面向领域分析的文献数据清洗策略研究

盛怡瑾，黄政，张学福
（中国农业科学院农业信息研究所，北京 100081）

摘要：为提高用于领域分析的文献数据质量，本文分析了文献数据的需求和特点，比较了常用的清洗方法和工具，并设计出一套清洗流程，用动物资源与育种领域的文献数据进行验证。结果表明，该流程科学有效，能够指导领域分析文献数据的清洗实践；同时在该流程指导下，可用多种工具实现优势互补，有助于提高后续领域分析的质量和效率。

关键词：文献数据；领域分析；数据清洗
中图分类号：TP391
DOI：10.3772/j.issn.1673-2286.2015.12.001

通过对某一学科或研究领域文献的相关特征如题名、关键词、作者、参考文献等进行动态监测和领域分析，可以了解和评价学科或研究领域的发展历史、现状和趋势。但从海量文献数据中通过多种来源和方法遴选出所需数据集时，还不能马上使用这些原始数据，因为其中存在诸多的数据质量问题。因此在分析文献数据之前，必须进行数据清洗，这是保证分析过程顺利进行和分析结果准确可靠的前提条件。据统计，在文献计量分析中，数据清洗所占的时间是全部工作量的80-90%[1]，这更证明了做好数据清洗工作的重要性。

1 基本理论

1.1 概念介绍

Rahm等将数据质量问题分为单数据源和多数据源问题两大类，每一类又可再分为模式层和实例层问题[2]（如图1所示）。

要提高数据质量，就必须要进行数据清洗。数据清洗的目的是检测和消除数据中的错误和不一致，以提高数据质量[3]。对模式层数据进行清洗主要是处理命名冲突和结构冲突，命名冲突是指同一名称代表不同的对象或者不同名称代表同一个对象。结构冲突是指不同数据源中，同一对象的表现方式不同。实例层数据清洗主要分为对属性清洗和对记录清洗[4]。对属性清洗包括对空值、异常值、不一致数据和嵌入值进行清洗；对记录清洗主要指的是对相似重复记录的清洗，即数据匹配。其中数据匹配是数据清洗研究领域的核心。数据匹配是将客观上表示现实世界同一实体的，但是由于在格式、拼写上有些差异而导致DBMS（Database Management System，数据库管理系统）不能正确识别的记录（也叫相似重复记录）进行识别的过程。

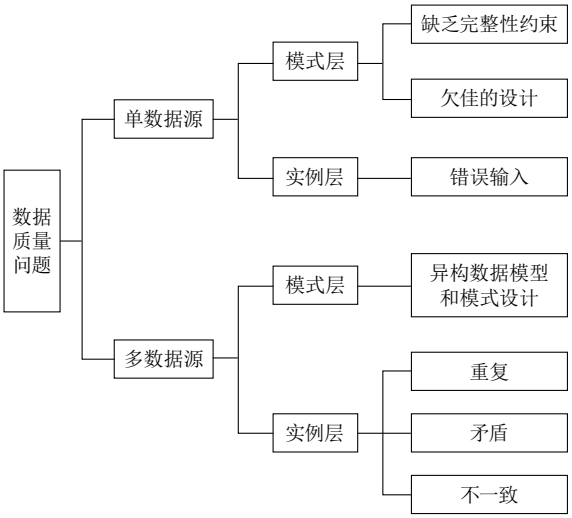

图1 数据质量问题的分类

1.2 面向领域分析的文献数据特点

首先，领域分析所涉及的文献数据量很大。尽管通过筛选构建了有针对性的数据集，但与常用的领域分析工具的承载量相比，还是存在着数据量过大难以导入的问题。而且，文献记录的字段值多为文本形式且字符串比较长，相对来说占内存比较大，对计算机的配置要求高。

其次，同源数据质量高。文献数据多收录在各个数据库中，如Web of Science、SpringerLink、EI等，在各数据库中文献数据格式都是一致的。而且，因为事先经过编辑的校对，文献数据的拼写错误和异常值很少。相对其他数据来说，质量很高。

第三，数据匹配问题突出。对于文献数据，同一作者和机构很可能有好几种不同的表达方式，或者是写法上的差异，如符号、缩写等；或者是粒度上的差异，如表述地址时描述的细致程度不同，使得同一实体拥有相似却不完全相同的表现形式。同时，在进行领域分析时，常常需要合并来自不同数据源的数据，这导致同一实体常有两个以上相似但不完全相同的记录，造成了数据匹配问题突出，使用之前必须对相似重复记录进行匹配。

