基于天涯论坛的用户发帖行为规律研究①

赵永亮，唐锡晋
（中国科学院 数学与系统科学研究院，北京，100190）

摘　要：互联网的快速发展使得网络成为数据的海洋，为了从网络中发现有用的信息，网络挖掘技术应运而生。本文利用爬虫程序从天涯论坛每日定时抓取数据，并存储到数据库和文件系统中；然后基于2012年全年的“天涯杂谈”板块的数据，研究用户的行为规律。数据分析结果表明节假日及周末用户的网上活动减少，用户的行为符合日常作息规律，有明显的日历效应；数据拟合发现点击量的分布可以用泊松分布与幂律分布的混合分布描述，而回复量的分布符合幂律分布。

关键词：网络挖掘；天涯论坛；用户行为；泊松分布；幂律分布

A Research of Pattern of Users' Posting Behavior Based on Tianya Club

Zhao Yongliang, Tang Xijin
(Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190)

Abstract: The fast development of Internet makes it an ocean of data, in order to find useful information, the web mining technology comes into being. This paper gets data from Tianya Club using spider program and stores it in files and database; then we investigate the pattern of users' behavior based on all the data of "Tianya By-talk" in 2012. The result is that users have less online activities on holidays and weekends, users' online behavior keeps to the pattern of people's daily routine. Furthermore, the distribution of the number of hits can be described by the mixed distribution of Poisson distribution and power law distribution, while the distribution of the number of replies is power law distribution.

Keywords: web mining; Tianya Club; user behavior; Poisson distribution; power law distribution

① 国家重点基础研究发展计划项目(2010CB731405)

1 引言

随着互联网的快速发展，网络成为数据的海洋。网络数据是典型的大数据，具有大数据的全体性、混杂性、复杂性等特点[1]。如何从网络数据中挖掘出有用的内容或规律成为当前数据挖掘、计算机应用、商业搜索、组织行为管理等相关领域的研究热点。网络挖掘包括网络内容挖掘、网络结构挖掘和用户行为挖掘等三种模式，其中，用户行为规律的挖掘具有重要的现实意义。张文等介绍了网络文本挖掘中的一种实用技术[2]；Ari Seifter 等利用 Google Trends 对流行病进行了研究，是用户行为挖掘较早的案例[3]；Fabio Celli 等对 Friendfeed 社交网络中的用户行为进行了挖掘[4]；李元等研究了新浪微博中热门微博的用户行为规律[5]；Li Jie Cui 等研究了天涯论坛帖子的热度计算以及一条热帖的回复量的分布规律[6]。

目前，天涯社区每月覆盖用户超过 2 亿，注册用户超过 8 000 万，是华语圈首屈一指的网络事件与网络名人聚焦平台[7]。对天涯论坛中用户行为规律的研究具有重要意义，例如掌握用户的行为规律，才能探测网络事件的发展与演变，引导网络事件向积极的方向发展，有利于构建和谐社会。

下面主要分为三个部分。第一部分介绍了天涯论坛数据的处理过程，包括数据的获取和存储设计。第二部分基于 2012 年“天涯杂谈”板块的数据，对用户行为进行了挖掘，研究了节假日、周末和发帖时间对用户行为的影响，以及点击量与回复量的分布。第三个部分对全文进行了总结，指出了本文的不足和待改进之处。

