**班级 011214**

**学号 01121390**

**西安电子科技大学**

本科毕业设计论文



**题 目**  **基于位置社交网络中的好友**

**推荐系统开发**

**学 院**  通信工程学院

**专 业**  通信工程

**学生姓名** 王长通

**导师姓名** 朱晓妍

毕业设计(论文)诚信声明书

本人声明：本人所提交的毕业论文《先进无线（光）通信试验室主页设计与实现》是本人在指导教师指导下独立研究、写作的成果，论文中所引用他人的无论何种方式发布的文字、研究成果，均在论文中加以说明；有关教师、同学和其他工作人员对本文的写作、修订提出过并为我在论文中加以采纳的意见、建议，均已在我的致谢词中加以说明并深致谢意。

本论文和资料若有不实之处，本人承担一切相关责任。

论 文 作 者： （签字） 时间：2016年5月23日

指导教师已阅： （签字） 时间：2016年5月23日

**毕业设计（论文）任务书**

学生姓名 王长通 学号 01121390 指导教师 朱晓妍 职称 副教授

学院 通信工程学院 专业 通信工程

题目名称 基于位置社交网络中的好友推荐系统开发

任务和要求

1. 了解国内外相关推荐算法，并完成推荐系统的编写；
2. 阅读和翻译不少于10000单词的相关外文文献；
3. 独立完成毕业论文工作，论文字数应不少于15000；
4. 要求论文结构合理，概念清楚，逻辑清晰，语言通顺以及文笔流畅；
5. 工作作风严谨，刻苦钻研，每周与指导老师至少一次沟通联系一次；
6. 按时参加论文答辩工作。

开始日期 2016年3月6日 完成日期 2016年5月23日

院长（签字） 2012年1月10日

毕业设计（论文）工作计划

学生姓名 王长通 学 号 01121390

指导教师 朱晓妍 职 称 副教授

学 院 通信工程学院 专 业 通信工程

题目名称 基于位置社交网络中的好友推荐系统开发

1. 毕业设计（论文）进度

|  |  |
| --- | --- |
| 起止时间 | 工作内容 |
| 2016.3.1-2016.4.1  2016.4.2-2016.4.9  2016.4.10-2016.5.15  2016.5.16-2016.5.23 | 准备相关参考资料，了解相关推荐算法，选出适合该系统的算法。学习android开发，为编写系统做准备。  找寻地理位置相关数据集，并对数据进行预处理。  具体代码编写，将推荐算法在系统中实现。研究不同算法效率，提出改进算法。  完善系统，撰写毕业论文，准备答辩。 |

1. 主要参考书目（资料）

[1] 詹尼士，《推荐系统》 人民邮电出版社，2013.

[2] 查九, 李振博, 徐桂琼. 基于组合相似度的优化协同过滤算法[J]. 计算机应用

与软件, 2014, 323-328.

[3] 李刚，《疯狂Android讲义》 中国工业出版社，2015.

[4] Liao, Ll, Patterson, D.J., H. Building Personal Maps from GPS Data. In proc. of *IJCAI MOO05*, Springer Press(2005), 249-265

1. 主要仪器设备及材料

硬件：计算机、Andriod手机、Internet等

软件：Eclipse、SQLite等

四、教师的指导安排情况（场地安排、指导方式等）

1、每周集中汇报、指导一次；

2、采用电话、邮件随时联系指导。

西安电子科技大学毕业设计（论文）中期检查表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **学 院** | 通信工程学院 | | **专 业** | 通信工程 | |
| **学生姓名** | 王长通 | **学 号** | 01121390 | **班 级** | 011214 |
| **导师姓名** | 朱晓妍 | **职 称** | 副教授 | **单 位** |  |
| **题目名称** | 基于位置社交网络中的好友推荐系统开发 | | | | |
| **检 查 内 容** | | **检 查 结 果** | | | |
| **题目是否更换及更换原因** | |  | | | |
| **学生出勤情况** | |  | | | |
| **进度评价 （完成总工作量的百分比）** | |  | | | |
| **质量评价、进度描述** | |  | | | |
| **总体评价 （按优、良、中、及格、不及格五挡评价）** | |  | | | |
| **存在的问题与建议** | |  | | | |
| **学院审核（盖章）** | |  | | | |

西 安 电 子 科 技 大 学

毕业设计（论文）成绩登记表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 学 院 | 通信工程学院 | | | 专 业 | 通信工程 | |
| 姓 名 | 王长通 | 学 号 | 01121390 | 成 绩 |  | |
| 题目名称 | 基于位置社交网络中的好友推荐系统开发 | | | | | |
| 指导教师 | 朱晓妍 |  | | 职 称 | 副教授 |  |
| 指导教师评语及对成绩的评定意见 | 签名 2012年6月16日 | | | | | |

|  |  |
| --- | --- |
| 评阅人评语及成绩评定意见 | 签名 2012年6月16日 |
| 答辩小组意见 | 签名 2012年6月16日 |
| 学院答辩委员会意见 | 签名 2012年6月16日 |

# **摘 要**

随着物质生活的不断丰富，人们在精神上的需求日益增加。能结识到情投意合的朋友也是精神世界的重要组成部分，但这在线下生活中往往是可欲而不可求得，那么帮助用户发现潜在好友，并向其进行推荐就成为设计本系统的主要目的。 我们知道兴趣类似的人们一般会出现在相似的场所，如音乐爱好者一般会经常出没KTV，篮球爱好者必然会经常出没于球场等等。本系统通过获取用户地理位置信息来推测用户兴趣进而挖掘用户相似度，并进行排序和推荐。智能手机的出现极大的方便了获取用户地理位置的环节，提高了该系统的可行性，该系统使用Android手机作为开发平台。为了数据的真实性和研究该系统所使用推荐算法的有效性，我们取得了主打签到功能的移动社交网络Brightkite的真实用户数据，将其导入到SQLite数据库中提高运算效率，并用其中真实的好友关系对不同算法的推荐结果进行评估。

本文首先介绍了本课题的研究背景和意义，然后介绍了协同过滤的概念和其中具体推荐算法的种类，如：Jaccard相似度算法，ConsineSimilar余弦相似度算法等。随后介绍了以上算法在本课题系统中的具体实现方法，并通过计算得出各种算法在该数据集中的表现，通过分析提出优化后的算法。最后是对本课题的展望与总结、本文的不足之处和对本课题可能存在的创新点。

**关键词： 相似度算法 位置信息 Android 数据库**

# ABSTRACT

With the wealth of material life, people's spiritual needs are increasing. To get to know an important part of the spiritual world is like-minded friends, but offline life is often desirable and can not be obtained, then the potential to help users find friends, and to make recommendations which would be the main purpose of the design of the system. We know that people with similar interests generally occur in similar places, such as music lovers usually frequented KTV, basketball fans will inevitably haunt the court and so on. The system by obtaining user location information to speculate further mining user interest similarity, sorted and recommendation. Smart phones which greatly facilitates the user to obtain location links and improve the feasibility of the system, which uses your Android phone as a development platform. In order to study the data the authenticity and validity of the system using the recommended algorithm, we have made mobile social network Brightkite main sign of real user data capabilities, import it into SQLite database to improve operational efficiency, with which the real relationship between friends recommended result of different algorithms for evaluation.

