文章编号:1000－2367(2015)03－0161－04 DOI:10.16366/j.cnki.1000－2367.2015.03.031

问答型社区主题响应时间分析

冯云芝,赵胜杰

(河南师范大学 计算机与信息工程学院,河南 新乡 453007)

摘 要:通过对知识型社区主题的响应时间进行分析,观察到主题的响应时间呈现近似幂率分布,大部分主题的响应时间发生在帖子发出后的 2 h 之内.基于该研究结果,可以模拟网络论坛中用户主题讨论的动态演化过程,并利用人们在虚拟环境下对新生事物关注程度逐渐降低这一规律开发出合适的学习推荐软件.

关键词:回复时间;主题响应;时间分析;论坛

中图分类号:TP393 **文献标志码**:A

随着互联网和 Web2.0 技术的快速发展,人们越来越多地通过网络进行信息交流.Web2.0 技术使得问题－回答类网站逐渐成为人们交流经验的主要平台.基于这类网站,人们通过发帖－回答问题的机制来获取知识.雅虎知识网站就是其中的一个典型代表.网站用户可以浏览其他用户提出的问题、搜索特定问题的答案、提出问题并等待解答,也可以围绕某个话题展开讨论等.问答类网站通过具有共同兴趣或专业知识的用户自发形成的社区进行信息的收集和经验的分享,从而向用户提供一个面向知识型社区的服务.社区成员通过发布、浏览、搜索问答类文档来发现知识.而随着网民数量的快速增长,人们在网上发布问题以及提供答案的数目呈几何指数增长,造成大量的问答型网站信息过载,进而严重影响知识获取的效率.因此,如何对这类网站内的信息进行过滤,向用户推荐合适的文档文件成为当前研究的主流问题.

目前的研究主要集中在如何寻找合适的专家来回答问题[1]或在社区网站上如何检索到高质量的答案向用户推荐[2]这两个方面.常用策略包括如下 3 种.第 1 种是对团队中的每个成员的兴趣进行归纳总结,形成群组兴趣,从而利用群组兴趣过滤拟推荐条目[3];与第 1 种方法相反,第 2 种方法首先利用个性化推荐生成关于每个成员的推荐列表,再对这些推荐列表进行聚合形成群组推荐列表[4].在这两类方法中,推荐机制利用组成员的重要性作为加权聚合过程中的权值,没有考虑问答类网站内成员在社区内的声誉对推荐结果的影响,造成推荐结果不准确.考虑到这一点,第 3 类方法从社区的角度出发研究推荐机制[5].首先,通过融入社区成员在收集和回答问题过程中形成的声誉以及收集到的文档协同生成社区属性;其次,社区成员在浏览特定问题时,通常也会对相关问题产生兴趣.传统的问答类系统主要使用关键词来进行匹配,没有考虑问题/答案之间的冗余及互补因素.而实际情况是用户可能需要通过查找更多相关的问题以获得更加完善的答案.因此,为了给社区成员提供和补充更加丰富的知识,面向社区的推荐机制考虑了文档之间的互补关系,通过问答类文档和社区属性的关联程度进行文档推荐.但是当前工作均没有考虑推荐时机对推荐系统的影响.而何时向用户推荐何内容是影响推荐系统成功与否的关键因素.考虑到这一点,本文鉴于网络社区聚集了绝大多数人们的日常上网行为及活动,对人们在论坛类网站上的行为及活动方式进行了研究.意义如下:首先,研究人们在论坛上的行为及活动方式有利于网络管理人员合理利用、统筹网络资源[6];其次,有利于更好地了解人们在虚拟网络环境下学习的行为特点[7],以便更好地开发出合适的学习推荐软件.

本文采用火车头采集器收集 BBS 网络论坛[8]的回复时间、回复内容.通过提取论坛成员参与讨论的时间以及他们回复的内容,研究人们在问答型论坛中话题响应时间的统计属性,主要步骤包括:首先使用网页搜集工具实时地获取 web 内容,然后对 web 内容进行过滤,提取有用信息,接下来将过滤后的结果存入数据

收稿日期:2014-10-30;**修回日期**:2015-03-23.

基金项目:国家自然科学基金(U1404602)

作者简介:冯云芝(1970－),女,河南新安人,河南师范大学实验师,研究方向为计算机应用,E-mail:yunzhif@126.com.

库中,最后对存入数据库的数据进行统计处理.

