DBLP 数据集

1. DBLP 介绍

DBLP 计算机科学参考书目（dblp computer science bibliography）提供有关主要计算机科学期刊和会议论文集的开放书目信息。

DBLP 最初于 1993 年在特里尔大学创建，现在由 Schloss Dagstuhl 运营和进一步开发。DBLP 的使命是通过免费提供高质量的书目元数据和出版物电子版链接来支持计算机科学研究人员的日常工作。

截至 2024 年 1 月，dblp 索引了超过 700 万份出版物，由超过 340 万作者出版。为此，dblp 索引了大约 55,000 多种期刊、超过 55,000 场会议和研讨会论文集以及超过 140,000 部专著。

DBLP 数据集每天会进行更新维护，提供最新的数据下载，同时会存储每个月稳定的数据记录方便进行科学研究。此外，DBLP 还提供了简单的 API 接口来获取少量的数据集。

2. DBLP 数据集

考虑到实验可重复性，本次实验采用 2024-10-01 发布的数据集：**dblp-2024-10-01.xml.gz**

2.1 来源

https://dblp.uni-trier.de/xml/

2.2 存储格式

使用 XML 格式存储

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE dblp SYSTEM "dblp.dtd">
<dblp>
  record 1
  ...
  record n
</dblp>
```

2.3 数据记录描述

字段（key）	描述（Description）
article key	主键
author	论文作者
title	论文标题
pages	页码
year	年份
volume	卷号
number	编号
ee	网页
url	来源

2.4 数据记录样例

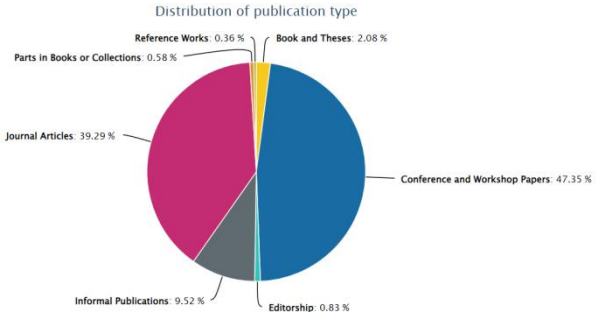
```
<article key="journals/cacm/Szalay08"
      mdate="2008-11-03">
  <author>Alexander S. Szalay</author>
  <title>Jim Gray, astronomer.</title>
  <pages>58-65</pages>
  <year>2008</year>
  <volume>51</volume>
  <journal>Commun. ACM</journal>
  <number>11</number>
  <ee>http://doi.acm.org/10.1145/
      1400214.1400231</ee>
  <url>db/journals/cacm/
      cacm51.html#Szalay08</url>
</article>
```

3. DBLP 数据集统计信息

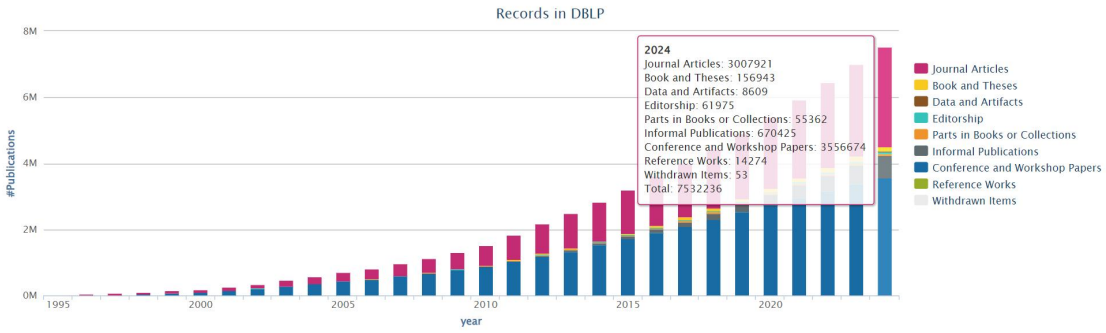
3.1 数据量

出版物类型（Publication Type）	数据量（Value）
Book and Theses	156366
Conference and Workshop Papers	3562625
Editorship	62275
Informal Publications	716314
Journal Articles	2955459
Parts in Books or Collections	43287
Reference Works	27365
Total	7523691

3.2 出版物类型分布



3.3 数据量年度统计



4. References

https://dblp.uni-trier.de/
https://dblp.uni-trier.de/xml/
https://dblp.uni-trier.de/xml/docu/dblpxml.pdf
https://dblp.uni-trier.de/xml/docu/dblpxmlreq.pdf
https://pypi.org/project/dblp-sax-parser/