文章编号：1003-0077(2014)03-0062-06

中文微博客的垃圾用户检测

李赫元[1,2]，俞晓明[1]，刘 悦[1]，程学旗[1]，程 工[3]

（1. 中国科学院 计算技术研究所，北京 100190；
2. 中国科学院大学，北京 100190；
3. 国家计算机网络应急技术处理协调中心，北京 100029）

摘 要：微博客的出现改变了我们获取信息的方式。然而，大量垃圾消息却此起彼伏，危害着微博的健康发展。该文研究了中文微博客中的垃圾用户检测问题。我们首先对垃圾用户的行为进行了分析，提出了基于用户图、用户资料、微博内容的 3 大类 7 种检测特征。随后，讨论了基于 SVM 分类器的垃圾用户检测方法。最后，我们对采集的微博数据进行了标注，并评价了分类器的效果。实验表明：分类器具有较高的准确率和召回率，该文提出的特征具有较好的区分度。

关键词：微博客；垃圾用户；检测

中图分类号：TP391 **文献标识码**：A

Research on Detecting Spammer in Micro-blogs

LI Heyuan [1,2], YU Xiaoming [1], LIU Yue [1], CHENG Xueqi [1], CHENG Gong[3]

(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100190, China;
3. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China)

Abstract: Micro-blogs changes the way people obtain information. However, Micro-blogs has been infiltrated by large amount of spam, which is a challenge to normal user. In this paper, we research on spam in Chinese Micro-blogs. We study the behavior of spam user and propose 7 new features for detecting them. Then, we describe how to apply features into detecting spammer via a SVM classifier. The experiment results indicate that the accuracy and recall of the proposed method is satisfactory.

Key words: Micro-blogs; spam; detection

1 引言

微博客（简称微博）是一种基于用户关系的信息分享、传播与获取平台。近几年，中文微博服务发展迅猛，截止 2012 年 5 月，新浪微博的注册用户已达 3 亿、每日发布的消息量超过 1 亿条[1]；腾讯微博的注册用户数也已超过 4 亿。微博的出现不仅改变了信息的传播方式，也改善了我们的生活质量。然而，微博上却充斥着炒作、营销、谣言等不良信息，困扰着微博的健康发展。如何对垃圾用户及其发布的垃圾消息进行识别，已经成为了亟待解决的问题。目前，相关研究工作主要集中在 Twitter 等英文微博中，中文微博与英文微博之间存在着一些较为显著的差异。

（1）评论模式与提及

在 Twitter 中，转发只显示原作者。Twitter 中的提及是用户之间的直接交互。针对这一点，国外学者提出了基于提及关系的检测方法。然而在中文微博中，转发的同时可以提及（@）原作者，使得两者难以区分。因此，利用提及关系的检测方法并不适用于中文微博。

（2）互粉行为

中文微博引入了“加 V”“人气用户”等概念。为

收稿日期：2013-03-28 **定稿日期**：2013-05-20

基金项目：国家自然科学基金(61100083)；国家 863 计划基金(2012AA011003)；国家 242 专项(2011F45, 2011F65)。

了提升自己的粉丝数量，新浪等微博中出现了大量的“互粉行为”，即主动关注别人并要求对方也关注自己。这一具有中文微博特色的现象，将影响“用户权威度”等在 Twitter 中有效的垃圾用户检测特征。

(3) 对待垃圾用户的态度

Twitter 中，官方开设了“spam 账号”接受举报。举报信息公开、透明、便于采集，许多学者都选用举报信息作为垃圾用户的范本。中文微博客虽然提供了举报功能，但信息是不公开的。因此，在中文微博客中我们无法自动取得垃圾用户的标注信息。

本文以中文微博客为重点，探索垃圾用户检测方法。本文的创新之处在于：(1) 关注中文微博中的“互粉”行为，并据此提出了新的用户图特征；(2) 研究了注册时间与垃圾用户行为的关系，并据此提出了近期活跃度特征；(3) 讨论中文微博开放平台中的应用，提出了应用来源黑、白名单特征。

本文余下的部分将如下安排：第 2 节讨论国内外的相关工作。第 3 节从用户图、用户资料、微博内容 3 个方面提出了 7 种新的检测特征。第 4 节首先介绍了数据的采集、标注，接着进行了分类器的训练、实验。通过对实验结果的分析，验证了特征的有效性。第 5 节对本文工作进行总结与展望。