2 数据获取与存储

本文采用网络挖掘技术对天涯论坛进行了挖掘。天涯论坛数据的处理过程如图 1 所示。数据的获取和存储过程每日自动完成。

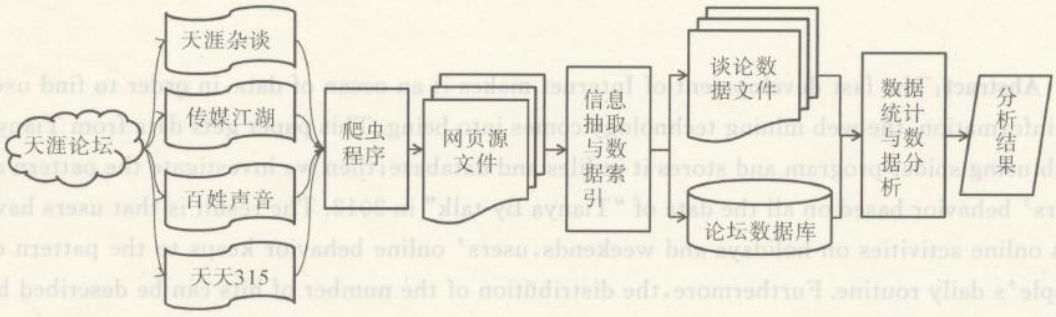

图 1　天涯论坛数据的处理过程

2.1 数据获取

天涯论坛的数据获取过程如图 1 的左部所示。数据的获取采用自行设计的爬虫程序每日自动从天涯论坛下载。该爬虫程序首先通过 web 页面收集，将天涯论坛中“天涯杂谈”等四个版块中的帖子更新信息页面定时收集到本地机器。之后通过 web 页面过滤、特征提取和信息

抽取，抽取原始页面中的帖子及其更新信息，存入数据库以及文件系统中[8]。

数据获取中有两个关键问题。第一个问题是自适应应对天涯论坛的改版，如 2013 年 1 月天涯论坛推出了新版；第二个问题是对缺失数据进行重新获取，保证数据的完整性。

2.2 数据存储

天涯论坛数据的存储采用 mysql 数据库与 xml 文件系统相结合的方式，兼有两者的优点。对于 mysql 数据库存储方式，主要包含两张数据表：post 和 postupdate。这两张数据表的结构如下所示：

post (*pID*, *title*, *author*, *content*, *link*, *date_posted*, *time_posted*, *boardID*)

postupdate (*pID*, *date_update*, *time_update*, *hits*, *replies*)

其中，post 数据表存放帖子信息。“pID”为自动递增主键，唯一标识一条帖子。其余字段分别存放帖子的标题、作者、首帖内容、原始链接地址、发表日期、发表时间和所属板块 ID 等。postupdate 数据表存放帖子的更新信息，以“pID”和“date_update”作为联合主键。各字段的含义分别为帖子的主键、帖子的更新日期、帖子的更新时间、点击量和回复量等。

对于 xml 文件系统存储方式，其存储的内容和 mysql 数据库相同。但是采用 xml 文件的形式，便于数据的管理和交换。而且采用两种存储方式相结合的方法，可以防止数据的丢失与破坏[9]。

3 用户行为规律分析

对于天涯论坛，个体用户的行为主要包括发帖、点击以及回复等，在宏观上就表现为发帖量、点击量与回复量等定量数据。本文将这些定量数据作为研究用户行为规律的指标，因此，也从宏观上揭示了用户的行为规律。具体的研究包括帖子的数量统计，节假日、周末和发帖时间对用户行为的影响，以及点击量与回复量的分布等。

本文的研究主要基于 2012 年“天涯杂谈”板块的所有数据。天涯论坛包括众多板块，例如，“天涯杂谈”、“百姓声音”、“煮酒论史”等。其中，“天涯杂谈”板块是一个和网络事件以及社会舆情高度相关，而且相对活跃的板块。因此本文的研究使用“天涯杂谈”板块的数据，这些数据包括每日的新发帖和更新帖，其中，新发帖指当日发表的帖子，更新帖指在当天被更新过的帖子，更新帖包括当日的新发帖，而发帖量指新发帖数量或更新帖数量。

3.1 帖子数量统计

对于 2012 年“天涯杂谈”板块的数据，统计每日的新发帖数量和更新帖数量，对于缺失数据，利用爬虫程序重新抓取，对于无法重新获取的数据，采用当月平均值代替当日值。经过数据预处理后得到 2012 年“天涯杂谈”板块的总的新发帖数量为 409 717 条，平均每日新发帖数量为 1 119 条；其中，每日新发帖数量的最大值为 1 927 条，最小值为 185 条。总的更新帖数量为 1 195 125 条，平均每日更新帖数量为 3 265 条；其中，每日更新帖数量的最大值为 4 755 条，最小值为 789 条。

