# 江苏大学计算机学院

# P2P软件开发课程报告

**课程实验、通过套接字实现兴趣相似推荐的p2p聊天室**

课程名称 P2P软件开发技术

学生姓名 曾 帅

专业班级 软件工程1601

学 号 3160608023

指导教师 李 致 远

开 课 系 软件工程

# 2019年6月20日

目录

[一、 大作业设计目的与要求](#_Toc28991_WPSOffice_Level1) [2](#_Toc28991_WPSOffice_Level1)

[二、 大作业设计内容](#_Toc32326_WPSOffice_Level1) [3](#_Toc32326_WPSOffice_Level1)

[三、 程序开发与运行环境](#_Toc14484_WPSOffice_Level1) [5](#_Toc14484_WPSOffice_Level1)

[四、 设计正文](#_Toc26710_WPSOffice_Level1) [6](#_Toc26710_WPSOffice_Level1)

[五、 实验运行结果及分析（不同情形的运行结果及详细分析）](#_Toc29245_WPSOffice_Level1) [12](#_Toc29245_WPSOffice_Level1)

[六、 总结与进一步改进设想](#_Toc10610_WPSOffice_Level1) [14](#_Toc10610_WPSOffice_Level1)

[七、 参考文献](#_Toc25547_WPSOffice_Level1) [15](#_Toc25547_WPSOffice_Level1)

# 大作业设计目的与要求

本系统最终实现一个基于P2P相关技术的聊天系统，该系统由多个聊天客户端和一个中心服务器组成，各个聊天客户端可以两两相互发送消息。目的在于设计GUI,多线程和套接字编程等。

主要涉及socket网络编程和多线程编程技术，采用的开发语言为java。为保证数据收发的可靠性，本系统采用面向连接的通信方式（TCP套接字）实现，具体采用流式套接字提供双向，可靠，有序且面向连接的数据流服务。其中，Socket类用于建立套接字连接，ServerSocket类用于监听套接字连接，OutputStream类用于实现网络套接字流的发送，InputStream类用于网络套接字流的接收。

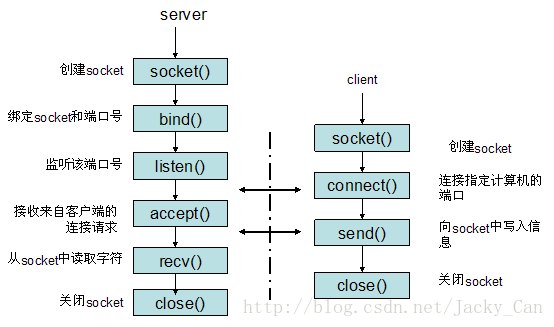
信息服务器需要不断的检测新的客户端发来的请求，并且为已经连接的客户端提供服务，所以需要不断的执行1.接收请求 2.解析请求 3.发送响应这三个操作，解析请求又会根据不同类型的请求发送不同的响应，收到的请求，服务器完成的操作，

客户端与服务器通信时，传递注册名和地址要保证传送的准确性和可靠，故选择TCP连接客户端和服务器，使用Socket对象，客户端与客户端之间进行通信时，要求实时性，不需要无比的准确，故选择UDP连接客户端与客户端，使用DatagramSocket对象进行通信，DatagramPacket为数据包

UDP为无连接传输，在DatagramSocket传送DatagramPacket时，只需要知道IP地址和端口号即可进行通信。

# 大作业设计内容

socket是在应用层和传输层之间的一个抽象层，它把TCP/IP层复杂的操作抽象为几个简单的接口供应用层调用以实现进程在网络中通信，socket是一种”打开—读/写—关闭”模式的实现，服务器和客户端各自维护一个”文件”，在建立连接打开后，可以向自己文件写入内容供对方读取或者读取对方内容，通讯结束时关闭文件；以使用TCP协议通讯的socket为例，其交互流程:Tcp协议：三次握手协议(服务端accept,客服端connect)，四次挥手(客服端close,服务端close)服务端accept阻塞的，等待着(多个)客服端的链接；大致的链接和通信流程如下：



获得网络中可以通信端点的IP和端口

假设有一个端点A，欲和P2P网络中其他端点通信，在通信之前，端点A必须首先把自己的IP和端口通知P2P网络中的其他每一个端点。其他每个端点收到这个信息后，就获得了端点A的IP和端口，随后向端点A反馈自己的IP和端口信息，使端点A也获得P2P网络中每个端点的IP和端口。

