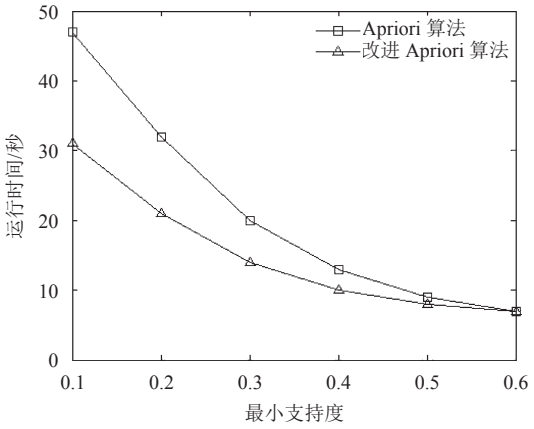

图 1 最小支持度与运行时间

从图 1 可以看出, 在相同的最小支持度下, 改进 Apriori 算法在运行速度上得到了明显性能提升.

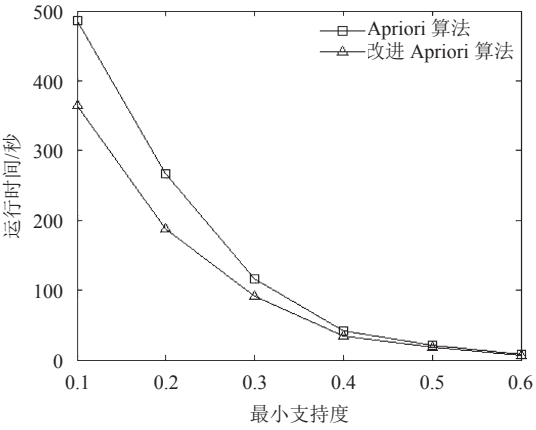

图 2 最小支持度与频繁项集数

从图 2 可以看出, 在相同的最小支持度下, 改进 Apriori 算法在探索频繁项集的过程中比经典 Apriori 算法冗余更少, 相应所占用空间也更小.

3 糖尿病预诊系统设计

3.1 糖尿病高危因素分析

本研究从 UCL 糖尿病数据集选取其中 8 个相关危险因素[10]进行分析, 分别是: 年龄, 有无高血压或高血脂病史, 身体质量指数 (BMI), 腰臀比 (WHR), 是否吸烟, 是否饮酒, 是否过度饮食和运动量是否达标. 将以上因素分别记为项 A 到项 H, 针对其中若干非布尔类型的数据预处理[11], 年龄大于 45 记为 1, BMI 大于 28 记为 1, 男性 WHR 大于 0.85、女性 WHR 大于 0.8 记为 1. 最后再加入预诊结果项 I, 将所有数据整理为事务数据库, 便于后续工作进行挖掘和分析.

3.2 系统架构设计

系统选用时下热门技术栈: RPC 框架 Dubbo, 微服务框架 Spring Boot, 消息中间件 RabbitMQ, 关系型数据库 MySQL, 以及作为缓存的 Redis.

Dubbo 是一款开源 RPC 框架, 它提供了三大核心能力: 面向接口的远程方法调用, 智能容错和负载均衡, 以及服务自动注册和发现. 该框架不仅实现了高性能、高可用性, 而且使用方便, 扩展性极佳[12].

Spring Boot 是 Java 领域知名的微服务框架, 微服务的目的在于化解整体架构服务的复杂性, 以简单快速的方式实现各个服务的部署和变更. 而 Spring Boot 提供了形式多样的库, 支持 JPA、RESTFul、Docker 等技术, 能够让配置、部署和监控变得简单方便[13].

RabbitMQ 基于 Erlang 语言编写, 用于在分布式系统中存储转发异步消息, 将彼此独立的计算机连接起来组成松耦合的系统, RabbitMQ 在易用性扩展性、高可用性等方面表现不俗[14].

MySQL 是一款由瑞典的公司开发并且广泛应用于中小型企业或组织的免费数据库, 基于 Linux 操作系统开发, MySQL 体积小、速度快、总体拥有成本低.

Redis 是一款基于内存的、可持久化的非关系型 Key-Value 存储系统, 它支持多种数据类型, 并支持原子性操作[15]. Redis 与其他 Cache 相比, 拥有更多的数据结构并支持更丰富的数据操作.