2　相关工作

Grier[2] 等研究了 Twitter 中包含 URL 的微博消息。统计表明，约 8% 的 URL 指向垃圾网页。研究还对垃圾账号进行了研究。只有 16% 的垃圾账号从注册之初就在发布垃圾 Tweet；余下 84% 的账号均是被盗用的，其行为特征是，账号注册于很久之前并已经被弃用，直到近期突然开始散布垃圾消息。

Wang[3] 等将垃圾用户的检测转化为机器学习的分类问题。该研究提出了 5 种检测特征，“用户权威度”、“重复 Tweet 率”、“含 URL 的 Tweet 比率”、“含提及(@)的 Tweet 比率”、“含话题(#)的 Tweet 比率”。统计结果表明，垃圾用户在上述各项指标中都略区别于普通用户，但多数特征不具有鲁棒性。应用上述特征构造的朴素贝叶斯分类器可以达到 91.7% 的准确率。

Song[4] 等从社交关系网的角度研究了 Twitter 中的垃圾用户。该研究提出了“用户距离”和“用户连通度”两个特征：若两个距离大于 4 的用户之间相互“提及”，则可认为是在传播垃圾信息；在用户距离相同的情况下，正常用户之间的连通度要强于垃圾账户之间的连通度。利用上述两个特征进行检测，可以达到 94.6% 的准确率。如第 1 节所述，在中文微博中，提及和评论混在了一起，因此该检测方法难以应用于中文微博中。

在国内的研究中，王宇[5] 等人对新浪微博中的“僵尸粉”进行了研究，总结出了“用户微博数”“用户是否包含简介”等 6 种具有区分度的特征。其中“用户昵称可疑度”等特征需要借助人工识别。该研究同样使用朴素贝叶斯算法训练分类器，准确率达到了 88%。

3　垃圾用户检测

3.1　检测特征

本节将从用户图、用户资料、微博内容三个方面，提出垃圾用户检测特征。在讨论相关研究中提出的 5 种检测特征的基础上，我们新提出了“纯粉丝度”“黑名单应用”“用户用字多样性”等 7 种新的检测特征。

3.1.1　用户图特征

微博中，用户的关系可以用有向图表示：出度表示“关注”，入度表示“粉丝”(被关注)；若用户彼此关注了对方，则称为“互粉”(互为粉丝的简称)。以图 1 为例，C 关注了 A，A 关注了 B；B 和 C 互粉。A 的粉丝是 C；B 的粉丝是 A 和 C；C 的粉丝是 B。

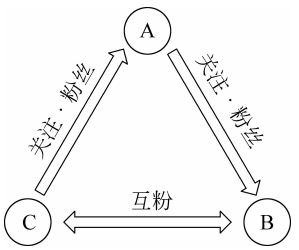

图 1　“关注”“粉丝”和“互粉”

用户权威度

在 Wang 等人的研究中[3]，定义了特征“用户权威度”，见式(1)。N_{follow} 表示用户 u 的粉丝数，N_{friend} 表示该用户的关注数。若用户既没有粉丝也没有关注，规定该用户的权威度为 0。

$$Authority(u)=\frac{N_{follow}}{N_{follow}+N_{friend}} \tag{1}$$

用户关注度

为了探讨垃圾用户、普通用户在主动关注方面有无差异，我们提出了“用户关注度”特征，如式(2)所示。N_{follow}是用户的粉丝数，N_{friend}是用户的关注数。直观地分析：垃圾用户会大量的关注别人，却很少得到别人的关注，因而该特征偏高。

$$FriendsRate(u)=\frac{N_{friend}}{N_{follow}+N_{friend}} \quad (2)$$

纯粉丝度

在中文微博中，用户之间的“互粉”可以提高双方的粉丝数，也会干扰“用户权威度”特征的区分度。为了避免这种情况，我们定义了纯粉丝度，见式(3)。N_{follow}依然表示用户 u 的粉丝数；分子部分为去除了互粉用户之后的“纯粉丝数”。该特征描述了粉丝质量。