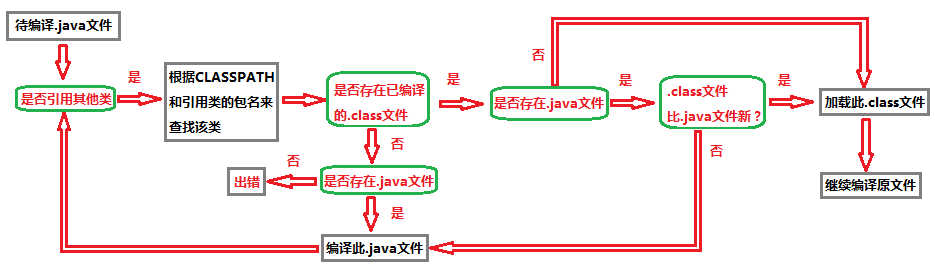
这里有两个技术可以完成端点A向其他端点通知其IP和端口的工作，一是广播技术，二是多播技术。

广播技术主要在局域网中使用，在局域网中的每一个端点（主机）都不得不接受并处理一个广播数据包。因此为了避免网络阻塞，路由器均限制广播数据包的通过。所以设计基于互联网的P2P程序不适合采用广播技术。

多播技术是一种允许一个或多个发送者（多播源）发送单一的数据包到多个接收者（一次的，同时的）的网络技术。 多播源把数据包发送到特定多播组，而只有属于该多播组的地址才能接收到数据包。多播可以大大的节省网络带宽，提高了数据传送效率。减少了主干网出现拥塞的可能性。多播组中的端点（主机）可以是在同一个物理网络， 也可以来自不同的物理网络（如果有多播路由器的支持）。因此，多播技术是我们的选择。

# 程序开发与运行环境

软件是一个能被执行的过程，能被正确执行才是一个合格的软件，一种开发语言要被执行就需要一个能被执行的环境做载体，执行之前需要进行编译，编译也就是检查软件的正确性。Java开发环境，即：使用eclipse编辑器



# 设计正文

因为当前 IPV4地址的缺乏 ,nat、防火墙的中介设备和不对称寻址建立起来的 p2p通信机制造成了地址访问的问题。

在 internet最初体系结构中，每个节点都有全球唯一的 ip地址，能够直接通信。可是随着节点的增多， ip地址使用紧张，他们需要中介设备如 nat连在一起。

私有网络中的节点可以直接连接到相同私有网络中的其他节点，也可以连接到全局地址空间中拥有全球唯一 ip地址的节点。。然而 nat通常只允许临时的向外连接申请，对于向内的申请会拒绝。这就造成了在 natA内网中的节点 A连接 natB内网中的节点 B时连接申请报到 natB时就被阻止了。此时我们需要的就是穿越技术。。。

总体来说穿越技术是利用一个公共服务器中转，使节点 A、 B都连接到中转服务器 S之后，通过 S中转A发送到 B的数据报或者是中转连接申请，，使 A、 B对于 natA和 natB来说都是向外申请。。。

1、中转数据报： A、 B都先向外与服务器 S建立接连，然后通过 S中转 A、 B之间的数据报。。

2、反向连接：当 A、 B都与 S建立了连接，并且只有一个节点在 nat之后（假设 A在 natA之后）。。当 B向 A申请连接时，申请背 natA拒绝。 B可以向 S提出申请要与 A建立连接，然后 S向 A发出指令，通知 A主动向 B申请建立连接。。

【 UDP打洞】

1、 A、B在同一个 nat之后：

用户 A让 S做介绍人来与 B建立对话

（1）A向 S发送一个消息请求与 B建立连接

（2）S使用 B的公共终端和私有终端响应 A

（3）同时 S也想 B发送 A的公共终端和私有终端，但是发送到公共终端的消息不一定能达到 B取决于 NAT是否支持“发夹”转化（回环转化）

（4） 如果 nat支持发夹转化的话，应用程序就可以免除私有和共有终端都要试图连接的复杂性。。

1. 不同 NAT后面的节点
2. 注册， A、 B都想服务器 S注册 natA安排了端口用作 A和 S对话使用，natB安排了端口用作 B和S对话使用，A向S的注册消息中报告了自己的私有终端这种情况下 A的公共终端，同理 B的私有终端和公共终端
3. A发送请求消息到 S，请求与 B建立连接，作为响应 S向 A发送了 B的私有终端和公共终端也向B发送了 A的私有和公共终端。
4. 既然 A、B处在不同的子网中，那么 A、B的私有终端是不能公共路由的，发送的消息肯能会发到自己子网中的 ip中（应为不同子网中的私有 ip可以相同）

