

SHANGHAIUNIVERSITY

毕业设计（论文）

**UNDERGRADUATEPROJECT(THESIS)**

**题目:针对某领域的数据分析和存储**

|  |  |
| --- | --- |
| **学院** | **计算机工程与科学学院** |
| **专业** | **计算机科学与技术** |
| **学号** | **0** |
| **学生姓名** | **0** |
| **指导教师** | **0** |
| **起讫日期** | **2018.01.29 – 2018.06.08** |

目录

[摘要 III](#_Toc513199145)

[ABSTRACT IV](#_Toc513199146)

[第1章 绪论 1](#_Toc513199147)

[§1.1 数据分析和存储的背景及意义 1](#_Toc513199148)

[§1.2 数据领域现状及存在的问题 1](#_Toc513199149)

[§1.2.1 数据分析存在的问题 1](#_Toc513199150)

[§1.2.2 数据分析的难点 1](#_Toc513199151)

[§1.2.3 数据存储存在的问题 1](#_Toc513199152)

[§1.2.4 数据存储的难点 1](#_Toc513199153)

[§1.3 本文研究内容及目标 1](#_Toc513199154)

[§1.3.1 研究内容 1](#_Toc513199155)

[§1.3.2 研究目标 1](#_Toc513199156)

[§1.4 本文组织结构 1](#_Toc513199157)

[第2章 数据分析 2](#_Toc513199158)

[§2.1 数据分析概述 2](#_Toc513199159)

[§2.1.1 数据分析的基本概念 2](#_Toc513199160)

[§2.1.2 数据分析方法 2](#_Toc513199161)

[§2.2 针对不同类型数据的分析方法 2](#_Toc513199162)

[§2.2.1 Txt数据 2](#_Toc513199163)

[§2.2.2 Excel格式 2](#_Toc513199164)

[§2.2.3 Xml/json格式 2](#_Toc513199165)

[§2.3 本章小结 2](#_Toc513199166)

[第3章 数据存储 3](#_Toc513199167)

[§3.1 多种数据库概述 3](#_Toc513199168)

[§3.2 关系型数据库 3](#_Toc513199169)

[§3.2.1 MySQL 3](#_Toc513199170)

[§3.3 对象型数据库 3](#_Toc513199171)

[§3.3.1 Db4o 3](#_Toc513199172)

[§3.4 图数据库 3](#_Toc513199173)

[§3.4.1 neo4j 3](#_Toc513199174)

[§3.5 文档型数据库 3](#_Toc513199175)

[§3.5.1 mongodb 3](#_Toc513199176)

[§3.6 不同数据库的性能对比 3](#_Toc513199177)

[§3.7 系统中的数据结构与其表现 3](#_Toc513199178)

[§3.8 本章小结 3](#_Toc513199179)

[第4章 食物项目系统的设计与实现 4](#_Toc513199180)

[§4.1 系统整体设计 4](#_Toc513199181)

[§4.1.1 开发环境简介 4](#_Toc513199182)

[§4.1.2 系统整体架构 4](#_Toc513199183)

[§4.2 数据库设计 4](#_Toc513199184)

[§4.3 功能模块设计 4](#_Toc513199185)

[§4.3.1 数据可视化 4](#_Toc513199186)

[§4.4 本章小结 4](#_Toc513199187)

[第5章 总结与展望 5](#_Toc513199188)

[§5.1 本文总结 5](#_Toc513199189)

[§5.1.1 本文的主要工作 5](#_Toc513199190)

[§5.1.2 本文的主要创新点 5](#_Toc513199191)

[§5.2 展望 5](#_Toc513199192)

[致谢 6](#_Toc513199193)

[参考文献 7](#_Toc513199194)

