微软学术图谱数据集（Microsoft Academic Graph, MAG）[[1]](#footnote-1)包含学术出版记录、文献引证关系记录，以及学者、发文机构、期刊、会议、学科领域等信息。本次练习所采用的数据来源于MAG数据集中“数据可视化”（Data Visualization）领域的学者及发文数据子集，具体包含的数据表及字段含义如表1所示。

**表 1 MAG数据集字段说明**

|  |  |  |  |
| --- | --- | --- | --- |
| **表** | **表含义** | **字段** | **字段含义** |
| author2paper | 学者文献发表记录表 | paper\_id | 文献id [[2]](#footnote-2) |
| author\_id | 学者id [[3]](#footnote-3) |
| author\_sequence\_number | 这位学者在这篇文献中的排名（如第2作者、第5作者等） |
| paper | 文献信息表 | paper\_id | 文献id |
| doi | 文献doi [[4]](#footnote-4) |
| doc\_type | 文献类型 |
| title | 文献标题 |
| pub\_year | 文献发表年份 |
| journal\_id | 文献所在期刊名称 |
| citation | 文献引证关系表 | citing\_id | 施引文献id |
| cited\_id | 被引文献id |

1. <https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/?from=http%3A%2F%2Fresearch.microsoft.com%2Fen-us%2Fprojects%2Fmag%2F> [↑](#footnote-ref-1)
2. ID是信息管理领域常见用词，它的全称是identification，即“唯一标识”。比如，车辆的“唯一标识”（ID）是VIN（Vehicle identification number，车辆唯一识别码）。 [↑](#footnote-ref-2)
3. 即完成这篇学术文献的作者ID。如果某篇学术文献有多位作者，那么author2paper数据表中将会有5行记录，其中paper\_id相同，但author\_id记录了五个作者的ID。 [↑](#footnote-ref-3)
4. DOI即数字对象唯一标识符（Digital Object Unique Identifier），是一套识别数字资源的机制，涵括的对象有视频、报告或书籍等等。它既有一套为资源命名的机制，也有一套将识别号解析为具体地址的协议。DOI的体现形式主要包括二维码、条形码、字符码、网络域名等，数字对象唯一性，是DOI的典型特征，也是数字时代的“身份证”号码。 [↑](#footnote-ref-4)