社交网络用户行为的体裁互文性分析

万亚平[1] 阳小华[1] 刘志明[1] 李 治[2] 张 娟[1]
(南华大学计算机科学与技术学院 衡阳 421001)[1] (北京空天技术研究所 北京 100074)[2]

摘 要 社交网络是基于用户之间共同的兴趣、爱好等构建的一种社会关系网络服务。社交网络中包含了大量的用户行为,研究这些行为对增强用户体验,增加用户使用粘性,提高资源的分享率以及其它一些服务的有效性都具有十分重要的意义。体裁互文性是语言学的基本概念,它与用户行为具有一些共性特征。实验表明,社交网络用户行为和体裁互文性具有相似性,利用体裁互文性研究用户的行为将更有利于信息共享、传播以及知识通信。

关键词 互文性,体裁,体裁互文性,用户行为
中图法分类号 TP399 **文献标识码** A

Analysis of Genre Intertexuality on Social Network User Behavior

WAN Ya-ping[1] YANG Xiao-hua[1] LIU Zhi-ming[1] LI Zhi[2] ZHANG Juan[1]
(School of Computer Science and Technology, University of South China, Hengyang 421001, China)[1]
(Beijing Aerospace Technology Institute, Beijing 100074, China)[2]

Abstract Social network is a kind of social relationships network service which is built based on users common interests and hobbies. Social network contains a large number of user behavior, and by analyzing these behavior it is very significant to enhance the user experience, increase user viscosity, improve resource sharing rate and effectiveness of some other services. Generic intertextuality is a basic concept in linguistics and it has some common characteristics with user behavior. The results show that the social network user behavior is similar to generic intertextuality. Studying user behaviors by generic intertextuality will be more conducive to information sharing, information dissemination and knowledge communication.

Keywords Intertexuality, Genre, Genre intertexuality, User behavior

1 引言

互联网技术的发展及 Web 2.0 的兴起使得社交网络[1,2]日益流行起来,并引起国内外人们广泛的关注与参与,社交网络的注册人数逐年增长,其内容不断更新,这为研究大规模社交网络提供了前所未有的真实的实验平台。例如,全世界影响力最大的社交网站 Facebook[3]拥有超过千万的注册用户。社交网络作为现实人际交往的扩展,其中蕴含的海量信息对人们工作和生活产生深刻的影响。传统的分类技术主要利用文本内容建立索引模型,但在社交网络中包含的各种结构化和非结构化数据(例如各种视频、音频、图形图像信息)使得很难利用传统的分类方法进行信息推荐或者获取。近些年来,计算语言学领域很多的文本分类研究者认识到按照体裁分类的重要性,并出现了一个重要的理论转向,即由重视内容的分类转而重视内容与体裁并重的研究。体裁已成为计算语言学、信息学以及情报学等领域的研究热点之一[4,5]。

体裁是一种共享概念模型的明确的形式化规范说明,可以被用来作为社交网络信息资源结构化定义的清晰描述,并且可以支持知识共享和重用。把体裁分类信息附加于社交网络信息交流和传播,可以显著改善"知识通信"效能[6]。然而,如何从社交语篇中识别、描述、预测、使用体裁是一项复杂而具有挑战性的工作。体裁是一定历史时期人类思维抽象归纳的结果,由于人类认识局限和它自身动态演变等因素,使得全面、准确、有效地概括表述庞大。研究类别复杂的体裁相当困难;其次,体裁分类并没有严格的逻辑推理和证明,所以其分类标准比较模糊,边界存在交叠,很难保障体裁类别的独立性和唯一性,分类的标准是相对的,而不是严格和绝对的;国内外一些学者虽然很早就已开始对体裁的研究[7,8],但整体上还处于探索阶段。即使有一些较好的应用,也还是针对特定领域,理论和方法很难被普及和广泛使用[9,10],这促使科研人员在现有的基础之上去开辟新的对体裁进行研究的理论和方法。

用户行为分析是目前社交网络一个重要研究子课题。如何挖掘和获取用户行为中蕴含的大量信息,特别是隐形信息,分析用户差异化需求,开发新的、更契合用户需求的服务和产品,成为社交网络生存发展关键。在社交网络的活动中,用户

本文受湖南省哲学社会科学基金(14YBA335),湖南省科技计划支撑项目(2013FJ3030),湖南省高校思想政治教育研究课题基金资助项目(2014[14C10]),湖南省教学改革研究项目(湘教通[2013]223号),南华大学重点学科建设项目,南华大学重点实验室建设项目,南华大学科技创新团队建设项目资助。

万亚平 男,博士,副教授,主要研究方向为信息检索、信息存储、分布式计算,E-mail:15197486128ypwan@aliyun.com;**阳小华** 男,博士,教授,主要研究方向为信息检索、可信计算;**刘志明** 男,博士,教授,主要研究方法为知识工程、信息检索、大数据分析。