针对某领域的数据分析和存储

# 摘要

XXX。

关键词：XXXX，XXX，XXX，XXX

XXXEnglish Title XXX

# ABSTRACT

Xxx

**Keywords:** xxx

# 绪论

本章主要描述了数据分析和存储的背景，意义。分析了相关课题的研究现状，提出了本文要研究的重点内容和目标。

## 数据分析和存储的背景及意义

## 数据领域现状及存在的问题

### 数据分析存在的问题

### 数据分析的难点

### 数据存储存在的问题

### 数据存储的难点

## 本文研究内容及目标

### 研究内容

### 研究目标

## 本文组织结构

# 数据分析

本章具体描述了数据分析：介绍了数据分析的基本概念，并引出针对不同类型数据的分析方法。

## 数据分析概述

### 数据分析的基本概念

### 数据分析方法

## 针对不同类型数据的分析方法

### Txt数据

### Excel格式

### Xml/json格式

## 本章小结

# 数据存储

本章是全文的重点章节，阐述和分析了数据存储方面的几种设计思路。首先简单介绍了关系型数据库，对象型数据库，图数据库，nosql数据库。随后分别举出例子，分析其特点和优劣，对比分析其性能。最后将系统数据存储的构建过程完整的进行了描述，并将部分系统功能运行的结果作了展示。

## 多种数据库概述

## 关系型数据库

### MySQL

## 对象型数据库

发展趋势：面向对象数据库->对象关系型数据库->对象代理数据库。

面向对象数据库：数据库与面向对象语言的结合。由以下四部分组成：

 对象模型（对象的特性，对象间的联系）

 对象描述语言：ODL（对象定义语言） OIF（对象交互格式）

 对象查询语言: OQL(对象查询语言)

 对象语言绑定：C++,JAVA 目的主要是实现ODL和ODMG缺乏的OML，利用语言灵活性。

面向对象模型：

 类层次结构（子类/超类；外延）

 对象（原子对象，结构对象，聚集对象）

 对象的状态（属性，联系（只支持二元联系））

对象语言绑定：

 原因1.面向对象特点决定的，绑定可以方便的加入面向对象的功能，由用户灵活的实现定义和操作对象。2.可以扩大在掌握一门以上语言的用户中的应用度。（为了开发者方便？是这个意思吗？）

 缺点：1.为了适应多种语言的特性，牺牲了一些良好的特性来满足对多语言绑定的支持。2.虽然有多语言的绑定，对一些普通用户来说应用依然是十分困难的。

面向对象数据库系统：（非常重要的面向对象数据库系统宣言）包括了：必备特性（复杂对象，对象标识，封装性，类型，继承性，重载，可扩充性，计算完备性，辅存管理，并发性，恢复，即时查询）。可选特性（多重继承，类型检查和推理，分布式，类型系统）。

面向对象数据库应用：

广泛应用于：多媒体，工程数据，地理信息等复杂信息管理领域。

针对这种复杂对象建模时，1.关系数据库需要为所有数据在每种尺度下建立自己的表，访问时，通过不同尺度的查询条件，查询不同的表来返回信息。往往同种数据在不同尺寸下的表会有大量数据冗余，这加大了同种数据一致性维护难度。而如果通过集成，可以很好的降低这一难度。2.其次，无法再关系型数据库中实现对数据的计算，如添加c=a\*b，c属性必须也要建立一个字段，这是不必要的。对象型可以在表上定义一个计算的函数或者方法来实现这类查询，这是关系型所不能支持的。3.一些结构关系复杂的数据类型无法再关系型中准出，二是通过基础的数据类型去记录地理数据，往往导致数据库表单的设计过于复杂，而且需要大量的join来实现复杂数据的表达，而对象型为了满足多尺度的要求，可以对统一对象创建多个类并根据类层次结构来进行管理。