本文针对国内王道论坛(WDBBS)数学模块下的"线性代数超强总结"以及国外 INFOCOM 2005 会议上与会人员交流的数据集进行样本统计.这两类主题分别代表人们线上线下的互动情况,贴近现实生活,同时样本比较充足,可以满足本研究的需要.

1 相关工作

网络社区是由共同兴趣、经验、需求和目标的用户自发形成的,感兴趣的主题在社区内进行共享和讨论.在社区中用户可以更高效地实现知识共享和信息交流[9-10].目前比较典型的虚拟社区主要包括:BBS 论坛、讨论版,博客,微博,微信,Facebook,Wretch 和雅虎.比如雅虎问答是一个基于发帖和回答问题的知识共享平台.它向用户提供知识型社区服务,有共同兴趣和专长的用户聚集在一起形成一个虚拟社区进行信息交流.每一个社区成员可以开展一些活动,比如收集文档信息、发布文章、发起一个话题进行讨论或代表社区回答问题等.当社区成员在浏览或搜索文档时,将那些他们感兴趣的相关文档进行收集,同时把那些认为对整个社区有帮助的文档进行发布.社区成员也可以代表他们所在的社区参与问题讨论,如果他们提供的答案被评为最佳答案,则这些信息将记录在社区平台上,同时提高相关用户在整个社区内的声誉值.

问答类网站为人们提供了一个交流信息的理想平台.本文通过对人们在该类网站上发帖行为进行分析,一方面可以帮助网络管理人员进行网络资源的有效整合,另一方面也可以向推荐系统提供有益的信息,比如何时向用户推荐相关内容用户更可能对这些内容进行浏览,而这一点在学习、广告类推荐系统中至关重要.

2 采集数据的方法及过程

本节描述数据收集及分析过程.首先,利用火车头采集器收集王道论坛数学模块下的"线性代数超强总结"的回复时间及内容,然后将收集的数据截取一部分构建数学模型并进行分析.此外,为了进一步扩大样本范围,本文还采用了 INFOCOM 2005 会议提供的开源数据集,该数据集提供了人们线下交流的情况.

2.1 使用火车头采集器

火车头采集器[11]如何搜集数据,取决于采集规则.要获取一个栏目里的所有内容,需要先将这个网页的网址采下来,该部分称为网址采集.程序按一定的规则抓取一定的列表页面,从中分析出网址,然后再去抓取这些网址内的网页里的内容.最后根据采集规则,对下载的内容进行分析,将标题内容等信息分离出来并保存.如果选择了下载图片等网络资源,程序会对采集到的数据进行分析,找出图片、资源等的下载地址并保存到本地.火车头采集器的工作流程如下:

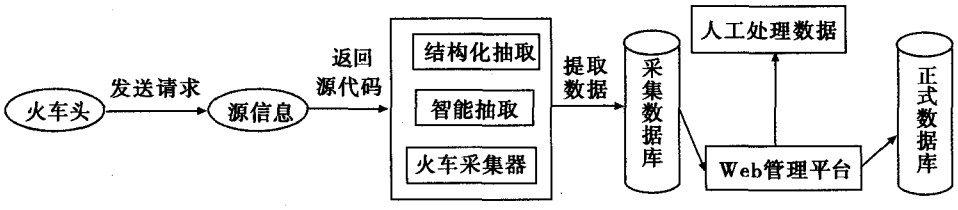


图 1 火车头采集器的工作流程

由图 1 可知,火车头采集器主要包括数据采集及数据发布两大功能.数据采集通过对网址和内容进行采集,获得抽样数据.数据发布功能负责将收集到的数据发布到网络论坛.可以基于 web 在线发布,也可以保存在数据库或本地文件内.具体来说,火车头采集器收集的数据在默认情况下保存在本地,具体处理包括以下几种情况:第 1 种是不做任何处理.因为数据本身是保存在数据库中(Access,db3,MYSQL 等等),如果只查看数据,直接用相关软件打开即可.第 2 种通过 web 将内容发布到网站.程序会模仿浏览器向网站发送数据,可以实现手工发布的效果.第 3 种是将数据存入数据库.通过 SQL 语句将数据导入到数据库中.第 4 种是将采集到的数据保存为本地文件.程序读取数据库里的数据,按一定格式保存为 SQL 或文本文件.

具体使用哪一种工作方式可根据实际需求决定. 比如在采集时先采集不发布，等将来有时间了再发布；或是采集与发布同时进行；或是先做发布配置再进行数据采集等等.

