|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 武汉大学logo  **开 题 报 告** | | | | |
|  | | | | |
| 基于Neo4j的图数据处理系统 | | | | |
|  | 研究生姓名 | ： | 陈小龙 |  |
| 学号 | ： | 2019282110194 |
|  |  |  |
|  | | | | |

目 录

[1 选题及背景意义 3](#_Toc55292593)

[1.1 关于目录修改后的排版技巧 3](#_Toc55292594)

[1.1.1 三级标题 3](#_Toc55292595)

[2国内外关于该课题的研究现状及趋势 3](#_Toc55292596)

[2.1 图形数据库 3](#_Toc55292597)

[2.2 图可视化应用 4](#_Toc55292598)

[3 研究计划 4](#_Toc55292599)

[3.1 研究目标 4](#_Toc55292600)

[3.2 拟突破的难题或攻克的难关 4](#_Toc55292601)

[3.3 实验方案或写作计划 4](#_Toc55292602)

[3.3 自己的创新及特色 4](#_Toc55292603)

[3.3.1 拟采用的研究思路（方法、技术路线、可行性论证等） 4](#_Toc55292604)

[参考文献 6](#_Toc55292605)

# 1 选题及背景意义

## 背景

随着物联网、社交网络、分布式计算等技术的高速发展，互联网技术不断的改变着每一个人的生活。同时，计算能力、存储空间、网络带宽的发展更是日新月异。这就导致了人类各行各业积累数据的速度大大加快，积累的数据量更是以PB计。在可以预见的未来，这一趋势还将加快。

在数据量暴增的情况下，传统的关系型数据库不论是容量还是单机性能已经越来越难以满足数据存储的要求；更重要的是互联网时代产生的数据中很大一部分是非结构化数据，传统关系型数据库的存储模式难以满足大数据存储的扩展性和可用性。于是，非关系型数据库应运而生。

非关系型数据库主要有如下特点：模式自由、满足最终一致性(非ACID)、分布式、水平可扩展等。NoSql数据库可以按数据模型分成4类：键值存储库（如：Redis）、列族数据库（如：Hbase）、文档数据库（如：MongoDB）、图形数据库（如：Neo4j）。其中图形数据库专门用来处理高度关联的数据，适用于推荐系统、社交网络、模式识别等领域。图形数据库具有高度的灵活性，并且支持复杂的图形算法，方便构建复杂的关系图。

在当今业界主流的图数据库中，Neo4j的社区最为成熟，应用也非常广泛。另外，Neo4j内置的算法库包含常用的图的算法，包括路径搜索算法，如：Shortest Path、Single Source Shortest Path；中心性算法，如：PageRank、Betweenness Centrality；社区发现算法，如：Strongly Connected Components、Label Propagation。上述算法在实际生产中有着广泛的应用，但与此同时，Neo4j支持的算法也存在着较大的局限。特别地，对于社区发现算法而言，近年来有了新的发展，很多优秀的思想和改进算法被提出；相较于现有的算法，新的算法在某些方面有着更强的表现，遗憾的是这些新的算法却不被Neo4j支持。除此之外，在数据可视化方面Neo4j的功能仍然比较单一，不够灵活，这也是Neo4j的一个较为明显的缺点。