每月的更新帖和新发帖数量如图 2 所示。从图 2 可以看出，每月的新发帖数量与更新帖数量之比大约为 1∶3。除了 1 月和 2 月的新发帖和更新帖数量较少外，从 3 月以后，每月的新发帖和更新帖的数量整体保持稳定，小幅度波动。这说明，“天涯杂谈”板块是一个用户活跃度高、

用户行为保持稳定的板块。

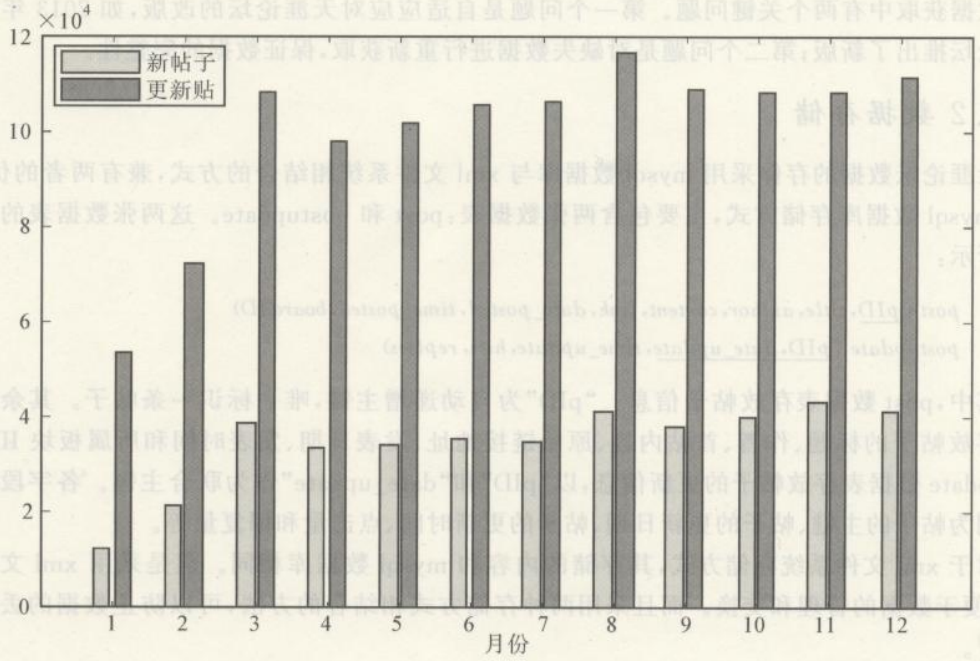

图 2　每月新发帖和更新帖数量

3.2 节假日对用户行为的影响

为了研究节假日对用户行为的影响，定义节假日的发帖率如公式(1)所示，发帖率反映了发帖量占平时的比重。

$$\text{节假日的发帖率 } r=\frac{\text{节假日平均日发帖量}}{\text{节假日前后 } m \text{ 天平均日发帖量}} \tag{1}$$

其中，m 为节假日的长度。

利用发帖率的定义，研究一些常见的法定节假日对用户行为的影响。考虑的节假日包括元旦(1月1日—1月3日)，春节(1月22日—1月28日)，清明节(4月2日—4月4日)，劳动节(4月29日—5月1日)以及中秋节和国庆节(9月30日—10月7日)，由于中秋节和国庆节在一起放假，所以放在一起考虑。各节假日的发帖率的计算结果如表1所示。

表 1　各节假日对应的发帖率

发帖率	元旦	春节	清明	劳动	中秋国庆	平均
新发帖	0.68	0.62	0.81	0.66	0.59	0.67
更新帖	0.58	0.71	0.94	0.83	0.76	0.76