$$RealFollow(u)=\frac{N_{follow}-N_{bi\text{-}follow}}{N_{follow}} \quad (3)$$

3.1.2 用户资料特征

在王宇等人的研究中[5]，提出了“用户简介”“微博域名”特征。本节将提出“用户头像特征”“近期活跃度”两个新的特征。

用户头像特征

在本研究中，我们对用户头像的图片进行了采集，并提出了“用户头像”特征，它识别用户是否使用了默认头像。若为默认头像，$g(u)=0$；若上传了头像，$g(u)=1$。

$$Gravartar(u)=g(u) \quad (4)$$

近期活跃度

用新账号发布垃圾信息很容易被识破。因此，垃圾用户更倾向于使用注册时间较长的“沉睡账号”散播垃圾信息。我们定义了“近期活跃度”指标，见式(5)。$N_{DuringDays}$是用户最近 100 条消息跨越的天数。$N_{CreateDays}$是截止采集当天，用户账号已存活的天数(注册天数)。对于突然活跃的沉睡账号，上述特征值会偏高；而对于一直活跃或已经很久不活跃的正常用户而言，这一特征值较低。

$$RecentActivity(u)=\frac{N_{DuringDays}}{N_{CreateDays}} \quad (5)$$

3.1.3 微博内容特征

基于内容过滤的技术在反垃圾邮件等领域取得了很好的效果。然而，这类方法需要大量的训练数据和标注样本。在国外的研究中，主要从“微博中是否含 URL”“微博是否大量重复”两个方面考虑内容特征。本节结合了中文微博客的特点，提出了“用字多样性”、“白名单应用率”、“黑名单应用率”三个新的检测特征。

不含 URL 微博比例

由于微博消息的长度限定在 140 字以内，垃圾用户通常会选用“在微博中附加链接”的方式推广垃圾信息。基于此，Benevenuto[6] 等提出了“不含 URL 微博比”这一特征，如式(6)所示。N_{All}是用户发布的微博数量，N_{NoURL}是在所发微博中不含有 URL 的数量。

$$NoURLRate(u)=\frac{N_{NoURL}}{N_{All}} \quad (6)$$

应用来源的白名单率和黑名单率

中文微博客推出了开放平台，提供了丰富多彩的应用，它们具有获取微博、发送微博、关注等操作权限。通过开放应用发布微博时，应用名称会显示在“消息来源”字段中。借助这一字段，我们对开放应用进行了研究，分为如下三类。

(1) 黑名单应用

为了降低维护成本、垃圾制造者会使用“皮皮时光机”等开放应用管理微博。在采集数据中，我们选取垃圾用户使用最多的 30 个应用作为黑名单。当然，使用黑名单发送的微博不一定都是垃圾信息，例如，很多用户会选用“皮皮时光机”定期地转发热门微博。

(2) 白名单应用

随着移动互联网的发展，越来越多的用户使用手机客户端发微博。但对垃圾用户而言，用手机客户端管理数百个账号的成本过大。因此，我们将手机客户端定义为白名单应用。

(3) 其他应用

并非全部的开放应用都符合上述两种分类。如“微博桌面”等应用，正常用户使用它们来发微博，垃圾用户通过“模拟点击”的方式操控账号。由于其不具有明显区分度，我们在特征研究中不使用这类应用。

在如上所述的分类基础上，我们定义了“白名单率”“黑名单率”两个特征：

$$WhiteList(u)=\frac{N_{WhiteList}}{N_{All}} \quad (7)$$

$$BlackList(u)=\frac{N_{BlackList}}{N_{All}} \quad (8)$$

N_{All}是用户 u 发布的微博总数。$N_{WhiteList}$是通过白名单发送的微博总数，$N_{BlackList}$是通过黑名单发送的微博总数。根据预期，垃圾用户具有较高的黑名单率和较低的白名单率。

微博相似度

在 Wang 的研究中[3]，使用了编辑距离计算微博的重复度，并以此作为检测特征。但在中文微博中，用户很少发布重复消息。我们应用余弦距离计算用户微博相似度，如式(9)所示。该特征计算了用户 u 所发布的 n 条微博之间的余弦相似度均值。

$$Sim(u)=\frac{1}{n}\sum_{i=0}^{n}\sum_{j=0}^{n}\frac{\vec{V}_i\cdot\vec{V}_j}{|\vec{V}_i|\cdot|\vec{V}_j|} \tag{9}$$

