

群组交流系统新功能开发

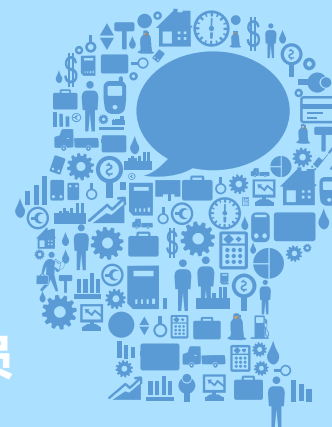
——基于QQ群组聊天记录分析

郝建锋



中央财经大学

狗熊会人才计划第二期学员



CONTENTS

1 背景介绍

互联网服务引领潮流，
社交类软件广受追捧。
新生代网民日益增长，
即时通讯功能需突破。



2 变量说明

全方位剖析现有信息，
多维度解读聊天数据。
划分整合文本型讯息，
提取搭建结构化变量。



4 商业应用

新发功能生动有趣，
吸引用户争相体验。
数据分析创造价值，
提高软件用户粘性。



3 数据分析

运用技术挖掘特征，
利用图表展现规律。
三大系统初步建成，
巧妙揭示群组奥秘。





背景介绍

SOLOMO 浪潮

John Doerr在2011年首次提出了“SoLoMo”的概念。

如今，基于SoLoMo的发展模式已被公认为是互联网行业的发展趋势。

Social

社交化——社交类网站和应用，
Social毫无疑问是当下乃至未来的潮流

Local

本地化——基于用户当时位置的
互联网服务，如大众点评等

Mobile

移动化 ——智能手机所支持的
各类移动互联网应用及软件



据艾瑞咨询发布数据，截止2016年12月底，中国移动社交网民超过**6**亿，

占总体移动网民的比例接近**90%**，同比增长率高于全球水平。超过一半为30岁以下的**新生代用户**，互联网的发展伴随着他们的成长，因此他们对互联网更加熟悉，也更愿意尝试移动社交的**新玩法与新功能**。新生代的大批加入以及同业竞争的不断增强，使得如何设计社交软件的新功能来吸引用户、提升用户粘性成了软件供应商面临的重要问题。

背景介绍



现状

如右图所示，群组交流系统一般包含**成员**、**聊天记录**和**群组功能**三个要素，典型应用QQ在这三个方面进行了许多功能拓展。

缺陷

通讯交流中极为重要的一环——“聊天记录”的功能开发还十分浅显。目前仅停留时间及发言统计上，这种纯频率统计的图表分析的趣味性与互动性都十分低下，并不能满足网络新生代们的好奇心，难以吸引用户。

本文

基于上述情况对QQ群组聊天记录进行分析来开发群组新功能，从而提升软件的竞争力，争取到更多有网络社交需求的用户。



QQ群组功能一览



真实的消息数据格式如屏幕所示，
数据虽有**固定格式**但需进一步处理。



本文对来自**三个不同类型**群组（朋友群、大学同学群和校园交易群）的有效文本数据**18798行**提取变量如表1.



变量说明

表1 “聊天记录” 数据变量说明表

变量名		变量类型	取值范围	备注
时间		日期型	2017-10-09 19:21:05 - 2017/10/22 23:45:35	包含日期、时间 分析时可用函数计算时间差
用户名		文本型	包含昵称、QQ号	由于涉及用户隐私，在本文分析中将昵称与QQ号换为游戏或漫画中的人物名称
聊天内容	图片/表情	文本型	三种格式 【详见备注】	图片、表情在导出的聊天数据中表现格式有以下三种： [图片] [表情] /托腮
	应用外链接	链接型	如： http://s.kugou.com/song.html?id=5DXAmdarAV2	包括网页分享和其他应用数据分享（如通过音乐软件分享的音乐）
	群应用	文本型	有固定格式 【详见备注】	如： [群签到] 请使用手机QQ进行查看。 [QQ红包] 请使用新版手机QQ进行查看。
	内容文本	文本型	—— ——	除图片/表情、应用外链接和群应用消息之外的群成员聊天文本

三大系统全景概览图



完成变量的构建之后，本文立足于对群组聊天记录的分析开发出三大系统——发言系统、称号系统和风格系统，各个系统的具体内涵如图所示。

成员称号：

冷场小王子：在他发言后半小时内无人发言

开聊能手：一段时间无人发言后首个发言并使得其他人加入讨论的人

表情达人：发送表情最多的人

斗图狂魔：发送图片最多的人

.....