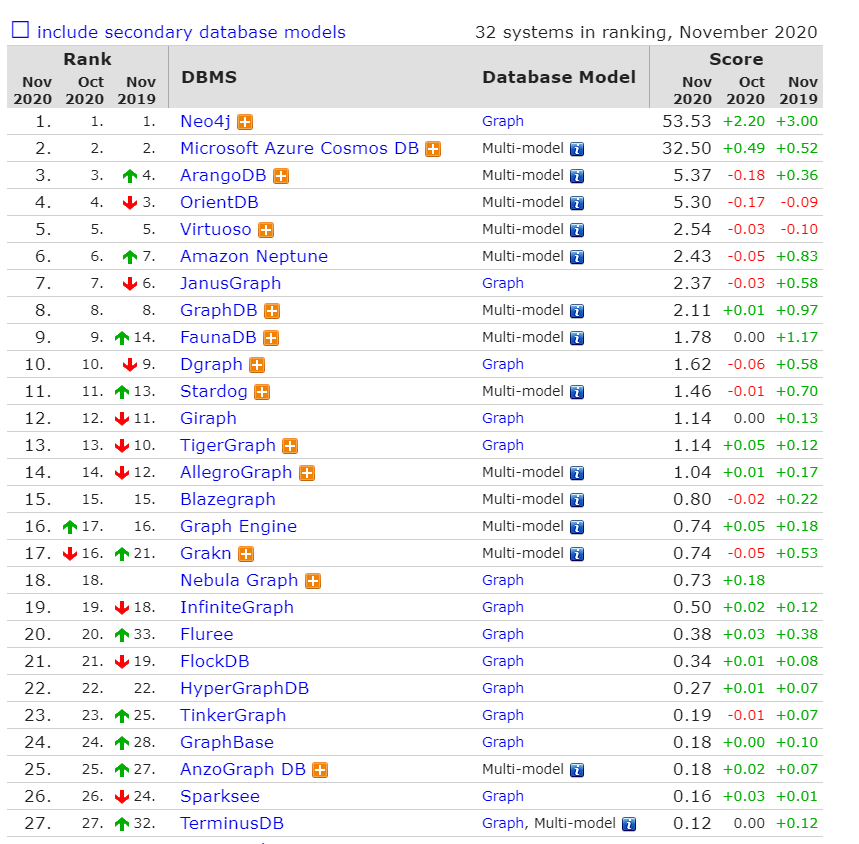
本课题旨在，基于Neo4j实现一个图数据处理系统，分为前端数据可视化、后端数据处理、图算法扩展。该系统的最终目标是扩展Neo4j的算法，并使用扩展的算法处理数据，最终在前端进行可视化展示。若上述目标实现，将提供一个图数据从存储到计算再到展示的一整套解决方案，并且在一定程度上解决Neo4j在数据可视化和算法扩展性方面存在的问题。