基于上述所选技术栈进行系统架构设计, 将系统划分为如图 3 所示的若干层面.

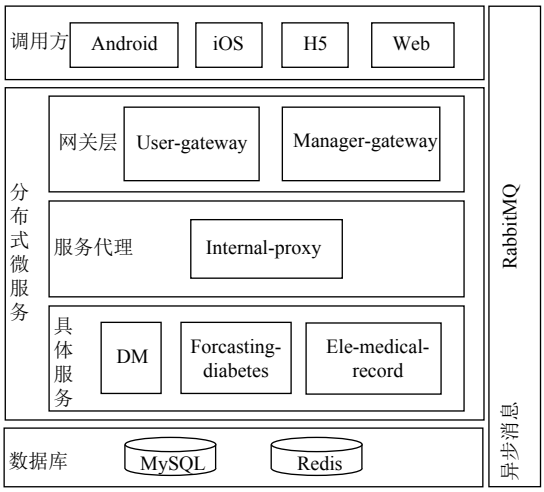

图 3 糖尿病预诊系统架构图

网关层: 作为统一请求入口, 处理权限认证及负载均衡等, 向外提供 RESTFul API, 并采用令牌桶算法实现 API 的动态限流.

服务代理层: 为提高系统扩展性和可复用性, 抽取公用服务接口, 由代理将请求路由至具体服务.

具体服务层: 具体实现三大核心功能, 包括关联规则挖掘、糖尿病预诊分析、电子病历处理.

预诊分析模块: 基于用户电子病历中的数据, 计算各项高危因素的指标, 并匹配满足置信度的关联规则, 分析糖尿病的患病概率.

电子病历模块: 为用户建立电子病历, 涵盖用户各项相关高危因素信息, 并针对特定项进行量化入库.

规则挖掘模块: 对于管理员设定的支持度和置信度, 基于所建事务数据库, 在后台挖掘满足支持度和置信度阈值的关联规则, 并将关联规则落库.

3.3 系统流程设计

系统具体使用流程如图 4 所示, 主要包括三大核心功能的使用流程: 糖尿病自查流程、电子病历录入流程和关联规则挖掘流程.

