Gender Classification Research on Web Forum Users' Posting Behaviors and Posting Contents

Yue Chen and Lijuan Bai

School of Management, Harbin Institute of Technology, Harbin, Heilongjiang, China
(chenyuehit@126.com), (baileejuan@126.com)

***Abstract*—The rise of Web2.0 makes more and more people participate in the network to exchange their views and information. In order to investigate whether gender differences exist in internet behaviors, this paper examined users' posting behaviors and language usages in network forum. This paper selected cars and stock forum message contents from a Chinese forum as research data. By hypothesis testing approach, this study confirms gender differences in posting behaviors. Text classification algorithms were used to investigate the differences existing in language usage between male and female users. Results showed that gender classification can be carried out by analyzing user's review content. By comparing several classification results, the SVM was found to have the highest classification efficiency.**

***Keywords*—Gender difference, user internet behavior, text mining, text classification, web forum**

网络论坛用户发帖行为及发帖内容的性别分类研究

陈越　白丽娟

哈尔滨工业大学管理学院，哈尔滨，黑龙江，中国

摘　要 Web2.0 的兴起使得越来越多的人们参与到网络交流中来表达意见并交换信息。为了探索男性和女性用户是否存在的网络行为差异，本文检验了不同性别用户的发帖行为及发帖语言运用中的差异。本文选取天涯汽车论坛和股票论坛的留言内容作为研究数据，以假设检验的方法对用户的发帖行为进行验证，发现不同性别用户存在的发帖行为差异。利用几种文本分类算法对评论内容进行分类分析，证明可以通过评论内容进行性别分类研究，并发现支持向量机是分类效率相对最高的分类算法。

关键词　性别差异，　网络用户行为，　文本挖掘，　文本分类，　网络论坛

1．引言

互联网的产生和发展改变了人们传统的交流沟通模式。更重要的，各种网络论坛的产生给人们提供了更多获取自己所需要的产品或其他信息的渠道，为来自不同地区、不同国家的人们的平等交流提供了平台。但是，在承认网络给人们的生活带来进步和发展的同时，不可忽视的是，网络毕竟是一个虚拟的世界，在电脑的另一端与你交流的用户到底是否是其所呈现的那样是不得而知的。通常，人们希望知道与自己交流的人的性别，以估计对方的反应和理解能力，以此采取恰当有效的沟通方式。通过对人类文化的研究发现，性别往往是与不同的社会规范，角色和沟通方式相关的。那么，在男性和女性用户中在网络论坛的发帖行为及语言使用中是否存在显著不同？如何通过对用户的发帖行为、发帖内容及评论或撰写文章内容等信息对作者性别进行判断并加以利用，值得学者对此进行深入探讨。

本文针对互联网中的中文论坛数据，以假设检验的方法对用户的发帖行为进行验证，发现了男性与女性用户的行为差异；运用文本挖掘方法对用户发表的评论内容进行分类分析，证明可以通过评论内容进行性别分类研究，并通过对比几种分类算法的分类效率找出了分类效率相对较好的分类算法。

2．文献综述

在网络环境下，男性和女性的差异不仅仅表现在网民数量上，更主要的表现在网民网络行为上。总结国内外学

者的研究，在网络环境下男性和女性的差异主要表现在:网络使用意图、电子商务购买行为差异和网络环境下的语言运用差异。

2.1 网络使用意图差异

Imhof 等人分已经证明男性和女性对网络技术的态度，使用网络时的心态存在着显著地差异[1]。Lindsay Shaw 和 Larry Gant 通过对网络用户信息的统计发现，在使用网络的频繁性确实存在着男性和女性差异。男性更频繁地登录网络，而且每天的登录时间也比女性登录时间要长[2]。但也有其他的学者证明在上网时间上并不存在性别差异。另一个热门的研究方向是男性和女性在使用网络时的使用意图不同。网络使用意图可以分为:交流，信息收集和娱乐二个方面，女性用户更多的是交流为动机，而男性网络用户则更多的是与信息收集和娱乐为目的的使用网络[3]。

