课程项目

1、项目介绍

在本项目中，你需要实现一个端到端的知识图谱分析系统，为图数据实现包括数据采集、数据转换、数据存储、数据分析和可视化的功能。所需要实现的模块大致如下：

1）、ETL: 该模块应当完成原始数据的读取，转换，并写入存储系统（Storage System）。

2）、Storage System: 该模块应当完成数据的存储，并为 Query Server 提供高性能的查询服务。在存储的过程中，应当考虑数据预处理、数据建模、索引等建立的合理性，以提高效率和系统稳定性。

3）、Query Server: 该模块为前端的数据可视化服务。应当为前端需要实现的针对知识图谱的检索提供实现。

4）、UI: 该模块为前端的数据可视化。需要开发针对知识图谱的查询（需要实现的查询在 3、评分点处详细介绍），并对查询的结果可视化为一张图，以实现对结果集的快速访问。

2、数据集

本项目中的知识图谱来源于学术论文关系数据集。你需要访问以下网址：https://www.aminer.cn/aminernetwork。数据集中包括论文、作者和联合撰写论文的信息。你需要让自己熟悉这个数据集。

Extraction and Mining of Academic Social Networks

Overview

Data Description

References

Overview

This dataset is designed for research purpose only.

The content of this data includes paper information, paper citation, author information and author collaboration. **2,092,356** papers and **8,024,869** citations between them are saved in the file AMiner-Paper.rar; **1,712,433** authors are saved in the file AMiner-Author.zip and **4,258,615** collaboration relationships are saved in the file AMiner-Coauthor.zip.

FileName	Node	*Number*	*Size*
AMiner-Paper.rar [download from mirror site]	Paper Citation	2,092,356 8,024,869	509 MB
AMiner-Author.zip [download from mirror site]	Author	1,712,433	167 MB
AMiner-Coauthor.zip [download from mirror site]	Collaboration	4,258,615	31.5 MB

Supplement: The relaionship between author id and paper id AMiner-Author2Paper.zip. The 1st column is index, the 2nd colum is auhor id, the 3rd column is paper id, the 4th column is author's position.

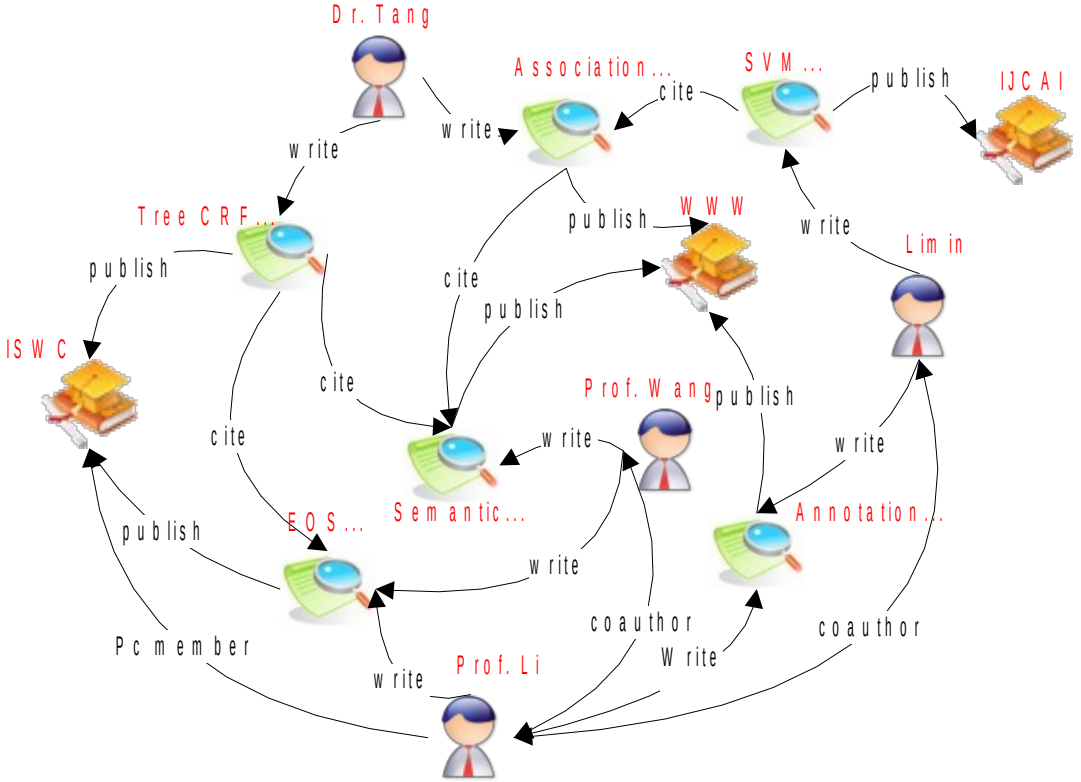
数据集分为三个数据，论文、作者和协作的关系，论文数据中包括论文 ID，论文标题、作者名称、所在单位、出版时间、出版物、摘要。作者数据中包括作者 ID，作者名称、所在单位、发表论文数量、总引用数、H 指数、研究兴趣等。协作数据包括协作作者 ID，协

作次数。你需要在 Storage System 中选择一个图数据库进行数据的存储，这将会把实体（至少包括但不限于 研究者-研究主题-研究机构-学术刊物/会议 等）和他们之间的关系进行关联，从而为前端需要实现的查询服务提供可能。

3、评分点

在本项目中你需要实现端到端的数据获取、清洗、存储、查询和可视化的数据流，其中在每一个模块中你可以发挥自己的能力对系统设计的关键点进行思考和优化。而每一个优化点都将使得评分有所提高。其中详细的评分规则如下：

1）、基本功能：你需要至少实现的功能为：数据集（或一个数据子集）的读取、导入，完成包括但不限于 研究者-研究主题-研究机构-学术刊物/会议 的关系存储，并完成至少一个基本查询。对查询的结果，需要以图的形式进行绘制并在前端展示，如下图所示：



2）、基本功能查询：你需要至少实现以下两个基本查询：

2.1、输入一个实体（如某作者 A 或研究主题词 A），查询其关联的所有关系和关联实体；

2.2、输入两个实体（如作者 A 和作者 B），查询其可能存在的多跳关系。其中多跳关系定义为，通过多条边链式的连接在一起（如合作论文等）。

3）、基本业务查询：你需要至少实现以下两个基本查询，并以交互的方式展现结果：

3.1、查询在某个领域中的关键作者和单位是什么，为什么？

3.2、查询在某个领域中的关键期刊/会议是什么，为什么？

4）、加分项：你可以考虑下面的加分项，并通过自己的思考对系统进行优化（在以下 4 个大类中选择 2 个即可）

4.1、ETL：如何支持更大规模的数据？如何更好地支持数据更新，而不是一次性导入？如何处理动态的数据增加和图的变化问题（例如会议/期刊的主题可能随着时间变化而变化）？

4.2、Storage System：如何更好地对图数据进行建模？如何提高查询性能？如何增强系统的可扩展性，是否是分布式系统？

4.3、Query Server：如何提高查询的性能？是否可以采用缓存？

4.4、可视化：如何让用户更好地查看检索的结果？是否支持沿着图进行进一步扩展？如何展示随着时间的变化？