（4） 当从 A发向 B的第一个消息到达 natA时， natA注意到这是一个新的外出会话， natA看到源地址是子网中地址，而目的地址是外网地址，所以 natA将从私有终端的外出会话转化到对应公共终端，这样 A的第一个到 B的公共终端的外出会话消息就在 natA上“打了一个洞”。新的 UDP会话由 A的私有网络上的终端和internet上的公共终端标识，同理 B也建立了对 A的私有、公共连接标识。

（5） 如果 A发向 B的公共终端的消息在 B发向 A的第一个消息穿过 B自己的 natB之前到达了 natB的话， natB会认为 A的内入消息是禁止的，丢弃 B的请求消息，但是 B的请求消息在 natB上为 A打了一洞，此时洞双向打开，通信可以进行下去了。。。

3、多级 NAT后面的节点：

（ 1） A、B都建立与 S的向外连接

（ 2）最终连接目的：

（ 3）但是在此时 A、B无法知道伪公众终端。

算法核心思想：在一个在线推荐系统中，当用户A需要个性化推荐时，可以先找到和他有相似兴趣的其它用户，然后把那些用户喜欢的、而用户A没有听说过的物品推荐给A，这种方法称为基于用户的协同过滤算法。

==>可以看出，这个算法主要包括两步：

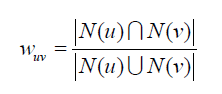
一、找到和目标用户兴趣相似的用户集合——计算两个用户的兴趣相似度

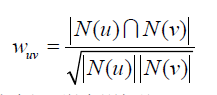
二、找到这个集合中的用户喜欢的，且目标用户没有听说过的物品推荐给目标用户——找出物品推荐

下面分别来看如何实现这两步：

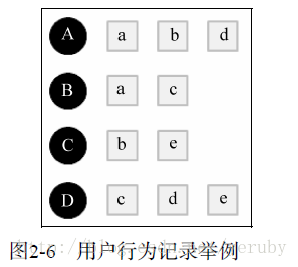
一、计算两个用户的兴趣相似度：

给定用户u和用户v，令N(u)表示用户u感兴趣的物品集合，N(v)表示用户v感兴趣的物品集合，那么可以通过Jaccard公式或者通过余弦相似度公式计算：

Jaccard公式

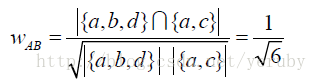
余弦相似度公式

举例：下图表示用户A对物品{a,b,d}有过行为，用户B对物品{a,c}有过行为

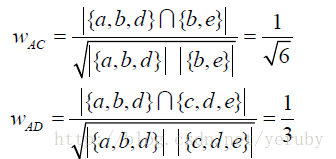


利用余弦相似度计算可得：

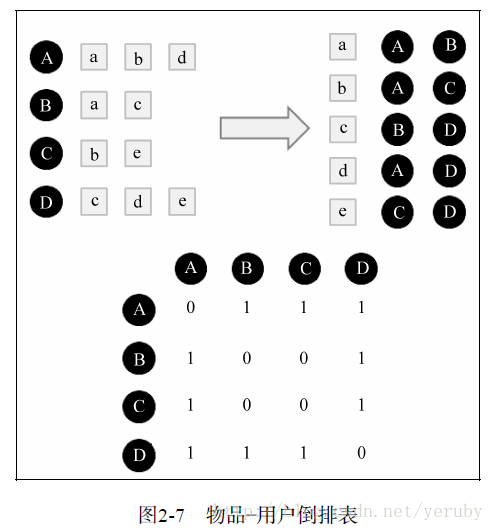
用户A和用户B的兴趣相似度为：



同理，



但是，需要注意的是，这种方法的时间复杂度是O(U^2)，因为我们需要计算每一对用户之间的相似度，事实上，很多用户相互之间并没有对同样的物品产生过行为，所以很多时候当分子为0的时候没有必要再去计算分母，所以这里可以优化：即首先计算出|N(u) 并 N(v)| != 0 的用户对(u,v)，然后对这种情况计算分母以得到两个用户的相似度。



