****

**研究生学位论文开题报告**

题 目： 基于Neo4j的图数据可视化展示系统

学 院： 计算机学院

专 业： 软件工程

学 号： 2019282110211

姓 名： 陆咏

导师姓名： 祝园园

导师职称： 副教授

2020年11月23日

目录

[1 选题背景及意义 1](#_Toc56968043)

[2 国内外关于该课题的研究现状及趋势 1](#_Toc56968044)

[2.1 图数据库 1](#_Toc56968045)

[2.2 图数据可视化应用 3](#_Toc56968046)

[2.2.1 VizCS社区搜索可视化系统 3](#_Toc56968047)

[2.2.2 AMiner科技情报大数据挖掘与服务系统平台 4](#_Toc56968048)

[2.2.3 瑞士联合银行集团数据沿袭和管理平台 5](#_Toc56968049)

[3 研究计划 6](#_Toc56968050)

[3.1 研究目标 6](#_Toc56968051)

[3.2 拟突破的难题或攻克的难关题 6](#_Toc56968052)

[3.3 自己的创新或特色 6](#_Toc56968053)

[3.4 实验方案或写作计划 6](#_Toc56968054)

[3.4.1拟采用的研究思路（方法、技术路线、可行性论证等） 6](#_Toc56968055)

[3.4.2 研究工作安排及进度 7](#_Toc56968056)

[参考文献 8](#_Toc56968057)

# 1 选题背景及意义

近年来，随着社交网络、零售电商、金融风控、生命科学等行业的快速发展，越来越多的数据以庞大而复杂的关系网的形式出现，需要处理的数据之间的关系随数据量呈几何级数增长，用户理解数据的难度也越来越高。因此，亟需一个可视化系统，帮助用户查看、筛选和处理图数据。

同时，随着数据量的增多和关联深度的增加，传统关系型数据库受制于检索时需要多个表之间连接操作，数据写入时也需考虑外键约束，从而导致较大的额外开销，产生严重的性能问题。因此，图数据可视化系统可以采用图数据库，借助其固有的数据索引结构和灵活的数据模型，提高数据查询和分析的速度，

本课题旨在构建一个通用的图数据可视化展示系统，借助Neo4j图数据库，实现大型图的图数据请求与显示、顶点与边的查询、图数据上传、图数据算法的调用和改进等功能，以满足图数据可视化的各类需求。

# 2 国内外关于该课题的研究现状及趋势

## 2.1 图数据库

当前市场上的图数据库种类繁多，根据2020年10月DB-Engine的图数据库排名，当前市场占有率较高的图数据库有Neo4j、JanusGraph、Dgraph、Giraph、TigerGraph等，如图2-1[1]、图2-2[2]所示。

