

群组交流系统新功能探究

——基于 QQ 群组聊天记录分析

一、背景介绍

美国著名的投资家 John Doerr 在 2011 年首次提出了“SoLoMo”¹ 的概念。他整合了当时互联网行业最具发展潜力的三个方向：Social（社交）、Local（本地化）和 Mobile（移动）。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。就中国市场来看，随着生活水平的日益提高、智能手机的不断普及，人们的社交需求开始趋于多样化，移动社交服务深受热捧。根据艾瑞咨询发布的数据显示，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。其中超过一半为 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。

从目前发展来看，好的社交应用应该在满足用户核心的社交需求的基础上，立足于自身的社交特色逐渐形成对用户具有吸引力的差异化需求价值。多种移动社交应用满足用户不同的需求，用户也会根据自己的意向来选择应用。根据艾瑞咨询报告，接近七成的用户会同时拥有两种以上社交软件，其中通讯社交（QQ、微信等）、兴趣社交（Keep、唱吧等）和图片社交（IN、QQ 空间等）三类应用拥有远超其他类型社交应用的用户数量，主打即时通讯的通讯社交软件的全体网民使用比例更是高达 91.8%，稳居第一，这类应用的主要代表有 QQ、微信和钉钉等。它们的主要服务是为用户提供便捷有趣的即时通讯平台，平台上的交流系统由个体互通与群组交流两部分组成。群组交流模式中成员间互动性更强，用户相互依存度² 更高，因此群组交流系统功能开发的好坏会直接影响到该应用的受欢迎程度。

¹ “So”代表社交类网站和应用，“Lo”代表基于用户当时位置的互联网服务，如大众点评等；“Mo”代表智能手机所支持的各种移动互联网应用。

² 依存：群组内有部分人喜欢此应用，所以其他人为了交流方便也选择使用此应用。

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，如图 1 所示，典型应用 QQ 在这三个方面进行了许多功能拓展，因此深受新生代网民喜爱。

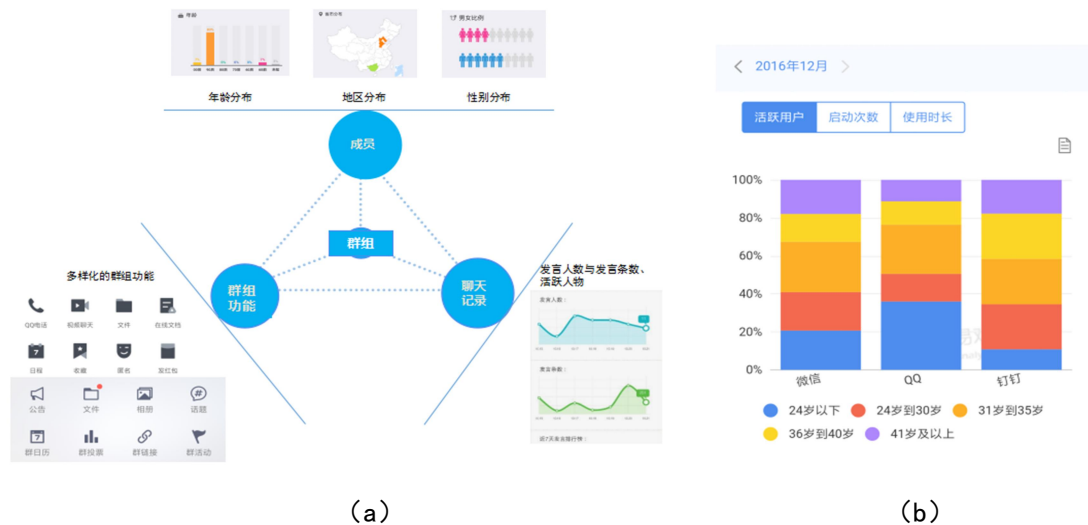


图 1 (a) QQ 群组功能一览 (b) 即时通讯软件活跃用户年龄对比³

但是可以看到的是，作为通讯交流过程中极为重要的一环——“聊天记录”在实际应用层面的探究还十分浅显。一条聊天记录可以分为时间、用户和内容三个部分。现阶段对聊天记录的分析仅仅停留在发言时间分布及成员活跃程度上，并没有对其内在信息进行深入挖掘，这种纯频率统计的静态图表分析的趣味性与互动性都十分低下，难以提升对用户的吸引力，也并不能满足网络新生代们的好奇心，他们渴望美观、有趣、可交互性强的新功能来满足他们日益增长的多样化社交需求。

基于以上现实情况，本文首先对发言时间进行描述性分析，探索所选群组的基本聊天情况；接着对群组聊天记录中的发言时间、用户及内容联合分析来完成一系列有趣称号⁴的建立；随后利用社交网络分析搭建成员互聊网络，找出彼此十分聊得来的小群体以及各群体的核心人物等；最后本文将利用搜狗细胞词库中的部分词典，根据发言内容的分词处理结果与各个词典进行匹配做出群组风格雷达图。经过上述步骤，本文大致建立起群组聊天记录分析的基本框架，后续再经过专业设计人员的修饰及交互式可视化处理，相信可以对企业开发群组交流新功能有一定借鉴与参考意义，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。

³ 数据来源：易观千帆 APP 数据。
⁴ 如：[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；[红包爸爸]发红包次数最多的人；[冷群小王子]在他发言后一段时间无人发言；[表情达人]发送表情图片最多的人。

二、变量说明

本次分析报告所使用的数据均来自即时通讯软件 QQ，共包含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。由于数据格式较为规范，可以提取出**发言时间**、**发言用户**和**聊天内容**三类变量。经过对聊天内容初步探索后，本文根据内容类型将其进一步划分为图片/表情、应用外链接、群应用和纯文本四个变量⁵。具体变量说明表如表 1 所示。

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 - 2017/10/24 18:45:35	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

⁵ 一条消息中可能混合包含四个变量，本文利用 R 软件通过正则表达式分离并提取出这些有固定格式的变量。