微博用字多样性

为了研究正常用户与垃圾用户在微博消息的用词(字)上有无差异，我们定义了用字多样性这个特征，见式(10)。由于微博消息具有长度短、用词不规范等特点，我们以字为最小分割单位，并在处理前删除消息中的 URL 链接。

$$WordDiversity(u)=\frac{cnt_d}{\sum_{i=0}^{n}len_i} \tag{10}$$

假设用户 u 一共发了 n 条微博，每条消息的长度记为 len_i，则全部消息的总长度为 $\sum_{i=0}^{n}len_i$。同时，统计这些消息中非重复的单字数，记为 cnt_d。

分类器与检测

垃圾用户的检测问题，可以视为一个分类(Classification)问题。假设微博用户的全集为 U，类别集合 $C=\{C_{spam},C_{normal}\}$，$C_{spam}$ 表示垃圾用户、C_{normal} 表示正常用户。垃圾用户的检测问题，即为求一个分类函数 F，将 U 中的微博用户影射到类别 C 上。

$$F:U\rightarrow C \tag{11}$$

上述影射函数 F 即代表了一个分类器，它可由机器学习算法习得。在本研究中，选用支持向量机(SVM)算法，它是一种有监督的机器学习算法，可以解决分类、回归等问题。对于分类问题，SVM 通过预定义核函数的非线性变换，将输入空间变换到一个高维空间，在后者中求广义的最优分类面[7]。常用的核函数主要有三种：多项式函数(Polynomial)、径向基函数(Radial Basis Function)和 Sigmoid 函数。

基于 SVM 分类器的垃圾用户检测流程如图 2 所示。

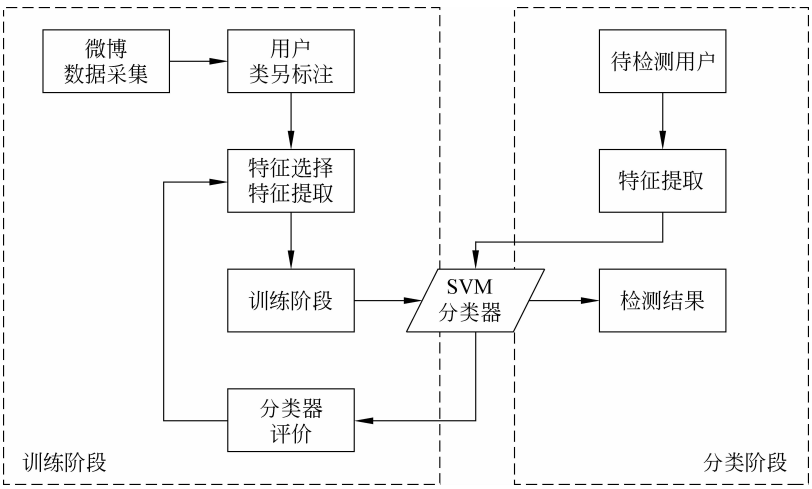

图 2　垃圾用户检测流程

检测可分为“训练”“分类”两个阶段。在训练阶段，我们对采集的微博用户数据进行标注，提取 3.1 节所述的检测特征。接着，用 SVM 训练分类器并对其效果进行评价，如有必要，将选择新的特征(集)并重新训练分类器。在检测阶段，提取待检测用户的特征，并使用训练阶段得到的 SVM 分类器进行分类，分类结果(C_{spam} 或 C_{normal})即为检测结果。

4　实验与分析

4.1　数据的采集与标注

我们使用 OAuth2 和 API 开发了新浪微博采集器，它的首要任务是搜集一些用户样本。微博的 API 提供了“public_timeline”接口，它会随机返回最新发言的 20 个用户及其微博信息。我们使用该接口，于 2011 年 12 月 20 日进行了用户数据的采集。

经过去重处理后，共包含 145 317 个微博用户、1 522 092 条微博。

上述采集只能获取部分活跃用户的最新微博信息，却并不包含用户的全部消息。此外，如 3.1 节所述，用户的表现行为与其注册时间有一定关联。为此，我们在 2012 年 6 月 23 日进行了二次采集。本次我们使用“user_timeline”接口，它根据第一次采集的用户 ID，获取用户头像等资料、粉丝数等信息，并抓取用户最新发布的 100 条微博。本轮采集后，共包含 125 683 个有效用户、4 657 811 条微博（新浪会定期清理垃圾账号，导致用户数减少）。