发言时间角度：

探查群组发言的**时间分布**特征

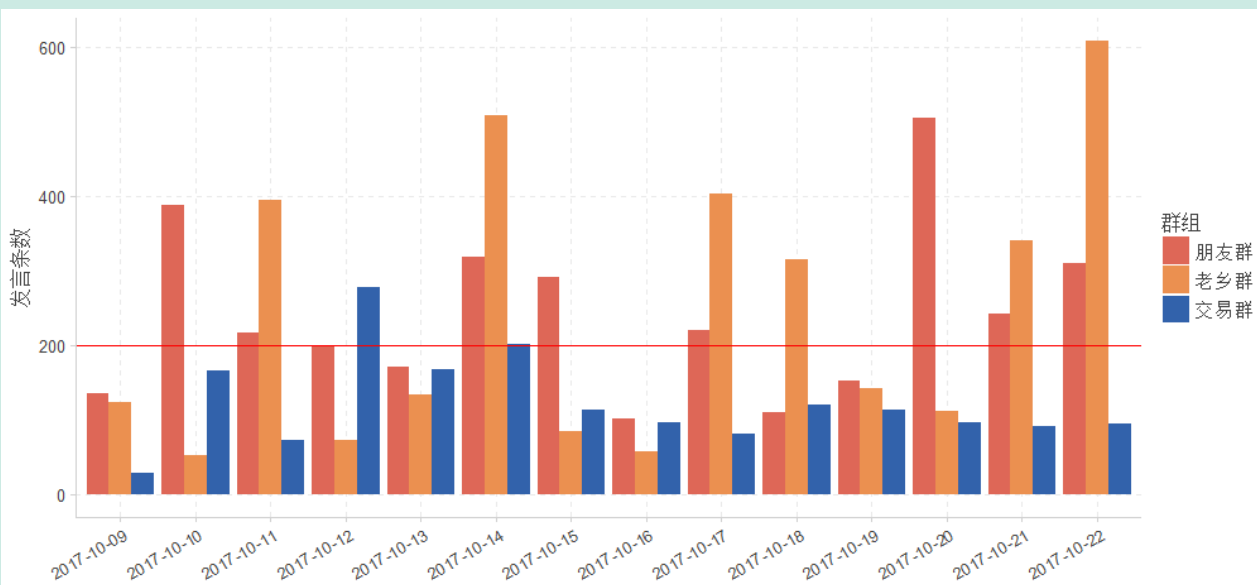
发言成员角度：

社交网络图：查看群组成员结构

人员更迭图：3天为周期群组活跃人员的变化

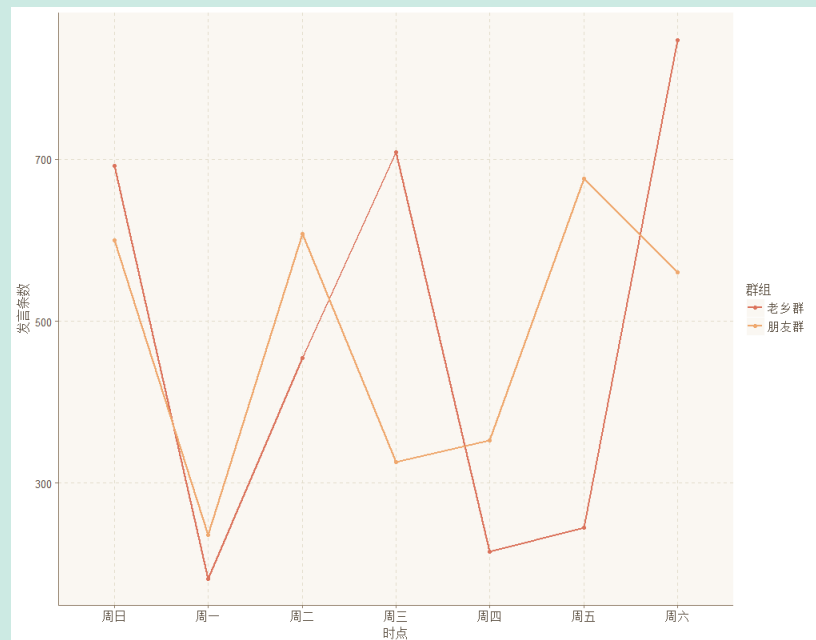
风格雷达图：

查看群组平时都在聊哪些方面内容



- 分析时段内交易群较其他两个群组交易群发言数量少。
- 老乡群在10月22日发言条数最多，超过了600条，看来那一天有某个话题引起了群成员的热烈讨论。
- 朋友圈和老乡群都在某几天发言条数很高，这是否涉及到星期内的分布呢？让我们一探究竟。