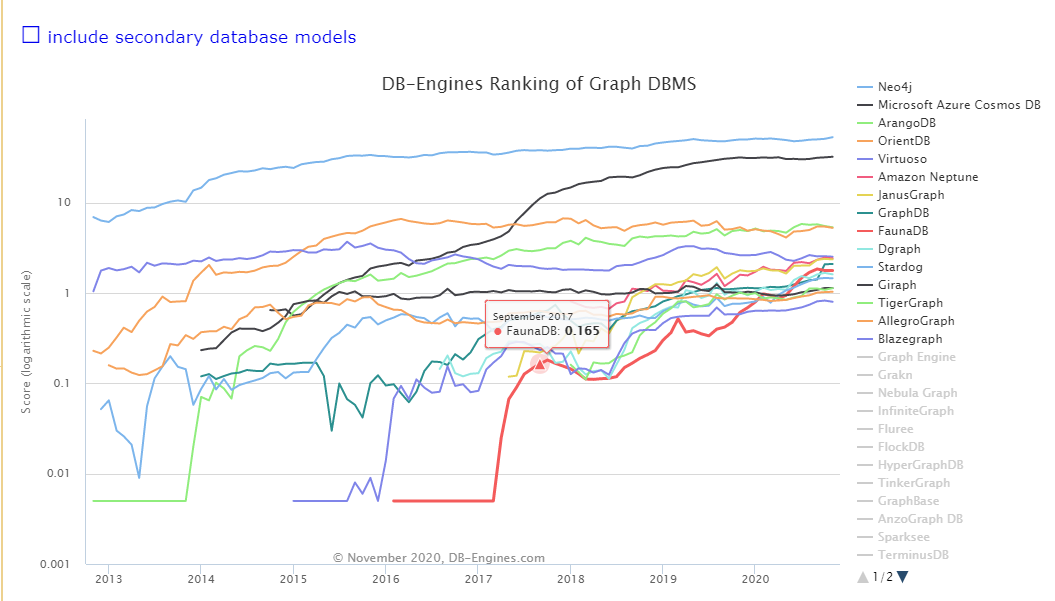
# 2国内外关于该课题的研究现状及趋势

## 2.1 图形数据库的发展现状

2.1.1 图数据库排名

根据db-engines.com的最新排名显示，当前在业界使用的主流图数据库中Neo4j占据榜首位置，市场占有率呈持续上趋势。





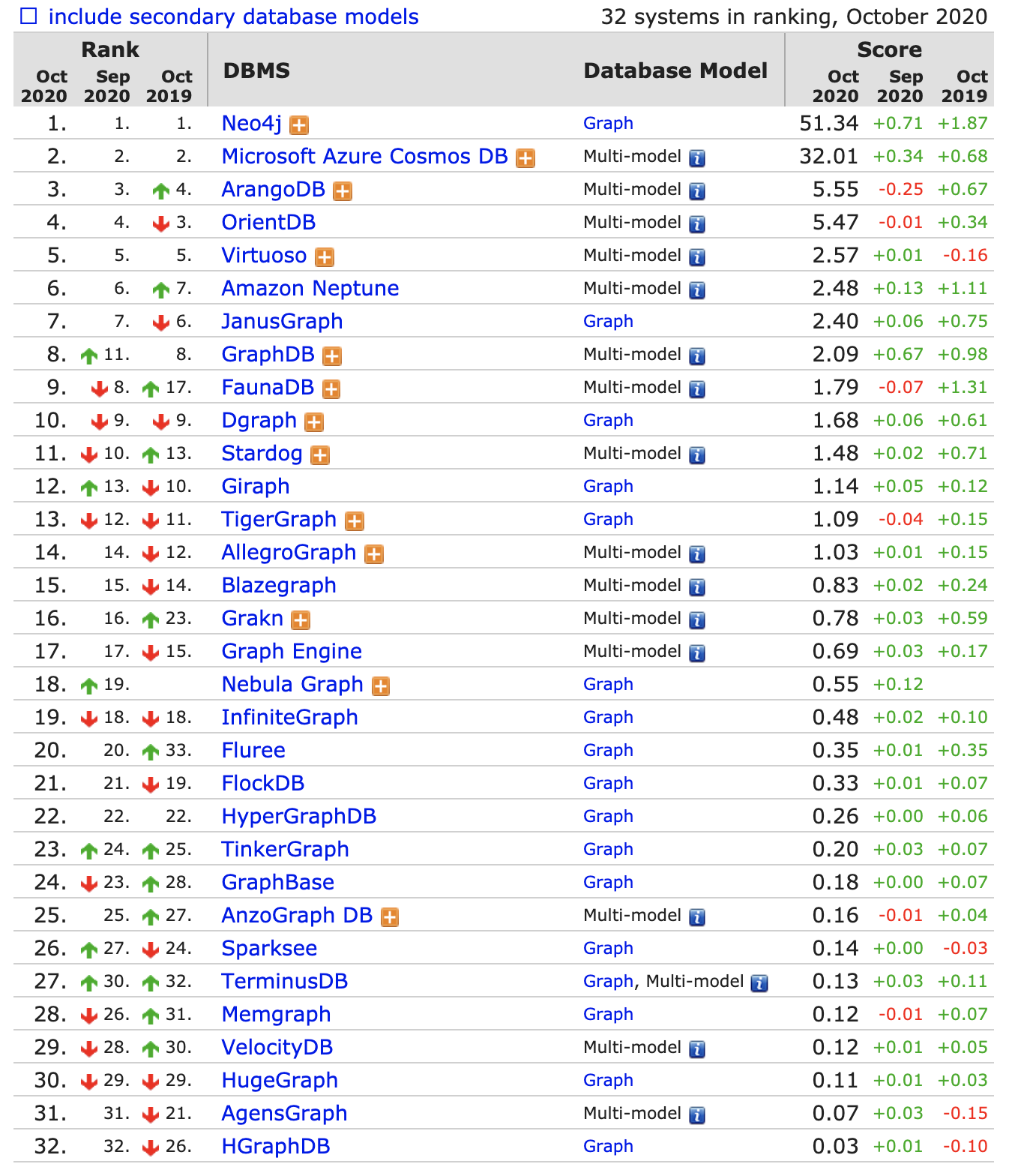
2.1.1 主流图数据库比较

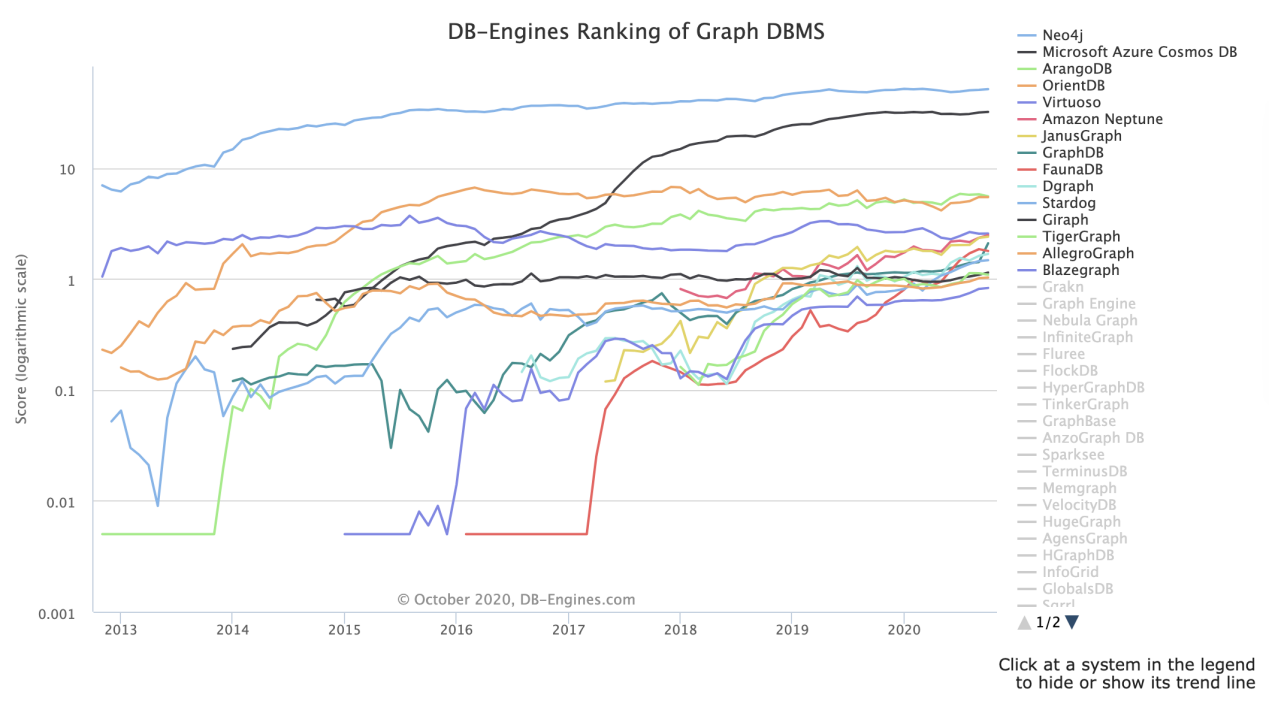
Neo4j长期处于图数据库的领头地位，分为企业版和社区版；其中社区版开源，只支持单机不支持分布式部署。Neo4j基于Java实现，兼容ACID特性，也支持其他编程语言，如Ruby和Python。Neo4J使用原生的图存储，以高度自由且规范的方式管理和存储数据。Neo4J可以以每秒一百万条的速度提供结果，因为数据中的链接部分或实体在物理上是已经相互连接的。

Azure Cosmos DB 是 Microsoft 提供的全球分布式**多**模型数据库服务，可通过 Cosmos DB 跨任意数量的全球 Azure 区域弹性且独立地缩放吞吐量和存储。Azure Cosmos DB通过自动缩放吞吐量、计算和存储来提供多主数据库功能。

OrientDB是第二代分布式图数据库,以混合数据模型为特点，它支持复制和分片，并以Apache2许可证提供开放源代码。ORIENTDB可以使用SQL作为查询语言之，OR使用身份验证、密码和静态数据加密等方式为所有机密数据提供安全保护。

根据db-engines.com https://db-engines.com/en/ranking/graph+dbms





## 2.2 Neo4j应用案例

2.2.1 可视化案例

AMiner https://www.aminer.cn/

AMiner技术细节 https://blog.csdn.net/u010161379/article/details/50760788

Jiang,Y.,Huang, X.,Cheng, H.,Yu, J. X.:VizCS: online searching and visualizing communities in dynamic graphs. In: ICDE, pp. 1585–1588 (2018)