This paper introduces the background and significance of this issue , and then introduced the recommendation algorithm related species , such as : Jaccard algorithm , ConsineSimilar algorithm. Then the algorithm described above in the subject system implementation methods, and by calculation algorithms performance in the data set , the analysis put forward by the optimized algorithm . Finally, on the subject of the outlook and summary , deficiencies and innovation on this topic may be present in this paper .

**Keywords : similarity algorithm location information Android**

**Database**

# 目 录

**第一章 绪论**............................................................................................................................ 1

1.1

1.2

# 第一章 绪 论

1.1 推荐系统研究的背景和意义

随着互联网的不断发展和生活节奏的加快，人们产生了越来越多的信息，这些信息往往是破碎的，混乱的。从这些零散的信息中挖掘出有价值的信息，会极大方便人们的日常生活，并获得巨大的商业价值，所以数据挖掘作为一门挖掘互联网信息的学科变得越来越引人关注。推荐系统便是在数据挖掘的基础上运行的，它成为一个相对独立的研究方向一般被认为始自1994年明尼苏达大学GroupLens 研究组推出的GroupLens系统。该系统有两大重要贡献：一是首次提出了基于协同过滤来完成推荐任务的思想，二是为推荐问题建立了一个形式化的模型。基于该模型的协同过滤推荐引领了之后推荐系统在今后十几年的发展方向。GroupLens所出的推荐算法实际上就是目前广泛使用的基于用户的协同过滤推荐算法。在之后的十几年中，其它一些著名的协同过滤算法被陆续提出，主要的有基于物品的协同过滤算法，基于矩阵分解的协同过滤算法等等。当然，基于其它方法而非协同过滤的推荐算法也在不断地发展，这些不同的推荐算法之间的取长补短、相互融合也成为一个重要的研究方向。

通过推荐算法进行好友推荐的系统极大的提高了用户认识潜在好友的效率，提供了难得的兴趣交友渠道。研究并实现推荐系统，使科技进步和算法发展直接可以提高人们的生活质量。

1.2 推荐系统的发展现状

推荐算法作为一门技术广泛应用在电子商户、大规模零售业和各种知识管理应用中。与此同时，随着各个学科研究人员的参与及方法的引入，推荐系统的研究得到迅猛发展。来自人工智能、信息检索、数据挖掘、安全与隐患以及商业与营销等各个领域的研究，都在为推荐系统提供了新的分析和方法。由于可以获得海量的数据，算法研究方面取得了巨大进步。

同时推荐系统的商业化也在不断的推进，在巨大的商业价值的推动下，各种有效的推荐算法层出不穷，大多数的电子商务平台尤其是网络购物平台，都不同

程度地集成了推荐算法。比较著名的有Facebook根据好友关系向用户推荐可能认识的好友、Amazon网络购物商城向用户推荐已购买货物周边商品、Google新闻个性化推荐等。国内如淘宝网会基于近期用户搜索向用户推荐商品、豆瓣网会根据用户近期的浏览条目推荐类似的内容、滴滴打车会根据用户和司机的地理位置推荐合适的司机等等。基于的信息Amazon 发布的数据显示，亚马逊网络书城的推荐算法为亚马逊每年贡献近三十个百分点的创收。

1.3 主要研究内容

本文介绍的系统基于用户所签到的地理位置信息进行数据进行挖掘，并计算用户间地理位置的相似度大小，向用户推荐潜在的好友。系统采用真实的数据集和好友关系，最终用三种方法实现好友推荐，在实现推荐后对每种算法得出的结果进行评价，并分析出每种算法在该数据集中表现出的特点，并提出作者自己的见解。

# 第二章 协同过滤和算法介绍

2.1 协同过滤

2.1.1 协同过滤的概念及一般步骤

在用户群过去的行为或意见的基础上，通过算法找出指定用户兴趣相投、拥有共同经验的用户群体，之后可以用该用户群体的喜好来推荐指定用户感兴趣的信息。

（1）收集可以代表用户兴趣的信息。包括影评网站使用评分的方式让用户对不同电影打分、本系统中用户在不同的地点签到，本这种方式被称为“主动评分”。另外一种是“被动评分”，是指系统根据用户的行为来代替用户完成评价，不需要用户自己录入或输入数据。比如电子商务网站根据购买记录向用户推荐可能需要的商品，搜索引擎根据用户的历史搜索记录向用户推荐可能感兴趣的新闻等。

（2）通过算法计算该用户和其他用户的相似度，找到其中相似度较高的其他用户作为最近邻集合。这一步骤有不同的算法实现，根据数据的不同、场景的不同它们的准确度各不相同，目前较多使用的相似度算法有Jaccard相似度、余弦相似度、Pearson相似度等。

（3）有了最近邻集合，就可以根据该集合中用户的历史数据对目标用户的兴趣进行预测，产生推荐结果。

2.1.2 优点

（1）能够过滤机器难以自动内容分析的信息，如艺术品，音乐等。

（2）共用其他人的经验，避免了内容分析的不完全或不精确，并且能够基于一些复杂的，难以表述的概念（如信息质量、个人品味）进行过滤。

（3）有推荐新信息的能力。可以发现内容上完全不相似的信息，用户对推荐信息的内容事先是预料不到的。可以发现用户潜在的但自己尚未发现的兴趣偏好。

（4）推荐个性化、自动化程度高。能够有效的利用其他相似用户的回馈信息。加快个性化学习的速度。

2.1.3 缺点

（1）新用户问题系统开始时推荐质量较差，即冷启动问题。

（2）新项目问题质量取决于历史数据集，对最新数据的增添时间不敏感。

（3）稀疏性问题，数据间联系不密切。

（4）系统延伸性问题，扩展性不好。

2.2 Jaccard相关系数

Jaccard相关系数强调两个用户数据并集的损益，用来比较样本集中的相似性和分散性的一个概率，其值等于样本交集与样本集并集的比值。给定用户u和用户v，其中N(u)表示用户u的评分项集合，N(v)表示用户v的评分项集合[1]。可通过公式（2-1）计算用户u和v的兴趣相似度。

（2-1）

Jaccard相关系数越接近1，两个集合重合度越高，两个数据集越相似。越接近0，两个集合重合度越低，两个数据集越不同。

2.3 ConsineSimilar余弦相关系数

ConsineSimilar余弦相关系数强调两个用户数据的方向一致性。余弦相关系数用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小。用户对 n个数据的评分可视为 n维向量，用户u和用户v的相似性即为相应两个n维 向量的夹角余弦[1]，可通过公式（2-2）计算用户u和v的兴趣相似度。