但最后还是因为“柔软性太差，复杂难用，成熟度低”没有撼动关系型的地位。

对象关系型数据库

关系数据库只能做到对复杂数据的存储和对简单数据的查询。而对象关系数据库的最大功能是能做到对复杂数据的查询且速度很快，具有可扩充性。

特点：

 扩充数据类型

 支持复杂对象

 支持集继承概念

 提供通用的规则系统

### Db4o

面向对象数据库的概念提出是很早的， “面向对象数据库系统”这一术语第一次出现于 1985 年。著名的研究项目包括：Encore-Ob/Server ( 布朗大学）， EXODUS（Wisconsin 大学）， IRIS （惠普）， ODE （ Bell 实验室）， ORION （MCC ） ，Vodak （GMD-IPSI）和 Zeitgeist （Texas Instruments）。其中以 ORION 项目发表的论文数为最多。最早的商品化 ODBMS 出现在 1986 年，是 Servio 公司（现在的 GemStone 公司）和 Ontos 公司推出的。后来（九十年代） Object Design （ ODI ）、 Versant 、 Objectivity 、 O2 Technology 、 Poet 、 Ibex 、 UniSQL 和 ADB MATISSE 等公司也加入了这个开拓行列。然而目前为止成熟的对象型数据库还是比较少，而且开源的更少。2004年，db4o作为免费的开源ODBMS发布，2005年，db4o首先将Native Queries 作为面向对象的数据访问API 来实现，该API完全依赖于编程语言（Java / C＃）本身。国内相对比较流行的也是db4o,但是这个项目也已经在XX年被官方停止维护。

db4o 提供两种运行模式，分别是本地模式和服务器模式。本地模式是指直接在程序里打开 db4o 数据库文件进行操作，而服务器模式则是客户端通过 IP 地址、端口以及授权口令来访问服务器。

## 图数据库

### 概述

作为NoSQL的一员大将，图数据库近几年被广泛的讨论，其使用者也逐渐增多。不得不说在研究过它的思想后，我也爱上了这种形象自然的建模思路和设计方法。

首先，让我们想想什么是图？图就是节点和联系的集合，其中比较主流的图模型主要是**属性图**，资源描述框架（RDF）三元图和超图。拿最流行的属性图举例，它主要有两个特性。一：包含节点和联系，节点有自己的属性。二：联系有名字和方向，也可以有属性。这个模型在我看来是如此的通用，以至于我认为生活中任何东西都可以用图来建模表示。既然它是一个符合人们思维习惯的建模思路，那它必然也可以抽象出来用于计算机世界里的数据存储。而我们要讨论的图数据库就是建立在这样的基础上的。图数据库使用节点，边/关系(relationship)和属性。节点代表实体，关系代表实体间的联系，属性是节点的相关信息。引用维基百科对图数据库的定义“In [computing](https://en.wikipedia.org/wiki/Computing), a graph database (GDB) is a [database](https://en.wikipedia.org/wiki/Database) that uses [graph structures](https://en.wikipedia.org/wiki/Graph_(data_structure)) for [semantic queries](https://en.wikipedia.org/wiki/Semantic_query) with [nodes](https://en.wikipedia.org/wiki/Node_(graph_theory)), [edges](https://en.wikipedia.org/wiki/Edge_(graph_theory)) and properties to represent and store data. A key concept of the system is the graph (or edge or relationship), which directly relates data items in the store.”

知道了什么是图数据库后，我们就可以对比分析图数据库建模的方法和特点。传统的关系建模流程大致是：草图，er图，映射成表。在过程中需要把数据间的联系以逻辑的形式存储下来。同时应该使用范式，约束等技术整理模型结构，减少数据库中数据冗余，增进数据。也就是规范化。如此，建模过程看上去已经完成了，但实际项目中，需求总是变化的，所以实际生活中常常会继续使用**反规范化技术**。也就是说，在实现关系型解决方案的过程中，为了适应关系模型，我们对物理模型强加了很多变化，如为了获得查询性能，人为制造重复数据和重复结构。这些变化使得概念模型和数据真实的物理布局之间产生了差异。这种概念上的失调，使得将业务需求的变化转化为底层的稳固的关系结构变得非常复杂和繁琐。同时，需求的大量变更，使得**迁移**花销巨大和风险极高。总之，在实际的项目过程中，我们需要经历：设计-规范化-反规范化的过程。

而图建模就去掉了使用复杂数据管理代码来规范化和反规范化数据的这一步骤。我们可以很轻松的将复杂的层次结构抽象出来，又因为其建模思路和人的思维习惯相符，整个过程非常的易理解而又高保真。虽然在建模初期，常见一个陷阱是错误的将本应是节点的数据用联系+联系的属性表示出来，反而丢失了节点本身的内容。但这种错误随着对图的熟悉会越来越少，这是一个设计上容易遇到的问题，但错误并不在图数据库。同样值得注意的一点是：不要小看图的天然可扩展性。相比关系型需要不断确认逻辑设计，图数据库在修改实体和联系上方便了太多，例如，当我们需要新增实体并添加于已有实体间的关系时，我们不需要修改本来实体的内容，只需要新增节点然后添加一条指向已有节点的边即可。

