|  |
| --- |
| 武汉大学logo  **研究生学位论文** |
| 题 目： 基于Neo4j图数据库的研究团体搜索系统实现  学 院： 计算机学院  专 业： 计算机科学与技术  学 号： 2019282110194  姓 名： 陈小龙  导师姓名： 祝园园  导师职称： 副教授 |

目 录

[1 选题及背景意义 1](#_Toc56093086)

[1.1 背景 1](#_Toc56093087)

[2国内外关于该课题的研究现状及趋势 1](#_Toc56093088)

[2.1 图形数据库的发展现状 2](#_Toc56093089)

[2.2 Neo4j商业案例 4](#_Toc56093090)

[3 研究计划 6](#_Toc56093091)

[3.1 研究目标及功能模块 6](#_Toc56093092)

[3.2 拟突破的难题 6](#_Toc56093093)

[3.3 实验方案 7](#_Toc56093094)

[参考文献 8](#_Toc56093095)

# 

# 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[1]、Memcached[2]）、列族数据库（如：Hbase[3]、BigTable[4]）、文档数据库（如：MongoDB[5]、Couchbase[6]）、图形数据库（如：Neo4j[7]、JanusGraph[8]、Hugugraph[9]）。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[10]，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，功能包括前端数据可视化、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，将提供一个相对完整的图数据从存储到计算再到展示的解决方案，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[11]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。