从表1可以看出，无论对于新发帖还是更新帖，发帖率均小于1，且最小值达到0.58。这说明节假日对用户行为有着重要的影响，节假日的发帖量大约为平时的2/3。

3.3 周末对用户行为的影响

同理，为了研究周末对用户行为的影响，定义周末的发帖率如公式(2)所示。其中，工作日

指周一到周五，周末包括当周的周六和周日。

$$\text{周末的发帖率 } r=\frac{\text{周末平均日发帖量}}{\text{工作日平均日发帖量}} \tag{2}$$

利用发帖率的定义，研究周末对用户行为的影响。将2012年每日发帖量以周为单位排开，计算每周的发帖率，然后取所有周的发帖率的平均值。主要分三种情况计算发帖率，即取所有周（情况1），去除放假调休所涉及的周（情况2）以及去除假期前后各1周（情况3）。各种情况下发帖率的计算结果如表2所示。

表2　周末对应的发帖率

发帖率	情况1	情况2	情况3	平均值
新发帖	0.83	0.80	0.75	0.79
更新帖	0.90	0.89	0.86	0.88

从表2可以看出，无论对于新发帖还是更新帖，在三种情况下发帖率都小于1，最小值达到0.75。由于情况3完全剔除了节假日对用户发帖量的影响，更能准确地表示周末对用户行为的影响。这说明周末对发帖量有着重要的影响，周末的发帖量大约为平时的4/5。

节假日与周末均可以归结为用户的非工作时间，用户在非工作时间的发帖率较低的原因可能是，网络用户在非工作时间内大多在休息，或者从事现实世界中的活动，而在网上的活动相对于平时偏少，从而导致发帖率偏低。这也从侧面反映了网络是用户缓解工作压力的一种方式。

3.4 发帖时间对用户行为的影响

为了研究发帖时间对用户行为的影响，对于2012年“天涯杂谈”板块的数据，本文统计了一天中每个小时的新发帖数量和更新帖数量，结果如图3所示。其中，横坐标为发帖时间，如“4”表示发帖时间为一天中的4点到5点之间。

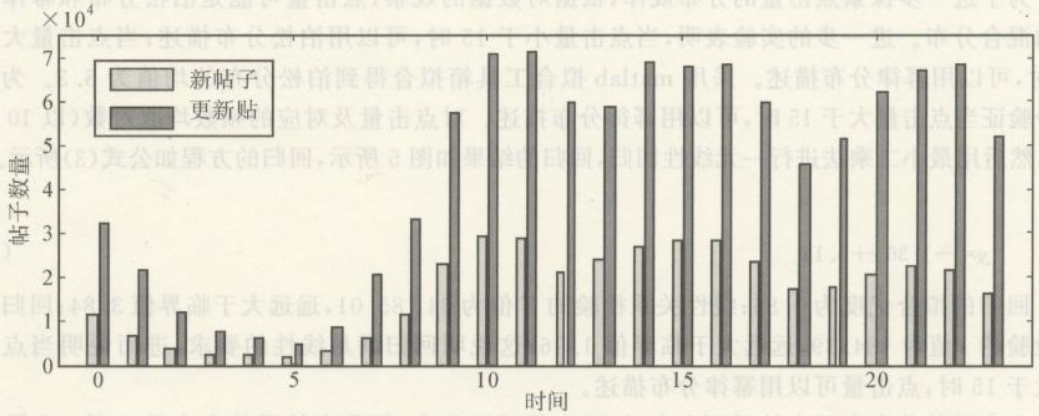

图3　每小时的新发帖和更新帖数量

从图3可以看出，新发帖数量和更新帖数量的变化保持一致。发帖量在夜间达到最低，在上午、下午以及晚上形成三个高峰，而在午餐以及晚餐的时候发帖量相对有所下降。这说明，从发帖时间来看，用户的发帖行为较符合人们的日常作息规律。

3.5 点击量的分布

点击量反映了用户对帖子的感兴趣程度，为了研究点击量所反映出的用户行为规律，本文以2012年“天涯杂谈”板块的数据为基础，统计出每个点击量所对应的每日新发帖的数量，统计的结果如图4所示。