这样分析了图建模的方式后，我们可以总结出图数据库的两大优势： 1.**灵活性**，可扩展性高。不必要一开始就把整个系统的结构设计出来，可以在开发的过程中低成本的增删节点和联系。2.**敏捷性**。图数据库不需要**schema**,所以它缺少以schema为导向的数据管理机制。他的管理通常作用于编程方式，利用测试来驱动数据模型和查询，以及依靠图来断言业务规则。它提供了数据模型的平稳演化，数据迁移和反规范化很少会成为问题，新的事实和新的组件会成为新节点和新联系。而优化关键性能访问模式通常需要在两个原本有中介连接的节点间引入一个直接联系。适用于敏捷开发。

那么，图在性能方面的表现足够好吗？让我们继续分析模型建好后的操作。

关系模型使用数据中的信息关联数据。例如，当需要查找所有电话号码包含区号“311”的用户时，它通过遍历整个表，在电话号码字段中查找字符串“311”。这在大型表中是一个耗时的过程，因此关系数据库提供了数据库索引的概念，以提高查询速度。

相比之下，图形数据库直接存储记录之间的关系。同样是查找某用户的电话号码，图数据库使用户有直接指向电话记录的指针，而不是通过用户id查找用户表里电话号码字段的内容。也就是说，选择一个用户后，可以通过指针直接得到对应的电话号码，而不需要搜索用户表来查找匹配的记录。这可以消除昂贵的连接（join）操作。这种优势在进行深度较大的搜索时更加明显。当系统需要查找和用户a的电话号码前六位相同的其他用户最近在某地的购买的物品的生产地的最普遍的姓时，关系型需要多次搜索匹配，而图数据库只需要查找节点，顺着边继续进入下一个深度即可。从技术上讲，图数据库共用了O（log（ n））+ O（1）时间，即大致相对于数据大小的对数。而关系版本将是多个 O（log（ n））查找。由此我们可以得出结论：图数据库具有良好的查询**性能**。因为哪怕在数据集增大的情况下，查询也总是和图的一部分相关，因此，每个查询的执行时间只和满足查询条件的那部分遍历的图的大小而不是整个图的大小成正比。（使用免索引链接，图数据库可以将复杂的join查询转化为快速的图遍历，节点和关系只有在结果迭代器需要访问它们的时候才会被遍历并返回，对于大规模深度遍历而言，这极大地提高了性能。因此无论数据集有多大，都可以维持毫秒级的性能。）

总结以上的分析。图数据库适用于处理**半结构化**的，紧密**关联且变动性大的**的数据。现实中，图数据库典型使用的领域如[语义网和RDF](http://en.wikipedia.org/wiki/Semantic_Web)、[LinkedData](http://en.wikipedia.org/wiki/Linked_Data)、[GIS](http://resources.esri.com/help/9.3/arcgisengine/dotnet/e084da94-d4f7-4da7-86ed-7df684ff2144.htm)、[基因分析](http://code.google.com/p/pygr/)、[社交网络数据建模](http://blog.neo4j.org/2009/09/social-networks-in-database-using-graph.html)、[推荐算法](http://wiki.github.com/tinkerpop/gremlin/linkeddata-sail#recommendation)，地理空间（基于位置的服务），网络拓扑以及[其他领域](http://wiki.neo4j.org/content/Neo4j_In_The_Wild)。不适用更新全部或某个子集的实体。