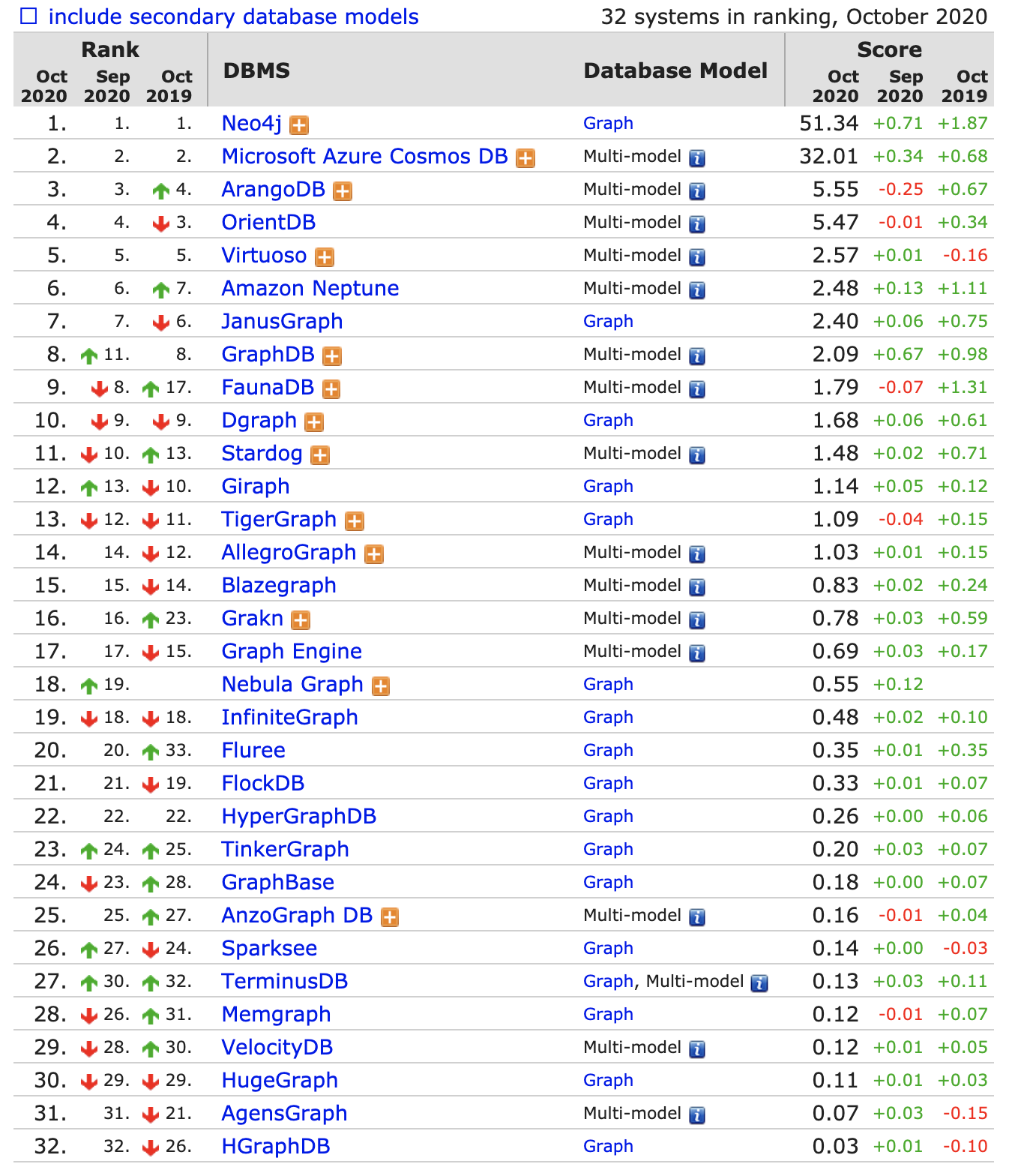


图1 2020年11月DB-Engine图数据库排名

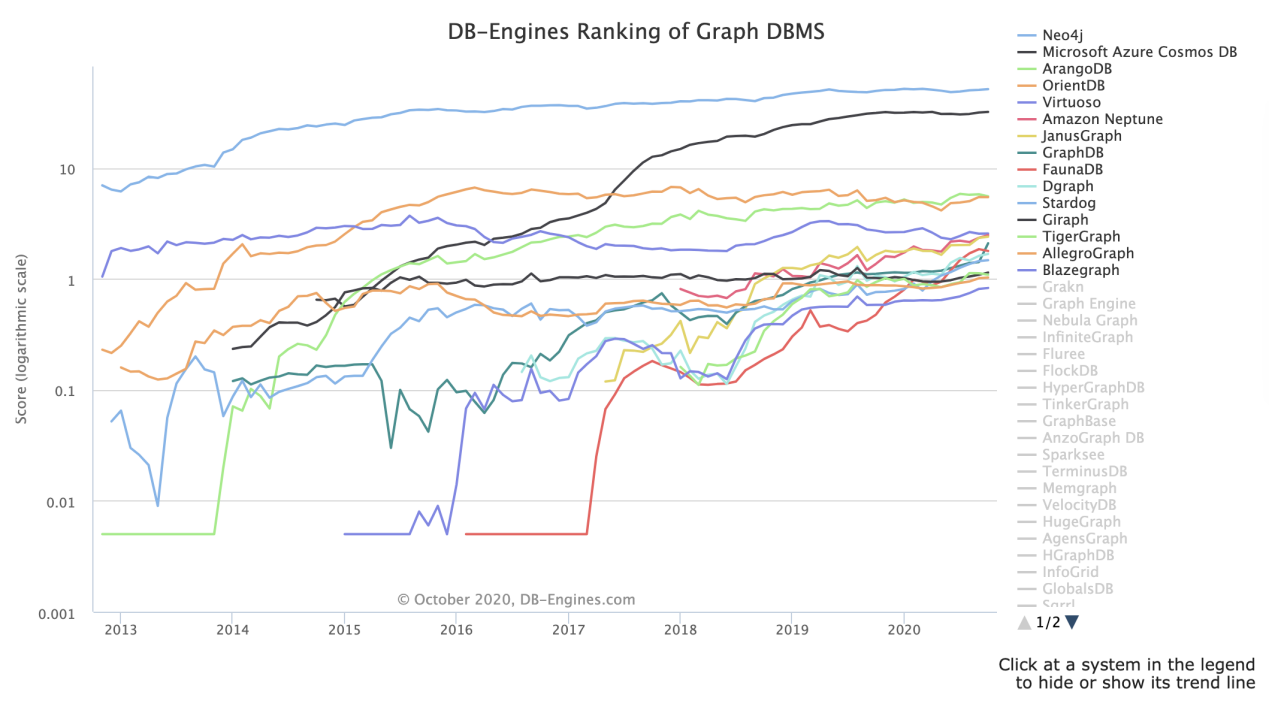


图2 2020年11月DB-Engine图数据库排名折线图

DB-Engine的数据库排名每月更新一次。图1中，截止2020年10月[Neo4j](https://db-engines.com/en/system/Neo4j)、Microsoft Azure Cosmos DB 、ArangoDB 三种数据库分别占据图数据库排行榜前三名；其中， Neo4j的分数为为51.34，Microsoft Azure Cosmos DB的分数为32.01，ArangoDB的分数为5.55。很显然，Neo4j相对于二三名具有压倒性优势，它的分数约是第二名的1.5倍，约是第三名的10倍；同时第二名的分数也几乎是第三名的6倍。

