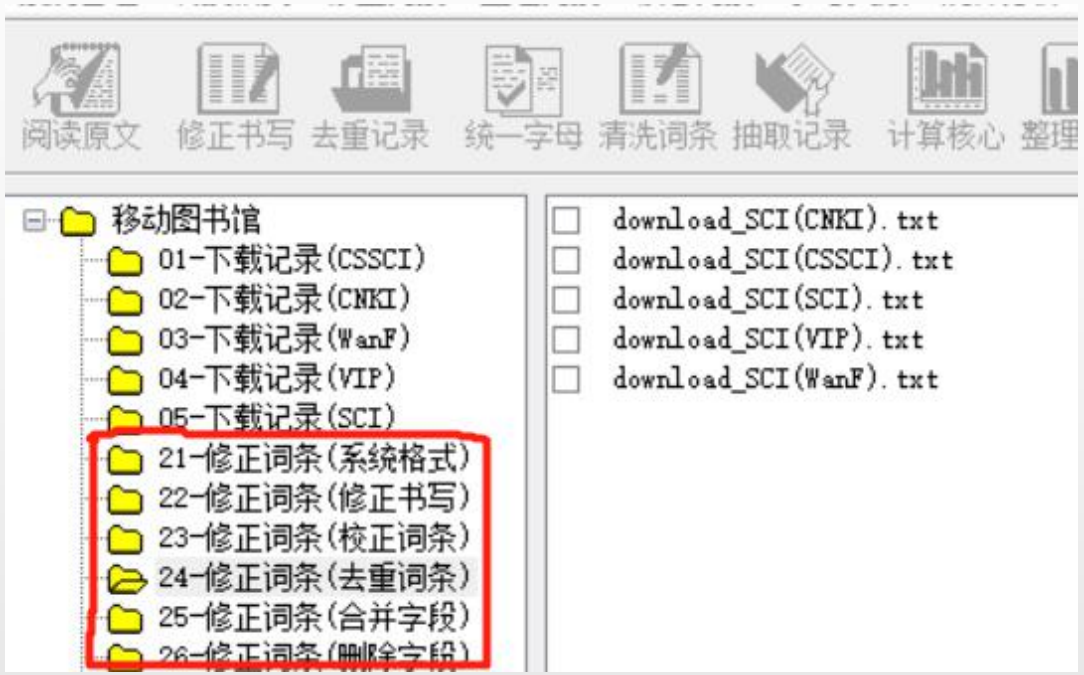
数据清洗与整理实践

- 数据来源：中国知网
- 检索词：主题=数据清洗
- 数据年份：2015年1月1日——2020年1月日
- 检索结果：2383条
- 清洗软件：Bibstats
- 检索时间：2020年4月26日