2.2　火车头采集器数据采集流程

利用火车头采集器软件对王道论坛网站下的“线性代数超强总结”的回复时间及内容进行了收集，数据采集具体流程如下：(a) 网址采集规则：该过程主要用来确定采集网址的开始地址及结束地址. 本文使用的首页及尾页地址分别是采集的首页地址：http://www.cskaoyan.com/thread-61978-1-1.html. 采集的尾页地址：http://www.cskaoyan.com/thread-61978-296-1.html. (b) 内容采集规则：该过程主要用来确定欲提取主题的内容. 本文主要对帖子的回复时间、回复内容以及回复时间间隔进行提取/计算. (c) 内容发布：最后将提取的数据存入到数据库中.

表 1　帖子的回复时间、内容、相邻帖的回复时间间隔

回复时间	回复内容	相邻的响应时间差	响应与主题的时间差
2011-05-12 19:13	以前同学下载的，不知道这上面还有没有了？	0 h 0 min	0 h 0 min
2011-05-12 20:01	谢谢 lz	0 h 48 min	0 h 48 min
…	…	…	…
2011-05-13 14:12	感谢楼主分享	0 h 13 min	18 h 59min
2011-05-13 14:22	看看	0 h 10min	19 h 9 min
2011-05-13 14:55	see see	0 h 33 min	19 h 42 min
2011-05-13 17:31	顶一下	2 h 39 min	22 h 21 min

表 1 显示了收集的部分帖子内容. 主要包含了搜集的数据条数、帖子的响应时间、帖子内容、相邻回复的响应时间差、每个响应与主题的时间间隔等部分.

3　回复时间分析的方法及步骤

本节主要介绍回复时间分析的整体框架. 分析回复时间的关键在于回复时间的获取、计算相邻的回复时间间隔以及每个回复时间与发帖的时间间隔，最后统计在某个时间段的回复条数. 实验用火车头采集器搜集的数据主要分为 3 个部分：回复时间、回复内容、相邻帖的回复间隔，本实验中主要是对相邻帖的回复时间进行了统计分析.

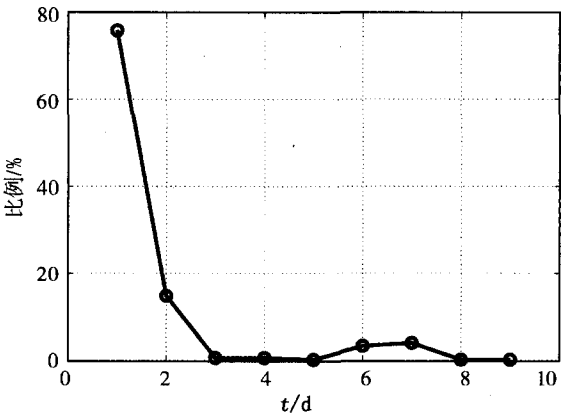


图2 相邻回复时间大于 1 d 的折线图(WDBBS)

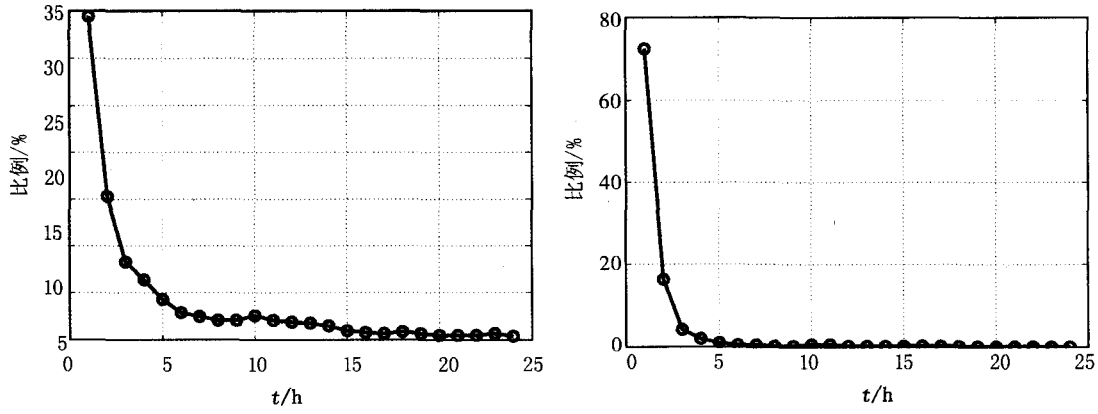
图 2 对相邻回复时间间隔大于一天的情况进行了显示. 可以看出相邻回复时间间隔呈现出一定的幂率分布. 在发帖的第一天内人们对帖子的关注程度比较高，参与人数比较多，人们对讨论内容表现出较高的兴趣. 第一天时间内人们回复的比例占总回复比例的 76.68%. 同时，同一个帖子相邻的回复时间之间也呈现一定的规律性，即当回帖人数较多时，人们对帖子的关注程度比较高，人们一起参与相关话题的讨论，此时人们的交流方式类似于断断续续的聊天方式，人们之间对相关话题讨论的活跃程度比较高. 由此可以得出，在论坛类网站上发帖的第一天时间内人们对帖子的关注程度比较高. 基于此，可以截取第一天内人们讨论话题的部分内容，研究人们在虚拟空间内进行讨论时的态度及倾向，进而，研究对人与人之间交流的语义进行分析，有利于舆情监测的开展[12].