1.3 数据需求及清洗任务

在用文献数据进行领域分析时，更多地会用到标志文献外部特征的篇名、作者和机构，以及标志文献内部特征的关键词。要求用于分析的数据必须做到相同作者、机构和关键词格式一致、表述方式相同；指向同一实体的记录具有唯一性。

文献数据的特点决定了同一来源的数据在模式上一致并且拼写错误和异常值很少，但存在关键词符号、大小写和单复数的不同（如2-breed和2-breeds）；作者和机构因词语位置排列、缩写和粒度等而形成相似重复（如Mick, David U.和Mick, D. U.；CAAS和Chinese AcadAgrSci）。多种来源的数据，首先模式不同；其次，不同来源中可能都包含了描述同一实体的记录，却因格式、写法等差异造成了相似重复记录。

所以在清洗工作中，对于单数据源，要对关键词进行标准化，对作者和机构进行匹配；对于多数据源，除了进行单数据源的清洗任务外，还要考虑对相似重复记录进行清洗。同时，因为文献数据量与分析工具承载量之间的矛盾，在清洗的时候，还需要保留重要数据，缩减数据集，从而在不影响分析结果的前提下保证工具正常运行。

虽然前人已对文献数据清洗方法进行了许多研究[5-7]，但这些方法彼此孤立，各自为营，并没有一个用于指导面向领域分析的文献数据清洗的策略和流程，使人们在面临众多方法和工具时不知所措。我们在下文中将常用的文献数据清洗方法进行对比分析，并提出一个清洗流程。

2 文献数据的清洗方法及工具

2.1 清洗方法

文献数据清洗方法包括模式转换方法、标准化方法和数据匹配方法。模式转换一般通过建立表字段映射规则、拆分规则和值合并规则或计算机转换函数进行处理。对关键词的标准化可以应用多种变换函数、格式函数、汇总分解函数和标准库函数去实现清洗[4]。

数据匹配方法分为分块、字段匹配和记录检测。对单数据源作者和机构进行匹配，会用到分块和字段匹配方法；对多数据源相似重复记录进行匹配，则一般会用到全部三个方法。

（1）分块

为了减少两两比较次数，可以采用分块技术，通过分块键将记录分成彼此独立的数据块，并假设不同的数据块中不存在能匹配的记录，分块键由单个或者多个属性组合生成。常用的分块方法有标准分块技术，基本近邻排序算法SNM（Sorted Neighborhood Method）、Bigram索引和聚类。