行为会影响信息传播，同时，在用户具备高密度和高粘度属性、真实身份信息较为完备，用户好友之间信任度较高的情况下，用户行为具备了较高的价值挖掘可能性。社区中大多数用户的信息行为总是表现出一定粘性，具有共同兴趣爱好的用户活动目标对象在体裁上具有更大相似性。

研究社交网络中用户的行为和体裁互文性的关系，对语篇基本体裁获取、体裁向量的形成以及体裁的识别都具有重要意义，这也是本文研究的主要目的。此外，研究体裁互文性和用户行为的共性特征，对社交网络社区关系挖掘、社会化推荐和信息传播都具有指导作用和功效。本文以电影社区为研究对象，将体裁互文性概念拓展到社交网络，尝试使用向量的形式表示体裁，以体裁为信息表示载体，以社交网络中语篇体裁互文性作为度量体裁相似性的依据，根据语篇空间体裁之间的互文强度来自动地识别语篇体裁，进而对语篇进行分类，通过实验研究用户行为和体裁互文性的潜在共性特征。归纳而言，本文的主要研究意义包括 3 点：其一，体裁的定义，使用向量定义和表示语篇的体裁；其二，将互文性概念引入到社交网络，利用互文性度量语篇体裁之间的关系；其三，对于体裁未知的语篇，使用社交网络用户行为来度量不同语篇体裁之间的相似关系。

2 相关工作

20 世纪末以来，计算语言学很多的文本分类研究者认识到按照体裁(genre)分类的重要性。体裁研究在语言学中历史悠久，然而比较全面的、系统的体裁研究却是现代语言学创立之后的事情。尤其是近些年，关于体裁的理论和应用研究已经逐步渗透到信息科学领域。目前国内外对体裁的研究主要集中在 3 个方面：体裁的识别、体裁的描述和体裁的应用。其中体裁应用的研究又体现在文本分类和信息检索上。

Swales 是最早研究体裁的学者之一，他最初将体裁描述为〈purpose，form〉二元组形式，认为体裁是由一组交际事件构成，这些事件都包含一些共同交际目标，这些目标由父语篇群体所识别，进而组成该体裁的基本原理。这些基本原理又影响语篇的组织结构，进而影响语篇的内容及形式(措辞)[11]。Swales 后来在另一篇论文中论述了体裁识别和交际目标之间的关系，指出交际目标被用作语篇分类的标准，对体裁的分析是由文本驱动和上下文驱动的过程[12]。

Kevin Crowston[8] 和 Orlikowski[13] 等人定义体裁元素为由〈content，form，purpose〉构成的三元组，认为体裁是多维概念，不仅考虑文档本身属性，也包括人类在各种活动中所扮演角色的性质。Kevin Crowston 从不同角度提出了实用的体裁分类方法，该方法不依赖于任何一个具体的概念视角，并且允许从各个概念层次上获得对体裁的总体认识和把握。

Rosso[14] 把体裁定义成〈function，style，form or content〉，认为体裁描述能表达出内容所不能(或不会)表达的东西。体裁标注者及目标用户需对体裁有共同的认知。Rosso 使用 257 个测试者、18 个体裁对 55 个新文档进行分类，最后的实验数据表明超过 70% 的测试者认同他对文档的体裁分类。

Inger Askehave 和 Anne Ellerup Nielsen[15] 基于电子文档提出了数字体裁(digital genre)概念。他们针对 Web 介质文本对传统的体裁模型进行了扩展和升级，研究表明导航模式的 Web 文档确实表明了交际目的，与此同时，在 Web 页面上的整体动作确实实现了处于阅读模式下 Web 文档的网际交际目的。他们认为数字体裁的定义不能和文档所在的媒介孤离开来。

随后英国布莱顿大学的学者 Santini 严格区分了“文本类型”(Text type)和体裁的差异，指出文本类型是表达交际目标的语篇模式，而体裁是社会共同认可的文本类别[16]，同时还提出了一个自动体裁分类模型。2011 年，Santini 在另一篇论著中表明体裁的识别或多或少基于人类认识的常规，对常规的认识使得人们能够重构或者推导文档产生时的上下文、目标及功能；并且基于主观贝叶斯方法和 if-then 推理规则提出了一个体裁识别模型，把体裁识别所分析的特征归纳为 4 类[17]。

在对体裁的应用方面，Freund[9]、Vidulin[18] 及 Rosso[19] 等都是利用体裁来表示文档目标。用体裁来表示信息检索查询目标是当前对搜索引擎进行优化的主流。2007 年 TGSE 研讨会(Towards Genre-Enabled Search Engines)专门讨论利用体裁改进搜索引擎。研究者已使用体裁来过滤搜索引擎返回结果[9,18]。Freund 的研究结果也表明文档体裁与用户当前任务关联关系能够有效改善信息检索质量。