图 3 对相邻回复时间小于 1 d 内的帖子在每个时间段的回复情况进行了统计. 从图中可以看出，帖子刚发出时人们的响应比较强烈，然而随着发帖时间的增长人们对帖子的关注程度逐渐降低. 有意思的是在发帖的 1 h 之内，后续的回复内容大多与前面的回复内容相同，参考价值不大. 由此，可以得出在网站论坛上搜集需要的信息时，可以把关注的时间集中在发帖的前几小时内，从而提高搜集信息的速度与有效性.

4　总结与展望

主要对网站论坛上的热点话题进行实时监控，并运用火车头采集器收集了王道论坛数学模块下“线性代

数超强总结"版块主题的回复时间和回复内容,并对回复时间进行了统计分析,研究人们在虚拟环境下的学习行为.通过对人们在虚拟环境下以学习为主题的回复响应时间分析,有助于在不同的时段向学习者推荐不同的学习资料,今后拟在以下几个方面进行深入研究:1)构建响应时间的数学模型,以提高响应时间的准确性;2)进一步扩充统计数据的来源,使所得结论更具有一般性.



(a) WDBBS (b) INFOCOM

图 3 相邻回复时间小于1 d 的折线图

参 考 文 献

[1] Liu D R, Chen Y H, Kao W C, et al. Integrating expert profile, reputation and link analysis for expert finding in questionanswering websites[J]. Inf Process Manage,2013(49):312-329.

[2] He T, Ming Z Y, Adriani M, et al. Discovering high quality answers in community question answering archives using a hierarchy of classifiers[J]. Information Sciences,2014,261:101-115.

[3] Kim J K, Kim H K, Oh H Y, et al. A group recommendation system for online communities[J]. Int J Inf Manage,2010,30(3):212219.

[4] Liu D R, Lai C H, Chen Y T. Document recommendations based on knowledge flows: a hybrid of personalized and group-based approaches[J]. J Am Soc Inform Sci Technol,2012,63(10):2100-2117.

[5] Liu D R, Chen Y H, Huang C K. QA document recommendations for communities of question-answering websites[J]. Knowledge-Based Systems,2014,57:146-160.

[6] 荣 波,夏正友,朱永真,等. BBS 在线复杂网络及其成员交互特性研究[J]. 复杂系统与复杂性科学,2009,6(4):57-65.

[7] 马天健. 浅析网络统计[J]. 甘肃科技,2009,25(13):27-30.

[8] 彭小川,毛小丹. BBS 群体特征的社会网络分析[J]. 青年研究,2004(4):39-44.

[9] 刘 萍,海本斋. 1 种移动自组织网络队列延时的计算方法[J]. 河南师范大学学报:自然科学版,2013,41(5):170-173.

[10] Alani H, Dasmahapatra S, Hara K O, et al. Identifying communities of practice through ontology network analysis[J]. IEEE Intell Syst,2003,18(2):18-25.

[11] 火车头采集器. 信息数据采集论坛[EB/OL]. [2014-04-10]. http://bbs.locoy.com/.

[12] 骆力明,尤佳鑫,孙 众. 基于 Android 系统的课堂记录与多元分析系统[J]. 河南师范大学学报:自然科学版,2015,43(1):146-151.

Response Time Analysis of Theme in Q&A Community

FENG Yunzhi, ZHAO Shengjie

(College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China)

Abstract: We study the response time of themes and observe one interesting phenomenon. That is, most questions receive replies within two hours and the response time shows a power law distribution approximately. The research result shows that we can model the dynamics of users and develop appropriate learning recommendation software by using the new characteristics.

Keywords: response time; theme response; time analysis; BBS