一种JSON和扩展JSON文档集合的模式提法方法

由于NoSQL面向文档数据库的日益普及，抽象-JSON文档正在成为数据表示的通用格式。这种流行的原因之一是它们在没有明确的数据模式的情况下处理大量数据的能力。但是，例如在数据检索、集成和分析任务期间，模式信息有时对应用程序至关重要。鉴于此背景，本文提出了一种方法——从存储在NoSQL面向文档的数据库或其他文件存储库的JSON或扩展的JSON文件集合中提取模式。考虑聚合操作以便为每个不同的结构获取模式，并且提出层级数据结构以对这些结构进行分组从而以JSON模式的格式生成全局模式。在真实数据集上进行的实验，如DBPedia和Foursquare，证明了生成的模式的准确性与相关工作相当甚至更优。

关键词：NoSQL；JSON；扩展的JSON；模式提取；JSON模式；面向文档的数据库

1. 绪论

大数据市场爆炸导致大型公司需要能够有效地存储和处理大量数据的数据库（DB）。鉴于此背景，传统的关系型数据库存在一些限制，因为它们优先考虑读写操作的强一致性而不是高可用性和水平扩展性，从而更好地处理增加的数据量【1】。

为应对关系型数据库的限制而出现的一系列新型数据库系统称为NoSQL DB【2】。基本上说，它们避免了传统ACID属性进行事务管理的开销，相反，提供了最终的数据一致性，强可用性和灵活性。NoSQL数据库的一个共性特征是它们是无模式的，即它们允许在不事先了解其结构的情况下存储数据【3】。例如，面向文档的数据库——NoSQL数据库类型之一，主要以JSON（JavaScript Object Notation）或扩展的JSON1格式（【1】，【4】）存储和检索具有简单和复杂属性的文档，并且文档不一定共享一个相同的结构。实际上，无模式的数据库避免了具有多个属性的记录问题，如果它们不统一，则没有值，允许每个记录仅包含必要的内容。

但是，缺少有关模式的信息是的难以执行数据集成过程，以及一些数据处理任务，例如数据检索、验证和分析【5】。然而缺少显式模式并不意味着完全没有模式，因为应用程序代码中通常存在访问数据库的隐式模式【6】。在NoSQL数据库的情境下，要了解数据收集模式在开发应用程序期间非常重要。例如，处理地理数据的几个应用程序（例如Foursquare），以JSON格式检索数据，但并没有为它们定义模式，这样由于不知道文档结构用户将难以查询这些数据。

基于此目的，本文提出了一种方法，旨在从JSON或扩展的JSON文档集合生成单个模式。生成的模式在JSON模式建议中定义，即建立作为指定JSON文档模式的标准【7】。所提出的方法，称为JSON Schema Discovery，主要用于帮助需要了解NoSQL面向文档的数据库模式的应用程序，用于多种用途，如NoSQL模式和数据集成。由于它们的流行，我们在这里强调NoSQL面向文档的数据库。但事实上，我们能够为它作为输入接收的任何JSON文档集合生成JSON模式。与相关的工作不同，我们的方法不仅为一组JSON文档生成模式，还可为一组扩展JSON文档生成模式，并定义了一个分层数据结构，其中包含对模式生成有用的多个元数据信息。这种分层结构由基于模型驱动工程（MDE）的过程操纵，该过程是用于处理数据模型转换的合适的软件开发技术【8】。此外，我们的解决方案可作为Web工具使用，允许以JSON Schema格式生成、持久化、查看和下载。

验证使用我们的方法生成的模式的质量的实验是在受控数据库以及相关工作使用的数据库上进行的。由于识别了模式的每个字段可可能数据类型，我们的方法生成的一些模式更加准确。这强调它是很有希望的。