（2-2）

ConsineSimilar余弦相关系数越接近1，两个向量方向越一致，两个数据集越相似。越接近0，两个向量方向越分歧，两个数据集越不同。

2.4 Pearson相关系数

Pearson相关系数强调两个用户数据一同变化移动的相同趋势。表示用户u的平均评分，表示用户v的平均评分[1]，可通过公式（2-3）计算用户u和v的兴趣相似度。

（2-3）

Pearson相关系数越接近1，两种变化趋势越相关，两个数据集越相似。越接近

-1，两种变化趋势越矛盾，两个数据集越不同。

2.5 其他推荐算法

除以上三种之外，其他的协同过滤算法还有很多，如基于云模型的协同过滤推荐算法[2]，基于矩阵划分和兴趣方差的协同过滤算法[3]，两阶段联合聚类协同过滤算法[4]等。

# 第三章 好友推荐系统实现

3.1 实现平台

3.1.1 Android系统现状

移动互联网热潮在全世界引起了巨大反响，移动互联网正改变着传统互联网格局，全世界的IT公司争相将业务重心向移动端转型。在这股浪潮之中Android系统也得到迅猛发展，它已经成为全球范围内最普及、占有率最高的移动操作系统。它具有集成硬件齐全、系统人性化、手机价格实惠、代码开源等优点。

2016年第一季度，Android在中国的占有率有77.7%，其主要竞争对手[iOS](https://zh.wikipedia.org/wiki/%E8%8B%B9%E6%9E%9C%E5%85%AC%E5%8F%B8)约占21.1%，在美国的占有率亦有65.5，稳稳的坐上智能手机系统的王者地位。Android系统已经不仅仅是一款手机的操作系统，它越来越广泛的被应用于平板电脑、可佩戴设备、电视、数码相机等设备上，这也造就了目前Android系统软件需求的快速增长，从大趋势上看，Android各种系统软件的需求将越来越大。

3.1.2 Android系统架构

Android系统的底层是建立在Linux系统之上的，它采用软件叠层（Software Stack）的方式进行构建。使得层与层之间相互分离，明确各层的分工。这种分工保证了层与层之间的低苟合，当下层发生改变的时候，上层应用程序无需做任何改变，图3.1为Android系统架构图。

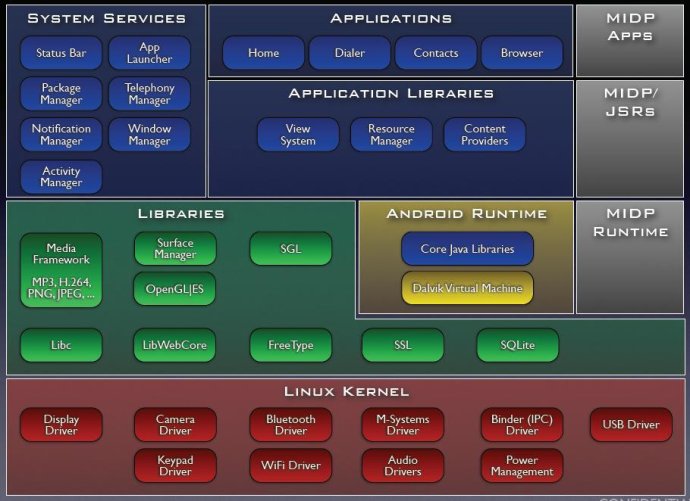


图3.1 Android系统架构图

（1）应用程序层（Application）：Android系统包含了一系列核心应用程序，包括电子邮件、短信SMS、日历、拨号器、地图、浏览器、联系人等。这些应用程序都是用Java语言编写。本书重点讲解如何编写Android系统上运行的应用程序，在程序分层上，与系统核心应用程序平级。

（2）应用程序框架层（Application Framework）：Android应用程序框架提供了大量的API供开发人员使用，Android应用程序的开发，就是调用这些API，根据需求实现功能。应用程序框架是应用程序的基础。为了软件的复用，任何一个应用程序都可以开发Android系统的功能模块，只要发布的时候遵循应用程序框架的规范，其它应用程序也可以使用这个功能模块。

（3）系统运行库层（Libraries）：Android系统运行库是用C/C++语言编写的，是一套被不同组件所使用的函数库组成的集合。一般来说，Android应用开发者无法直接调用这套函数库，都是通过它上层的应用程序框架提供的API来对这些函数库进行调用。

（4）Android运行时：Android运行时由两部分完成：Android核心库和Dalvik虚拟机。其中核心库集提供了Java语言核心库所能使用的绝大部分功能，Dalvik虚拟机负责运行Android应用程序。虽然Android应用程序通过Java语言编写，而每个Java程序都会在Java虚拟机JVM内运行，但是Android系统毕竟是运行在移动设备上的，由于硬件的限制， Android应用程序并不使用Java的虚拟机JVM来运行程序，而是使用自己独立的虚拟机Dalvik VM，它针对多个同时高效运行的虚拟机进行了优化。每个Android应用程序都运行在单独的一个Dalvik虚拟机内，因此Android系统可以方便对应用程序进行隔离。

（5）Linux内核：Android系统是基于Linux2.6之上建立的操作系统，它的Linux内核为Android系统提供了安全性、内存管理、进程管理、网络协议栈、驱动模型等核心系统服务。Linux内核帮助Android系统实现了底层硬件与上层软件之间的抽象。

