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，当它们不统一