从图2的趋势图中可以看到，Neo4j发展一直很强势，流行度长期稳定增长；另外，Microsoft Azure Cosmos DB发展速度也很快，一路超越众多对手，目前市场占有率排名第二。毫无疑问，在目前的主流图数据库中，Neo4j、Cosmos DB以极大的优势占有着市场第一二名的位置，短期内这一趋势仍然会持续，同时，一些小众的图数据库也在不断的发展之中，值得人们持续关注。

2.1.2 主流图数据库的特点与优劣

Neo4j 是目前最流行的图形数据库，支持完整的事务，在属性图中，图是由顶点（Vertex），边（Edge）和属性（Property）组成的，顶点和边都可以设置属性，顶点也称作节点，边也称作关系，每个节点和关系都可以由一个或多个属性。Neo4j创建的图是用顶点和边构建一个有向图，其查询语言cypher已经成为事实上的标准。Neo4j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的优点是很容易表示图数据金额半结构化的数据；它采用原生的图模型，检索遍历数据的速度很快，很容易检索相邻的节点和关系且不需要复杂的连接操作；Neo4j CQL查询语言命令是人性化的可读格式，非常容易学习。Neo4j的缺点也很明显，它并非完全开源没不方便分布式部署，另外，由于Neo4j 采用原生图存储缺乏分片存储机制，图数据结构导致写入性能不高，实时性读写较弱，大数据量导入不够高效且无法有效存储极大的图(含千亿节点)。

Azure Cosmos DB 是 Microsoft 提供的全球分布式多模型数据库服务。Cosmos DB 2010年立项，经过7年研发，于2017年5月正式发布的云数据库服务，该数据库服务支持图数据、列存储、键值存储和文档数据库等多种数据模型，同时支持强一致性和最终一致性，以及介于之间的有边界一致性（ Bounded Staleness）、事物一致性（Session）与单调一致性（ConsistentPrefix）。Azure Cosmos DB的优势是它支持多种数据模型，支持弹性存储扩展，并且提供了五种数据一致性。它的缺点是使用成本比较高。

JanusGraph 是Titan 1.0.0版本的延续，Titan 是从2012年开始开发，到2016年停止维护的一个分布式图数据库。JanusGraph 继承了 Titan 的全部功能并做了进一步的改进，支持 Hadoop 2和 Tinkerpop 3.2.3，采用 Gremlin 图查询语言。JanusGraph最大的优点在于支持实时、数千用户并发遍历图和分析查询图的功能；它采用分布式架构架，可以自由的扩展集群节点的， JanusGraph 可以存储包含数千亿个节点和边的图。

JanusGraph的缺点是依赖底层存储，并且学习成本相对较高，不提供数据导入与备份[12]。

HugeGraph 是一款面向分析型，支持批量操作的图数据库系统，它能够与大数据平台无缝集成，有效解决海量图数据的存储、查询和关联分析需求。HugeGraph 支持 HBase 和 Cassandra 等常见的分布式系统作为其存储引擎来实现水平扩展。HugeGraph 可以与 Spark GraphX 进行链接，借助Spark GraphX 图分析算法（如PageRank、Connected Components、Triangle Count等）对HugeGraph的数据进行分析挖掘。HugeGraph的优势是，它由百度自研，2016年项目启动，开源社区新功能迭代更新快速，国内社区成熟；并且支持Spark GraphX，提供数据导入服务，支持的数据规模在千亿级以上[13]。

总的来看，多存储模式的图数据库(如：Azure Cosmos DB)技术是目前发展的一个主要趋势，以 JenusGraph 为代表的 NoSQL 存储的分布式图数据也日趋成熟，但由于其存储和查询严重分离，性能上还有很大提升空间。以 Neo4j 为代表的 native 图数据库，主要特点是查询节点或者边时，不需要借助于其他索引(如B+树索引)，而是直接指针指向下一度的物理地址;但是Neo4j的十字链表结构对硬件(内存、磁盘)性能依赖较大存在性能瓶颈。

## 2.2 Neo4j商业案例

2.2.1 NBC新闻欺诈检测[14]

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。最后，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。

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。

之后，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。但是，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是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；虽然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。团队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。