08

数据清洗与整理实践

格式转换与修正词条



08

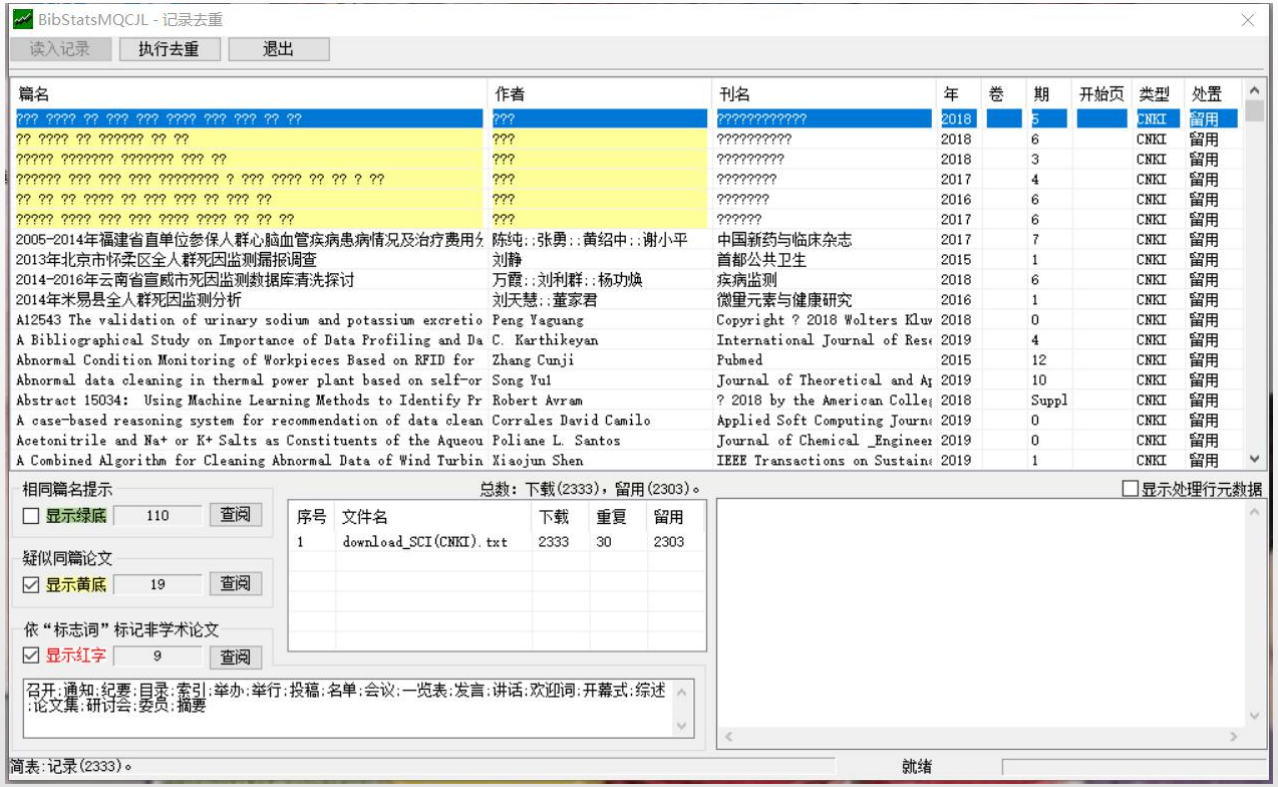
数据清洗与整理实践

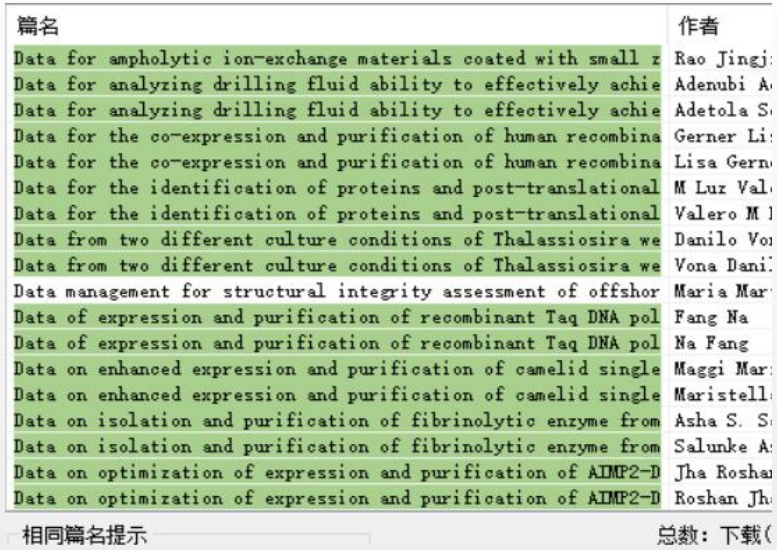
校正词条

08

数据清洗与整理实践

- 执行去重之后留用了2303篇



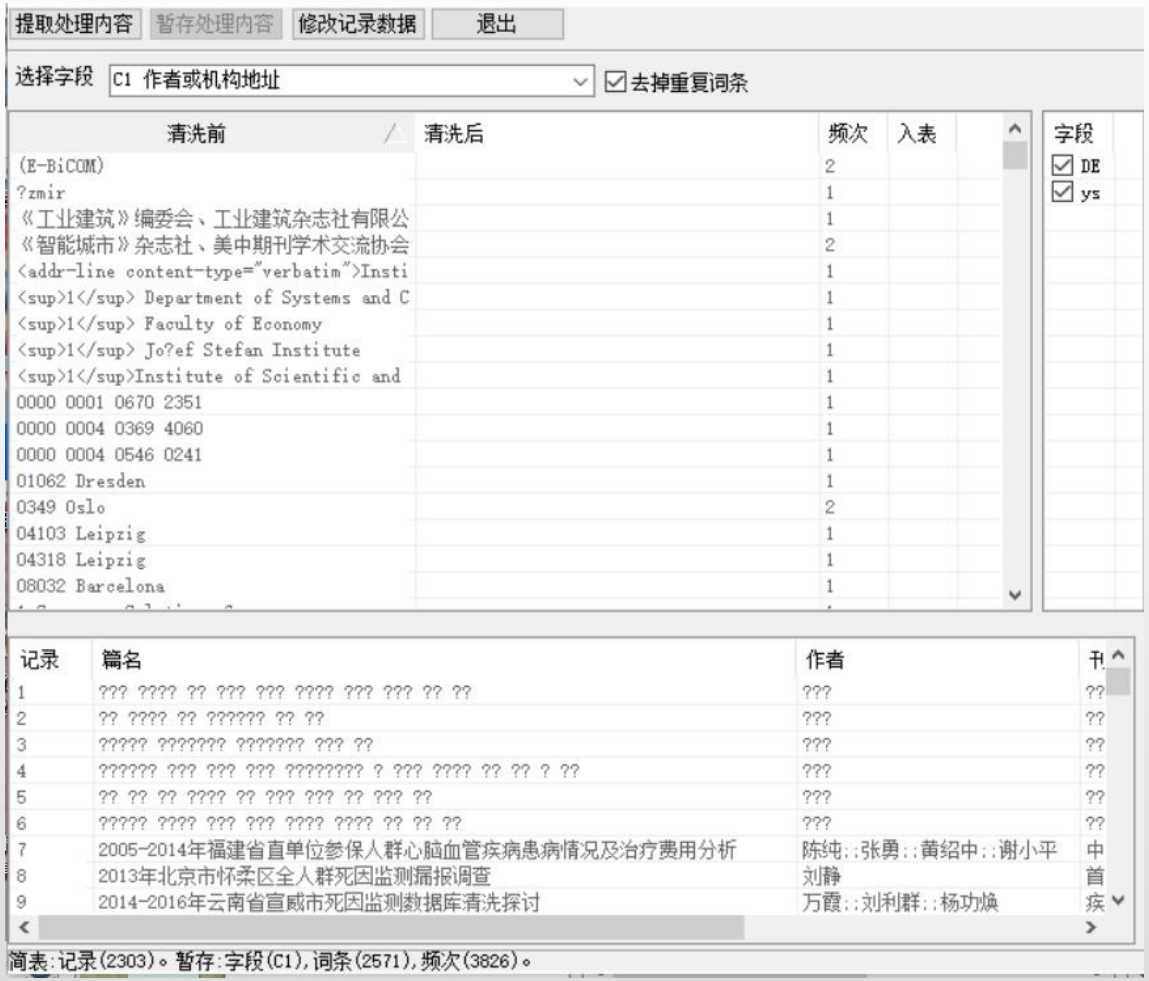


重复数据

08

数据清洗与整理实践

- 清洗词条：提取处理内容-修改数据记录



08

数据清洗与整理实践

- 清洗结果

清洗前	清洗后
AU--词条（4214）	词条（4008）
DE作者关键词--词条（10985）	词条（7533）
C1作者名或机构名--词条（5234）	词条（2571）

小结

现今企业的成功和社会的进步，越来越依赖于数据和对其所做的分析．为了获得竞争优势即使是小企业也会投入时间和精力来收集和分析数据．很多大公司都部署了自己的云服务平台，国内比较著名的有百度云、阿里云、等．但是如果一味地将精力投入到对数据所做的分析而不关注数据本身，很可能产生灾难性的后果．统计表明，美国企业中1%～30%的数据存在各类错误和误差，医疗数据库中13.6%～81%的关键数据不完整或陈旧。