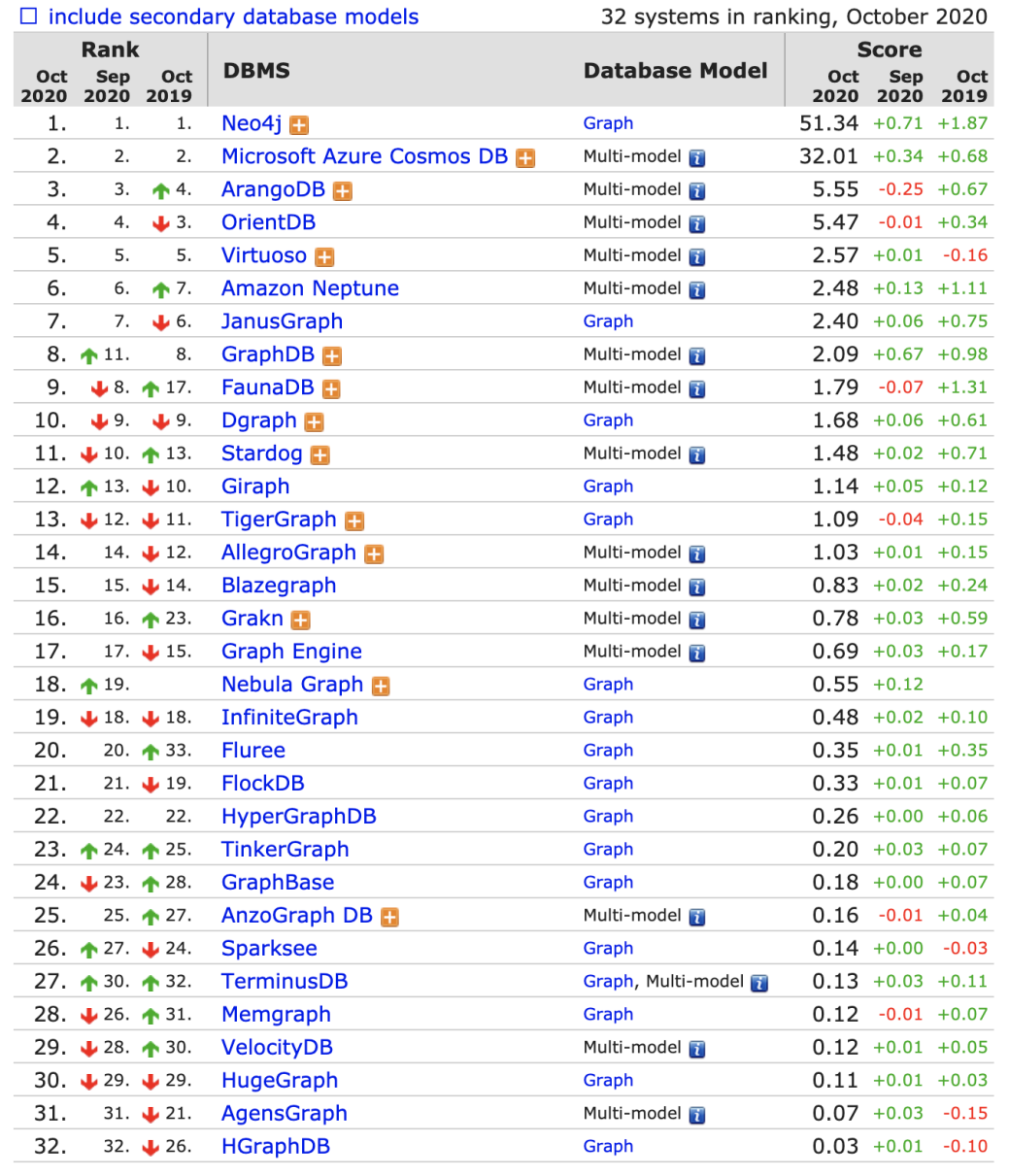


图2-1 2020年10月DB-Engine图数据库排名

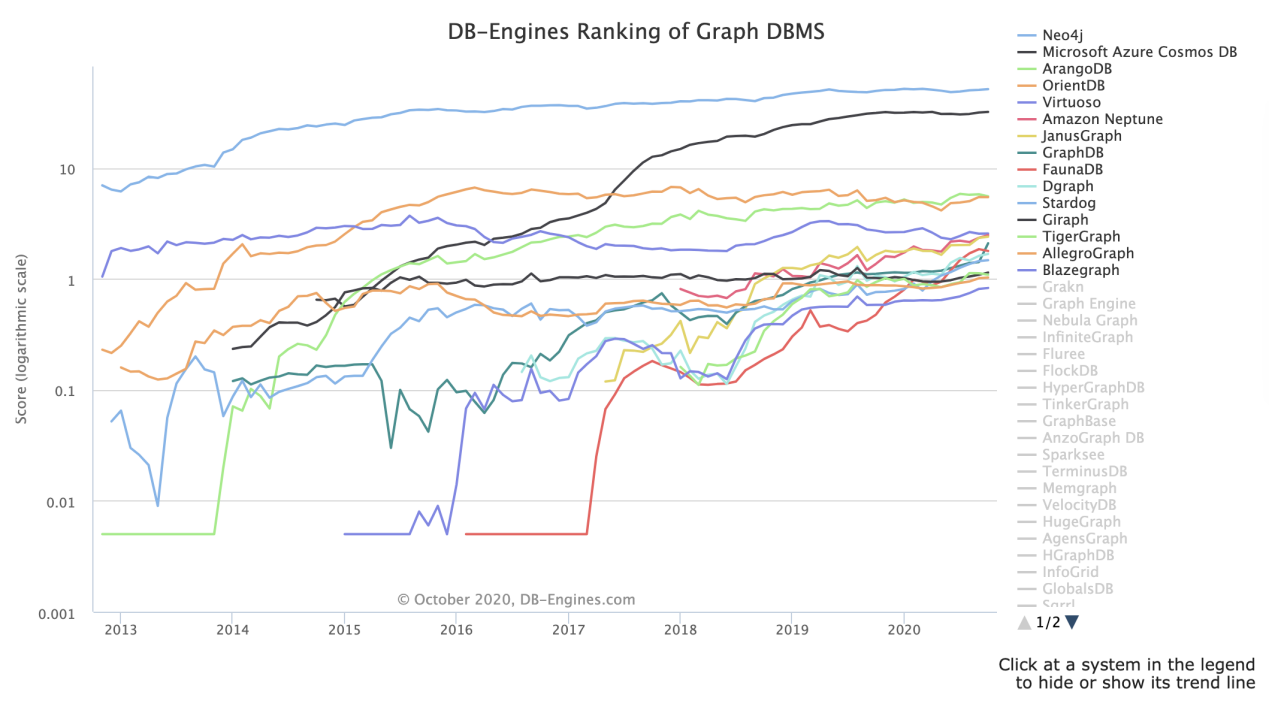


图2-2 2020年10月DB-Engine图数据库排名折线图

Neo4j长期稳居图数据库市场占有率第一，社群活跃，服务稳定，功能强大，主要有以下几个优点：(1)使用原生图（Native Graph）存储和处理数据，提供最优化的关系遍历执行效率，比关系数据库的表连接快上千倍；(2)完全ACID兼容、保证数据一致性，因此同样适用于事务型（OLTP）和分析型（OLAP）应用；(3)基于因果集群(Causal Clustering)的分布式数据库，提供高可用性、故障切换、数据冗余和可扩展的吞吐量；(4)丰富的语言驱动，包括Java, JavaScript, Python, .Net, GO, C/C++, R等，各类工具齐全[3]。

JanusGraph是Linux基金会下的一个开源分布式图数据库，采用第三方存储（Apache Cassandra、Apache HBase、Bigtable、Berkeley DB等）作为底层存储；通过与大数据平台（Apache Spark，Apache Giraph，Apache Hadoop等）集成，支持全局图数据的分析、报告和ETL；通过外部索引存储（Elasticsearch，Solr，Lucene等）支持地理、数字范围和全文搜索。JanusGraph可以配合多种工具一起使用，具有很高的扩展性，但是由于需要开发者自行组合，使用时的技术门槛较高，图数据可视化也需要借助外部系统才能实现。

Dgraph是基于GO语言开发的开源分布式图数据库，使用原生图存储数据，性能优异；可以自动组成集群，运维部署非常简单；原生支持全文检索、正则表达式、地理位置检索、可视化等功能。但是存在不支持多重边、与大数据生态兼容不足、一个集群只支持一个图等问题。

Giraph是以Apache Hadoop为基础开发的迭代图处理框架，由Pregel模型[4]改进而来，其系统架构和计算模型与Pregel模型一致，主要用于大型图的分布式计算问题，例如，目前Facebook使用Giraph分析由用户及其关系形成的社交网络。