<https://www.comp.hkbu.edu.hk/~xinhuang/publications/video/VizCS.mp4>

2.2.2 NBC新闻

2016年美国大选期间，俄罗斯水军渗透了美国的网络空间。NBC新闻在调查发现中遇到了两个挑战，一个是检测水军的推特账号，另一个是通过数据分析出水军账号的行为模式。他们使用Neo4j详细调查了成百上千的推特账号，并以此发现了俄罗斯水军的行为策略。

NBC新闻是全世界新闻媒体行业的的领导者，它的一个团队试着弄清克里姆林宫支持的水军如何利用推特影响美国的政治。但是检测水军到底如何影响2016年美国总统大选面临很大的困难，一方面由于网络战的隐蔽性，例如匿名信息、虚假身份等；另一方面社交媒体数据巨大的体量也进一步增加了分析的难度。

2017年11月，美国众议院常设特选委员会情报局发布了2,752个克里姆林宫支持的Twitter水军帐户。他们宣称，俄罗斯机构假扮美国公民、新闻组织、政治团体，并且设置假的账户传播错误信息、煽动分裂。随后推特冻结并且删除了这些账号。

如何恢复并且分析这些数据？水军网络如何运作？水军如何影响美国的社交网络并且改变公众的观念？这些问题激发了公众的兴趣，如果没有合适的工具去恢复和分析数据，将很难找到答案。

最后NBC选择了Neo4j，Neo4j完美的适用于处理大量的图数据，他们的第一个任务时去尽可能的恢复丢失的数据。研究人员从备份数据恢复数据，并且独立分组以便分析选举期间的推特。这个数据库由454个推特账户的202973条推特构成；这些数据只是原始数据的一部分，但是足以让研究人员找出水军账号是如何运作的。

图包含了实体之间的关系，实体包括：推特、用户、推特标签、链接等。图的算法基于实体之间的连接关系，揭示了实体在网络中的重要性。通过社区检测算法找出频繁与他人互动沟通的用户，这些用户有影响力并且会放大其他水军的影响；通过pagerank算法识别出最有影响力的账号。

他们发现水军网络只有很小的核心来产生最初的推文，其余的是大量的转发；只有25%的推文是原创的，其余的都是转发。水军利用公共的标签和回复有名的账号来扩大他们的影响力，增加关注度。

研究人员发现了三类水军账户，一类编造成典型的美国人，另一类模仿新闻网站，第三类会假装是某个政治组织。研究人员称，所有这些账号都被俄罗斯的机构控制，Neo4j揭示了这些虚假账户在网络中如何协作。

NBC用Neo4j构建了一个包含20万推特的数据集，通过使用Neo4j分析，他们揭示了俄罗斯如何模仿美国人，进而影响正常人，并且吸引大量粉丝，最终影响美国政治。通过用Neo4j构建“关系引擎”，政府或者社交平台就能在闹事的水军产生影响之前采取行动维护社交秩序。

# 3 研究计划

## 3.1 研究目标

实现了什么功能/模块

图展示、顶点/边查询、算法调用、图数据上传及展示、...

## 3.2 拟突破的难题或攻克的难关

Challenges 主要解决的问题

数据量、算法封装及调用、显示速度/性能、.... + 某一算法改进？

## 3.3 实验方案或写作计划

## 3.3 自己的创新及特色

### 3.3.1 拟采用的研究思路（方法、技术路线、可行性论证等）

Vue + neo4j + java + ...

前端模块：vue、数据展示、请求&调用

后端模块：数据库部署、数据建模、建立OGM对象图映射、对节点的增删改查（批量操作）

其他：数据量、算法封装及调用、显示速度/性能、.... + 某一算法改进？

3.3.2 研究工作安排及进度

# 参考文献

[1] 李遥.一键排版人工智能引擎研究:[硕士学位论文].上海:华西理工大学，2015.

[2] 田彬.一键排版PaperYes的理论基础.北京:机械出版社，2007.11.

[3] 陈炜煜，李建彬.毕业论文格式要求.特区经济，2013，（2）:221-222.

[4] 樊伟.论项目时间管理.现代商贸工业，2010，（5）:46-47.

[5] 郝利华.JC公司基于PDM系统的研发项目进度管理研究: [硕士学位论文]. 济南:山东大学，2015.

[6] 何国伟.管理质量的三要素:效果、效率及可靠性.质量与可靠性，2007，（4）:4-6.