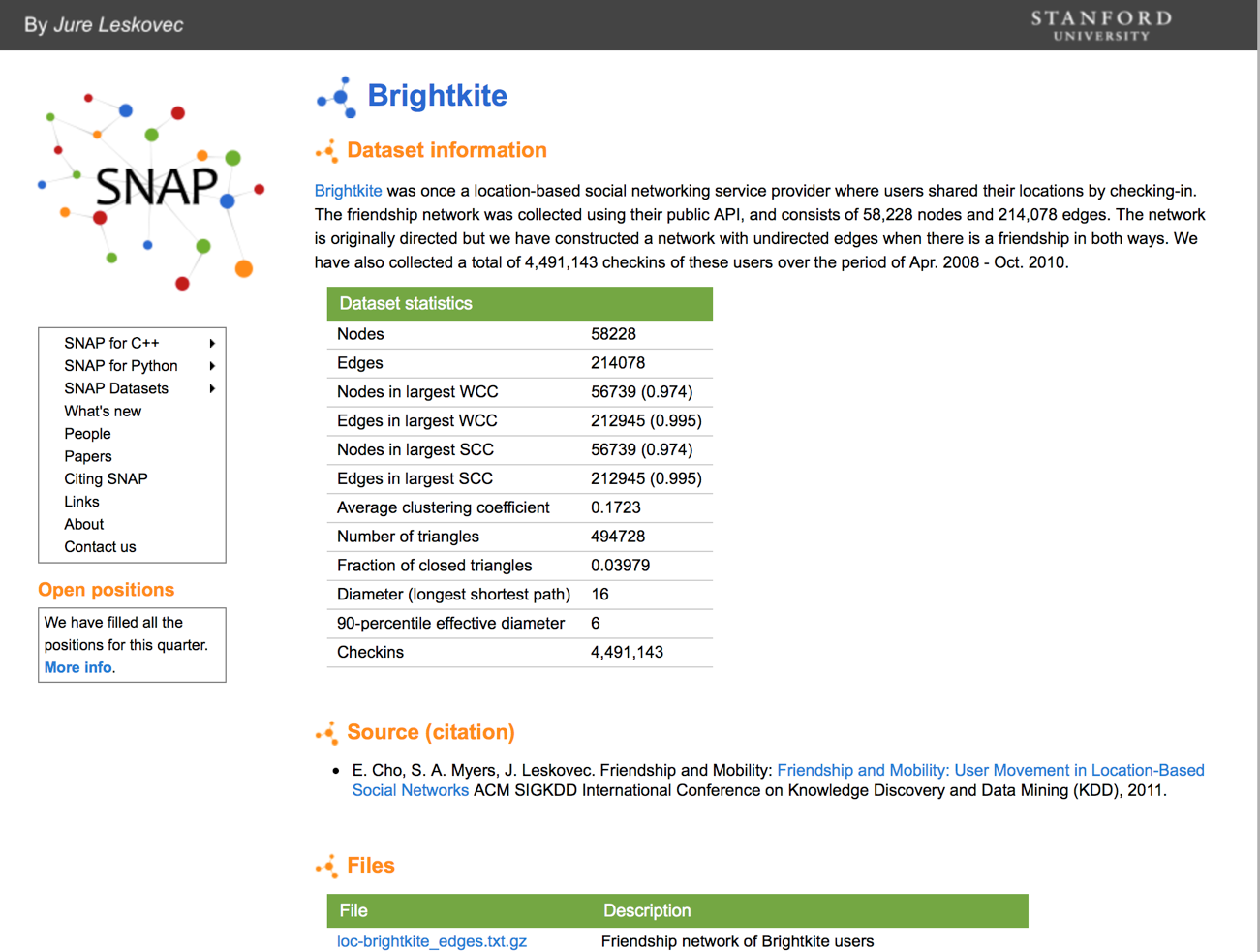
3.1.3 Android程序开发

Android系统基于面向对象的Java语言开发，Java语言作为服务器端编程的语言已经取得了极大的成功，JavaEE平台发展得非常成熟，而且一直是电信、移动、银行、证券、电子商务应用的首选平台。但在客户端应用开发方面，Java语言一直表现不佳，虽然Java既有AWT/Swing界面开发库，也有SWT/JFace界面开发库。但对于客户端应用开发人员而言，大多不愿意使用Java语言。Andorid系统的出现改变了这种局面，Java面向对象语言的特性将给Android开发带来极大魅力，Java也将在客户端应用开发上大展拳脚。

3.2 数据集预处理

3.2.2 数据集获取

数据集的真实性能暴露出系统运行时可能出现的问题，还有助于我们对不同算法推荐结果进行评估。在斯坦福网络平台上得到了Brightkite社交网络的真实用户数据集。



3.2.3 数据集处理

（1）去除无效位置数据，如图3.2中的第二行。

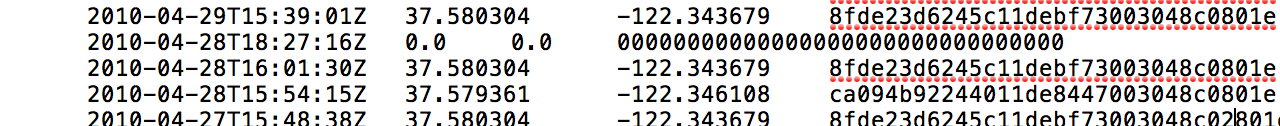


图3.2 无效用户数据

通过代码将为0的数据舍去，程序如图3.3。

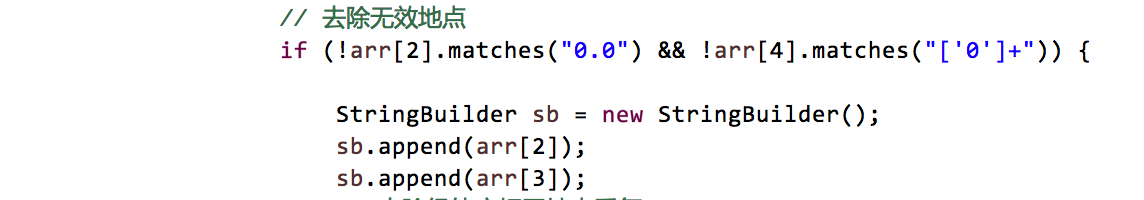


图3.3 舍去为0数据

（2）将地理代码进行编号，便于直观观察并提高系统运行效率，如图3.4。

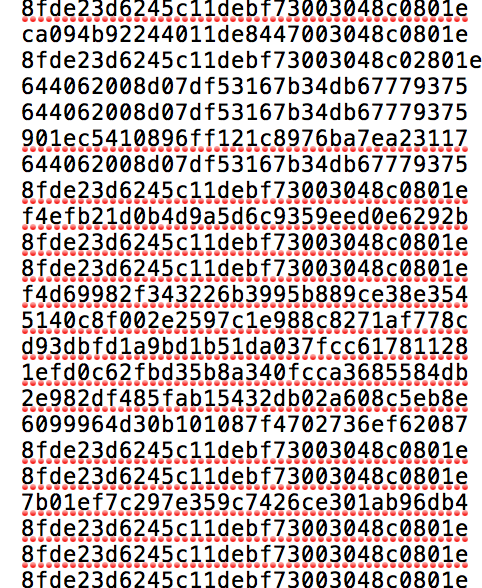


图3.4 繁杂的地理编码

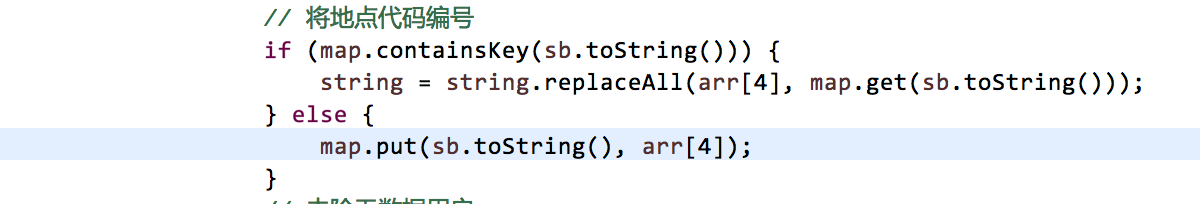
若地理位置编码已出现，返回第一次出现时的编号；若地理位置未出现，储存该编码并返回新编号，程序如图3.5。