Amauri Lopes 等人[20] 针对音乐信号提出一个自动体裁分类的策略。该技术把音乐信号分成 21.3ms/帧，每一帧抽取出 4 个特征，每个特征值处理超过 1s 的分析片段。顺着每个分析片段的特征统计结果被用来确定由这些特征集合构成的向量，这些特征刻画了各自的片段，进而分类程序使用这些向量去划分体裁类别。Amauri Lopes 通过该过程对音乐信号进行了体裁分类。

总体而言，国外对体裁的研究大多集中在利用体裁的内部特征对文本进行分类。一些学者即使认识到了体裁具有交际功能，但是在对体裁概念的定义和使用上，还存在模糊和局限性，所采用的方法不具有普遍效应。而国内体裁研究大多集中在语言学领域，对体裁进行分类的研究还处于全面探索的初级阶段，技术还不成熟。

3 基本概念

由于所涉及到的一些基本概念来自于语言学、符号学等其他学科，因此在此有必要简述其义。

3.1 社交网络语篇

1967 年，哈佛大学的心理学教授 Stanley Milgram 创立的六度分割理论被认为是社交网络的理论基础。按照六度分隔理论，网络上每个用户的社交圈都不断放大，最后就可构成一个大型的社交网络。社交网络与传统 Web 网络的最大不同之处在于：传统的 Web 网络的主体是内容信息，依靠内容信息组织在一起，呈现给用户；而社交网络的主体是人，依靠人与人之间的朋友关系组织在一起。因此，社交网络语篇更多反映了人的社会属性，包括行为特性。

语篇通常指一系列连续的话段或句子构成的语言整体。语篇就其长短而言并没有一定的模式，也就是说有时一个句子就可以成为一完整的语篇，如：“Hello World!”。有时即使是一个完整的段落也不能构成一个语篇，成为语篇的关键在于某个句子或段落或篇章是否是一个完整的语言整体。Halliday 和 Hasan 认为，最好把语篇看作是语义单位(semantic unit)，即不是形式单位，而是意义单位，语篇分析又称话语分析。衔接与连贯是语篇特征的两个重要内容。衔接体现在语

篇的表层结构上，通过语法手段和词汇手段来实现，使作者所欲表达的意图贯通整个语篇，达到交际目的；语篇的连贯指交际行为之间的统一关系，语篇连贯依赖于语篇产生时的语境及语篇使用者的语用知识。仅有语言成分之间的衔接并不能构成完整的语篇，主要在于各衔接成分是否是一个意义的整体。语篇研究之所以出现是因为句子语法已不能说明自然语言的许多现象。

本文中社交网络语篇是指在社交网络中共享概念模型的各种信息，包含文本、声音（音乐等）、图像、视频（电影等）等各种格式和非格式数据的泛语篇概念。

3.2 语言学体裁概念

体裁（Genre）起源于传统的语言学范畴，是计算语言学和修辞学的交叉研究课题。体裁与写作风格、句法分析联系紧密，对文章的写作有着明显的制约和规范作用。

体裁的分类标准模糊的边界存在交叠。标准模糊是由于文学家在划分体裁时，往往采用的是含混的字眼，其标准很难有一个定量的衡度，分类的标准是相对的。边界交叠是由于文体交融演变的结果，不少作者为了求新求变，往往有意融合两种以上体裁的写法，以增进文章的可读性，导致了文体之间往往相互包融、相互渗透，这已经成为公认的客观事实。比如，报告文学介于文学体与新闻体之间；科学小品介于科学体与文学体之间。由于它具有知识的一般特征，因此体裁具有知识经验性。

体裁显现出某些相对稳定的特征，并以此区别于其他文章体裁。文本体裁主要从篇幅角度着手，讨论文本形式的整体特征，它抽象归纳了文章的互文特征信息，并反映了从文字表达事物对象、内容含义之外的一般化特征。这些特征信息对于文章一般是相对固定的，并且在通篇文章中都会保持前后一致性。对于某种特定的体裁，在一段时期内，文体的结构、语言、叙述等呈现出相当的稳定性。例如，诗歌作为一种最古老的体裁之一，它语言精炼，讲究音韵和谐，饱含想象和热烈感情，这些诗歌的最基本特征直到现在也没有改变。所以体裁具有相对稳定性。

