大数据治理中数据清洗方法的思考

陈世鹏

（北京锐安科技有限公司 北京 100000）

[摘 要]在各类大数据项目中，数据清洗是数据挖掘和应用的重要保障。脏数据会导致不可靠输出，如何高效去掉脏数据、提升数据质量、对异常数据进行修复是数据治理运营的重要环节。本文通过对数据质量问题及数据清洗方法的分析与总结，提出基于NLP算法对某类大数据行业的数据进行清洗和质量提升的方法。

[关键词]数据清洗；NLP；数据质量；数据挖掘

【DOI】10.12252/j.issn.2096-627X.2020.07.955

0 引言

随着移动互联网的发展，人类正处于一个信息爆炸的时代，面对海量的数据，人们仍然经常抱怨“数据丰富，信息贫乏”，因此怎么组织和存储数据，才能挖掘提炼数据价值、高效的获取所需的信息，是目前行业迫切关注的问题[2]。在某些大数据行业，相较于一般企业的单一或几种数据源，它汇聚了各种类型的多源异构数据，但由于数据复杂的来源，受到源头的环境或录入、传输的影响等，数据在采集和导入过程中容易引入脏数据。由于没有统一的标准规定，只能依靠数据分析师对采集的数据进行分析和清洗，从而提升数据质量，据统计，数据清洗在一个项目生命周期中占据了相当比例的时长，因此数据清洗的方法和效率显得尤为重要[1]。

1 数据质量问题

大数据相关很多行业的数据典型特征是多数据源，由于各业务系统在功能需求和设计上的不同，经常会出现数据不一致、字段命名冲突、属性值和结构冲突等问题。例如空缺值问题，这类数据主要是一些应该有的信息的缺失，影响对数据的分析挖掘；不一致问题，这类问题主要是对字段取值没有约束条件或约束条件过于简单，没有进行逻辑校验，表现在字段取值超出了规定范围。作为对数据清洗结果的检验标准，综合各参考文献，数据质量主要有如下10个维度[3]评判标准：数据规范、数据完整性、数据重复检测、数据准确性、一致性和同步、及时性和可用性、数据覆盖、可理解性和数据衰变。

2 数据清洗方法

针对不同的数据质量问题，对应有不同的数据清洗方法，常见的清洗方法有：

1）缺失数据处理

缺失值问题在一种普遍现象，常见的处理方法是忽略对数据分析挖掘无价值的特征值；对于一些具有连续变量值属性的特征变量，可以采用均值、插值或回归分析等进行填充缺失值。基于不完备数据分析的思想，通过定义约束容差集合的差异度，从集合角度判断不完备数据对象的总体差异程度，并以聚类的结果为基础可进行缺失数据的填补[4]；针对一元回归方法的预测进度不够，可以利用多元回归[5]，配以适当惩罚函数防止过拟合，在某些场景下对缺失值进行很好的估计。

2）相似重复数据处理

将个别字段差异或完全一致的记录称为相似重复记录，在多数据源数据集成时该现象普遍存在。重复数据不仅导致数据的冗余，浪费网络带宽和存储空间，还提供用户很多相似信息，对用户有误导作用。针对相似重复记录，邻近排序算法是常用的检测方法，还有各种改进排序算法等，例如多趟排序、优先权队列等，还有基于机器学习进行相似度聚类来进行重复检测。

3）逻辑错误数据处理

数据逻辑错误指特征变量与其属性值不符，违背了业务规则或逻辑。一般都是使用业务规则对逻辑错误进行检测，在具体的应用领域，根据领域知识制定约束规则，例如规定特征变量取值的有效范围、正则化匹配等来对数据的逻辑和有效性进行检验。

4）不一致数据处理

不一致数据的常见问题有特征值的表示方法不统一、量纲不一致等，给数据分析带来了挑战。目前，常用消除数据不一致的方法有排序、融合和基于规则等方法。

5）异常数据处理

异常数据指不符合一般规律的数据对象[2]。其检测一种是基于统计学的方法，即采用数理统计方法获取数据的总体分布特征，利用箱线图等判断异常点；一种是基于距离的方法，例如KNN等机器学习方法，将数据按距离划分成不同的层，根据定义的距离计算各数据点到中心点的距离来判断数据是否存在异常。

3 基于NLP算法的数据规则制定和清洗

在某些具有特殊业务应用的大数据行业中，上述常用的数据清洗方法难以起到立竿见影的效果。由于来源的多数据源特性，字段名称、属性值和结构都可能存在冲突，特征变量的名称比较混乱，无固定标准，命名不统一。例如，不同表中使用相同字段表示不同属性，不同名称的字段表示相同的属性。命名规则和标准的缺乏是该类行业数据清洗和治理的一大难点。因此这类数据清洗需要从结构层和数据层两方面进行处理。

随着人工智能的发展，自然语言识别（NLP）也有飞速的发展，充分利用前言的NLP算法，对特征变量命名进行规范，结合传统的数据清洗方案，可以有效提升类似大数据行业的数据质量。具体流程如下：

1）整合语料库，形成初始数据元集合。目前有许多开源的语料库，由命名实体构成，命名实体一般分为3大类（实体类、时间类和数字类）和7小类（人名、地名、组织机构名、时间、日期、货币和百分比）。可对此进行整理，将命名实体进一步泛化，收集形成初始数据元集合。