- ✓ 周六、周日是聊天高峰期，这刚好是学生党的放假时间；
- ✓ 明显注意到两个群组聊天数随星期分布的形态与变化趋同，在一周中都有三四天聊得很happy，而且峰值在一周中分布均匀，看来聊天也很讲究劳逸结合啊！





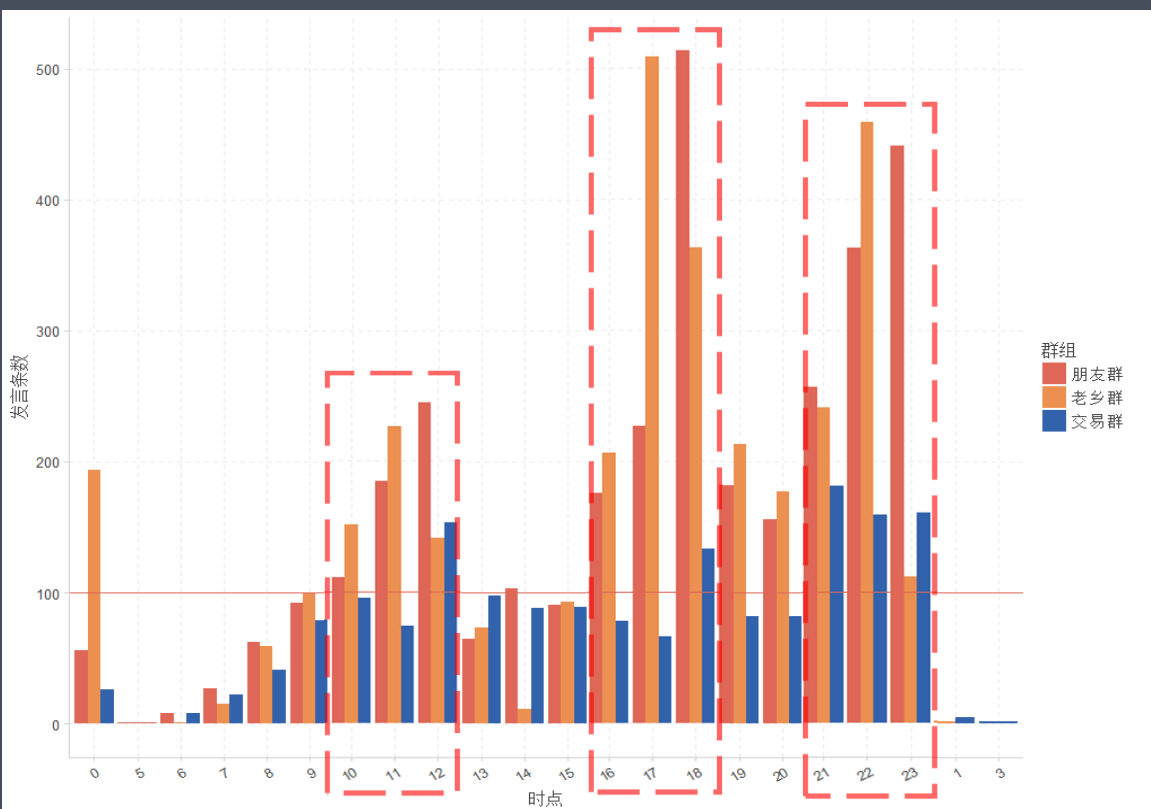
发言系统

日内分布

学生式分布

由图可以看到在一天中两个群组的聊天峰值出现时间几乎相同，都是**中午、傍晚以及晚上十点后**。

这不难理解，中午和傍晚刚好是学生们的下课时间，而晚间十点后学生一般是完成学习回到宿舍开始上网闲聊，这样的聊天分布是典型的大学生群组聊天时间分布。