2.2 网络环境下的语言运用差异

国外学者对在线交流系统中不同用户的交流内容进行了研究发现，在不同性别的用户之间在语言表达和文字运用中确实存在着差异[4,5]。以最具代表性的 Susan Herring 的研究为例，她通过对网络论坛内容的研究，发现交流中确实存在着性别差异，并且将这些不同的原因分为:表达形式和价值系统。表达形式主要是指用户在网络留言中所使用的词语，内容长短，语气等。男性网络用户的留言内容往往包含一些挑战性，对抗性的语言；女性用户的表达形式则趋于温和的，正面一些。但还有一部分用户的表达形式在环境和交流对象的不同还是会发生变化。价值系统主要指的是对于不同性别的网络用户对于在网络环境下的礼貌或粗鲁的定义不同的，即男性和女性用户在网络交流中存在着不同的价值观念，这些观念可能在现实世界中并没有太明确的体现，而在网络中则有较为明显的差异[6]。

通过对网络使用意图、网络购买行为和网络语言运用等方面的分析发现，在网络环境下确实存在着性别差异，这种差异并不是单方面的是渗透在用户网络行为的方面，从开始接触网络的初衷到网络行为都受到性别的影响，特别是对于网络环境下语言运用的差异也证明了通过文本分类的方法区分不同性别的用户的思想是可以实现的。

3．研究方法及数据

通过对天涯论坛数据的分析中文环境下男性和女性是否同样存在发帖行为和内容差异。以下载的网络论坛数据为基础，应用统计学方法，文本挖掘等方法分析中文网络环境下的男性和女性之间的差异。

结合已有的网络论坛的规模，影响力等因素考虑最后选择“天涯论坛-汽车论坛版面”和“天涯论坛-股市论谈”作为论文数据来源。下载的主要内容包括：发帖内容、发帖时间、发帖用户信息等，用户信息包括性别，年龄，用户现居地以及家乡等信息。获取了 2008 年 7 月 20 日至 2012 年 11 月 9 日的汽车论坛帖子 207811 条，用户信息 18253 条，股票论坛帖子 3956378，用户总数为 52686。最后随机选择了男性和女性用户发表的汽车评论和股票评论各 3000 条作为分析样本（总数 3000*2*2=12000 条）。

4．网络行为差异研究

4.1 研究设计

通过研究性别与语言的研究发现此类研究主要由两个理论进行支持。第一种是社会分工理论。在社会中角色的不同会有不同的表现模式，根据研究，男性和女性属于不同的社会语言文化，这种不同会表现在男性和女性语言使用中。第二种是身份理论。虽然现在已经没有非常严重的重男轻女的现象。但是不可否认男性在社会中占有主导地位，女性往往处于被支配和配合的角色。这种社会地位的差异在男性和女性的语言表达中表现得非常明显。基于以上学者的研究和文章数据特点提出以下几点假设：

H1：男性的评论内容与女性相比更长。该假设是在设分理论的基础上提出的，男性在交流中往往处于主导地位，而假设这种主导的体现也包括使用更多的内容进行表达。

H2：男性与女性相比更多的进行论坛留言交流。从身份理论和以前相关研究发现，男性在交流中表现出更多的竞争性，语气更为强硬而且与女性相比更频繁地发起话题等。

H3：男性发帖时间一般较晚于女性。根据社会角色和自身健康等条件考虑，男性相对女性更喜欢熬夜，女性在白天进行家务劳动或更操心家务等所以比较少进行熬夜进行发帖评论。

以上三个假设从不同角度对不同性别用户在网络环境中的发贴行为进行假设，通过统计检验这些假设的可靠性，并发现不同用户的发帖规律等。

4.2 结果分析

为了验证以上三个假设，以用户性别为预测变量，分别提取了汽车论坛和股票论坛的男性和女性各 3000 条数据进行研究分析。通过 T 检验，发现在评论内容，评论频率以及评论时间三个假设中在汽车论坛和股票论坛中存在的规律如下表所示：