综上阐述，体裁具有社会公认性、历史演变性、知识经验性、相对稳定性的特点。

3.3 互文性和体裁互文性

互文性（Intertextuality）概念最早是由法国符号学家朱丽娅·克里斯蒂娃在《符号学：意义分析研究》一书中提出的[24]，其认为任何文本都是其它文本的吸收和转化，它们相互参照，彼此牵连，形成一个潜力无限的开放网络。费尔克兰福将语篇的互文性分为语篇表层的互文性和语篇深层的互文性，并将语篇深层的互文性称为体裁互文性（Interdiscursivity）[21]。辛斌随之提出了体裁互文性（Generic Intertextuality）的概念，认为体裁互文性是指在一个语篇中不同风格（style）、语域（register）或体裁（genre）的混合交融[22]。体裁互文性蕴含着用户构造语篇过程中想要表达的深层次交互和传播目的，它是不同风格、语域或体裁的混合交融。这表明一个具体语篇的体裁是由多个基本体裁构成，可以表示成向量的形式，其维度为语篇空间所有的语篇基本体裁的数量。

体裁互文性由于不在字面上体现，其研究存在一定的困难。体裁与词汇、句法、像素、音节等语篇的具体表现细节相比，具有无清楚标记的隐蔽性、高度的概括性、层次的复杂性和视角的多样性等特点，甚至体裁本身的概念也有模糊不清的地方。所以，廓清体裁概念及建立体裁的形式化定义是语篇体裁互文性量化研究的基础。姜怡和姜欣[23]等人以典籍文本《茶经》和《续茶经》为试验对象，使用向量空间模型（VSM）度量文本互文性，在此基础上提出了改进方法，即扩展项之间的相似度计算和采用序列模型，并给出了基于文本互文性度量的文本翻译索引方法。

4 语篇体裁及其形式化描述

4.1 基本体裁的稳定性与具体语篇体裁的多样性

体裁形式永远处于一个动态的嬗变过程，体裁又是一个历史范畴，有其相对的独立性和内在的发展逻辑。体裁的兴起、发展及演化受历史环境的制约，体现时代特征、人们的生活方式和行为特征。体裁反映着语篇发展过程中的各种稳定倾向，体裁运动中的某些稳定因素在传承过程中得以保留，这些因素可以保证体裁在一定历史时期的一致性。这种存在于语篇内部的稳定因素，俄国学者 V. Propp 和瑞士研究者 M. Lüthi 最早称之为“本质性特征”。所以，在一定历史范畴内，体裁具有一定稳定不变的本质性特征，我们将这些特征把握的体裁称为“基本体裁”。Fowler 把具体语篇涉及的不同体裁分为“类”（Kind）和“式”（Mode）。类指的是通常意义上的体裁，由一组特定的语义和形式特征来定义；式与类有对应关系，但它们只是有其对应类的部分特征。对类与式的区分表明：整体上属于某一个体裁的语篇可以（而且往往）具有其它体裁的特征。Fowler 定义的“式”实际含有基本体裁的意味，体裁互文性造成某一个语篇的体裁具有多种基本体裁的特征。例如，在我国语文教学中，文章的体裁共计分为 7 小类：记叙文、议论文、说明文、诗歌、散文、小说、戏剧。可以认为纯小说或者纯诗歌是一种基本体裁，而报告文学则可能是由这 7 种基本体裁中的若干种构成的。前者就是基本体裁，而后者则是语篇体裁，基本体裁构成了语篇体裁当中的基向量。再如，查找电影网站豆瓣（www. douban. com）可以看到，其定义了总共 40 多种基本体裁，如悬疑、动作、爱情、科幻等。而任意一部电影都是由这些基本体裁所构成，如《让子弹飞》这部电影的语篇体裁包括剧情、喜剧、动作、西部等 4 个基本体裁。

4.2 体裁向量

由上所述，体裁存在本质性特征，由本质性特征规定的体裁为“基本体裁”，语篇体裁是基本体裁的交叉混合。社交网络中语篇的体裁由多种基本体裁组成。如一篇小说的基向量可以从题材、流派、表现手法、篇幅等多角度划分，那么这部小说的体裁可以是：神话小说＋浪漫主义小说＋叙述体＋长篇小说等，语篇体裁是基本体裁的混合。同样，以电影语篇为例，同一部电影可能具有科幻、恐怖、言情等多种体裁，也是多种基本体裁的交叉混合。

本文对电影社区当中语篇体裁的研究也是采用向量空间模型，将基本体裁定义成向量空间的基向量，用向量表示语篇的体裁。

定义 1 若社交网络语篇空间 V 中存在 n 个线性无关的向量 $a_1, a_2, \cdots, a_n$，使得 V 中的任何元素都可由它们表出，则称 $a_1, a_2, \cdots, a_n$ 为 V 的一组基，称这组基中的分量为 V 中基本体裁，基所含向量的个数 n 称为语篇空间 V 中的维数，记为 $\dim V$，并称 V 为 n 维语篇空间。