为了将二次采集的数据用于分类器，随机抽取了 3 000 个用户作为实验的数据集。为了获取客观、准确的标注结果，开发了标注平台，并邀请 3 位评价者对实验进行随机、交叉标注。标注遵循如下标准。

1）垃圾用户

标注值为－1，其行为特征为：转发大量广告、炒作消息；发布的微博具有明显的商业意图（如商品推介）；发送或转发大量低质微博（如心灵鸡汤）。

2）正常用户

标注值为 1，其行为特征为：发布较为贴近生活的微博（如聚会、心情等）；在转发微博时，包含了个人的见解、评价（转发加评论）；与他人存在较为真实的互动（如相互@、评论）。

3）不确定用户

标注值为 0，若评价者认为账号难以区分，可将其标注为“不确定用户”。该类用户可能同时具有垃圾用户、正常用户的部分行为。

在标注者完成标注后，我们对数据进行了筛选与清理：首先，选择出至少 2 位评价者给出一致标注值的用户账号；其次，去除标注结果为“不确定”的账号。经过上述处理后，共剩余 2 471 个用户，本研究使用它们作为训练、测试数据。

4.2 实验结果

本研究使用 LIBSVM[8] 软件训练分类器。在效果的评价方面，选用准确率（Accuracy）、召回率（Recall）和 F 值。为了说明指标在本研究中的意义，考虑如表 1 所示的混淆矩阵。

表 1 混淆矩阵

		预测情况	
		分类成垃圾用户	分类成正常用户
真实情况	实际是垃圾用户	A	C
	实际是正常用户	B	D

其中，准确率（*Accuracy*）描述了分类器将垃圾用户、正常用户正常分类的百分比。

$$Accuracy = \frac{A+D}{A+B+C+D} \tag{12}$$

召回率（*Recall*）表明了检测出的垃圾用户中，真实垃圾用户的百分比。

$$Recall = \frac{A}{A+C} \tag{13}$$

F 值则综合考虑了准确率和召回率。

$$F = \frac{2 \cdot Accuracy \cdot Recall}{Accuracy + Recall} \tag{14}$$

在明确了指标之后，我们从标注数据集上提取出第 3 节提出的各种特征，并采用 10 折交叉验证的策略，对分类器进行训练、验证。表 2 记录了两组实验结果：F_ALL 和 F_OPTIMAL。

表 2 分类器实验结果

实验（ID）	准确率（Accuracy）/%	召回率（Recall）/%	F 值（F Measure）/%
F_ALL	93.46	97.64	95.51
F_OPTIMAL	94.40	97.71	96.03

在第一组实验，F_ALL 中，我们选择了第 3 节提出的全部 12 个特征训练分类器。实验结果如表 2 第 1 行所示：分类器的准确率达到 93.46%，召回率为 97.64%。

在进行该实验的过程中，我们发现部分特征具有“负效果”，会降低分类器的准确率。为了找出最优的特征组合，我们使用 Wrapper[9] 策略对 12 种特征进行选择：首先，求出特征组合的幂集，它共包含 $2^{12+1}=8\,192$ 个特征组合；其次，使用上述每一种特征组合训练分类器，对用户进行检测，计算分类结果的 F 值；最后，选出 F 值最高的特征组合，作为最优特征组合。最优组合共包含 7 个特征，如表 3 第 1 列所示。

为了验证最优组合中不同特征的贡献，我们单独使用每一特征训练分类器，并计算其 F 值，如表 3 第 2 列所示。从表中不难发现：应用黑名单率、纯粉丝度等本文提出的特征排名较为靠前，说明其具有较好的区分度。

在上述研究的基础上，我们进行了第二组实验。使用如表 3 所述的最优特征组合作为特征集合，训练、测试分类器。实验结果见表 2 的第 2 行，F_OPTIMAL。与第一组实验对比，准确率提升到了

94.4%，召回率为 97.71%。

表 3　最优特征组合

特征名称 (Feature Name)	F 值 (F Measure)
不含 URL 的微博百分比	0.807 3
应用黑名单率	0.634 8
纯粉丝度	0.584 6
近期活跃度	0.428 0
应用白名单率	0.387 3
微博域名	0.130 4
用户简介	0.115 1

