

文章编号:1007-6735(2013)05-0420-05

基于网页浏览的用户兴趣度研究

许国迎, 张 宁

(上海理工大学 管理学院, 上海 200093)

摘要: 根据高校校园网用户的网页浏览记录,着重分析用户在浏览网页时的不同行为兴趣特征.首先利用文本分类技术,分析提取出用户的兴趣行为关键词.其次在兴趣行为关键词的基础上,选取了若干典型个体用户,统计出这些典型个体用户网页浏览记录中对于每个兴趣关键词的访问量随时间变化的关系.最后运用归一化的方法对典型个体用户的兴趣度进行量化研究,从而实现了对于用户网页浏览过程中不同兴趣行为的度量以及进行相互比较的目标.

关键词: 人类动力学; 网页浏览; 归一化; 兴趣度

中图分类号: N 94; TP 274 **文献标志码:** A

Users' Interests Based on Users' Web Browser Log

XU Guo-ying, ZHANG Ning

(Business School, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Many different users' behavior characteristics were emphatically analyzed according to the college campus network users' web browser log. Some kinds of interest keywords were extracted from it to show the different users' interests. Then typical users were selected in order to calculate the amount of the users' access which is changing with time on the basis of the interest keywords in classification. By using the normalization method, the quantified different users' degrees of interests were displayed and compared with.

Key words: *human dynamics; web browser log; normalization; degree of interest*

当今时代,网络的发展日益成为人们生活的重要组成部分.随之而来的,网络中的巨大数据资源亦得到了广大学者越来越多的重视.因此,许多学者致力于数据挖掘技术与人类动力学研究的结合,并应用于实践当中.

学者 Barabasi 曾在《Nature》上提出了一个基于决策的优先权模型^[1],自此开启了人类动力学方面的热烈讨论.随后, Vazquez 又得到了对该模型的精

确解^[2],这些模型都是从排队论的任务模型角度来研究人类动力学的.其实,生活中的人类行为纷繁复杂,何止是完成任务这类行为.比如说兴趣爱好往往也是人类的一种重要行为去向,有学者就结合 QQ 群的聊天记录^[3],对单个用户和群里所有用户发言的时间间隔进行了研究,证明了群体间这种网络即时沟通行为具有人类动力学特征.更有文献从人类行为、复杂网络和信息挖掘的角度给后续的研究和

收稿日期: 2013-01-15

基金项目: 国家自然科学基金资助项目(70971089);上海市重点学科建设资助项目(S30501)

第一作者: 许国迎(1987-),男,硕士研究生.研究方向:信息网络优化. E-mail: usstxgy@163.com

通讯作者: 张 宁(1956-),女,教授.研究方向:系统分析与集成. E-mail: zhangning@usst.edu.cn

学习提供了不可多得的重要资料^[4]。

现阶段,个性化推荐已经成为了热门的研究方向,特别体现在网络购物中,各商家都希望能够获取用户最真实的兴趣所在,并为之进行准确合适的个性化信息推荐。这些问题都需要深入挖掘用户兴趣,也蕴含了巨大的商业价值。由此,不难看出,如何得到用户的兴趣取向以及兴趣的度量方式成为了至关重要的课题。

当前挖掘用户兴趣行为的方式有两种,一种是单纯从用户行为的历史信息中发现其中所隐藏的规律,另一种是基于浏览内容和行为相结合的方式^[5]来研究用户的兴趣行为。事实上单纯从一个方面来分析用户的兴趣是不够的,应该从各个角度,不同层面来建立用户的兴趣簇。因此本文根据用户的网页浏览记录,利用文本分类技术提取出若干兴趣关键词,并进行分类统计分析。在得到用户访问量的基础上,运用归一化的方法,实现用户兴趣的度量和相互比较的目标。

1 用户兴趣的挖掘

1.1 兴趣的聚类分析

聚类分析就是将一组对象集合按照相似性分成若干类别,目的是使得同一类别的对象之间的相似度最大,而不同类别的对象间相似度最小。聚类的思想源于很多学科,如数学、计算机科学、生物学、统计学和经济学等。在不同的领域里,这种技术都被用于描述数据、衡量数据源之间的相似性,并把数据源分类到不同的簇中。特别是在商业领域,经常会通过聚类分析来发现不同类型的客户群,进而刻画不同客户群的特征,从而可以更好地帮助商家了解自己的客户,向客户提供更好的服务。聚类分析的算法^[6]主要包括层次聚类法、基于密度的方法、平面划分方法、基于网格的方法和基于模型的方法。