2）针对来源的多数据源，对其特征变量属性，选择合适的NLP算法来进行匹配，识别命名实体。如果该命名实体存包含于1中的数据元集合中，可用该数据元进行命名；如果不存在，则对数据元集合进行扩充，并用新的命名实体进行命名。如此可对每张来源表统一命名规范，并对数据元的引用频次进行统计，根据业务需求可对数据元进行泛化扩充。

3）在形成统一的数据规则后，然后根据数据质量问题，定义数据转换规则和有效性校验规则，采取对应的数据清洗方法对数据进行清洗，提升数据质量。

4）根据规范后的结构和数据制定格转标准，对数据进行转换入库。

4 结语

目前，国外开发的数据清洗工具有很多，功能多样，各有优劣。数据质量问题也受到了各行各业的高度关注。本文分析总结了数据质量和质量问题的处理方法，并针对特定大数据行业的数据特点，提出了基于NLP算法的规则制定和清洗方案。当然，NLP算法也有多种，怎样选择合适的算法才能制定最佳的标准也需要进一步研究。

参考文献

[1]郭智慧，周傲英.数据质量和数据清洗研究综述[J].软件学报，2002.11

[2]宋金玉，陈爽.数据质量及数据清洗方法[J].实践与应用，2013.10第4卷第5期

[3]McFilvray D.数据质量工程实践[M]刁兴春，曹建军，张健美，等译.北京：电子工业出版社，2010

[4]武森，冯小东，单志广.基于不完备数据聚类的缺失数据填补方法[J].计算机学报，2012，35（8）：1726-1738

[5]张建中，方正，熊拥军，等.基于SNN数据清洗算法的优化 [J].中南大学学报：自然科学版，2010，41（6）：2240-2245

[6]庞雄文，姚占林，李拥军.大数据量的高效重复记录检测方法 [J].华中科技大学学报：自然科学版，2010，38（2）：8-11

幼儿园课堂教学有效提问的研究

陈 颖

（广西玉林市玉东新区第一幼儿园 广西 玉林 537000）

[摘 要]幼儿园是学前教育的具体教育场所，它主要是针对3到6周岁的幼儿，幼儿园的主要任务就是帮助幼儿来摆脱家庭教育所带来的限制，让幼儿的身体和智力可以得到有效的开发，帮助他们提供正确的成长环境，是整个教育体系的基础，可以初步的对幼儿进行素质上的培养。在幼儿园的课堂学习过程中，提问是主要手段，教师需要从这方面进行具体的研究。本篇文章通过对幼儿园课堂教学提问中存在的问题进行阐述，分析幼儿园课堂有效提问的意义，并且就如何实现有效提问进行探讨。

[关键词]幼儿园；课堂教学；有效提问

【DOI】10.12252/j.issn.2096-627X.2020.07.956

引言

幼儿园主要以语言、科学、艺术、健康和社会五个领域构成，教学内容主要以游戏为主，在新时代下，教师需要加强教学与生活之间的联系，增强幼儿在学习中的主动性，构建和谐的课堂氛围。在幼儿园的课程结构体系上，它主要是由课程目标、课程内容、课程活动以及课程评价四个要素工程，它具有一定的基础性，是对幼儿的一种启蒙性培养，通过最浅显的一些观念进行品质和能力上的渗透。在幼儿园课堂教学中，由于幼儿在年龄以及性格上的特点，如何与他们产生有效的互动这是教师的主要研究方向，就课堂提问来说，它的设计内容、提问时间规划等方面都有着一定的欠缺，需要教师采取有效的措施看优化。

1、幼儿园课堂教学提问中存在的问题

1.1提问内容的问题

对于现阶段的幼儿园课堂教学来说，很多教师都没有认识到提问的重要性，在问题内容的设计上没有很强的目的性，没有对幼儿进行一定的引导，往往是教师随口的提问，也没有对幼儿的回答进行分析和评价。其次，有些内容在设计的过程中过于直白，没有给幼儿思考的空间，无法形成思维模式上的培养，就是单纯的回答问题，无法真正的实现课堂提问的作用。实际上对于幼儿来说，这个阶段的思维塑造和习惯养成是非常重要的，如果内容过于简单，不利于幼儿个性化的培养。

1.2提问时机的问题

对于幼儿来说，他们在课堂上的注意力并不集中，没有集体的意识，做任何事情都是随着自己的心情，有些教师在提问上没有注意时机的重要性，没有注意对幼儿情绪上的调动，也没有注意提问过程中的层次性，这就导致有些幼儿根本无法对教师的问题进行理解，不会实现课堂提问的效果。

1.3提问方法的问题

有些教师在对幼儿进行提问的方法设计上过于单一，往往就是采用一问一答的形式，没有设计一些有趣的提问环境，也没有一定的创新性，很难激发幼儿的兴趣，而且有些不同的幼儿有着不同的性格特点，如果采用统一的提问方式，许多幼儿会不敢表达自己真正的想法。

2、幼儿园课堂有效提问的意义

2.1集中幼儿的注意力

对于幼儿园阶段的学习来说，教师需要在课堂教学中加强与幼儿的互动，吸引他们的注意力，提高课堂提问的有效性就是在侧面上激发他们的学习兴趣，使他们的思维方式处于一个高度集中的状态，让幼儿可以带着求知欲进入到学习中，感受