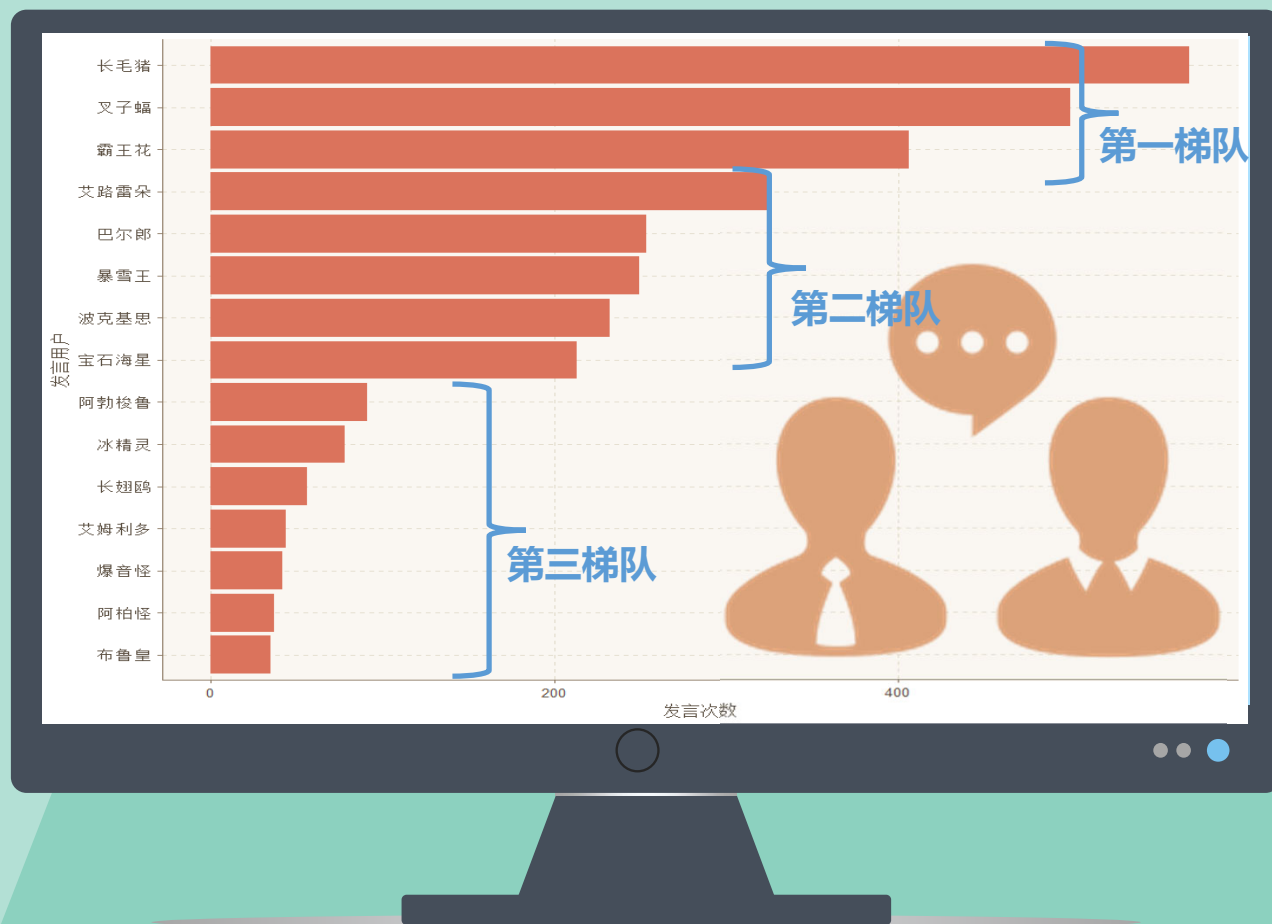


发言系统

谁是话痨?

从图中可以清晰地看到大学朋友群中平日里发言较多的小伙伴，在该群中，发言用户根据发言次数大致分为三个梯队，其中长毛猪是发言次数最多的那一位——**话痨!!!**

仅看发言次数只能知道谁说的最多，接下来本文进一步的分析来查看成员间的结构。





发言系统

社交网络图

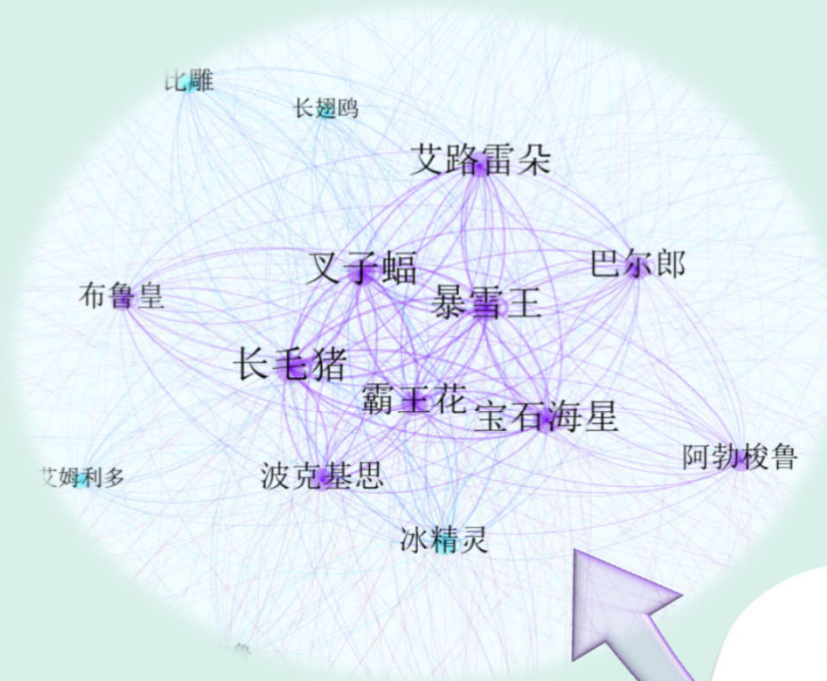
本文通过对用户连续对话次数中群成员两两出现的次数进行统计最终画出了部分群组成员之间的社交网络图。