层次聚类算法,又称系统聚类法,是被广泛应用的算法之一。虽然复杂度较高,不适合大数据的计算,但操作步骤简单方便。本文利用层次聚类的方法,深入分析群体用户的网页浏览记录,并从中得到相关的用户兴趣类。

1.2 兴趣的分类标准

在得到用户兴趣簇之后,综合利用文本分类技术提取出用户的兴趣关键词。首先对用户浏览过的页面进行内容分析,并根据主题信息对页面再进行

聚类分析。在聚类的过程中除了考虑页面内容的相近程度外,还辅以页面路径进行归类判断,从而得到网页页面的兴趣簇。最后,本文为了较为准确地反映用户的真实兴趣,将用户的上网行为分为18类:搜索引擎、教育、新闻门户、论坛博客、交友聊天、娱乐、网上购物、生活相关、游戏、体育、电影音乐、web邮件、文学、财经、求职招聘、房产装修、股票交易和军事。鉴于所选对象为高校师生,因此这18类兴趣关键词可基本代表了这一特定群体的主流兴趣取向。

1.3 用户兴趣度模型

兴趣度,就是用来衡量人们对某事物的感兴趣程度。个性化推荐系统对于这种抽象的概念,一般是通过模型的方法,先给出相应的兴趣度定义,再加以数据的实证研究。目前主要有传统的基于浏览内容的兴趣度模型^[7-8]、基于用户浏览行为的兴趣度模型^[9]和动态变化的用户兴趣模型,这3种模型各有所长,代表了研究水平的不断深入和提高。

基于浏览内容的用户兴趣模型一般是考虑用户在某一兴趣类中的访问频繁度,即兴趣类页面集的页面总数或用户会话总数。如果某一兴趣类的页面总数最多,那么对该类兴趣的倾向程度也就最高。然而,这种方法的不足之处是将用户访问的所有页面等同地看待,没有分出主次轻重。事实上,每张页面所包含的信息并非总是单一,用户对页面中不同内容的兴趣程度也有所区别。

基于用户浏览行为的模型主要是分析用户的行为模式,并结合用户的浏览内容,挖掘用户的兴趣及给出相应的度量标准和计算方法。因为用户的需求不同,从而用户的各种浏览行为也就体现出不同的兴趣倾向。一般来说,用户浏览行为有页面标记行为(包括增加书签、删除书签、保持页面和打印页面行为)和页面操作行为(包括复制、粘贴、剪切、拉动滚动条、点击链接和移动鼠标行为)。这种模型引入了用户的浏览行为作为兴趣度的变量,从而可以更准确地度量用户的兴趣度。

不难发现,上述两种兴趣度模型都是假设用户的兴趣不变,但实际生活中用户的兴趣却是动态变化的,既可以是逐渐增加,愈加感兴趣,也可以是一时兴起,继而逐渐淡忘。因此,有学者提出了动态变化的用户兴趣度模型,如蒋翀等^[10]建立基于线性衰减的用户兴趣度模型,单蓉^[11]建立了基于遗忘机制的用户兴趣度模型,两者分别选用不同的数学方法来分析和量化用户动态变化的兴趣。

1.4 用户兴趣度计算公式

本文中用户兴趣度是指用户对某一兴趣关键词的感兴趣程度,是对兴趣的一个量化指标.综合分析各种兴趣度模型的优缺点后,结合研究数据的特点,本文选择归一化的方法来度量用户的兴趣度.计算公式为

$$L_i = \frac{x_i - \min(x_1, x_2, \dots, x_n)}{\max(x_1, x_2, \dots, x_n) - \min(x_1, x_2, \dots, x_n)}$$

$$i = 1, 2, 3, \dots, n$$

其中, L_i 表示用户对第 i 个兴趣关键词的兴趣度, x_i 指用户对第 i 个兴趣关键词的访问量, $\max(x_1, x_2, \dots, x_n)$ 表示访问量的最大值, $\min(x_1, x_2, \dots, x_n)$ 为访问量的最小值.

2 典型用户的兴趣模式

对每个兴趣关键词,选取典型个体用户进行网页浏览记录的统计分析.对于上述 18 个关键词,本文选取了 8 位典型用户的网页浏览记录加以分析,为表述方便,分别用英文字母 A, B, ..., H 表示.统计出的用户访问量变化规律如图 1~8 所示.