图3.5 将繁杂的编码转换为直观的编号

（3）排除没有数据的用户，程序如图3.6。

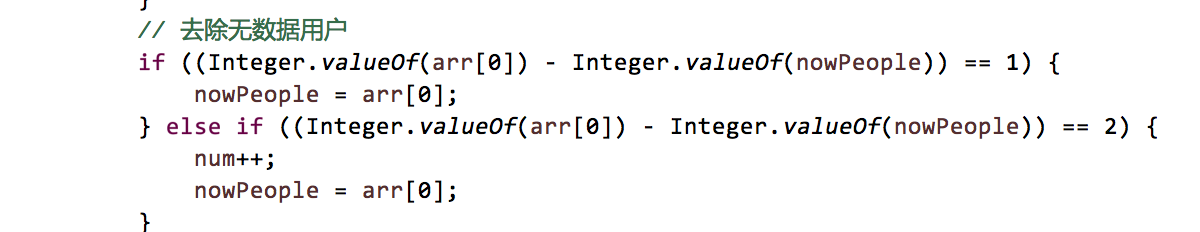


图3.6 排除签到次数为0的用户

处理后的数据集就可以进行下一步，存入数据库。

3.3 SQLite数据库

3.3.1 SQLite数据库介绍

SQLite是一个开源嵌入式的关系型数据库，最先发布于2000年，它提供给应用程序一个方便的方式去管理数据，而不用去负担专用的大型的数据库管理系统。SQLite由于高度轻便，易于使用，紧凑简洁，高效和稳定可信赖，正变得越来越流行[5]。SQLite遵循SQL语言并支持多种事物。Android中集成有SQLite数据库，它的特点是存储在本地，方便程序快速高效的读取等。

