# 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、**引用结构抽取**

获取待研究数据论文的所有施引文献全文， 进行预处理，得到待研究数据论文在施引文献中的引用位置信息及引用上下文数据，同时抽取待研究数据论文的摘要和致谢数据；

步骤2、**基于引用位置判断的引用深度映射**

定义五种引用深度层级类别。采用穷举法建立受控词表，将步骤1获取的数据论文在施引文献中的位置信息自动化归一判断，将引用位置信息初步映射至已定义引用深度对应级别。

步骤3、**面向引用语义分析的内容向量化表征**

将步骤1中抽取到的施引文献引用上下文作为输入，使用scibert预训练模型得到其语义特征向量表示；

步骤4、**基于引用语义的引用深度判断**

设计神经网络分类模型根据步骤3得到的样本向量表示将施引文献从引用语境的角度映射至步骤2定义的引用深度层级。。同时对施引文献中的摘要及致谢进行目标数据集名称、数据集收集的机构作者等关键词进行匹配。

步骤5、**数据复用性判断赋值**

将步骤2得到的基于引用位置的数据引用深度、步骤4所得到的基于引用语义情况引用深度及摘要致谢部分所反映的引用深度三个角度的引用深度转化为可量化的强度指标，并按照不同的权重进行加权【有了语义，最后的判定还是归类，还是进一步有权重深度】，最终形成数据论文复用深度测度值。

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** | **引文位置关键词受控词表** | **量化强度（单位：1）** |
| 背景介绍 | 仅提及该数据作为背景的介绍 | Introduction, Background, Motivation, Literature review | 1 |
| 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 | Conclusion | 2 |
| 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 | Method, Methodology, Data collection, Data preprocessing, Data augmentation, Dataset | 4 |
| 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） | Experiments, Results, Evaluation | 4 |
| 数据扩充 | 在原有数据集的基础上进一步扩充数据 |  | 5 |