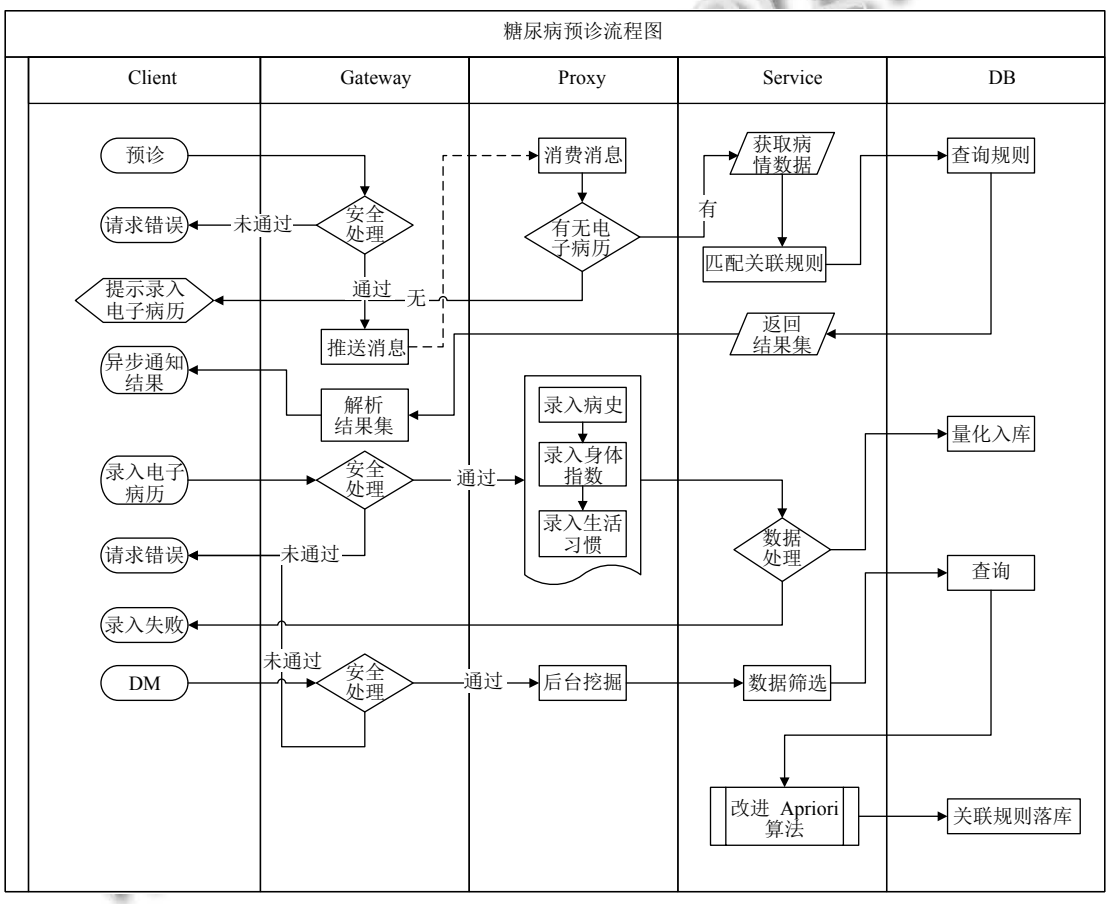

图 4 糖尿病预诊系统流程图

① 当用户提交自查请求, 网关层会对请求做权限认证、接口限流令牌校验、安全处理等. 随后会生产一条异步消息推送至 RabbitMQ, 由代理层消费消息并路由至具体服务, 实现规则匹配和异步结果返回.

② 当用户电子病历记录为空时, 医护人员录入用户各项相关数据. 数据提交后, 代理层调用具体服务处理数据, 生成电子病历并存储入库.

③ 当管理员提交关联规则挖掘请求时, 网关层对权限进行校验, 随后代理层路由至具体服务, 使用改进 Apriori 算法在后台对数据集进行筛选和挖掘. 在关联规则落库后, 以站内信和其他特定方式通知管理员关联规则的挖掘结果.

3.4 系统核心配置和 UI 设计

HIS 子系统服务提供者配置文件 provide.xml 核心

内容如下, 其中包括暴露的服务接口及注册地址:

```
<dubbo:application logger="slf4j"
  name="${dubbo.application-name}"/>
<dubbo:protocol name="dubbo" charset="UTF-8"
  port="${dubbo.port}" serialization="java"/>
<dubbo:registry username="root"
  address="${dubbo.registry-address}"/>
<!-- 声明需要暴露的服务接口 -->
<dubbo:service version="${dubbo.service-version}"
  ref="hisRemoteService"
  interface="com.heren.ois.rpc.
  hisInterface.service.HisRemoteService"/>
```

服务消费者配置文件 consumer.xml 核心内容如下, 同样包含其注册地址等信息:

```
<!-- 服务接口 -->
<dubbo:reference id="hisRemoteService"
  version="${dubbo.service-version}"
  interface="com.heren.ois.rpc.
  hisInterface.service.HisRemoteService"/>
```

接口 HisRemoteService 定义如下, 其中包含根据病历号或姓名性别查找患者、根据患者查找电子病历和诊断结果等方法:

```
public interface HisRemoteService {
  Patient findPatientBy(String MedicalNo);
  List<Patient> findListByNameAndGender(
    String name, Integer gender
  );
  Map<String, List<MedicalRecord>>
  findMedicalRecordListBy(Patient patient);
  Map<String, List<Diagnosis>>
    findDiagnosisListBy(Patient patient);
}
```

系统用雷达图展示各指标危险临界点与自身指标情况; 用饼图展示糖尿病患者某指标异常的比例; 最后辅以诊断分析和医嘱建议等文字, 效果如图 5 所示.

3.5 预诊结果分析

截取部分电子病历的核心数据, 如图 6 所示, 其中年龄、病史、BMI、WHR、吸烟、饮酒、过度饮食、运动量达标为八项相关因素.

将八项高危因素按照 3.1 节规则量化, 得到如图 7 所示的项 A 到项 H. 运算得出结果即项 I, 可以发现与真实诊断结果无异.