（2）字段匹配

文献数据中常用的字段匹配方法有基于字符和基于token两种，分别从字符和词的层面上判断字符串相似性，其对比如表1所示。

（3）记录检测

多年来，主流的记录检测方法一直是概率匹配模型，近年来，对机器学习方法的研究越来越多。各方法的特点对比如表2所示。

2.2 清洗工具

进行领域分析的工具如TDA、CiteSpace、Scimat、

表1 基于字符和基于token的字段匹配方法特点对比

	基于字符的字段匹配方法				基于token的字段匹配方法	
适用	在字符水平效果很好，对拼写错误、位置差异或单词缩写等造成的相似重复有好的识别效果				在单词水平效果很好，比如可以解决因书写习惯不同而产生的意义相同但词组内单词顺序不同的两字符串的匹配问题[5]	
算法	Jaro-Winkler[6]	编辑距离[7]	Smith-Waterman算法[8]	q-gram算法[9]	cosine相似度与TF.IDF相结合[10]	q-grams与 TF.IDF相结合[11]
时间复杂度	$O(\|s_1\|+\|s_2\|)$	$O(\|s_1\|\cdot\|s_2\|)$	$O(\|s_1\|\cdot\|s_2\|)$	$O(\max\{\|s_1\|,\|s_2\|\})$	-	-
特点	添加前缀范围因子，增加字符串起始部分的权重	考虑对单个字符的最小操作数量（插入、删除、替换）；复杂度高	对开头和结尾的不一致比较宽容；计算复杂；矩阵在计算其值时具有对称性，可减轻运算量	计算两字符串中任意连续q个字母的有序组合q-gram向量间的距离来判断相似度；灵活；计算迅速	TF.IDF方法对单词i在字段中赋予权重， cosine相似度通过权重向量计相似度	TF.IDF方法对q-gram赋予权重，cosine相似度通过权重向量计相似度
衡量标准	Jaro-Winkler距离的值越小越不相似	编辑距离越大，两字符串越不相似	-	q-gram向量间的距离越小，相似度越高	cosine相似度越大，两字符串越相似	cosine相似度越大，两字符串越相似
适用情况	短字符串（如姓的比较）	误植（键入/拼写）错误；对长度差异很大的两字符串进行快速过滤	长字符串；字符串有缩写、缺失和少量语法不同的情况	格式不一致；误植；常见拼写错误；比编辑距离更适合处理略长一点的字符串	对字符串中单词的位置不敏感；因书写习惯不同而产生的意义相同但词组内单词顺序不同的两字符串	因单词的插入和删除而形式不同但意义相同的情况；拼写错误
不适用情况	长字符串；前缀相同但意义不同的两字符串	字符串与其缩写形式的匹配	字符串顺序颠倒的情况	-	拼写错误	-

表2 记录检测方法特点对比

	特点	分类	简介
概率匹配模型	贝叶斯推断问题；一般情况下，$P(x\mid M)$，$P(x\mid U)$，$P(U)$，$P(M)$的分布都难以得知，要做出条件独立性假设，独立性假设不满足时可以用EM模型，已发展了无监督的EM模型	基于最小错误的贝叶斯决策规则	将x错误分配到M和U所蒙受的损失或者付出的代价是相等
		基于最小代价的贝叶斯决策规则	根据犯错误造成的后果严重性，两类错误权重是不同的，本类方法目标是使付出的总代价最小
		有拒绝域的决策	当似然比接近阈值时，容易判断错误，需要专家进行人工判断；需要合理地确定拒绝域的阈值
机器学习	聚类，决策树，人工神经网络[12]，遗传算法等[13]	有监督的机器学习方法	-
		半监督的机器学习方法	减少了训练数据的数量，节省了很多人力也提高了准确性
		无监督的机器学习方法[14]	不需要或者需要很少的训练数据
距离	领域无关，难点在于对于对阈值的确定，适用于训练数据完全匮乏，人工输入也不可能	将记录看成长字符串	简单；丢失了很多统计信息
		综合每个字段的距离计算结果	-
规则	需要数据集有规律可循；需要专家定义规则；需要规则制定者对整个数据集和相关领域的特点非常了解[15]	-	需要找到数据之间的关联规则，当数据量很大，数据规律性不强且研究人员缺乏对数据的领域知识时表现不好

Sci2等都会自带一些清洗功能，但特点和侧重都不同，对比如表3。

研究中还比较常用的工具如VosViewer，基本没有数据清洗功能。通过以上表格我们看到，在现有比较常用的文献分析工具中，数据清洗功能并不能满足需求，在实际工作中有时还需要用到一些专有的数据清洗工具，其特点对比如表4。

在实际研究工作中，应该根据数据集特点和分析需要，选择一个或者几个工具进行数据清洗，从而达到更好的清洗效果。

表3 文献分析工具清洗功能对比

	TDA	CiteSpace	SciMat	Sci2
数据清洗功能	可用List cleanup进行字段匹配；用thesauri进行基于规则的数据匹配	对数据按照学科分类、机构、被引频次等进行筛选；对arXiv，CNKI，CSSCI等9个文献数据库的数据进行模式转换成为软件能处理的格式	规范词的大小写、单复数，模式转换，对作者进行数据匹配	文本数据预处理，如去除来自ISI的完全重复数据等；对记录排序筛选
具体操作	List cleanup使用字段匹配方法；thesauri使用事先设定好的叙词表（规则）	内置数据库和数据转换器（模式转换）	用编辑距离进行作者数据匹配	根据ISI数据的ID号去除完全重复的数据
可处理格式	能够对多种文献数据库格式的数据进行处理；处理后可以生成与输入数据相似的格式	以ISI数据库的文本数据格式为标准，输入的数据必须是以“download**.txt”命名的文档	只有三种输入格式；采用java开发，代码规范，扩展性好；内置数据库用于文献元数据的管理	可处理txt、csv、Pajek、TreeML等；对ISI数据库的txt进行首行处理后才能识别
优点	强大的数据清洗功能，普遍	-	开源，便于按需求开发；数据处理量较大；优秀的数据管理功能	-
不足	不开源；需要用户对list cleanup的判断界面进行手工审核，耗时；自带叙词表不能满足所有需求；处理的数据量不大	基本不具备数据处理功能；数据处理量不超过2万条	不能对机构进行清洗	只可对数据进行简单的处理