TigerGraph使用原生并行图系统，将图表示为一个计算模型，实现了大规模并行和快速计算，可以进行实时图分析和深度关联多步分析任务。

本课题的目的是构建一个通用的图数据可视化展示系统，实现大型图的图数据请求与显示、顶点与边的查询、图数据上传、扩展图算法的调用等功能，以图的遍历、查询、图算法调用和图的可视化展示为主，图数据的更新较少。综合考虑以上几种图数据库的特点，最终决定选择使用Neo4j图数据库。

## 2.2 图数据可视化应用

当前，不管是学术界还是工业界，图数据可视化都有着广泛的应用需求。

### 2.2.1 VizCS社区搜索可视化系统

VizCS （Visualization system for Community Search）是由香港中文大学与香港浸会大学的学者于2018年搭建的社区搜索可视化系统[5]，使用TCP（Triangle Connectivity Preserved）索引维护动态图，支持运行k-truss、k-core、quasi-clique、稠密子图等社区搜索算法，并且用户可以通过上传文件更新图数据，同时借助JavaScript库D3.js对运行结果进行可视化展示，如图2-3所示。

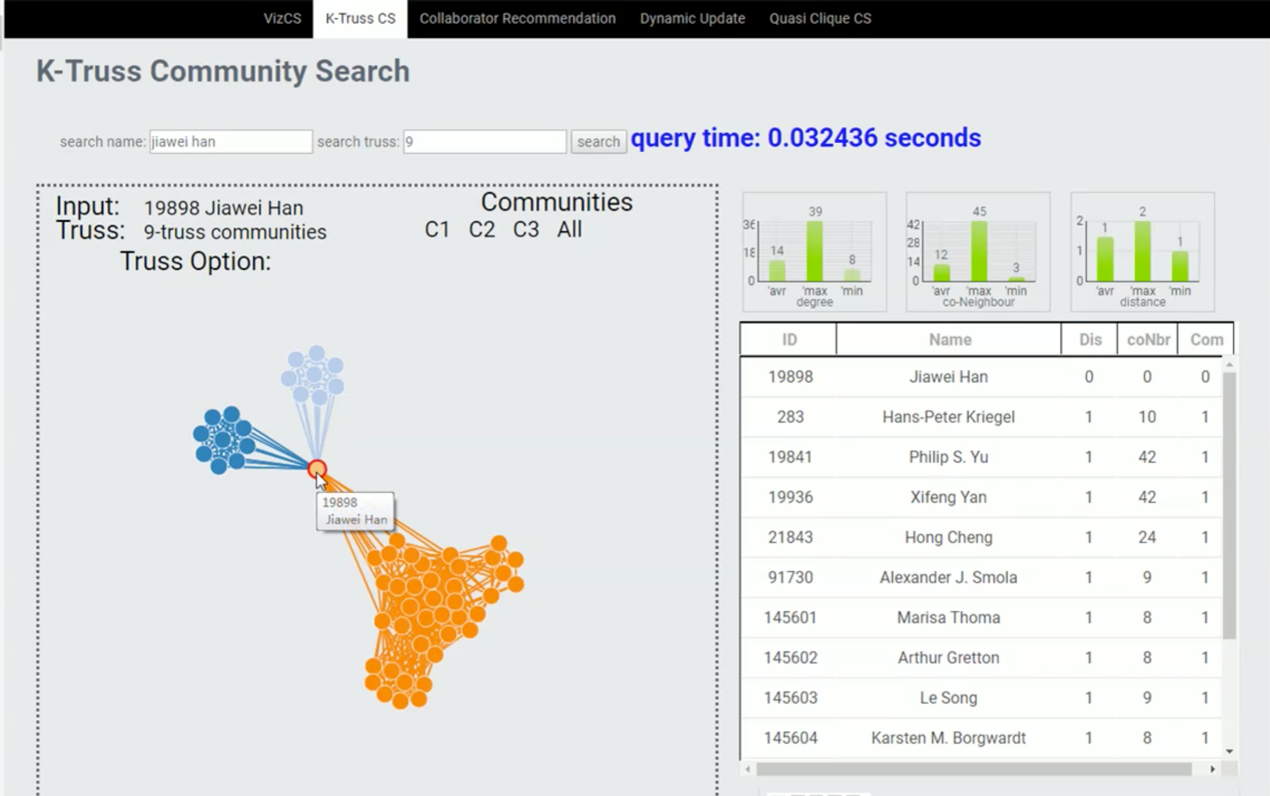


图2-3 VizCS k-truss社区搜索算法的可视化展示

### 2.2.2 AMiner科技情报大数据挖掘与服务系统平台

AMiner是由清华大学唐杰教授率领团队建立的科技情报大数据挖掘与服务系统平台，利用数据挖掘和社会网络分析与挖掘技术，提供研究者语义信息抽取、面向话题的专家搜索、权威机构搜索、话题发现和趋势分析、基于话题的社会影响力分析、研究者社会网络关系识别、即时社会关系图搜索、研究者能力图谱、审稿人推荐在内的众多功能[6]，提供了溯源树、AI发展河流图、机器学习趋势图等多种图数据可视化实现，具有很高的学术价值和实用性。AI发展河流图[7]见图2-4。