数据质量问题会使基于其的分析和研究毫无意义甚至还会产生灾难性的后果，在美国由于数据错误引起的医疗事故每年使98000名患者丧生，因此对于数据的清洗和整理工作就显得愈发重要和必要。

参考文献

[1]王曰芬,章成志,张蓓蓓,吴婷婷.数据清洗研究综述[J].现代图书情报技术,2007(12):50-56.
[2]陈孟婕. 数据质量管理与数据清洗技术的研究与应用[D].北京邮电大学,2013.
[3]叶鸥,张璟,李军怀.中文数据清洗研究综述[J].计算机工程与应用,2012,48(14):121-129.
[4]赵一凡,卞良,丛昕.数据清洗方法研究综述[J].软件导刊,2017,16(12):222-224.
[5]杨辅祥,刘云超,段智华.数据清理综述[J].计算机应用研究,2002(03):3-5.
[6]郭志懋,周傲英.数据质量和数据清洗研究综述[J].软件学报,2002(11):2076-2082.
[7]杜小勇,陈跃国,范举,卢卫.数据整理——大数据治理的关键技术[J].大数据,2019,5(03):13-22.
[8]潘玮,牟冬梅,李茵,刘鹏.关键词共现方法识别领域研究热点过程中的数据清洗方法[J].图书情报工作,2017,61(07):111-117.


谢谢
THANK YOU