表4 常用数据清洗工具特点对比

	DataStage	Kettle	Google Refine	Febrl
数据清洗功能	ETL工具，抽取、清洗、转换、装载	ETL工具，抽取、清洗、转换、装载	数据匹配；具备excel对数据的处理功能	数据标准化，分区，字段匹配，记录匹配
可处理格式	包括txt、XML、企业应用程序SAP、Siebel等；几乎所有数据库系统	包括txt、csv、XML、excel、Access等	Txt，网络数据	CSV，COL，TAB
优点	几乎支持所有编码；并行运行能力强；能处理大量数据，清洗功能强大	开源；数据转换过程灵活；能满足一般的ETL需求	强大的文本聚类能力	开源；既可进行数据标准化，又可以进行数据匹配；整合了26种不同的字段匹配方法
不足	价格昂贵；运行时占用资源较多	对异构数据库的处理能力不强；抽取速度和处理能力逊色于DataStage	能处理的数据类型少	能处理的数据格式少；没有固定的清洗流程

3 清洗流程制定

用于领域分析的文献数据特点和领域分析对数据质量的要求，决定了清洗时不用照搬一般数据的清洗流程，而是可以有针对性地进行取舍，以求得到更精确的结果。

在进行文献数据清洗时，先要进行需求分析，通过分析数据来源、特点和领域分析的细分对象确定清洗对象，并结合清洗目标确定清洗任务。分析数据来源主要是确定单数据源或多数据源，因为单数据源和多数据源存在的问题和解决的方法是不同的；分析数据特点主要是明确数据集中会存在哪些脏数据；分析领域分析的细分对象是为了确定哪些属性对后续的分析很重要，从而避免将过多精力用在不重要的属性清洗上。清洗目标是根据领域分析工作的需要，对数据进行清洗后所期望实现的结果。

清洗任务确定后，需要进行工具选择。通过对工具的功能和接口性质进行综合考量，选择能够完成清洗任务的工具或工具组合。工具功能指的是工具是否具有完成清洗任务的功能模块或算法，接口性质指的是工具的导入导出格式和承载数据量，在组合使用多种工具时应重点考虑工具的接口性质。

最后进行数据清洗，单数据源数据依次进行词标准化和作者与机构匹配，多数据源要进行模式转换，对词标准化，在属性层面进行作者与机构匹配后还需要在记录层面进行数据匹配，如图2所示。

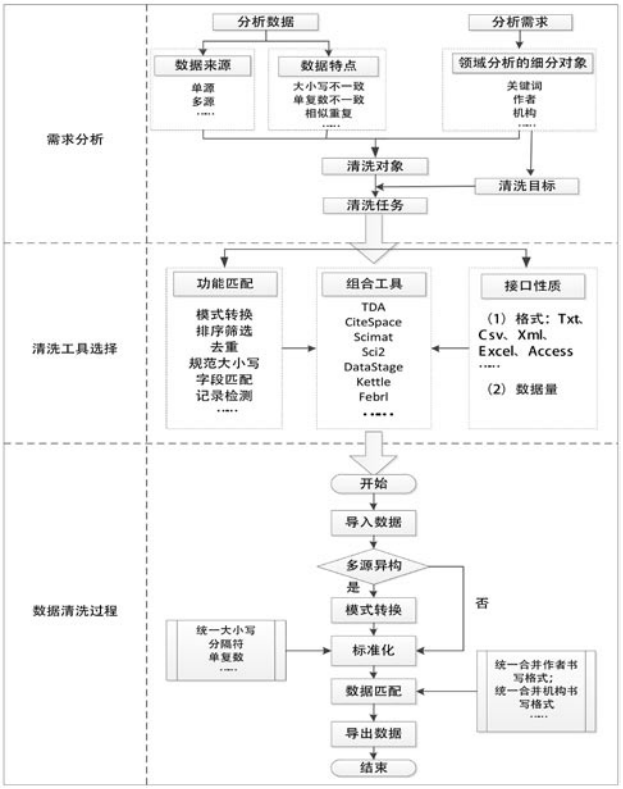

图2 清洗流程图

综上，需要借助于其他可以处理ISI格式并且处理量较大的工具来完成数据标准化和数据匹配工作，进而面向特定领域分析需求，基于文献集特征测度方法动态构建并缩减数据集，设计导出符合CiteSpace处理格式的数据。