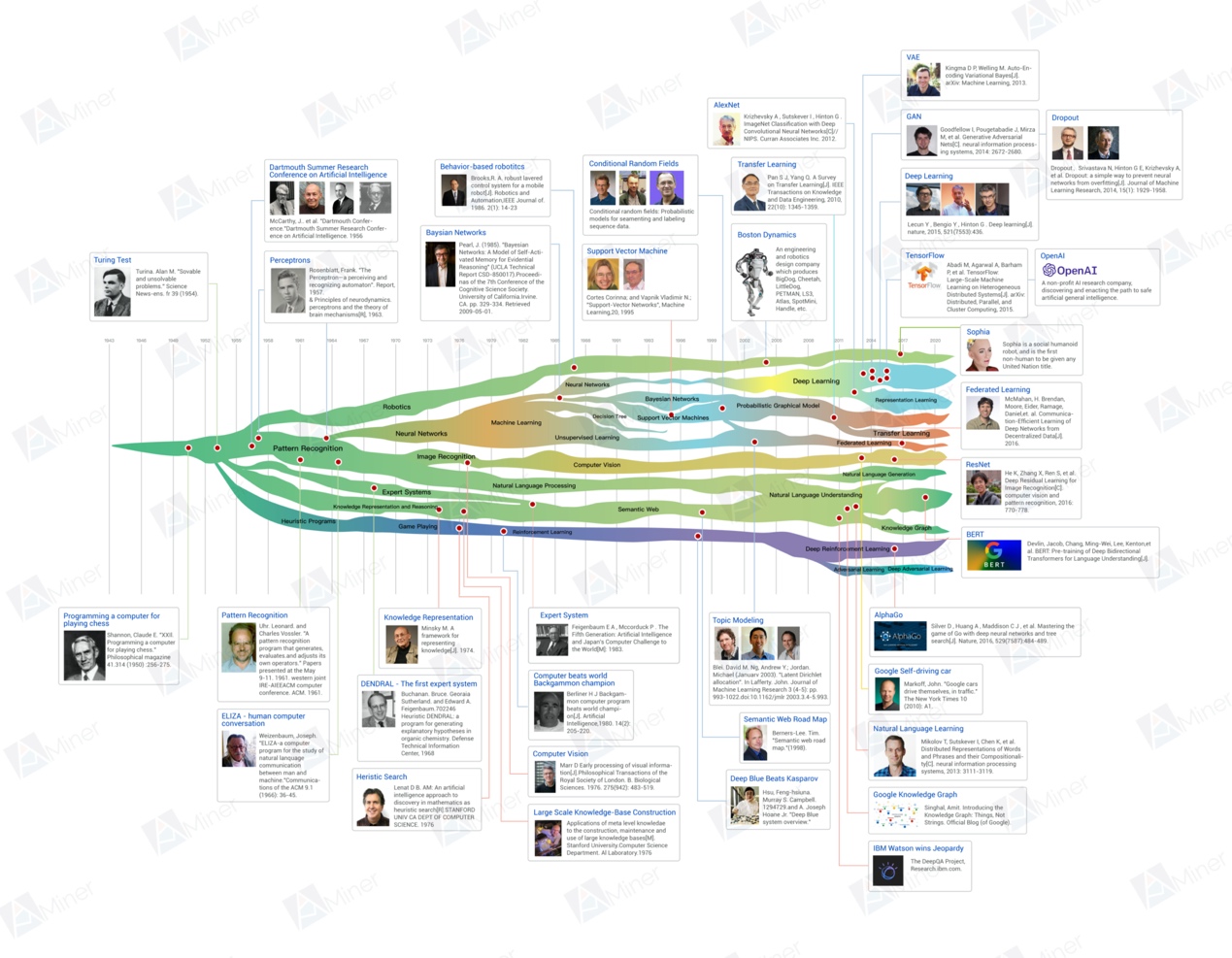


图2-4 AI发展河流图

### 2.2.3 瑞士联合银行集团数据沿袭和管理平台

2007年全球金融危机爆发后，巴塞尔银行监管委员会发布新规，要求银行提供透明的风险报告数据流，这需要大规模的数据管理和详细的数据沿袭。瑞士联合银行集团作为全球最大的金融机构之一，面对大量亟待处理的金融数据，使用了Neo4j图数据库构建了其数据沿袭和管理平台。

数据沿袭跨越数十个层次的实体和依赖关系，传统关系型数据库难以处理，而Neo4j可以以图形格式存储全银行所有行为的所有数据，进而借助JavaScript库D3.js将数据转化为直观的沿袭图。而且，Neo4j不太依赖预定义的模式或索引，能够应对业务分析需求的不断变化，并支持将整个沿袭数据导出到Excel，兼顾了实时响应时间和灵活性。

基于可视化图数据的数据沿袭和管理平台帮助瑞士联合银行集团在时间紧迫、资源有限的情况下满足了监管的要求[8]。

# 3 研究计划

## 3.1 研究目标

本课题的研究目标是构建一个通用的图数据可视化展示系统，以满足图数据可视化的各类需求。主要包括图数据请求与显示、顶点与边的查询、图数据上传、图数据算法的调用和改进等模块。

图数据请求与显示主要负责从数据库获取图数据并进行可视化的展示；顶点与边的查询是对已可视化的图数据的顶点或边进行更进一步的具体查询和展示；图数据上传实现了自主上传图数据，并对其进行展示；图数据算法的调用和改进支持在图上运行外部图数据算法和Neo4j数据库内置的图数据算法，并展示其运行结果。

## 3.2 拟突破的难题或攻克的难关题

本课题主要面临两个难点。

第一，在图数据算法的调用和改进模块，除了使用Neo4j数据库的API调用其内置的图数据算法之外，还有必要支持其他经典算法，如何高效地调用外部的图数据算法是课题的首个挑战。