1）评论内容：

表 1 汽车和股票论坛评论内容长度 T 检验结果

数据	T 值	P 值
汽车论坛数据	1.046	0.295(>0.05)
股票论坛数据	-3.995	0.000(<0.05)

检验结果表明：汽车论坛发帖内容长度并不存在性别差异，股票论坛中男性的评论内容平均长度显著长于女性评论内容长度。

2）评论频率：

表 2 汽车和股票论坛发帖频率 T 检验结果

数据	T 值	P 值
汽车论坛数据	-5.500	.000(<0.05)
股票论坛数据	-4.031	.000(<0.05)

检验结果表明：在汽车和股票两个板块中男性更加频繁地进行评论行为。

3）评论时间：

表 3 汽车和股票论坛发帖时间 T 检验结果

数据	T 值	P 值
汽车论坛数据	5.728	.000(<0.05)
股票论坛数据	-.222	.824(>0.05)

检验结果表明：汽车评论数据的检验结果表明男性发帖时间晚于女性，但该规律并不适用于股票论坛，即在股票论坛评论中检验不显著。

通过对比汽车和股票论坛的数据发现，对于网络环境下用户的行为其实并没有非常大的差异。在评论内容长度和发帖时间两个研究中都得出的结果在汽车和股票论坛中并不一致，即表明对于评论内容和发帖时间的研究并不存在非常显著的性别差异，通过研究长度和时间并不能得出其发帖用户的性别。发帖频率的验证结果表明，男性更多的参与发帖，这表明在网络环境下男性的参与意愿更强烈一些，该结果也与国外相关研究的结果一致。

5．性别分类研究

对于中文网络环境下男性和女性差异的研究，除了分析男性和女性网络行为的差异外，还包括从男性和女性语言使用差异方面进行研究。通过对男性和女性用户在论坛发表的内容进行文本分析发现，男性和女性在语言表达，词汇使用方面的差异。

为了去除论坛数量的多少对研究结果的影响，在进行分本分析研究的时候分析提取了男性和女性 3000 条论坛数据作为研究的基础数据。并且为了研究评论内容的长度对于分类结果的影响程度，对这 6000 条数据根据其内容的长短进行了分类，如下：

表 4 汽车论坛数据

数据长度	>0	>10	>30	>60	>90	>120	>150
男性	3000	2773	2023	1427	1065	838	680
女性	3000	2738	1969	1376	1070	865	718
总数	6000	5511	3992	2803	2135	1703	1398

表 5 股票论坛数据

数据长度	>0	>10	>30	>60	>90	>120	>150
男性	3000	2679	1759	1198	908	726	596
女性	3000	2675	1791	1147	851	618	483
总数	6000	5354	3550	2345	1759	1344	1079

通过对这些数据在应用 CFS+BestFirst 方法进行特征词提取，通过几种常用的文本分类算法：朴素贝叶斯，支持向量机，lingpipe 以及统计分类方法罗杰斯特回归等方法进行分类研究。

在进行特征提取的前提下的分析结果如下：

表 6 汽车论坛特征提取后的分类准确率

数据(长度)	>0	>10	>30	>60	>90	>120	>150
朴素贝叶斯	60.52%	61.66%	64.25%	66.68%	66.37%	68.47%	69.99%
lingpipe	67.83%	69%	69.75%	73.67%	74%	75%	82.5%
支持向量机	73.91%	73.91%	73.19%	81.88%	82.61%	83.33%	82.61%
罗杰斯特回归	56.3%	56.5%	59.9%	61%	61.1%	63.5%	65.6%