图形解读

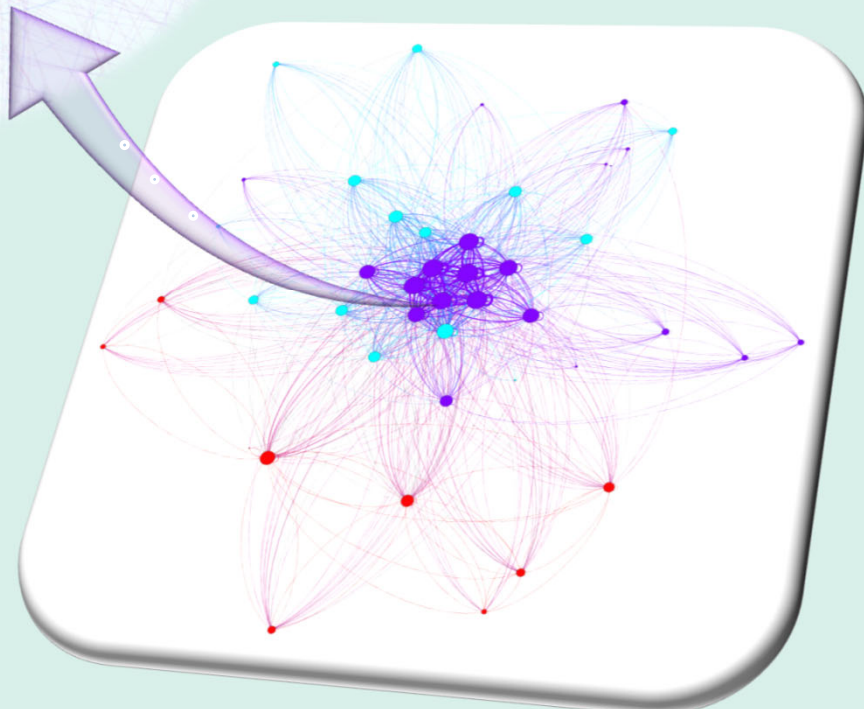
右下角图中**节点越大**代表与他人交流次数越多；
同一颜色的节点代表平时经常一起聊天的小群体。

核心人物

中间紫色的大节点所代表的人员基本与所有人多有交谈，并且发言次数最多，本文对图形进行了局部放大，可以看到平时聊的最开心的要数这几位了，他们称得上是本群的核心人物。



 与前文话唠分析
所得人员一致





衡量一个群活跃与否的指标除了发言次数还有**活跃程度**，即每日发言人数，本文依据活跃程度做了以下研究：



活跃度：群中每天发言的人数

周期：两天为一个周期

旧面孔：本周期与上周期皆有发言

新面孔：本周期发言且上周未发言

根据构建的指标将数据划分为**七个周期**，每个周期的**活跃人数及新旧面孔数量**如图所示。

从10月9号后的14天，绿色代表旧面孔，蓝色代表新面孔，绿色加蓝色为总的发言人数。

- ① 发言人数呈现先增加后减少的分布形态，即**活跃度先增后降**；
- ② 活跃度升高的日子里新面孔不断涌现，活跃度降低的日子里新面孔不再出现，即**活跃度与新面孔率呈反向关系**。



发言系统



目前想法是：

1、构建出开聊能手、冷场小王子.....等指标后在PPT上以颁奖的形式将这些奖项颁给获奖人

2、排列出表情使用频率，图形大小代表使用次数（直接以原表情形式，如



3、排列出群功能（群签到、红包等）使用频率，展现方式如2

风格系统

因为内容涉及隐私不便直接展示，但可以通过对内容的分析展现不涉及隐私的结果

目前想法有两个：

- 1、根据发言内容进行情感分析；
- 2、找一些像游戏、明星之类的常讨论的词典进行词典匹配后统计频数画出雷达图查看群组风格



商业应用

待填写