当 n 维语篇空间 V 的一组基 $a_1, a_2, \cdots, a_n$ 固定后，空间内的任何一个元素（语篇）可以表示成：

$$T=g_1 \cdot a_1+g_2 \cdot a_2+\cdots+g_n \cdot a_n \quad (1)$$

即 T 可以由$(g_1,g_2,\cdots,g_n)$表示，由此给出定义 2 如下：

定义 2 给定社交网络语篇空间 V，$G=\{a_1,a_2,\cdots,a_n\}$为 V 的基本体裁集合，语篇 T 的体裁 Tg 是一个 n 元向量组$(tg_1,tg_2,\cdots,tg_n)$，即

$$(tg_1,tg_2,\cdots,tg_n)=(g_1,g_2,\cdots,g_n)\begin{pmatrix}a_1\\a_2\\\vdots\\a_n\end{pmatrix} \quad (2)$$

其中，$tg_i\in[0,1]$，$i=1,2,\cdots,n$。如果 T 没有基本体裁 a_i 的特征，则 $tg_i=0$；如果 T 具有基本体裁 a_i 的特征，则 $tg_i>0$。

给定语篇空间 V，$\forall T_i$、T_j，其中 $i,j=1,2,\cdots,n$，对 T_i 和 T_j 之间体裁互文性度量用向量之间距离表示，距离越小，则认为 T_i 和 T_j 的体裁相似度越高。

假设 $\forall T_i$、T_j 体裁向量分别为：

$T_{gi}=(t_{gi,1},t_{gi,2},t_{gi,3},\cdots,t_{gi,n})$

$T_{gj}=(t_{gj,1},t_{gj,2},t_{gj,3},\cdots,t_{gj,n})$

其中，n 为基本体裁的个数。

T_1 和 T_2 的体裁互文性关系度量为 $sim(T_{gi},T_{gj})$，具体见式(3)。

$$sim(T_{gi},T_{gj})=\cos(T_{gi},T_{gj})=\frac{\sum_{k=1}^{n}(tg_{i,k}\times tg_{j,k})}{\sqrt{\sum_{k=1}^{n}tg_{i,k}^2}\times\sqrt{\sum_{k=1}^{n}tg_{j,k}^2}} \quad (3)$$

式中，使用余弦向量度量语篇间体裁的互文强度。夹角余弦值愈大，则语篇之间的夹角愈小，就可以认为语篇 T1 和语篇 T2 的体裁互文性越强。

用向量来形式化描述体裁，对于体裁已知的语篇，采用传统的向量空间模型来计算体裁互文性。而对于那些体裁未知的语篇，通过分析体裁互文性和用户行为的潜在共性特征，使用用户行为来度量语篇体裁互文性，因此引出下面的定义。

定义 3 如果语篇在用户行为中共现次数超过阈值，导致语篇之间存在的体裁混合交融，称为语篇体裁互文性。

性质 1 多个语篇在社区用户行为中共现的次数可以量化为语篇间体裁互文关系强度。

上述定义的前提是电影社区语篇体裁互文性和用户行为具有共性关系。为此，以下实验将验证这一结论的正确性。

5 实验及结果分析

5.1 实验方法设计

本文选取电影和资源网站豆瓣(http://movie.douban.com)的数据作为依据，依次选择不同的用户标签、用户数以及用户行为数据进行统计和分析。这些下载的数据使用 MySql 保存在本地，其统计数据结构如图 1 所示。

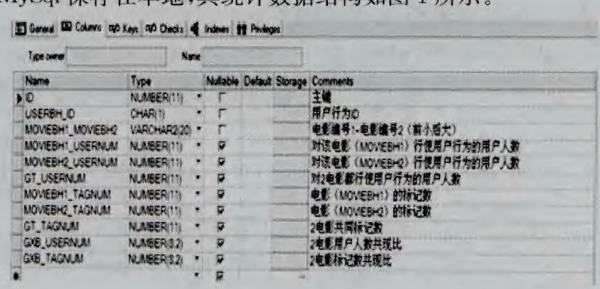

图 1 实验设计数据结构

具体实验方案设计如下：

1. 实验目的：统计豆瓣网上任意两部电影被共同看过的人数与共同的标签数，验证体裁互文性和用户行为的潜在共性关系。

2. 实验对象：豆瓣网络上的电影数据。

3. 实验步骤：现已得到豆瓣网上的原始电影数据 Movies、用户数据 Users、用户对电影的行为数据 UserBehavior-Movie(主要是 4 种用户行为数据：在看、看过、想看、影评)。

预处理(去噪)过程如下。

(1)筛选用户。要筛选出“看过”一定量的电影并有一定电影偏好的用户。

先求出每个用户“看过”的 15 种不同类型(对应着系统标签)的电影的数量。假设每个用户的这 15 种类型里最大值为 n，15 种类型数量的平均值为 m，过滤出平均值 $m>5$，最大值 n 减去平均值 m 的结果大于 5 的用户得到用户集 *Users*2，集合数量为 108(平均值和差值取 5 是经过多次实验得到的，这样得到的用户才能达到一定数量且有一定质量)。