3.3.2 数据库中表的设计

根据系统功能需求，本系统设计了5个数据表。分别是：

（1）AllData：存储用户的全部地点数据

（2）FriendData：存储用户的好友关系

（3）PeopleCountData：存储用户在不同地点的签到次数

（4）PeopleData：存储用户的总签到次数

（5）ConsequenceData：存储最终算法评价，方便查看

3.3.3 存入数据库

将文本格式的数据集存入SQLite数据库中以方便读写，以AllData存入过程为例，数据以空格为间隔被切分存入，程序如图3.7。

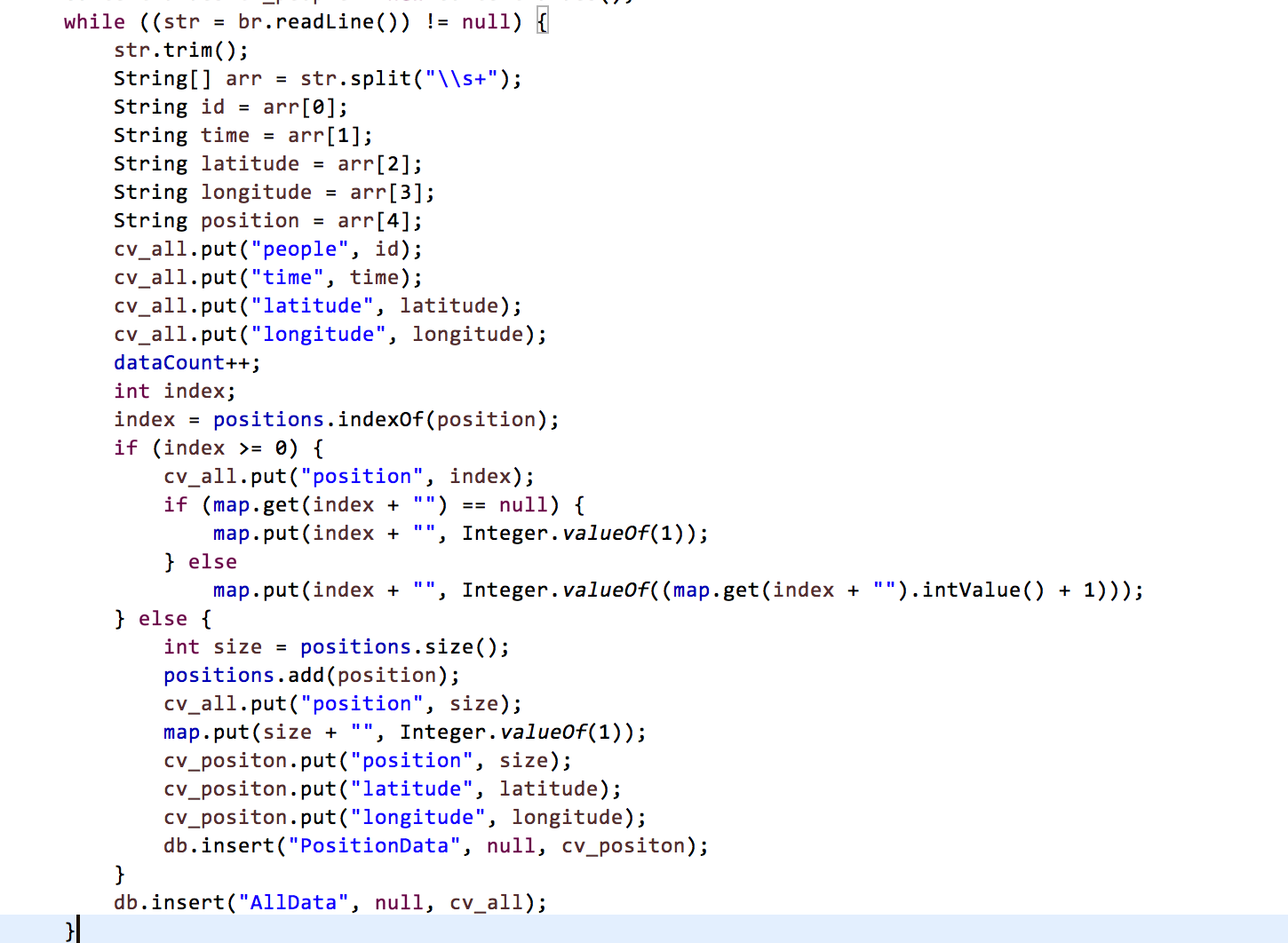


图3.7 AllData录入代码

3.4 算法实现

3.4.1 计算Jaccard相关系数

程序实现如图3.8。



图3.8 Jaccard相关系数计算

首先计算登入用户签到总数，再计算与其他用户签到总数之和、与其他用户相同的签到总数，最后通过相除关系得到结果。

3.4.1 计算Cosine余弦相关系数

程序实现如图3.9。

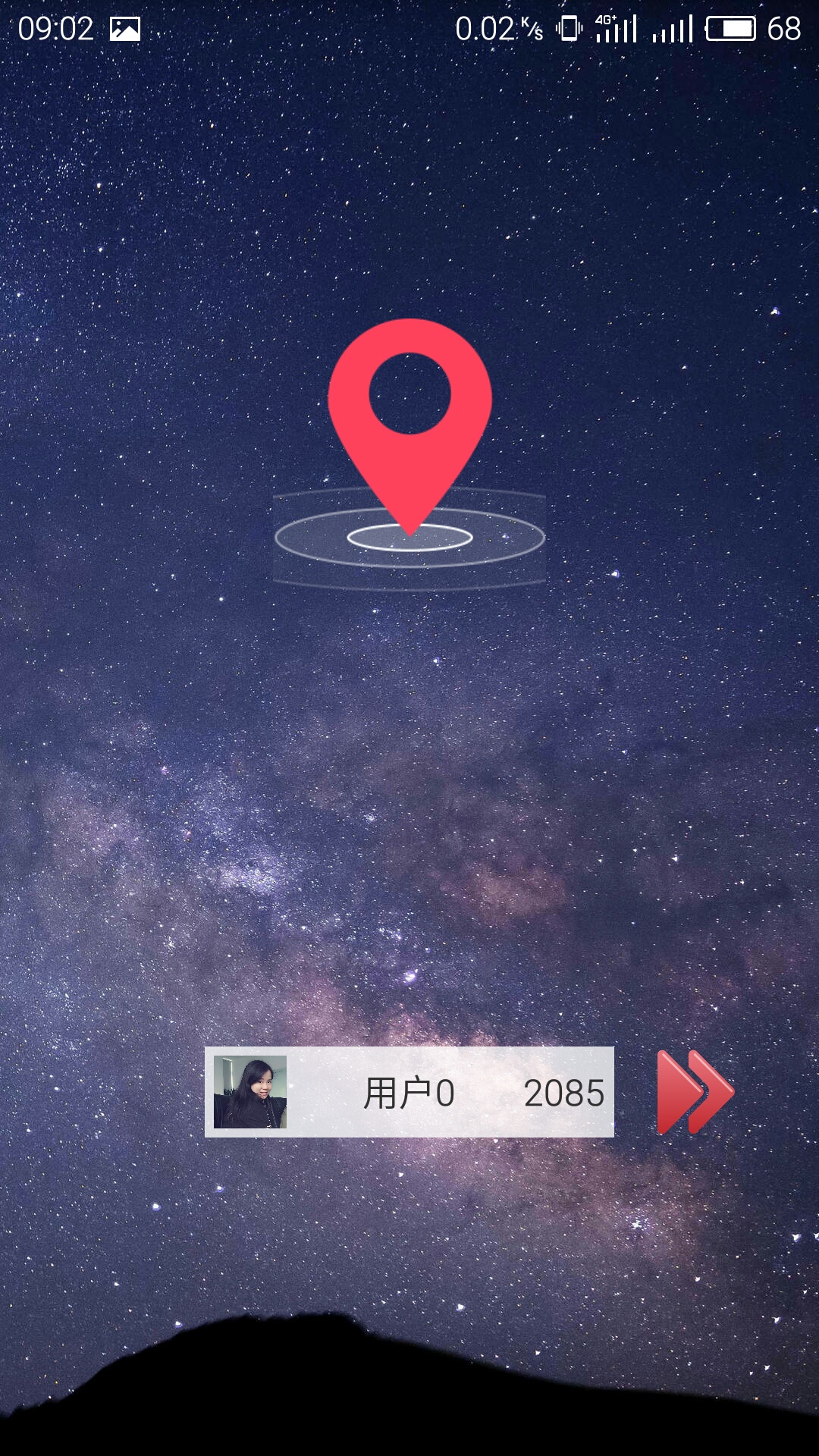


图3.9 Cosine余弦相关系数计算

首先获取登入用户地点签到表，用Map类型对象存储，与其他用户比对后进行添加或修改。最后用Iterator遍历该表，计算出余弦相关系数公式的分子和分母，相除运算后得到结果。

3.5 系统界面

系统界面主要分为登陆界面和好友推荐界面。登陆界面显示所有用户已签到地点数量来提供参考，使用者可以选择任意用户并进行登陆。如图3.10

  
图3.10 登陆界面

登陆后进入好友推荐界面，上方指示器提示用户当时所使用算法，该界面用三种算法实现好友推荐，分别为Jaccard、Cosine和作者提出的混合Mix算法，好友推荐结果根据相似度的大小进行由大到小依次排序，并显示用户名称和相似度具体数值方便分析，如图3.11。

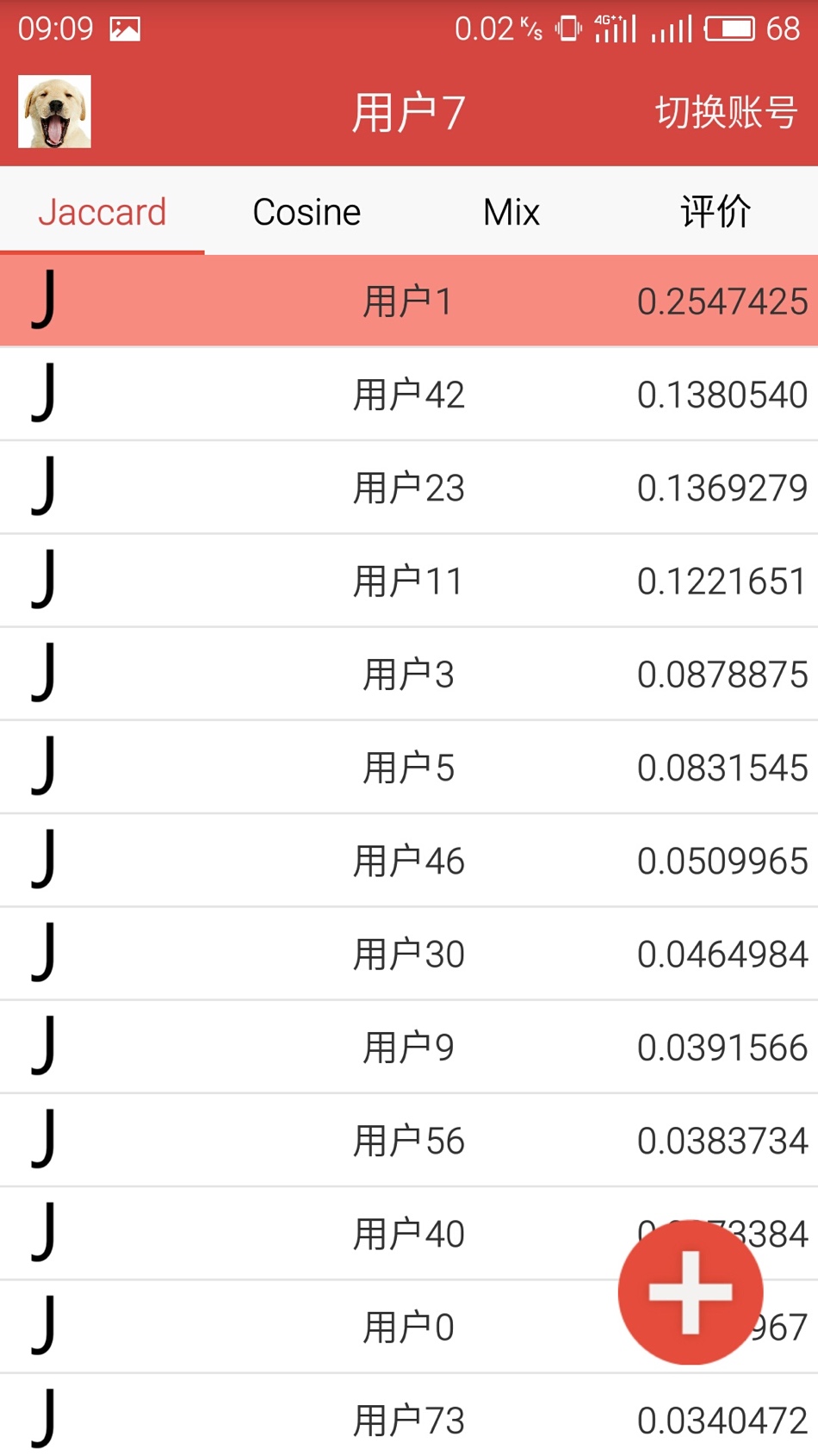


图3.11 好友推荐界面

这里我们选取用户7作为登陆用户，可见在Jaccard算法中，用户1和用户7的相似系数最高，达到了0.254725，系统将首先向用户7推荐用户1。排名第二的是用户42，Jaccard相似系数达到了0.1380540，系统将用户42排在第二个推荐位，以此类推。基于用户好友关系数据集，也就是FriendData数据表中的数据，好友推荐界面同时对三种算法进行了评估，如图3.12。