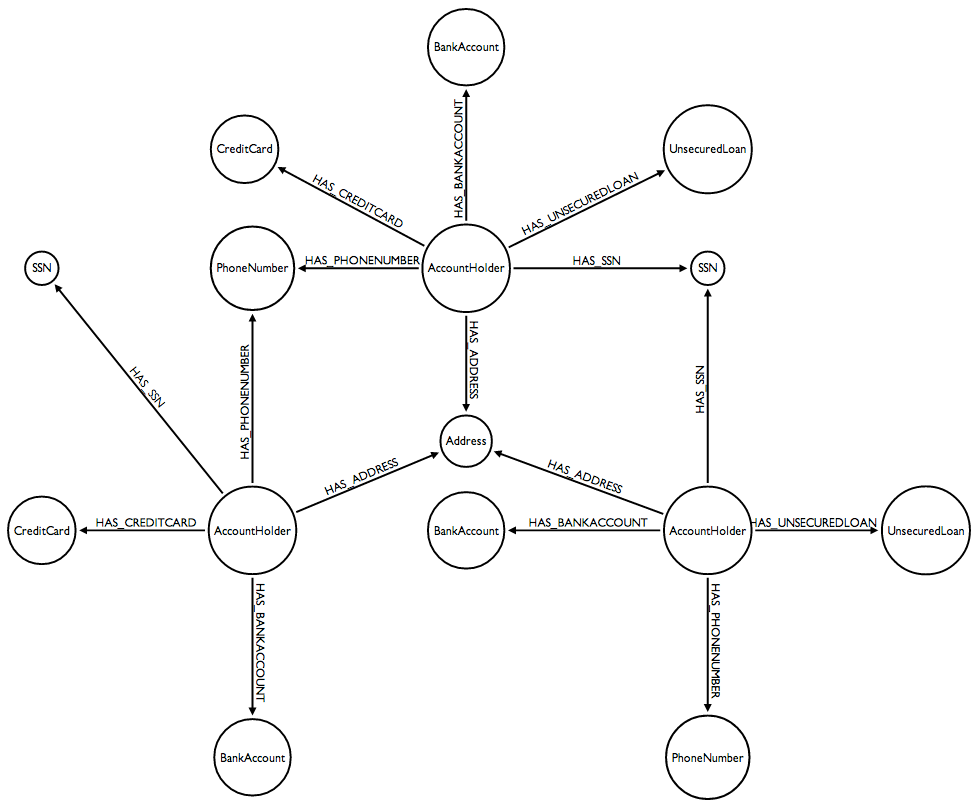
他们发现水军网络中只有数量很小的核心账号来编辑最初的推文，原创推文大概只占推文总量的25%，其余的都是转发推文。水军利用公共的标签和回复有名气的账号来扩大他们的影响力，增加关注度。研究人员发现了三类水军账户，一类伪装成典型的美国人，另一类模仿新闻网站，第三类会假装是某个政治组织。研究人员称，所有这些账号都被俄罗斯的机构控制，Neo4j揭示了这些虚假账户在社交网络中的协作模式。NBC用Neo4j构建了一个包含20万推特的数据集，通过使用Neo4j分析，他们揭示了俄罗斯如何模仿美国人，进而影响正常人，并且吸引大量粉丝，最终影响美国政治。

这个项目的意义在于，通过用Neo4j构建“关系引擎”，政府或者社交平台就能在恶意的水军影响舆情之前，采取行动维护社交秩序。

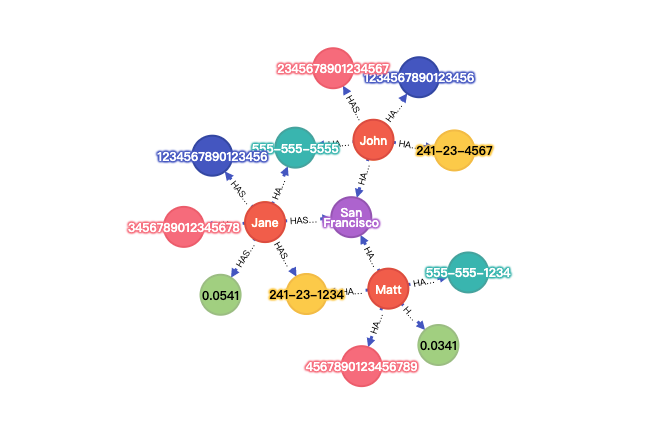
2.2.2 银行欺诈监测[15]