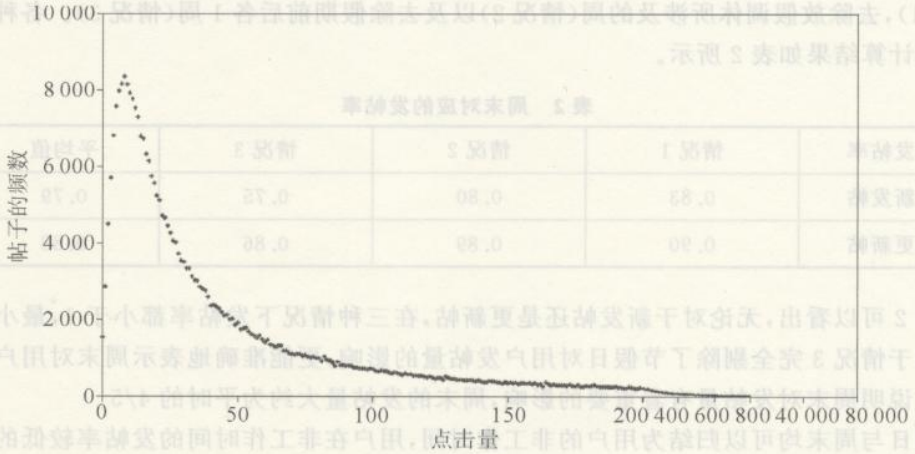

图4 点击量的分布图

从图4可以看出，当点击量从0到8变化时，对应的帖子频数快速上升；当点击量为8时，帖子频数最大，达到8 351条；当点击量大于8时，随着点击量的增加，帖子频数不断下降，但是下降的速率减慢。点击量为0到130之间的帖子占总帖数的80%；当点击量大于2 000时，对应的帖子频数均小于10，该部分帖子仅占总帖数的1.64%。这说明只有少数帖子的点击量很高，绝大部分帖子的点击量都很低，缺乏用户关注。

为了进一步探索点击量的分布规律，根据对数据的观察，点击量可能是泊松分布和幂律分布的混合分布。进一步的实验表明，当点击量小于15时，可以用泊松分布描述；当点击量大于15时，可以用幂律分布描述。采用matlab拟合工具箱拟合得到泊松分布的均值为8.3。为进一步验证当点击量大于15时，可以用幂律分布描述。对点击量及对应的频数均取对数(以10为底)，然后用最小二乘法进行一元线性回归，回归的结果如图5所示，回归的方程如公式(3)所示。

$$y=-1.36x+5.16 \tag{3}$$

回归的拟合优度为0.86；线性关系检验的F值为34 185.01，远远大于临界值3.84；回归系数检验的t值为184.89，远远大于临界值1.96；这说明回归满足线性的要求，进而说明当点击量大于15时，点击量可以用幂律分布描述。

点击量的分布表明在帖子刚发表时，用户是随机点击，但是当帖子的点击量达到一定数目时，用户就会抱着围观的心态来看帖，从而导致这种混合分布的产生。

3.6 回复量的分布

回复量反映了用户对帖子讨论的激烈程度，为了研究回复量所反映出的用户行为规律，本文基于2012年“天涯杂谈”板块的数据，统计了每个回复量所对应的每日新发帖的数量。回复量共有692个离散值，除了一个异常值51 960外，其余取值范围为[0,4 984]；对应的帖子频数的取值范围为[1,163 714]，特别是回复量为0的帖子达到163 714条，约占总帖数的1/3。回