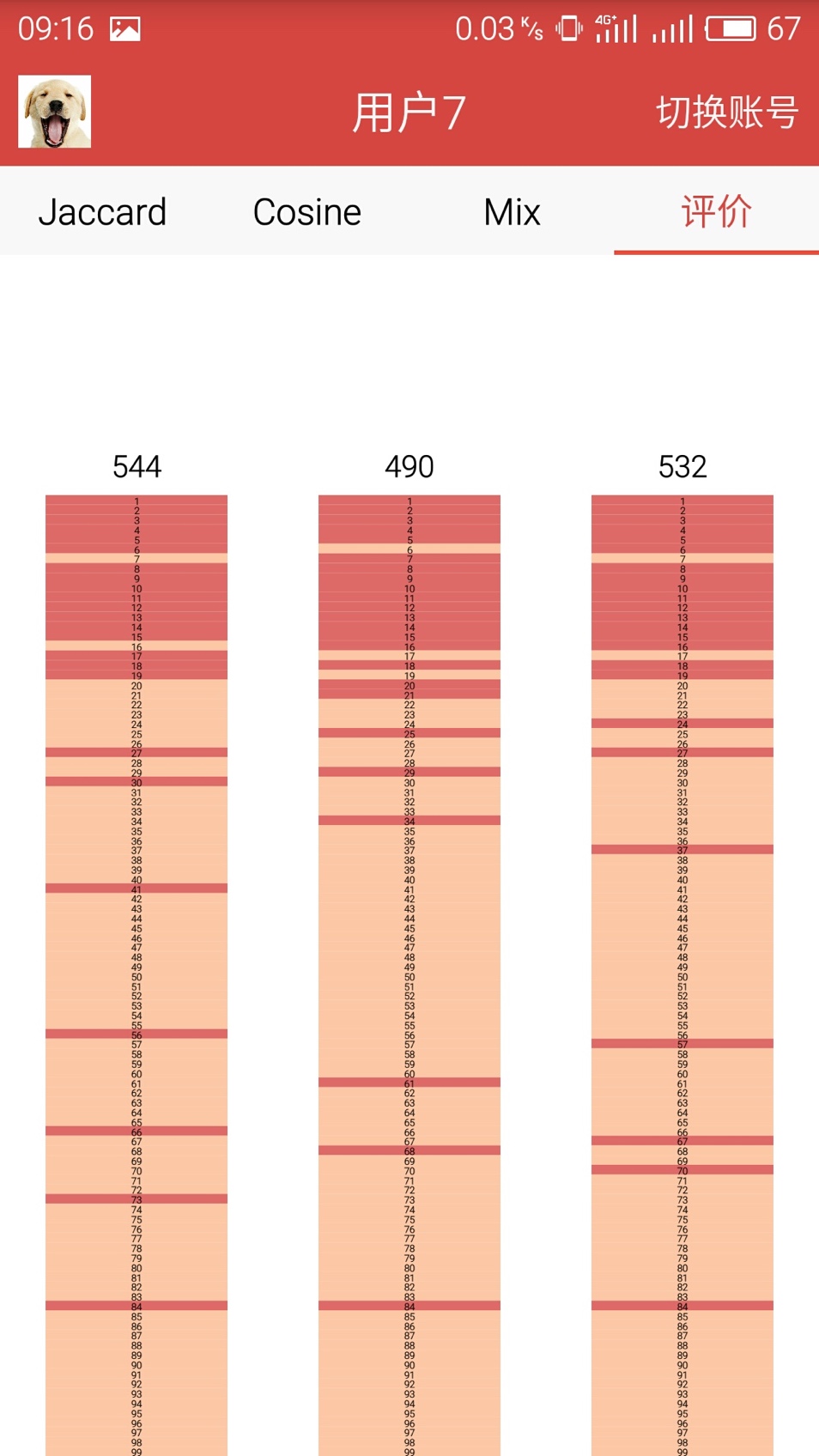


图3.12 算法结果评估

从左至右三个柱形统计图分别代表Jaccrad相关系数，Cosine相关系数，Mix相关系数推荐结果。每一个柱形图右99个线段组成，代表其他99位用户。数字1至99代表系统推荐顺序，对于用户7，线段1即代表用户1，线段2即代表用户42。红色线段代表该用户是用户7的真实好友，推荐成功；橙色线段代表该用户不是用户7的真实好友，推荐失败。所以红色线段越靠近上方说明好友的推荐顺位越高，推荐效果越好。每个柱形图上方的数字代表该用户真实好友的系统推荐顺位之和，该数字越小表示推荐越准确。对于用户7，显然Cosine相关系数算法推荐更为准确。

使用者可通过点击右上角切换用户，完成用户切换，查看其它用户推荐结果。

# 第四章 算法分析和改进

4.1 从公式上分析

Jaccrad相关系数是两个数据集重合部分和两个数据并集的比值，见式（2.1）。优点是杜绝某一用户由于数据量过多导致在其他用户推荐靠前的现象，但是缺点也很明显，当其中一个用户的数据量较少而另外一个用户数据量较多时，Jaccard相关系数结果往往会很低，但这两个用户很可能有潜在的好友关系，造成推荐误差。

Cosine余弦相关系数则很好的补足了这一点，它只关注用户数据向量的方向一致性，见式（2.2）。有点是当某一用户数据量数过少时，仍可以向其推荐有类似数据量向量方向的用户。缺点是它忽略了用户数据量总数，当用户数据量过多时，Cosine余弦相关系数往往偏小，但此时有可能两个用户数据量重合度并不低，造成推荐误差。

4.2 系统数据集验证

为了验证我们的判断，对该数据集上的所有用户的两种算法推荐进行优劣判断，计算两种算法推荐结果分别胜出时，登陆用户的平均数据量。来获取数据量对算法结果的影响。根据之前的判断，当用户数据量较多时，Jaccrad相关系数占优；当用户数据量较少时，Cosine余弦相关系数占优，实际结果如图4.1。