(2)筛选电影。筛选电影分两步：

第一步：筛选出用户集 *Users*2“看过”的所有电影集合 1。

第二步：筛选出电影集合 1 中被用户集合 *Users*2“看过”次数超过一定值的电影，得到电影集合 *Movies*2，集合数量为 85。

(3)根据 *Movies*2 求集合中任意两部电影被共同看过(想看、在看、影评)的人数和共同的标签数。求被共同看过(想看、在看、影评)的人数时并不是限定在先前得到的用户集合 *Users*2 中的用户，而是面向集合 *Users* 中的所有用户。因为求 *Users*2 集合的目的只是为了得到集合 *Movies*2，求 *Movie*2 任意两部电影被共同看过(想看、在看、影评)的人数并不需要把用户限定在 *Users*2 中，如果限定在 *Users*2 中，则得到的数据量将会很少而达不到我们的目的。从而统计得到任意两部电影被共同看过(想看、在看、影评)的人数和共同的标签数集合 *MovieUserTag*。

(4)数据预处理。将任意两部电影被共同看过(想看、在看、影评)的人数作为横坐标，两部电影共同的标签数作为纵坐标进行绘图。但是如果将某些离散点作为坐标点，得到的将会是一个看不出规律的曲线。例如，如果两部电影被看过(想看、在看、影评)的人都多，很可能是由于这两部电影是大片，宣传做得好的效果，而不是人们对某些类型电影偏好的驱动。将得到数据进行分段求平均值处理，比如被共同看过的人数在 40～45 为一段，35～40 为一段，一直到 0 到 5 为一段。这样得到步长为 5 的 9 个平均点，分别计算看过分步长统计、想看分步长统计、在看分步长统计、影评分步长统计。整个实验结果如图 2—图 4 所示。

5.2 结果分析

图 2 显示了任意两部电影被共同用户看过的行为统计和这两部电影具有同样多个体裁标签之间的关系。整个曲线总的趋势呈线性递增，只是在 25～30 这个区间出现一个拐点，但是整个曲线的态势还是和实验预想吻合的，即在电影社区中，若两部电影被共同用户看过的数量越多，则这两部电影体裁具有更大的相似度，进而表明这两部电影体裁具有较强的互文关系。

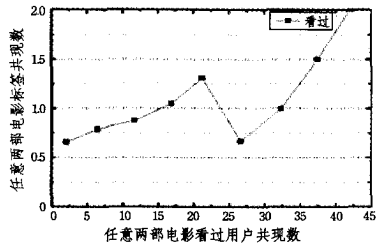

图 2　用户看过行为和电影标签关系图

图 3 显示了用户想看行为和体裁共现的关系曲线。由图 3 可知，用户想看行为和电影标签共现数规律曲线亦呈线性关系，只是图 3 中曲线线性增加趋势平缓，并且整个曲线当中也存在拐点。这其中一个主要原因是想看行为统计数据是二次计算的结果。在此之前系统已经对电影做了一个推荐，推荐算法的差异性和简单性都会使得在此基础上的统计数据出现一定的误差。

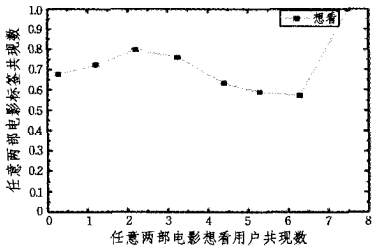

图 3　用户想看行为和电影标签的关系图

图 4 显示了用户影评行为和电影标签的关系。随着任意两部电影影评用户共现数目的增加，这两部电影的标签数也呈线性增加。只是在用户共现数为 7 个左右时，出现了一点波动，但是随之整个曲线还是呈上升趋势。这表明用户的影评行为也符合我们的设想，两部电影被共同用户评论的数量越多，这两部电影体裁越相似，体裁互文关系愈强。

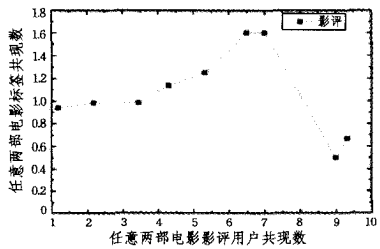

图 4　用户影评行为和电影标签关系图

另外，作者还收集了用户在看行为的统计数据，但是由于所获得的数据量太少使得曲线特征分散，不符合设计的需求，因此没有对其数据进行绘图和分析。

由以上可以看出，用户的看过行为数据最符合我们基本的设计思想。这是由于系统看过行为的统计数据最多，数据量最大。而统计学基本思想就是大量观察、平均分析、平衡分析等。样本的容量越大，样本的误差越小，也就是说，当所研究的现象越复杂、差异越大时，样本量要求越大；当要求的精度越高，可推断性要求越高时，样本量越大。想看和影评行为数据也基本符合要求，而用户在看行为由于所获数据量过少而没有看出规律。但是可以设想的是，如可以获得更多的电影网站在看行为数据，则在看行为也应该符合整体的统计规律和特征。