第二，由于大型图的数据量巨大，系统进行可视化显示时的性能和系统能够支持的数据量都面临着考验。

## 3.3 自己的创新或特色

本课题的主要有三个创新之处。

第一，本课题实现的图数据可视化展示系统，在其可以使用的算法方面，具有极高的扩展性，能够调用任何封装好的外部图数据算法和Neo4j数据库内置的所有算法。

第二，本课题计划在调用部分外部图数据算法时，对算法做出针对性的改进。

第三，本系统将支持自主上传图数据，以最大程度地满足图可视化的需求。

## 3.4 实验方案或写作计划

### 3.4.1拟采用的研究思路（方法、技术路线、可行性论证等）

本课题首先借助C++语言，处理里耶秦简原始数据，并将其导入Neo4j数据库，作为初始数据；然后，通过Vue框架，构建一个可视化的页面；再使用neovis.js JavaScript库，在页面获取Neo4j数据库中的图数据并进行可视化的展示，再进一步实现对该图的顶点与边的查询。在此基础上，实现图数据上传模块和图数据算法的调用和改进模块。图数据上传模块负责实现文件上传功能，将文件导入Neo4j数据库，并在页面展示图数据。图数据算法的调用和改进模块能够调用外部图数据算法和Neo4j数据库内置的图数据算法，在图上运行算法并在页面对运行结果进行可视化的展示。

本课题主要借助Neo4j数据库、Vue框架、JavaScript语言和C++语言实现。Neo4j是当前使用最为广泛的图数据库，功能强大、服务稳定、社群活跃，能够高效地处理图数据，并内置了多项图数据算法。Vue是一套用于构建用户界面的渐进式框架，易用、灵活、高效，具备丰富的生态系统，借助它可以轻松地完成组件化开发。

本课题期望构建一个通用的图数据可视化展示系统，以满足图数据可视化的各类需求。

### 3.4.2 研究工作安排及进度

2020.10.23—2020.10.30 接受学位论文任务，协商选题；

2020.10.31—2020.11.18 学位论文前期调研、文献资料查阅等准备工作；

2020.11.19—2020.11.25 完成学位论文开题报告；

2020.11.26—2020.12.15 文献阅读及整理，调研；

2020.12.16—2021.01.10 设计项目算法及项目整体框架；

2021.01.11—2021.01.23 完成初始数据和数据库准备工作；

2021.01.24—2021.02.20 实现可视化页面，实现顶点与边的查询模块；

2021.02.21—2021.03.20 实现图数据上传和图数据算法的调用和改进模块；

2021.03.21—2021.04.10 进一步优化相关图数据算法；

2021.04.11—2021.04.30 撰写毕业设计论文初稿；

2021.05.01—2021.05.09 修改初稿，完成毕业设计论文终稿；

2021.05.10—2021.05.14 完成毕业设计成果验收和提交；

2021.05.15—2021.05.21 整理材料，修改论文，完成毕业设计答辩。

# 参考文献

[1] DB-Engines. DB-Engines Ranking of Graph DBMS[DB/OL]. <https://db-engines.com/en/ranking/graph+dbms>, 2020-11-05.

[2] DB-Engines. DB-Engines Ranking - Trend of Graph DBMS Popularity [DB/OL]. <https://db-engines.com/en/ranking_trend/graph+dbms>, 2020-11-05.

[3] Neo4j图数据库. Neo4j作为图数据库有哪些特色和优势？[EB/OL]. <https://zhuanlan.zhihu.com/p/125741905>, 2020-11-05.

[4] Grzegorz Malewicz, Matthew H. Austern, Aart J.C Bik, et al. Pregel: a system for large-scale graph processing[A]. Proceedings of the 2010 ACM SIGMOD International Conference on Management of data[C]. New York: ACM, 2010, pp.135-146.

[5] Jiang Y, Huang X, Cheng H, et al. VizCS: Online Searching and Visualizing Communities in Dynamic Graphs[A]. 2018 IEEE 34th International Conference on Data Engineering (ICDE)[C]. Paris: IEEE, 2018, pp.1585-1588.

[6] AMiner. AMiner – AI赋能科技情报挖掘[DB/OL]. <https://www.aminer.cn>, 2020-11-10.

[7] Jie Tang, Jing Zhang, Limin Yao, et al. ArnetMiner: Extraction and Mining of Academic Social Networks[A]. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New York: ACM, 2008, pp.990-998.

[8] Neo4j. Data Lineage Tool Improves Risk Management, Drives Compliance[EB/OL]. <https://neo4j.com/case-studies/ubs-case-study>, 2020-11-15.