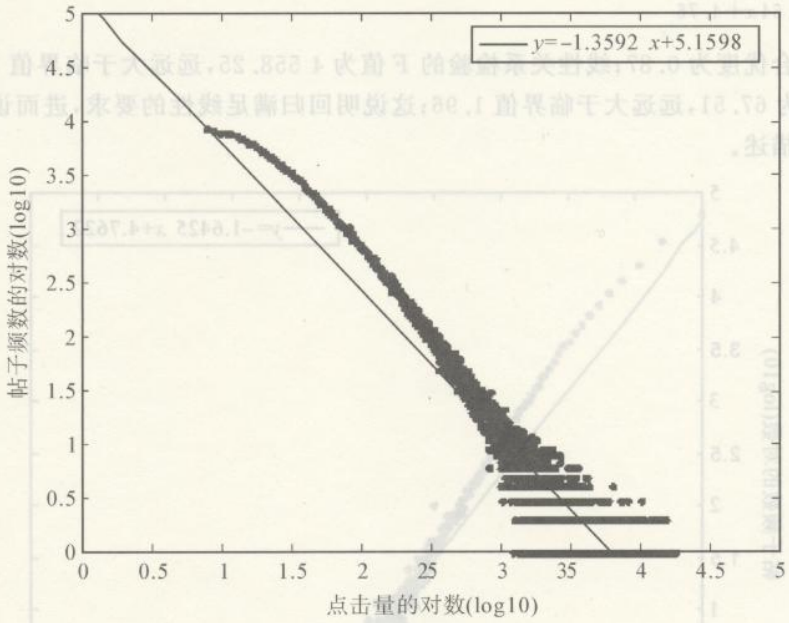

图 5　取对数后点击量的分布图

复量的统计结果如图 6 所示。

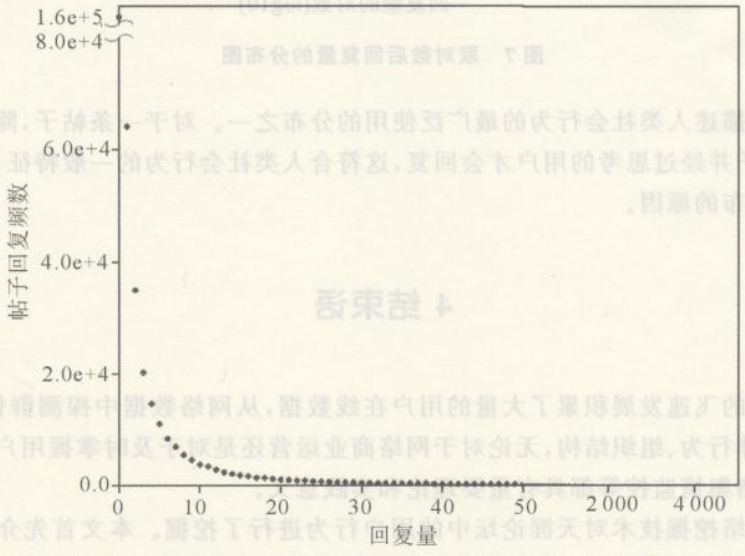

图 6　回复量的分布图

从图 6 可以看出，随着回复量的增加，其对应的频数不断下降，但是下降的速度不断减缓；回复量为 0 到 5 之间的帖子占总帖数的 80%；当回复量大于 300 时，其对应的频数均小于 10，该部分帖子仅占总帖数的 0.17%。这说明只有极少数帖子的回复量很高，绝大部分帖子的回复量都很低，不被用户关注与讨论。

为进一步探索回复量的分布规律。根据对数据的观察，回复量可能满足幂律分布。为进一步验证，对回复量及对应的频数均取对数(以 10 为底)，然后用最小二乘法进行一元线性回归，回归的结果如图 7 所示，回归的方程如公式(4)所示。

$$y=-1.64x+4.76 \tag{4}$$

回归的拟合优度为 0.87；线性关系检验的 F 值为 4 558.25，远远大于临界值 3.84；回归系数检验的 t 值为 67.51，远远大于临界值 1.96；这说明回归满足线性的要求，进而说明回复量可以用幂律分布描述。