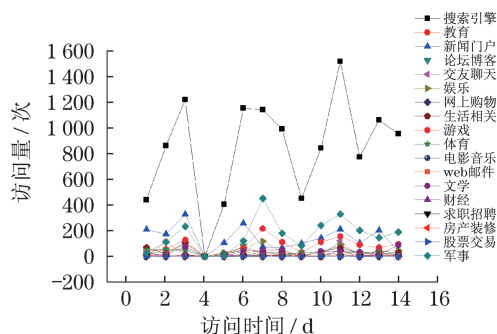


图1 用户A的网页浏览记录(搜索引擎)

Fig.1 Web browser log of the user A (search engine)

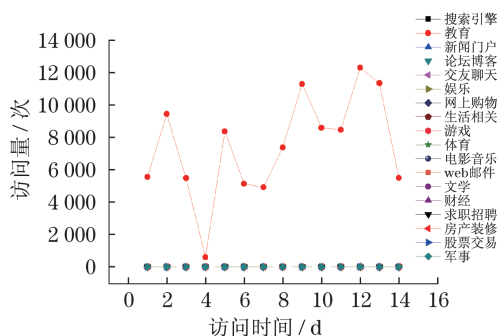


图2 用户B的网页浏览记录(教育)

Fig.2 Web browser log of the user B (education)

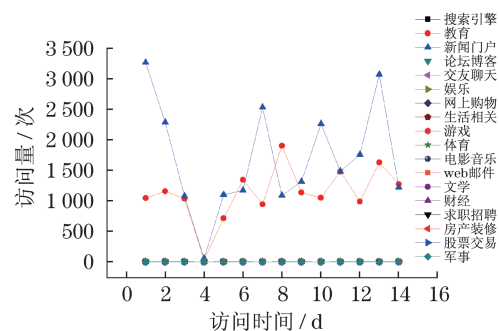


图3 用户C的网页浏览记录(新闻门户)

Fig.3 Web browser log of the user C (news portal)

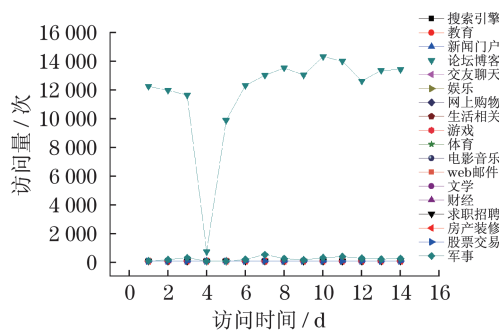


图4 用户D的网页浏览记录(论坛博客)

Fig.4 Web browser log of the user D (forum blog)

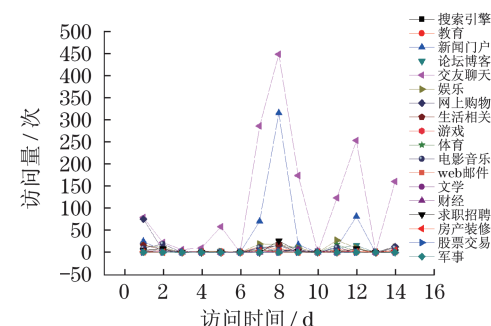


图5 用户E的网页浏览记录(交友聊天)

Fig.5 Web browser log of the user E (online chatting)

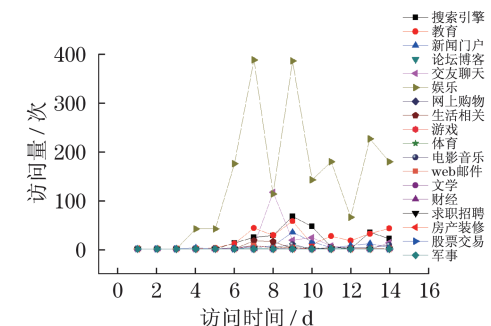


图6 用户F的网页浏览记录(娱乐)

Fig.6 Web browser log of the user F (entertainment)

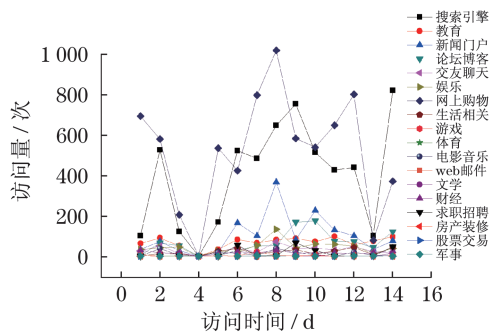


图7 用户G的网页浏览记录(网上购物)