由上述实验可以基本得出一个结论：社交网络和语言学语篇之间的交涉具有共同特征，社交网络当中的用户行为和体裁互文性具有极大的隐性共性特征。换言之，语篇体裁越相似，社交网络当中用户的行为就越相似，因此可以用用户社交行为度量体裁互文性。

结束语　随着网络渗透率的提升和网民对于网络应用的深入，社交网站用户规模将会得到进一步扩大，越来越多的用户会将更多现实生活中的人际关系延伸到网络。本文的研究表明，社交网络的用户在信息行为中具备了很高的用户黏度，用户总是聚焦在自己感兴趣的语篇体裁上，大多数用户感兴趣的语篇集在体裁上也具有更大的相似度。由于使用目的不同，用户表现出的行为也有所不同。总体而言，社区用户在社交网站上的行为较为分散，但是语篇体裁互文性所体现的交流沟通和在某一个固定时间对特定语篇体裁的关注依然是用户行为的中心。

参 考 文 献

[1] 孔维泽，刘奕群，张敏，等. 问答社区中回答质量的评价方法研究[J]. 中文信息学报，2011，25(1)：3-8

[2] 王玉祥，乔秀全，李晓峰，等. 上下文感知的移动社交网络服务选择机制研究[J]. 计算机学报，2010，33(11)：2126-2135

[3] Facebook[OL]. http://www.facebook.com

[4] 方莺飞，林鸿飞，杨志豪，等. 中文文本体裁的自动分类机制[J]. 中文信息学报，2006，20(2)：24-32

[5] Stein B，Eissen S M. Retrieval Models for Genre Classification[J]. Journal of Information Systems，2008，20(1)：93-119

[6] 陆汝钤. 知识科学及其研究前沿[EB/OL]. http://www.iipl.fudan.edu.cn/research/ks.html，

[7] Shepherd M，Watters C. The Evolution of Cybergenres[C]// Proceedings of the Thirty-First Annual Hawaii International Conference on System Sciences. 1998：97-110

[8] Crowston K，Kwasnik B H. A Framework for Creating a Faceted Classification for Genres：Addressing Issues of Multidimensionality[C]// Proceedings of the 37th Hawaii International Conference on System Sciences. 2004：1-9

[9] Freund L S. Exploiting task-document relations in support of information retrieval in the workplace[D]. Toronto：Faculty of Information Studies，University of Toronto，2008

[10] Vidulin V，Luštrek M，Gams M. Using Genres to Improve Search Engines[C]// Proceedings of the International Workshop "Towards Genre-Enabled Search Engines：The Impact of NLP". Borovets，Bulgaria. September，2007

[11] Swales M，M J. Genre Analysis：English in Academic and Research Settings[M]. Cambridge：Cambridge University Press，1990

[12] Askehave I，Swales M. Genre Identification and Comminicative Purpose：A Problem and a Possible Solution[J]. Applied Linguistics，2001，22(2)：195-212

[13] Orlikowski W J，Yates J. Genre repertoire：The structuring of communicative practices in organizations[J]. Administrative Sciences，1994，33(1)：541-574

[14] Rosso M A. User-Based Identification of Web Genres[J]. Journal of the American Society for Information Science and Technology，2008，59(7)：1053-1072

[15] Askehave I，Nielsen A E. Digital genres：a challenge to traditional genre theory[J]. Information Technology & People，2005，18(2)：120-141

（下转第 288 页）

MSMRC 协议利用主用户的活动规律来设计可靠链路时长，充分考虑了主用户的活动因素可能对链路中断和节点失效造成的影响。

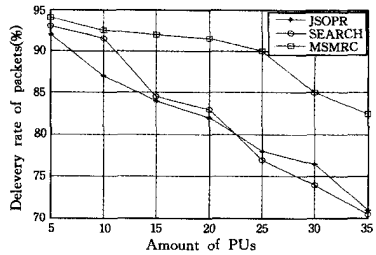

图 3　不同主用户数下的认知用户数据包投递率

不同主用户数量下的 3 种路由开销情况如图 4 所示。此处的路由开销以链路和节点失效时所需的路由重构分组数来表示。随着主用户数量的增加，平均路由开销都呈现上升趋势。相较于其它两种协议，MSMRC 协议的性能稍好一些，说明当通信连接事件频繁发生时，预先构造备份路由比传输过程中临时构造路由能够减轻网络负载和路由开销。