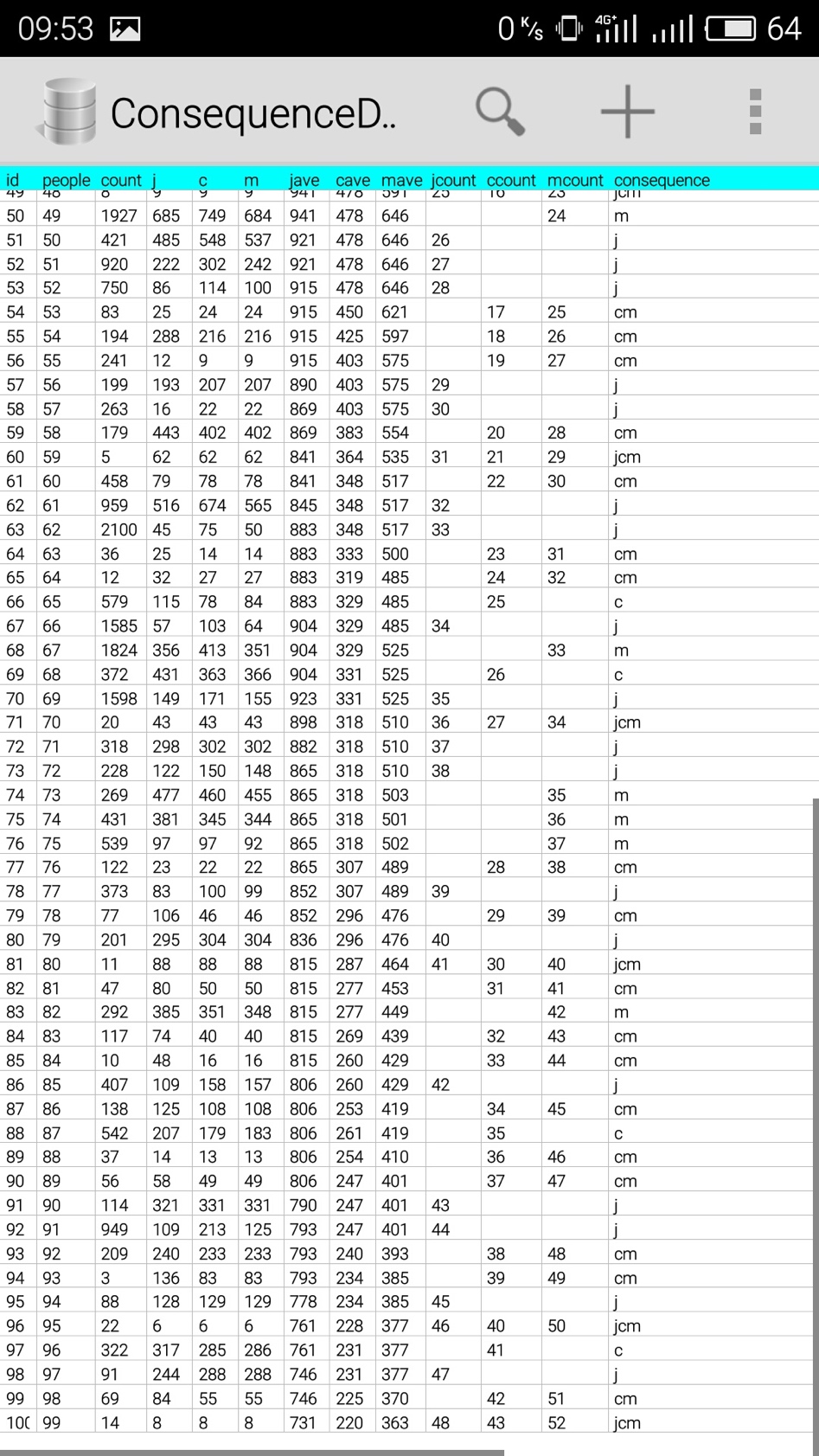


图4.1 全部用户的算法评价

其中people代表当前用户，count代表该用户数据量总数，j代表Jaccard相关系数，c代表Cosine相关系数，m代表下文作者提出的Mix相关系数，jave、cave、mave分别代表直到运算到该用户时，三种算法胜出时用户数据量的平均值，它反映了用户数据量对三种算法的影响，jcount、ccount、mcount分别代表三种算法的胜出数，该值越大表示算法越优秀，consequence代表对于当前用户的胜出算法。

由此可见，当Jaccard相关系数胜出时，用户的平均数据量为731；当Cosine相关系数胜出时，用户的平均数据量为220。实际结果验证了上文对公式的分析，且二者特性较明显。

4.3 提出优化算法

既然我们的分析结果被证实，即当用户数据量较多时，Jaccrad相关系数占优；当用户数据量较少时，Cosine余弦相关系数占优。那么我们就可以结合两种算法提出一种优化的组合算法，保留两种算法的优势，并且互相填补彼此的劣势。Mix算法就是在这种思想下诞生。

我们假设用户数据量为变量x，J为Jaccrad相关系数，C为Cosine相关系数。当x增大时，J所占比例增大，当x减小时，C所占比例增大。鉴于实际验证中x的数据量影响比重较大，所以x采用3次指数。设a为三次方系数，M为Mix相关系数，则公式如下。

（4-1）

系数a值得确定要依赖于两种算法胜出时的平均用户数据量，即Jaccard相关系数胜出时的730和Cosine相关系数胜出时的220，我们可以做初步的判断，当用户数据量为400以下时，Cosine更为出色。即当用户量为400时，C应有优势，考虑到Cosine相关系数普遍数值大于Jaccard相关系数，相同用户大概在8倍左右，故当x=450时，a取3效果最佳（经测试验证）。经计算，a的取值为4.6875\*，带入公式算出Mix相关系数的值，如图4.2。

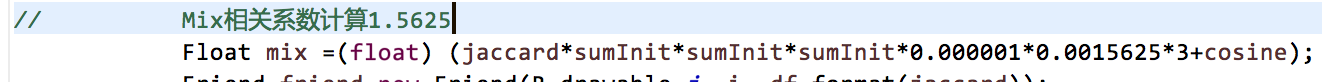


图4.2 Mix相关系数计算

得出结果后进行算法评估，如图4.1所示，Mix相关算法在52个用户中占有，在三种算法中胜出。

# 第五章 结论

5.1 本文总结

5.2 可能的创新与不足

5.2.1 可能的创新

5.2.2 存在的不足

**致 谢**

**参考文献**

[1] 查九, 李振博, 徐桂琼. 基于组合相似度的优化协同过滤算法[J]. 计算机应用与软件,

2014, 323-328.

[2] 张光卫, 李德毅, 李鹏, 等. 基于云模型的协同过滤推荐算法 [J][J]. 软件学报, 2007,

18(10): 2403-2411.

[3] 潘红艳, 林鸿飞, 赵晶. 基于矩阵划分和兴趣方差的协同过滤算法[J]. 情报学报,

2006, 25(1): 49-54.

[4] 吴湖, 王永吉, 王哲, 等. 两阶段联合聚类协同过滤算法[J]. 软件学报, 2010, 21(5):

1042-1054.

[5] Owens M, Allen G. SQLite[M]. Apress LP, 2010,1.