Fig.7 Web browser log of the user G (online shopping)

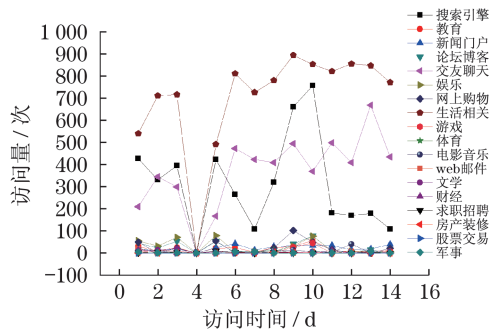


图8 用户H的网页浏览记录(生活相关)

Fig.8 Web browser log of the user H (life-related)

这8幅图显示的是典型用户在14 d时间内的访问量曲线图,以天为单位,分别以上述18类兴趣关键词作为统计对象,统计出用户对这些兴趣关键词的访问量.从图中可以看出,每个典型用户都有自己特有的兴趣倾向面,访问曲线也呈现出形态各异的波动性.如从用户A的浏览记录里发现该用户更多的是使用搜索引擎网站,常用的有百度、谷歌等热门搜索引擎网站;用户B和D表现出对某一个兴趣关键词有极大兴趣,而对其它兴趣类则很少浏览的特点.因此,在个性化推荐系统中,像此类兴趣单一型的用户,可以从群体用户中显著地分离开来.另外,可设置个性化信息推送系统,迎合该类特定用户的兴趣,有针对性地推荐信息,一定会达到事半功倍的效果.用户C、E、F、G和H则都表现出对某几类兴趣关键词的较高关注.具体来说,用户C关注新闻,用户E喜欢交友聊天,用户F更多的是浏览娱乐信息,用户G在网上购物时,倾向于浏览与网络购物相关的信息,用户H喜欢与生活相关的内容.这些性格多样、兴趣广泛的用户在群体用户中占有相当大的比重,可以对该类用户作多样性信息推荐.曲线图的优势在于直观,易分析出典型用户的最大兴趣关注领域,但对于其它兴趣关键词之间的区别如何、

它们之间又有什么联系,无法从图中直接得知.为此,用户兴趣度的量化就显得十分必要,这也是本文兴趣度模型的重点所在.

3 典型用户的兴趣度计算

针对典型个体用户的网页浏览记录,运用兴趣度计算公式分别进行归一化处理,具体结果如表1所示.

表1 典型个体用户的兴趣度

Tab.1 Degree of typical individual user's interest								
兴趣关键词	A	B	C	D	E	F	G	H
搜索引擎	1.000	0.000	0.000	0.000	0.020	0.121	0.773	0.441
教育	0.027	1.000	0.664	0.000	0.028	0.138	0.132	0.027
新闻门户	0.172	0.000	1.000	0.000	0.330	0.043	0.197	0.034
论坛博客	0.041	0.000	0.000	1.000	0.041	0.005	0.125	0.033
交友聊天	0.019	0.000	0.000	0.000	1.000	0.095	0.061	0.530
娱乐	0.055	0.000	0.002	0.000	0.077	1.000	0.086	0.040
网上购物	0.010	0.000	0.000	0.000	0.073	0.001	1.000	0.034
生活相关	0.046	0.000	0.000	0.000	0.033	0.019	0.039	1.000
游戏	0.100	0.000	0.000	0.000	0.005	0.000	0.006	0.012
体育	0.035	0.000	0.000	0.000	0.020	0.000	0.001	0.004
电影音乐	0.009	0.000	0.000	0.000	0.044	0.019	0.024	0.013
web邮件	0.003	0.000	0.000	0.000	0.000	0.006	0.001	0.003
文学	0.040	0.000	0.000	0.000	0.006	0.000	0.003	0.007
财经	0.005	0.000	0.000	0.000	0.011	0.000	0.033	0.005
求职招聘	0.001	0.000	0.000	0.000	0.031	0.000	0.028	0.002
房产装修	0.007	0.000	0.000	0.000	0.009	0.000	0.001	0.001
股票交易	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
军事	0.198	0.000	0.000	0.000	0.000	0.000	0.001	0.003