银行和保险公司每年因欺诈而损失数十亿美元。传统的欺诈检测方法在最小化这些损失方面起着重要作用。但是，越来越老练的欺诈者已经通过合作以及利用各种其他构造虚假身份的手段，开发了多种方法来逃避欺诈监测。尽管没有任何欺诈预防措施是完美的，但如果我们将目光超越单个数据点，将所有有关系的数据链接起来，就可以从这些链接的关系中发现重要的线索，以达到更加准确的欺诈检测率。

尽管每一桩欺诈案例在细节上有所不同但是，它们在内在模式上有着共通之处。图形数据库提供了一种新的发现欺诈事件和其他复杂骗局的新方法，具有很高的准确性，能够实时阻止高级欺诈方案。它通过使用图形数据库运行适当的实体链接分析查询，并在客户和帐户生命周期的关键阶段（如在创建帐户时、余额达到阈值时、支票退票时）运行检查来增强现有欺诈检测功能以支持环检测，通过与正确类型的事件相关联的实时图形做，可以帮助银行识别出可能的欺诈事件。



## 图3 Bank Fraud Graph Data Model



## 图4 Bank Fraud Graph Data Model

上图数据模型表示数据实际上如何显示在图形数据库中，并说明了如何通过简单地遍历图形就可以找到圆环(欺诈环)。Neo4j的Cypher语言提供了一种简单的语义，用于检测图中的环，对于上述模型可以实时查询内存中的连接。

# 3 研究计划

## 3.1 研究目标及功能模块

3.1.1 目标

基于Neo4j构建一个图数据处理系统，包括图数据导入、存储、增删改查，以及扩展Neo4j的算法（如：社区发现）。

3.1.2 功能模块

数据建模：基于原始数据的网络关系确定实体（节点）和关系（边），以及实体和边的属性。然后根据确定的实体和关系构建对象与图的映射模型，作为对Neo4j中图数据操作的基础。同样的数据可以有多种建模方式，不同的建模方式对应不同的对象与图的映射，并且直接关系具体的代码实现。好的模型可以大大提高数据处理效率，所以数据建模模块是最基础也是极重要的一步。

数据预处理：按预先建立的图的模型，处理原始数据；去除不合规范的记录，并且把原始数据改造成Neo4j可以直接识别的格式。

模型实现：使用编程语言实现先前建立的数据模型。具体来说是建立对象和节点、边的映射关系。不同的节点类型与边的类型分别映射为不同的对象，对象的中包含节点或边的属性映射。

系统RESTful API接口：RESTful API是目前最成熟的一套互联网应用程序的API设计理论。REST（Representational State Transfer）表述性状态转换，REST指的是一组架构约束条件和原则。 如果一个架构符合REST的约束条件和原则，我们就称它为RESTful架构。此模块基于restful协议开发服务接口，为前端提供统一的服务调用以及交互方式；不同的接口提供不同的数据处理服务，程序和数据库的交互对客户端透明，客户端只需要对接口发起网络请求就可以实现对数据库中的数据进行不同的操作，并获取操作结果。

算法扩展(优化)：使用实验室的图相关的算法（如：社区发现）扩展Neo4j算法库，并整合API接口与扩展算法对外提供服务；实现使用扩展算法高效处理图数据。

## 3.2 拟突破的难题

由于原始图数据格式的不同，首先要解决的问题是数据的预处理，以规范数据格式，方便数据导入。对于大量数据的预处理是一个可以预见的问题。当前数据清洗的工具，比如hadoop、sql、python等，均有各自的使用场景，需要找到一个简单可行的方案进行数据预处理，这是一个待解决的问题。

另一方面，由于图数据建模的灵活性，如何找到一个恰当的数据模型来描述数据集，并不是一件容易的事情。往往同一个数据集使用不同的建模方式，带来的效率完全不同。模型的建立不仅要考虑是否符合业务需求，还要考虑后续开发的难度。理想的情况是找到一个统一的模型，以满足不同的场景，这个问题还需要进一步的探索。

该课题最大的困难在于如何优化图的算法，以及将优化的算法和系统整合，并对外提供服务。当前社区检测的算法应用非常广泛，比如个性化推荐、欺诈检测等。典型的社区检测算法，如：Louvain、Label Propagation、Weakly Connected Components都存在各自的不足。如何找到一个新的，在某些性能指标上超过现有的算法的社区检测算法，并将优化的算法和系统整合需要进一步调研，主要问题包括：使用何种优化算法、使用什么方式整合优化的算法、如何保证性能、以及如何调用。

