# 一种数据论文复用深度的评估方法

## 背景技术：

随着科学研究向“第四范式”即数据密集型科学的方向发展，学术界对研究可重复性与对数据共享的要求增加，科学数据共享的重要性受到广泛关注。数据论文作为三种主要数据出版模式之一（其它两种模式分别为独立的数据出版、作为论文辅助资料的数据出版），对数据来源、收集方法、数据结构、处理方法等内容进行描述，数据论文在促进科学数据共享方面发挥着重要作用，可以加强科学研究的可靠性、透明度和合作性，从而推动科学领域的发展。

当谈及现有的数据复用评价方法时，许多方法通常依赖于传统的引文分析指标，如引用次数和引用期刊的影响因子，这样的方法将所有的引用行为视为同等重要，而实际上，不同的引用行为可能反映了对数据的不同程度的复用。例如，对于一篇数据论文的引用，将其作为背景介绍与对数据结果的深入分析是截然不同的情况。在前者中，引用与数据论文的数据部分可能关系不大，而在后者中，引用者需要对数据论文中承载的数据进行一定的了解和分析，这明显反映出了后者对数据的更深层次的复用。因此，为了更准确地反映数据被复用的深度，需要对引用行为的意图进行更细致的分类。目前的引用行为意图分类研究主要关注一般的学术文献或科技文献，而没有专门针对数据论文进行深入的分类。

同时，值得注意的是，以往关于引文内容分析的研究数据往往需要从文献原文中提取引用信息，这些文献原文多数为来自于特定的期刊或出版机构的固定标准结构化文本，然而在实际情况中，一篇文献的引用可能来源于各种各样的期刊和出版物，文献的原文形式可能迥异，有些以不同的结构化文本格式呈现，而有些可能只存在于PDF文档中，这种多样性增加了对引用位置和引用上下文进行准确提取的难度。

## 发明内容

针对背景技术部分所提及的现有方案所存在的不足和缺陷，我们提出了一种数据论文引用意图的评价方法，旨在更好地反映数据被复用的程度。并整合各个必要路径的解决方案，处理不同来源的学术文献原文，设计出一个较为完备的数据论文复用深度的自动化评估系统。

为解决上述技术问题，本发明采用以下技术方案：

步骤1、 数据获取和预处理：

获取待研究数据论文的所有施引文献全文PDF，并进行严格的数据预处理和清洗，以去除噪声和不必要的信息。从中提取待研究数据论文在施引文献中的被引用位置（即施引文献引用部分内容对应的章节标题）、引用上下文和数据论文的摘要。

步骤2、 标注样本收集和扩展：

抽取待研究数据论文的施引文献的样本，并根据其数据引用行为类别进行人工标注。确保标注样本的数量足够多，并考虑使用半监督学习方法来扩展标注数据集，以提高训练效果。

步骤3、 向量表示的生成：

将步骤1中抽取到的引用位置、引用上下文和摘要作为模型输入，选择适用于学术文献任务的自然语言处理预训练模型SciBert生成文本的向量表示。

步骤4、 聚类分析：

根据步骤3生成的文本向量表示以及步骤2中标注的标本，选择适当的聚类算法（如K均值聚类、层次聚类或DBSCAN），以聚类的方式分析数据引用行为。

步骤5、 结果解释和可视化：

最后，对步骤4得到的结果进行解释和可视化，以帮助数据发布者理解不同的数据引用行为类别以及它们在数据论文中的分布，确保结果能够有效地反映出数据的复用深度。

其中，步骤1的详细步骤分解如下：

S1.1、 选定待研究数据论文的施引文献，并提取其元数据，包括文献标题、文献doi、文献作者名、期刊名、期刊出版年份、出版页码信息组成六元组，并抽取文献的摘要；

S1.2、 根据所提取的元数据在数据库中检索施引文献全文PDF并保存，并建立索引表跟踪每一篇文献的处理情况；

S1.3、 使用格式转化模块将文献PDF源文件转化为带有文中和文末引用标记的文本文件；

S1.4、 根据步骤S1.1中提取到的六元组元数据在步骤S1.3得到的带标记的xml文件末尾参考文献列表处匹配对应的引用项；

S1.5、 根据步骤S1.4中匹配的引用项寻找正文中的引用位置，对引用上下文和引用位置章节标题进行文本清洗和规范化并返回；

步骤2中的数据引用行为类别是根据现有的数据论文的引用行为总结出来的，分为以下几个类别，每个类别行为的复用深度是随着顺序依次递增的：

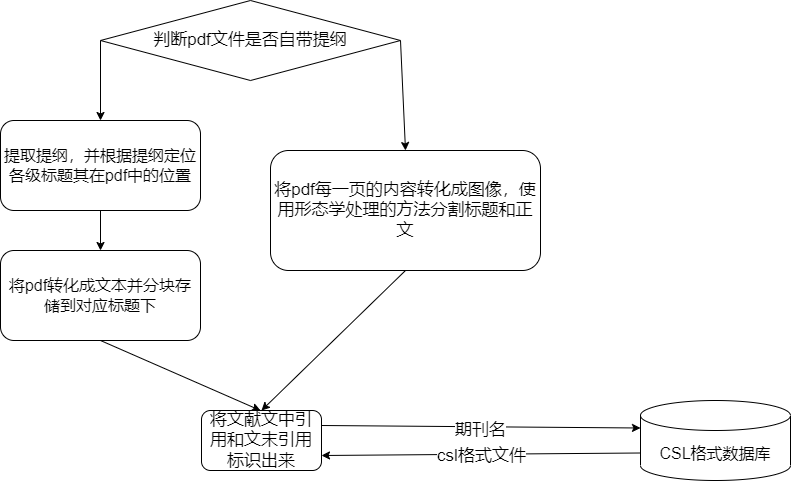
|  |  |
| --- | --- |
| **行为类别** | **描述** |
| 背景介绍 | 仅提及该数据作为背景的介绍 |
| 结论重用 | 包括简单得出的统计性的结果，对数据论文的结果进行引用 |
| 方法重用 | 借鉴了数据论文收集或者处理数据的方法，并运用到自己的数据收集或实验中 |
| 实验重用 | 使用了数据进行实验，进行了深度的处理并得出一定的结论；也可以是对自己实验结果的验证或者对比（如很多算法研究员喜欢用benchmark data对自己算法的性能进行验证） |
| 数据扩充 | 在原有数据集的基础上进一步扩充数据 |

步骤S1.3中的格式转化模块，首先判断PDF文件是否自带提纲（即包含章节标题及其具体页码、页中位置的目录），如果自带提纲，则根据提纲里的标题条目确定各标题的具体位置，将PDF转化成文本并分块存储到各级标题下；

如果PDF文件没有提纲，则将PDF文件按页码转化成图像文件，使用图像形态学的方法对文字进行分块，确定不同文字块中的字体大小从而区分正文和标题并提取标题，对文本进行分块存储。

在对文本进行分块提取后，根据文献的期刊名从存储了各学术期刊CSL格式文件的数据库中提取对应的CSL格式文件，并依据格式文件中的引用格式识别分块后文本中的文中引用位置和文末引用条目。

格式转化模块部分的设计图如下所示：



整体的技术路线如下所示：