### 图数据库实例之neo4j

在图数据库领域，现在比较流行的数据库有：[Neo4j](http://www.neo4j.org/) - 开源的Java属性图形模型。[AllegroGraph](http://www.franz.com/agraph/allegrograph/)，闭源，RDF-QuadStore。[Sones](http://www.sones.com/) - 闭源，关注于.NET。[Virtuoso](http://www.openlinksw.com/) - 闭源，关注于RDF。[HyergraphDB](http://www.kobrix.com/hgdb.jsp) - 开源的Java超图模型。其他的例如 [InfoGrid](http://infogrid.org/)、Filament、[FlockDB](http://github.com/twitter/flockdb)等。实际开发中，我使用的数据库是Neo4j。Neo4j是一款强健的，可伸缩的，用Java实现、完全兼容ACID的高性能图数据库。数据以一种针对图形网络进行过优化的格式保存在磁盘上。Neo4j的内核是一种极快的图形引擎，具有数据库产品期望的所有特性，如恢复、两阶段提交、符合XA等。自2003年起，Neo4j就已经被作为24/7的产品使用。Neo4j既可作为无需任何管理开销的内嵌数据库使用；也可以作为单独的服务器使用，在这种使用场景下，它提供了广泛使用的[REST](http://wiki.neo4j.org/content/Getting_Started_REST)接口，能够方便地集成到基于[PHP](http://github.com/onewheelgood/Neo4J-REST-PHP-API-client)、.NET和JavaScript的环境里。开发者可以通过[Java-API](http://api.neo4j.org/current/)直接与图形模型交互，随着开源的发展，目前它也支持[JRuby/Ruby](http://github.com/andreasronge/neo4j)、[Scala](http://github.com/ept/neo4j-scala-template)、[Python](http://components.neo4j.org/neo4j.py/)、[Clojure](http://github.com/JulianMorrison/neo4j-clojure)等其他语言。而它的这两个版本都可以做成集群，从而（使用主从复制）实现高可用性和水平读扩展，甚至我们可以将两个版本进行混合，在数据库级执行。

Neo4j的几个特点如：使用多种算法支持路径搜索，使用键值和关系进行索引，为读操作进行优化，支持事务（用 Java api），使用 Gremlin图形遍历语言，支持 Groovy脚本，支持在线备份，高级监控及高可靠性支持使用 AGPL/商业许可。

接下来我们尝试**将图数据库与场景结合起来**。在过程中继续分析其优劣。

根据拿到的资料，我们简单拿方便面多维度属性和定量描述举例。具体的场景如下：有品牌A,B,C生产了口味a,b,c,d等n种方便面，每种方便面有自己表层物品属性，如编号，生产日期，规格，风味等。厂商为了更好的分析食品口味，研究市场对口味的喜好，请了m名测试员从色泽，光滑性，复水性，韧性，黏性，耐泡性等多个角度进行打分。同时，在口味方面，厂商进行了细化研究，通过多种评价标准，如强度等级，麻度等级，斯科维尔指数，描述词，根据方便面在不同时间的表现进行评级，得出了味道的时间强度评价和时间频率记录。当食物投放市场后，厂商进行了反馈调查，拿到了o位用户针对n种方便面的喜恶数据。

分析以上场景，我认为大致可以分为三个模型。模型一：方便面的属性记录。这里主要包括表层属性，分数情况。模型二：测试员分析过程中的打分记录。这里数据模型简单，数据量大。需要注意的是如何把时间因素记录得当。模型三：用户的反馈记录。这里可以为后续的推荐功能，产量设计提供数据支持。

分析完数据模型后，我们可以结合图数据库的特点来看如何设计数据存储模型。在以上提到的食物领域，我认为源数据的主要特点是结构简单，数据量大，变动性高。那么，个人比较遗憾的发现，图数据库最大的优点——联系在这个场景里并没有很好的得到表现。因为食物之间并不会有很多的事件交集，显然无法像社交网络那样完美的体现图数据库中联系一等公民的地位。

那么，这就意味着图数据库不适合了吗？我认为也不是这样的。虽然事物间的联系不多，但从以上描述中可以清晰看到一个食物的属性是很多的。特别是在口味这里，作为食物最重要的特性，我们可以很明显的看到厂商在这里用了多种角度分析，在味觉粒度和时间粒度上都做了细化。也就是说，一个物品的属性之间是有联系的。这样无疑也可以让图数据库发挥特长。而且，在反馈分析这里，我认为图数据库是非常适合的。不同人对不同事物有不同的喜好。这个过程可以看出有3部分的联系。人和食物（是否喜爱），人和人（口味是否类似），食物和食物（是否是统一风格大类）值得注意的是，我们在介绍图数据库时提到过它擅长的领域，正如当时分析的，做推荐系统图数据库具有天然的优势，它带来的对性能的提高不容小觑。