## 3.3 实验方案

3.3.1 创新及特色

首先，从具体模块看，该项目会整合改进的图算法，整合完成之后会提高系统对图数据的计算能力，还可以增加Neo4j原生算法库所没有的算法。这种模式提供了一种功能扩展的模式，方便后期对系统功能的进一步优化与扩展。

其次，从项目整体来看，它实现了从图数据清洗、建模，到图数据的存储和操作，再到图算法的扩展调用，最后到图数据前端的可视化展示的一整套流程。综上所述，项目提供了一套相对完整的图数据处理方案，并留有进一步扩展优化的空间。

3.3.2 拟采用的研究工具

数据清洗工具拟采用Hadoop平台的MapReduce模型或者使用Python数据处理相关的开源包（如：[pandas](https://github.com/AI-10/Data-processing/blob/master/%E6%95%B0%E6%8D%AE%E5%88%86%E6%9E%90%E5%9F%BA%E7%A1%80---%E5%B7%A5%E5%85%B7%EF%BC%9Apandas.ipynb)），具体采用何种方式需要进一步调研，取决于数据量和数据格式。

数据建模工具拟采用PowerDesigner。PowerDesigner是目前数据建模业界的领头羊。功能包括：完整的集成模型，和面向包含IT为中心的、非IT为中心的差异化建模诉求。支持非常强大的元数据信息库和各种不同格式的输出。PowerDesigner拥有一个人性化的界面，简单易懂的帮助文档，可以实现快速建模。

模型实现拟采用Java实现；系统API接口拟采用SpringBoot 2.3.0.RELEASE框架进行开发。SpringBoot是全新开源的轻量级框架。它基于Spring4.0设计，继承了Spring框架原有的优秀特性，并通过简化配置来进一步简化了Spring应用的整个搭建和开发过程。

Neo4j使用社区版4.0版本，数据库部署环境拟采用ubuntu18.04.1x86\_64/4核/8G内存/200GSSD硬盘,测试阶段使用低配版服务器1核2G。

图算法拟采用C++实现，Java和C++的相互调用拟采用HTTP（restful）协议。

3.3.3 研究工作进度

2020.10.23—2020.10.30 接受学位论文任务，协商选题；

2020.10.31—2020.11.13 学位论文前期调研、文献资料查阅等准备工作；

2020.11.13—2020.11.24 完成学位论文开题报告；

2020.11.24—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—2021.01.25 完成数据库安装部署，以及数据建模，数据导入工作；

2021.01.26—2021.02.20 完成后端API设计开发，具体包括节点的查询、删除、新增等；

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]Redis Labs.Redis官网[EB/OL].https://redis.io/,2015-06.

[2]Dormando.memcached官网[EB/OL].https://memcached.org/,2015-08-03.

[3]Apache Hbase.The Apache Software Foundation[EB/OL].https://hbase.apache.org/,2007.

[4]Fay Chang,Jeffrey Dean,Sanjay Ghemawat.Bigtable: A Distributed Storage System for Structured Data[EB/OL].https://research.google/pubs/pub27898/,2006.

[5]Eliot Horowitz.The database for modern applications[EB/OL].https://www.mongodb.com/cloud/atlas,2020.

[6]Couchbase.couchbase官网[EB/OL].https://www.couchbase.com/,2011.

[7]Neo4j.Neo4j官网[EB/OL].https://neo4j.com/,2020.

[8]janusgraph.Distributed, open source, massively scalable graph database[EB/OL].https://janusgraph.org/,2017.

[9]王二铁.百度安全开源大规模图数据库HugeGraph[EB/OL].https://zhuanlan.zhihu.com/p/41240429,2018-08-03.

[10]Neo4j.the Neo4j Graph Data Science library[EB/OL].https://neo4j.com/docs/graph-data-science/current/algorithms/,2020.

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

[12]Jermy Li.图数据库选型比较：Neo4j、JanusGraph、HugeGraph[EB/OL].https://blog.csdn.net/javeme/article/details/105000288,2020-03-20.

[13]Jermy Li.图数据库选型比较[EB/OL].https://blog.csdn.net/javeme/article/details/105000288,2020-03-20.

[14]Will Lyon.NBC News Analyzes Hundreds of Thousands of Russian Troll Tweets Using Neo4j[EB/OL].https://neo4j.com/case-studies/nbc-news/,2019.

[15]kbastani.Bank Fraud Detection by Kenny Bastani[EB/OL].https://github.com/neo4j-contrib/gists/blob/master/other/BankFraudDetection.adoc,2013-10-06.