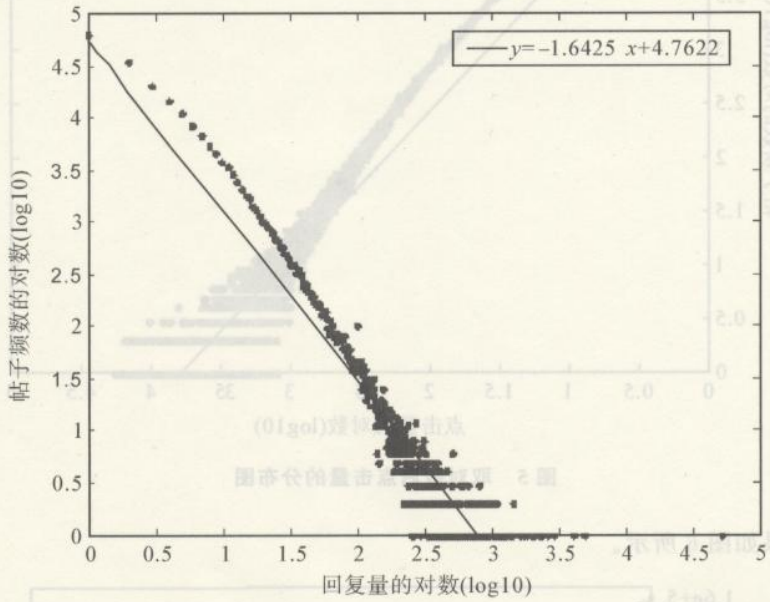

图 7　取对数后回复量的分布图

幂律分布是描述人类社会行为的最广泛使用的分布之一。对于一条帖子，除了网络水军之外，只有看过帖子并经过思考的用户才会回复，这符合人类社会行为的一般特征，这也解释了回复量满足幂律分布的原因。

4 结束语

互联网技术的飞速发展积累了大量的用户在线数据，从网络数据中探测群体行为规律，挖掘观点，感知群体行为、组织结构，无论对于网络商业运营还是对于及时掌握用户的兴趣或行为的变化规律，网络舆情监控等都具有重要理论和实践意义。

本文利用网络挖掘技术对天涯论坛中的用户行为进行了挖掘。本文首先介绍了天涯论坛数据的处理过程，包括数据的获取和存储设计。然后基于 2012 年“天涯杂谈”板块的数据，对用户行为进行了挖掘，研究了节假日、周末和发帖时间对用户行为的影响，以及点击量与回复量的分布规律等。

本文的研究也存在一些不足和待改进之处。首先，本文只是对用户行为规律进行了初步研究，还需要进行深入挖掘，例如，研究更新帖中的分布规律、点击量的分布和回复量的分布之间的关系、帖子的沉没时间分布等。其次，对天涯论坛数据的挖掘，除了对用户行为的挖掘外，还应对帖子内容进行挖掘，曹丽娜等利用 LDA 算法从帖子中提取出主题[10]，除此之外，还要进行更深入的基于内容的分析。

参考文献

[1]维克托·迈克-舍恩伯格,肯尼迪·库克耶.大数据时代——生活、工作与思维的大变革[M].浙江人民出版社,2013.

[2]张文,唐锡晋,吉田武稔.AIS——基于文本挖掘的增强型 Web 信息处理技术[J].系统工程理论与实践,2010,30(1):96－99.

[3]Ari S,Alison S,Kate G,et al. The utility of "Google Trends" for epidemiological[J]. Geospatial Health. 2010,4(2):135－137.

[4]Fabio C,Barbara P,Luca R,et al. Social network data and practices:the case of friendfeed[C]. International Conference on Social Computing,Behavioral Modeling & Prediction,2010.

[5]李元.热门微博中用户行为与信息传播特性研究[D].中国科学院大学硕士学位论文,2013.

[6]Li J C,Hui H,Wei L. Research on hot issues and evolutionary trends in network forums[J]. International Journal of u- and e- Service,Science and Technology,2013,6(2).

[7]http://help.tianya.cn/about/history/2011/06/02/166666.shtml.

[8]张泽代,唐锡晋.面向天涯论坛的 Web 挖掘的初步研究[A].系统科学与管理科学新理论、新方法、新技术及应用——第十一届全国青年系统科学与管理科学学术会议暨第七届物流系统工程学术研讨会论文集[C].武汉理工大学出版社,2011,199－204.

[9]张泽代.基于天涯论坛的网络挖掘系统——天涯论坛视点 1.0 设计与实现[D].中国科学院研究生院硕士学位论文,2012.

[10]Cao L N,Tang X J. Prevailing Trends Detection of Public Opinions Based on Tianya Forum[C]. The 14th International Conference on Intelligent Data Engineering and Automated Learning,2013.