表 7 汽车论坛特征提取后的分类 F 值

数据(长度)	>0	>10	>30	>60	>90	>120	>150
朴素贝叶斯	0.6664	0.669	0.6726	0.6732	0.7079	0.7348	0.7407
lingpipe	0.6712	0.7048	0.6873	0.7358	0.7417	0.7525	0.8325
支持向量机	0.7313	0.7049	0.71312	0.7474	0.820	0.8099	0.8154
罗杰斯特回归	0.6826	0.689	0.5172	0.4853	0.7036	0.6449	0.7304

表 8 股票论坛特征提取后的分类准确率

数据(长度)	>0	>10	>30	>60	>90	>120	>150
朴素贝叶斯	53.13%	53.53%	55.3%	55.74%	56.45%	55.95%	58.48%
LINGPIPE	60%	61.1%	61.7%	61.7%	62.4%	61.8%	62%
支持向量机	62.53%	63.91%	66.31%	69.73%	73.05%	71.58%	73.86%
罗杰斯特回归	56%	55.8%	54.6%	57.1%	57%	58%	59%

表 9 股票论坛特征提取后的分类 F 值

数据(长度)	>0	>10	>30	>60	>90	>120	>150
朴素贝叶斯	0.2149	0.2429	0.305	0.3553	0.3861	0.4915	0.478
LINGPIPE	0.6058	0.6211	0.6289	0.6373	0.6403	0.6487	0.6564
支持向量机	0.5369	0.512	0.5414	0.6336	0.684	0.7724	0.7936
罗杰斯特回归	0.647	0.6563	0.308	0.64	0.625	0.7038	0.71

从汽车论坛和股票论坛的文本分类研究中发现，不论用哪种分类方法都可以对内容进行分类研究，通过比较几种分类方法的结果，发现支持向量机的分类结果最好，其次是罗杰斯特回归以及 LINGPIPE，最次的是朴素贝叶斯方法。

6．结论

本文通过假设验证及文本分类的方法对中文网络论坛数据进行分类研究，发现男性和女性在网络发帖行为及中文使用等方面确实存在差异。对于发帖行为研究主要是从发帖频率，发帖时间和内容长度进行分析，统计检验结果证明在评论长度和发帖时间上不存在明显的性别差异，在发帖频率上其差异性更为显著。对于发帖内容的研究中，以文本分类方法为研究主要方法结合特征提取和几种分类算法，对用户论坛评论内容进行文本分类研究。对比几种分类算法分类结果发现，以评论内容为依据进行性别分类是可行的，即表明在不同性别用户之间确实不存着词语使用的不同。

参考文献(References)

[1] M. Imhof, R. Vollmeyer and C. Beierlein, “Computer Use and The Gender Gap: The Issue of Access, Use, Motivation and Performance”, *Computers in human behavior*, vol.23, pp.2823-2837,2007

[2] L. Shaw and L. Gant, “Users Divided? Exploring the Gender Gap in Internet Use”, *Cyberpsychology and Behavior*, Vol.6, no.5, pp.517-527, 2002.

[3] E. Garbarino and M. Strahilevitz, “Gender differences in the Perceived Risk of Buying Online and The Effects of Receiving a Site Recommendation”, *Journal of Business Research*, Vol.54, pp. 768- 775,2004.

[4] A. Fox D. Bukatko, M. Hallahan and Mary Crawford, “The Medium Makes a Difference Gender Similarities and Differences in Instant Messaging”, *Journal of Language and Social Psychology*, Vol. 4, no.26, pp. 389-307, 2007.

[5] A. Flanagin, V. Tiyaamornwong, J. O'Connor and D. Seibold, “Computer-mediated Group Work: The Interaction of Member Sex and Anonymity”, *Communication Research*, Vol. 29, pp. 66-93, 2002.

[6] S. Herring, “Posting in Different Voice: Gender and Ethics in Computer-mediated Communication”, *Philosophical Perspectives on Computer-mediated Communication*, pp.115-145, 1996.