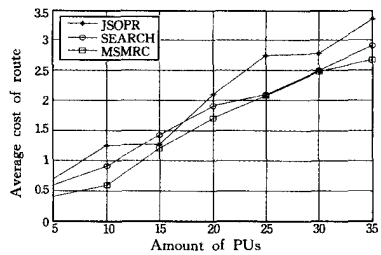

图 4　不同主用户数下的平均路由开销

链路修复成功率如图 5 所示。随着主用户的不断增加，由于采用了主路由和备份路由机制，MSMRC 协议具有较好的链路保持。而 JSORP 协议对主用户活动没有应对策略，SEARCH 协议不能提供可靠的路由修复机制，因此链路修复能力相对较差。

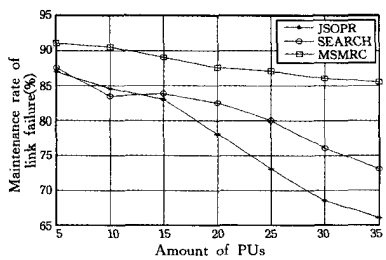

图 5　PU 数目与链路修复成功率之间的关系

结束语　本文提出一种适用于 CRAHNs 的可靠路由协议 MSMRC。该协议与传统路由相比，在路由发现和保持方面显著提升路由可靠性的同时，也能将数据包交付率保持在一个较高的水平上。MSMRC 的路由开销也处在一定可接受的范围内。但是，由于 MSMRC 在构造主路由和备份路由时，需要中继节点的合作而产生一定的节点计算开销和能量消耗。因此，在未来的研究中将重点分析这类因素对链路中断的影响程度，从而设计出更符合实际要求的路由策略。

参 考 文 献

[1] Akyildiz I F, Lee W-Y, Chowdhury K R. CRAHNs: Cognitive radio Ad Hoc networks[J]. Ad Hoc Networks, 2009, 7(5): 810-836

[2] 冯光升，郑晨，王慧强，等. 认知无线网络的认知能力保障方法研究综述[J]. 计算机科学，2014，41(5)：8-13，19

[3] Chowdhury K R, Felice M D. Search: A routing protocol for mobile cognitive radio Ad-Hoc networks[J]. Computer Communications, 2009, 32(18): 1983-1997

[4] Chowdhury K R, Akyildiz I F. CRP: A routing protocol for cognitive radio ad hoc networks[J]. IEEE Journal on Selected Areas in Communications, 2011, 29(4): 794-804

[5] Wang Jyu-wei, Adriman R. Analysis of Cognitive Radio Networks with Imperfect Sensing and Backup Channels[C]// Proceedings of Innovative Mobile and Internet Services in Ubiquitous Computing(IMIS). 2013: 626-631

[6] Shin D, Kim J, Ko Y-B. A hybrid topology based multicast routing for cognitive radio ad hoc networks[C]// Proceedings of Computing, Communication and Networking Technologies (ICCCNT), 2014. 2014: 1-5

[7] Al-Rawi H A A, Yau K L A, Mohamad H, et al. A reinforcement learning-based routing scheme for cognitive radio ad hoc networks[C]// Proceedings of Wireless and Mobile Networking Conference(WMNC). 2014: 1-8

[8] Lee J-J, Lim J. Cognitive routing for multi-hop mobile cognitive radio ad hoc networks[J]. Journal of Communications and Networks, 2014, 16(2): 155-161

[9] Samar A, Mustafa E. Metric-based taxonomy of routing protocols for cognitive radio ad hoc networks[J]. Journal of Network and Computer Applications, 2014, 40(7): 151-163

[10] Cheng Geng, Liu Wei, Li Yun-zhao, et al. Joint on-demand routing and spectrum assignment in cognitive radio networks[C]// Proceedings of Communications. 2007: 6499-6503

(上接第 272 页)

[16] Santini M. Automatic Genre Identification: Towards a Flexible Classification Scheme[C]// Proceedings of Future Directions in Information Access. BCS IRSG Symposium, 2007

[17] Santini M. Cross-Testing a Genre Classification Model for the Web[J]. Text, Speech and Language Technology, 2011, 42(3): 87-128

[18] Vidulin V, Luštrek M, Gams M. Using Genres to Improve Search Engines[C]// Proceedings of the International Workshop "Towards Genre-Enabled Search Engines: The Impact of NLP". Borovets, Bulgaria. September, 2007

[19] Rosso M A. User-Based Identification of Web Genre[J]. Journal of the American Society for Information Science and Technology, 2008, 59(7): 1053-1072

[20] Barbedo J G A, Lopes A. Automatic Genre Classification of Musical Signals[J]. EURASIP Journal on Advances in Signal Processing, 2007(1): 1-12

[21] Fairclough N. Discourse, Social Change[M]. Cambridge: Polity Press, 1992

[22] 辛斌. 语篇互文性的语用分析[J]. 外语研究，2000(3)：14-16

[23] 姜怡，姜欣，方淼. 基于互文性度量的文本翻译索引[J]. 计算机工程与设计，2010，31(15)：3490-3493

[24] 朱立元. 现代西方美学史[M]. 上海：上海文艺出版社，1993