有了归一化的度量结果后,就可以比较清晰地看出每一个用户的兴趣关键词之间的区别和联系.如用户A对搜索引擎的兴趣度是1.000,充分说明了该用户对搜索引擎网站的兴趣度是最高的,对教育类信息的兴趣度是0.021,而对游戏类信息的兴趣度是0.100,新闻类和军事类信息的兴趣度都是大于0.100.由此可以推断,用户A不仅倾向于用搜索引擎网站,还对新闻类和军事类信息相当关注,并且关注程度是高于游戏类信息的.不妨大胆预测,这是一位朝气蓬勃、喜欢军事的年轻男性.综合所选取的8位典型用户,也只有该用户表现出对军事信息最高的兴趣,非常具有代表意义.若是在个性化推荐系统中,这样的用户就可以作为典型的用户类型,个性化地为其推送军事相关的信息.用户B则是非常

个性化的典型个体用户,在他的网页浏览记录中,几乎全部浏览的是教育类信息,可见其对教育领域的关注程度是非常高的,甚至可以大胆推测,该用户极有可能是一位教师,非常关注国家的教育事业.用户C浏览新闻信息的兴趣度为1.000,并且浏览教育信息的兴趣度是0.664,除了这两类兴趣关键词外,甚少浏览其它兴趣关键词的信息,说明了该用户是不仅关注新闻,尤其是关注教育领域新闻的人,他的兴趣倾向也会更多地偏向于教育领域.从用户聚类分析的角度,用户B和C可以说是归于一类.用户D浏览论坛博客新闻兴趣度为1.000,其它兴趣关键词为0.000,充分展现了这是一个喜欢交友,喜欢网络聊天的用户,可以想见,他对交友类信息会比他人更为敏感和关注.用户E则是一个兴趣较为广泛的人,表现出喜欢看新闻、听音乐、交友聊天和网上购物,而且兴趣度相差不大.对于其他用户,也可以通过类似的比较方法,发现不同用户的不同兴趣倾向,以及进行相互之间的纵向比较.

可以展望,用上述兴趣度计算方法,可以很好地定位互联网中成千上万用户的兴趣倾向及兴趣度,对网站的设计和建立个性化推荐系统有一定的参考意义.

4 结束语

从实证角度利用兴趣聚类方法对网页浏览日志中群体用户的兴趣进行挖掘,分析提取出群体用户的18类兴趣关键词,并运用统计学方法,对8位典型用户的网页浏览记录进行分析,统计各自的兴趣关键词访问量.然后根据兴趣度的计算公式,给出了用户兴趣的度量方法,得到了理论模型和实证结果.

文章提供了一个寻找用户兴趣倾向、度量用户兴趣度的方法,从而为人类动力学研究提供了一个切实可行的思路,并且对个性化推荐系统的研究也具有较好的指导意义.

参考文献:

- [1] Barabasi A L. The origin of bursts and heavy tails in human dynamics [J]. Nature, 2005, 435 (7039): 207 - 211.
- [2] Vazquez A. Exact results for the Barabasi model of human dynamics [J]. Physical Review Letters, 2005, 95 (24): 248710.
- [3] 罗芳,杨建梅,李志宏. QQ群消息中的人类动力学研究[J]. 华南理工大学学报, 2011, 13(4): 14 - 19.
- [4] 汪秉宏,周涛,周昌松. 人类行为,复杂网络及信息挖掘的统计物理研究[J]. 上海理工大学学报, 2012, 34 (2): 103 - 117.
- [5] 赵银春,付关友,朱征宇. 基于Web浏览内容和行为相结合的用户兴趣挖掘[J]. 计算机工程, 2005, 31 (12): 93 - 94.
- [6] Han J W, Kamber M. 数据挖掘概念与技术[M]. 北京:机械工业出版社, 2001.
- [7] 吕佳. 基于兴趣度的web用户访问模式分析[J]. 计算机工程与设计, 2007, 28(10): 2403 - 2407.
- [8] 郭岩. 网络日志中用户兴趣的挖掘及利用[D]. 北京:中国科学院计算技术研究所, 2004.
- [9] 王微微,夏秀峰,李晓明. 一种基于用户行为的兴趣度建模[J]. 计算机工程与应用, 2012, 48(8): 148 - 151.
- [10] 蒋翀,费洪晓. 基于线性衰减的用户兴趣建模[J]. 计算机系统应用, 2010, 19(6): 140 - 143.
- [11] 单蓉. 用户兴趣模型的更新与遗忘机制研究[J]. 微型电脑应用, 2011, 27(7): 10 - 11.

(编辑:丁红艺)