在 3.1.2 节，我们提出了“用户头像特征”，但在最特征优组合中却不包含该特征。相反，前人提出的“简介”“微博域名”等特征却具有较好的效果。为此，我们对这三种特征进行了统计研究，结果如图 3 所示。在全部测试数据中，只有约 3%的用户使用了默认头像，特征数据的不均衡导致它失去了应有的区分度。相反地，“简介”“微博域名”等特征的分布相对均衡，具有一定的区分度。

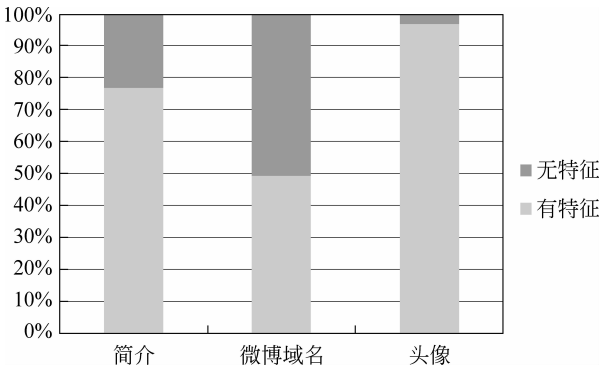

图 3　用户资料特征的统计特征

4.3　特征分析

用户图特征

由 4.2 节实验可知，本文提出的“纯粉丝度”比前人提出的“用户权威度”更能显著区分垃圾用户。在中文微博中，许多用户选择了“互粉”“刷粉”等不正当手段来提高自身人气。纯粉丝度能很好地过滤掉“互粉”导致的“假人气”，因而具有更好的区分度。另一方面，为了更好地伪装自己，垃圾账号之间往往会互粉，形成错综复杂的关系网，造成受到高度关注的假象。因此，本文提出的“用户关注度”特征并没有取得预期的效果。

用户资料特征

本文提出的“近期活跃度”特征效果良好。垃圾用户的确会使用早期注册的“沉睡账号”发布垃圾消息。而本文提出的“用户头像”特征效果不佳。如 4.2 节所述，用户头像的分布极不均匀，致使该特征很难表现出应有的区分度。

微博内容特征

本文提出的“应用白名单率”“应用黑名单率”均具有很好的区分度。首先，手机客户端的应用门槛和管理成本阻碍了垃圾用户。其次，确实有大量的垃圾微博是通过黑名单中的应用传播的。此外，传统的“微博相似度”效果不佳。本文提出的“微博用字多样性”考虑了用户的微博用词习惯，取得了较好的效果：正常用户的微博话题广泛、用词随意，多样性较高；垃圾用户传播的信息较为单一，用词单调。

5　总结和展望

本文研究了中文微博客中垃圾用户的检测问题。研究从用户图、用户资料、微博内容三个方面提出了 7 种新的垃圾用户检测特征。利用上述特征训练的 SVM 分类器，取得了较好的准确率和召回率。实验表明，本文提出的“纯粉丝度”“用户近期活跃度”等 5 个特征具有良好的区分效果。

在实验与研究中，我们也遇到了一些问题：(1)对采集数据的标注依靠人工判别，工作量巨大。有必要寻找一种更好的实验数据标注方法。(2)在本文中，垃圾用户分类器的召回率较为理想，但分类器的准确率只有 94%，仍有一定上升空间。我们将在未来的工作中对上述问题进行更为深入的探索与研究。

参考文献

[1] 新浪科技. 新浪微博用户数超 3 亿 [EB/OL]. 2012-05-16. http://is.gd/Qfn4Z9.

[2] Grier C, Thomas K, Paxson V, et al. @spam: The Underground on 140 Characters or Less [C]//Proceedings of the 17th ACM Conference on Computer and Communications Security (CCS 2010). New York, US, 2010: 27-37.

[3] Wang A. Don't follow me: Spam detection in Twitter [C]//Proceedings of the International Conference on Security and Cryptography. Athens, Greece, 2011: 142-151.

[4] Song J, Lee S, Kim J. Spam Filtering in Twitter Using Sender-ReceiverRelationship [M]. Berlin, German: Springer, 2006: 301-317.

(下转第 74 页)