图 5 糖尿病预诊分析界面

性别	年龄	高血压高血脂病史	bmi	whr	吸烟	饮酒	过度饮食	运动量达标	患糖尿病
男	32	无	25.1	0.74	是	否	否	是	否
女	35	有	28.2	0.82	否	否	是	否	是
男	46	无	32.2	0.91	是	是	是	否	是
男	29	有	24.6	0.81	否	是	否	是	否
女	24	无	23.5	0.71	否	否	是	是	否

图 6 电子病历部分数据

A	B	C	D	F	F	G	H	I
0	0	0	0	1	0	0	1	0
0	1	1	1	0	0	1	0	1
1	0	1	1	1	1	1	0	1
0	1	0	0	0	1	0	1	0
0	0	0	0	0	0	1	1	0

图 7 量化后的数据

4 结束语

针对 Apriori 经典算法存在的缺陷, 本研究进行了改进, 并应用于糖尿病与其高危因素间的关联规则挖掘. 通过实验对算法进行对比, 结果表明改进 Apriori 算法性能得到了大幅度提高. 基于以上工作, 本研究设计了一款糖尿病预诊分析系统, 随着挖掘样本数量的逐步增加其准确率也逐步提升. 此系统为用户自诊和医护人员辅助诊断提供了更加便捷的方式.

参考文献

1 Yıldırım EG, Karahoca A, Uçar T. Dosage planning for diabetes patients using data mining methods. Procedia Computer Science, 2011, 3: 1374–1380. [doi: 10.1016/j.procs.2011.01.018]

2 李武成, 王官权, 金科. 2 型糖尿病并发高血压的危险因素分析. 实用医学杂志, 2010, 26(17): 3180–3181. [doi: 10.3969/j.issn.1006-5725.2010.17.045]

3 董宁. 基于数据挖掘的 Apriori 算法研究与改进. 自动化与仪器仪表, 2016, (9): 232–234.

4 张伟科. 一种改进的 AprioriTid 算法. 沈阳工业大学学报, 2016, 38(3): 314–318.

5 苗苗苗, 王玉英. 基于矩阵压缩的 Apriori 算法改进的研究. 计算机工程与应用, 2013, 49(1): 159–162. [doi: 10.3778/j.issn.1002-8331.1107-0579]
6 王蒙, 邹书蓉, 方睿. 一种基于矩阵的 Apriori 改进算法. 信息技术, 2018, (3): 150–154, 158.
7 周发超, 王志坚, 叶枫, 等. 关联规则挖掘算法 Apriori 的研究改进. 计算机科学与探索, 2015, 9(9): 1075–1083.
8 李超, 余昭平. 基于矩阵的 Apriori 算法改进. 计算机工程, 2006, 32(23): 68–69. [doi: 10.3969/j.issn.1000-3428.2006.23.024]
9 杨志刚, 何月顺. 基于压缩事务矩阵相乘的 Apriori 改进算法. 中国新技术新产品, 2010, (6): 57–58. [doi: 10.3969/j.issn.1673-9957.2010.06.045]
10 冯玉欣, 赵希兵. 2 型糖尿病家系发病特征与危险因素调查. 糖尿病新世界, 2017, 20(12): 19–20.
11 韦哲, 叶广健. 一种 Apriori 改进算法在 2 型糖尿病危险因素分析中的应用. 电子测试, 2015, (17): 63–65, 84. [doi: 10.3969/j.issn.1000-8519.2015.17.026]
12 陈晓栋. 基于 Dubbo 分布式框架的信用卡无卡大额分期系统设计. 信息与电脑, 2017, (7): 132–135. [doi: 10.3969/j.issn.1003-9767.2017.07.051]
13 温晓丽, 苏浩伟, 陈欢, 等. 基于 SpringBoot 微服务架构的城市一卡通手机充值支撑系统研究. 电子产品世界, 2017, 24(10): 59–62.
14 鱼朝伟, 詹舒波. 基于 RabbitMQ 的异步全双工消息总线的实现. 软件, 2016, 37(2): 139–146. [doi: 10.3969/j.issn.1003-6970.2016.02.032]
15 曾超宇, 李金香. Redis 在高速缓存系统中的应用. 微型机与应用, 2013, 32(12): 11–13. [doi: 10.3969/j.issn.1674-7720.2013.12.004]