4 实证分析

以“动物资源与育种领域动态监测与演化分析”研究为例，构建了58,000余篇文献数据集，并选择用CiteSpace作为领域分析的工具，按照上文提出的清洗流程逐步进行清洗工作。

4.1 需求分析

数据来自于Web of Science的文本格式数据，属于单数据源数据。其主要存在如下问题：关键词大小写和单复数不一致；机构缩写和全称、粒度等造成的数据匹配问题；受单机内存以及CiteSpace工具的数据承载量限制，只能处理2万余条数据，数据集无法导入使用；CiteSpace对数据格式要求严格，清洗能力较差。

4.2 工具选择

据前期调研分析，SciMat可以处理大量的ISI数据，其强大的专门针对文献数据的数据清洗功能可以满足大部分需要，而且，其开源性允许对导出功能进行编程改进以便缩减数据集，故将SciMat作为初步的数据清洗工具；由于SciMat对机构的清洗功能较差，故选择可以单独清洗机构的TDA对机构进行再次清洗，因为TDA能处理的数据量不大，因此需要缩减从SciMat中导出的数据量，并且将格式转化成TDA能处理的ISI文本或特定格式的EXCEL。

4.3 数据清洗

SciMat自带SQLite数据库工具，通过SQL语句多

表查询，可以对数据进行简单的清洗。本实例数据来源于同一平台，不需要转换模式。运用SciMat软件的词自动归一功能，可以自动实现文献关键词的单复数自动分组，对98,617个词进行单复数归一操作，操作后得到92,008个词，操作前后部分示例如表5所示，耗时近20分种。利用所设计的数据集导出接口，将按机构出现频次排序的2万余条数据以ISI格式导出。

表5 关键词清洗前后对比实例

关键词清洗前	关键词清洗后
(co)variance-components	(CO)VARIANCE-COMPONENTS
(co)-variance-components	
(co)variance-component	
buffalo	BUFFALO
buffaloes	
buffalos	
22-khz-call	22-KHZ-CALL
22-khz-calls	

机构相似重复实例如表6所示，使用TDA的List cleanup对其进行清洗，List cleanup使用模糊字段匹配的算法对机构进行相似度的计算，给出的参考结果如图3所示，这个结果需要人工审核并进行调整，将误判为相似重复机构的记录移除出去，或者调整每一组记录的标准名称，完成2万余条机构清洗共需要700min，因为介入了人工调整，清洗结果准确率较高。

表6 机构相似重复情况

序号	机构
1	Chinese AcadAgrSci, Lanzhou InstHusb&PharmaceutSci, Lanzhou 730050, Gansu, Peoples R China.
2	Chinese AcadAgrSci, Lanzhou InstHusb&PharmaceutSci, Lanzhou 730050, Peoples R China.
3	Chinese AcadAgrSci, Lanzhou InstHusb&PharmaceutSci, Key Lab Yak Breeding Engn Gansu Prov, Lanzhou 730050, Peoples R China.
4	CAAS, Inst Lanzhou AnimSci& Vet Pharmaceut, Lab Vet Pharmaceut, Lanzhou 730050, Peoples R China.
5	InstAnimSci& Vet Pharmaceut CAAS, Lab Vet Pharmaceut, Lanzhou 730050, Peoples R China.