针对此优化，需要2步：

（1）建立物品到用户的倒查表T，表示该物品被哪些用户产生过行为；

（2）根据倒查表T，建立用户相似度矩阵W：在T中，对于每一个物品i，设其对应的用户为j,k，在W中，更新相应的元素值，w[j][k]=w[j][k]+1，w[k][j]=w[k][j]+1，以此类推，扫描完倒查表T中的所有物品后，就可以得到最终的用户相似度矩阵W，这里的W是余弦相似度中的分子部分，然后将W除以分母可以得到最终的用户兴趣相似度。

得到用户相似度后，就可以进行第二步了。

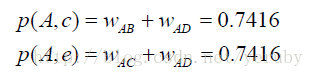
二、给用户推荐和他兴趣最相似的K个用户喜欢的物品。

公式如下：

IMG_256

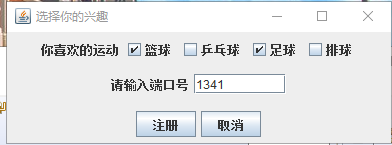
其中，p(u,i)表示用户u对物品i的感兴趣程度，S(u,k)表示和用户u兴趣最接近的K个用户，N(i)表示对物品i有过行为的用户集合，Wuv表示用户u和用户v的兴趣相似度，Rvi表示用户v对物品i的兴趣（这里简化，所有的Rvi都等于1）。

根据UserCF算法，可以算出，用户A对物品c、e的兴趣是：



# 实验运行结果及分析（不同情形的运行结果及详细分析）

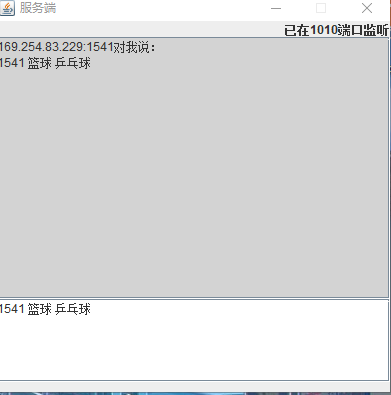
首先运行服务端，服务端只是收集用户的兴趣爱好等基本信息，根据用户所发的基本信息，在服务端拥有用户协同过滤的推荐接口，客户端通过调用接口信息得到与之相匹配的其他用户的相似度



运行客户端，对用户的基本注册，由于没有使用数据库，所以只能通过基本的信息注入发给服务端，注册成功即到客户端页面



客户端通过获取推荐用户得到与自己相匹配的用户会显示到第二栏的文本框里。



服务端对客户端发送的用户基本信息进行收集注册，通过客户端的发送信息进行实时监控，而且服务端的信息收集目前只能对用户的基本信息推荐

# 总结与进一步改进设想

通过这次聊天程序课程设计，又有了许多收获。

聊天软件，是我之前一直想完成的一个小软件，但没有机会去做，这次有机会做，我决定把它做好，做的像QQ一样。刚开始，对于网络通信、数据库连接及操作这部分，我之前没练习过，因此，对于整体的设计都很难把握，我参考了部分书籍，大概了解了其原理，之后就是确定需求，虽然我们对QQ都很熟悉，也都基本了解其大概需求，但在实际设计时，很多需求方面东西都是看不见的，必须自己查资料、思考、练习才能发掘。然后就是总体设计及人员分工，这一步也很关键，如何协调每个人，如何发挥每个人的优势，这需要很多工作。

在整体设计完成后，我考虑先开发出简单的聊天软件，然后逐步细化，因此，在详细设计时，我简化了一些东西，先开发出一个基本原型，用以验证技术并进一步明确需求。然后，对部分技术进行改进和细化，最后，再次基础上不断的迭代进行，由于我的水平有限，我最初的设计并不一定是好的设计，只有不断的试验和改进，才能开发出好的软件，当然，前期的整体架构设计非常重要，这将很大程度上决定软件的质量和适应需求变更的能力。总之，在试验与改进中，我学到了很多东西，不光是技术，还有合作。

# 参考文献

1.石彦芳，李丹.《Oracle数据库应用与开发》.机械工业出版社，2013

2.耿祥义，张跃平.《Java面向对象程序设计》.清华大学出版社，2010

3.张海藩.《软件工程导论》(第5版).北京.清华大学出版社，2008

4.刘新.《Java开发技术大全》.北京.清华大学出版社，2009

5.明日科技 《Java经典编程300例》清华大学出版社 2012

6.梁建全 《你必须知道的261个Java语言问题》 人民邮电大学 2009