同时，正如之前提到的，本情景下，需求变更快，源数据的变动性高，换句话说就是对数据模型的可扩展性有很高的要求。如果整体模型太过僵化，那么食物每增加一种评价方式，每发生一次某属性的突然增删，测试员每提供不完全统一的数据，数据库每涉及到迁移（无论大小），无疑都是一次痛苦的重新设计，系统的维护成本升高不说，万一由于操作者（开发者）经验问题，很可能导致冗余数据/丢失数据/模型臃肿等恶性情况的发生。这大大增大了项目的风险，也有很大概率降低系统的性能。

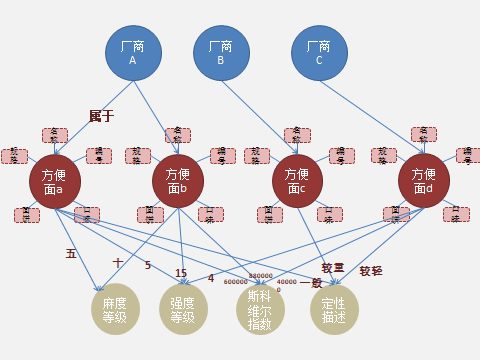
考虑到如上的可能性，图数据库瞬间变得非常亲切可人了。它天生良好的可扩展性无疑非常适合于需求变更快的敏捷开发模式。它没有那些复杂的，恼人的物理模型，概念模型转换。建立表结构是如此的简单，在原有表上新增属性和联系只需新增节点并画上线即可。实在是非常的高效方便。

再其次，在数据的查找方面，图数据库更有的可说。介于它的查询只发生在图的局部，不需要遍历整个库，而且由于它联系分明，在深层次的数据查询方面比传统的好上太多，传统关系型查询性能随深度增加是指数级减小的，而对于图数据库，这基本是线性的，当深度大于3时，其中的差距实在需要引起注意。

然而，虽然我这么喜爱这种“又聪明又懒惰”的图数据库建模方式，总体来说，它也还是有表现不佳的地方。这里主要体现在对时间流的掌控上，图的亮点在于实体间的联系，而时间是流式的，每个时间都有新的不同的数据，个人认为，这个变化的过程用图数据库来记录实在是不太合适的，图数据库表示时间一般是两种方法，时间轴数和链表。放在本例的场景下都不是很合适。

这里画出图数据库的模型设计：

模型一：



## 文档型数据库

### mongodb

## 不同数据库的性能对比

## 系统中的数据结构与其表现

## 本章小结

# 食物项目系统的设计与实现

本章主要介绍食物项目系统的设计与实现，包括系统整体设计、数据库设计以及系统内各功能模块的设计与界面展示。

## 系统整体设计

### 开发环境简介

### 系统整体架构

## 数据库设计

## 功能模块设计

### 数据可视化

## 本章小结

# 总结与展望

本章对全文的主要工作和创新点作了总结，并提出需要进一步研究和改进之处。

## 本文总结

### 本文的主要工作

### 本文的主要创新点

## 展望

# 致谢

XXX。

# 参考文献

1. 林鸿飞,杨元生.用户兴趣模型的表示和更新机制.计算机研究与发展，200239(7):838-842.
2. 胡学联,潘金贵,李俊,张灵玲. 一个个性化的信息搜集Agent的设计与实现.软件报,2001,12(7):1074-1079.
3. FragoudisD.User Modeling in Information Discovery:An overview.InProceedings of Advanced Course on Artificial Intelligence,1999(ACAI99), July,1999, Greece.
4. Balabanovie M, Shoharn Y.Learning Information Retrieval Agents:Experiments with Auto-mated Web BrowsingIn:Proceedings of the AAAISpring Symposium Series on Information Gathering from Heterogeneous, Distributed Environments, March, 1995:13-18.
5. 附录：部分源程序清单