图3 TDA机构清洗示意图

经过两种工具优势互补和组合使用，得到了用于领域分析的高质量数据，很好地完成了清洗任务。

5 结论与展望

数据清洗在文献分析工作中起着举足轻重的作用，其在整个分析过程中所占的工作时间和工作量都是比例最大的。现有方法和工具在满足多样的数据清洗需求方面各有特点和不足，基于此，本文设计合理的清洗流程并使用工具组合进行改进以达到更好的清洗目的和效果。组合使用的过程中，需要特别了解各种工具所需要的数据格式和数据量要求，并加强工具间的数据衔接。未来研究中，可以考虑根据数据需要，对开源工具的数据清洗算法进一步优化，以增强其功能和表现；对于数据匹配任务，要引入如人工神经网络、遗传算法等机器学习算法，从而在节省人力的基础上取得高准确率的结果；增强接口技术，开发能够衔接多种工具的中间件，减少单独开发的重复浪费；还要通过分布式架构和算法优化，将运行时间控制到更小的区间，从而为领域分析提供质量保证。

参考文献

[1] 张晋辉,刘清. 基于推理机的SCI地址字段数据清洗方法设计[J]. 情报科学,2010(05):741-746.

[2]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.

[3] ERahm, H HDo. Data cleaning: Problems and current approaches[J]. IEEE DATA ENGINEERING BULLETIN, 2000, 23(4): 3-13.

[4]李明. 数据清洗技术在文本挖掘中的应用[D].南京:南京理工大学,2008.

[5] RBaxter, PChristen, TChurches. A comparison of fast blocking

methods for record linkage[J].KDD WORKSHOPS, 2003: 25-27.

[6] W EWinkler. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage[J].PROCEEDINGS OF THE, 1990:8.

[7] V ILevenshtein. Binary codes capable of correcting deletions, insertions and reversals[J].SOVIET PHYSICS DOKLADY, 1966, 10(10): 707-710.

[8] T FSmith, Waterman M S. Identification of common molecular subsequences[J]. JOURNAL OF MOLECULAR BIOLOGY, 1981 (1): 195-197.

[9] 邱越峰,田增平,季文赟,周傲英. 一种高效的检测相似重复记录的方法[J]. 计算机学报,2001(01):69-77.

[10] WWCohen. Integration of heterogeneous databases without common domains using queries based on textual similarity[J].ACM SIGMOD RECORD,1998, 27(2): 201-212.

[11] LGravano, P GIpeirotis, NKoudas, et al. Text joins in an RDBMS for web data integration[C]. Proceedings of the 12th international conference on World Wide Web. New York:ACM, 2003: 90-101.

[12] D RWilson. Beyond probabilistic record linkage: Using neural networks and complex features to improve genealogical record linkage[C].Neural Networks (IJCNN),USA:IEEE, 2011: 9-14.

[13] G.P.Hettiarachchi, N.N.Hettiarachchi,D.S. Hettiarachchi,et al. Next generation data classification and linkage: Role of probabilistic models and artificial intelligence[C]. 2014 Global Humanitarian Technology Conference (GHTC), USA:IEEE,2014: 569-576.

[14] PChristen. Automatic training example selection for scalable unsupervised record linkage[M]. Berlin :Springer, 2008: 511-518.

[15] SPorwal, DVora. A Comparative Analysis of Data Cleaning Approaches to Dirty Data[J]. INTERNATIONAL JOURNAL OF COMPUTER APPLICATIONS, 2013, 62(17): 30-34.

作者简介

盛怡瑾，女，1988年生，中国农业科学院农业信息研究所,硕士研究生，研究方向：信息资源管理，E-mail：shengyijin_syj@163.com。
黄政，男，1992年生，中国农业科学院农业信息研究所,硕士研究生，研究方向：信息资源管理，E-mail：17888802420@163.com。
张学福，男，1966年生，中国农业科学院农业信息研究所,研究员，研究方向：农业知识组织与可视化分析，通讯作者，E-mail：zhangxuefu@caas.cn。

Research on Domain Analysis Oriented Literature Data Cleaning Strategy

SHENG YiJin, HUANG Zheng, ZHANG XueFu
(Agricultural Information Institute of CAAS, Beijing 100081, China)

Abstract: To improve the quality of literature data used in the domain analysis, we analyze the needs and characteristics of literature data, compare the common used cleaning methods and tools, and then design a cleaning process. We also use literature data in the field of animal resources and breeding to illustrate. It shows that this cleaning process is scientific and effective, and capable of directing the cleaning practice. Meanwhile, under the guidance of the cleaning process, carrying out combination of tools can complement each other to further improve the quality and efficiency of subsequent domain analysis.

Keywords: Literature Data; DomainAnalysis; Data Cleaning

（收稿日期：2015-12-08）