

个性化搜索用户兴趣更新学习及评价研究

宋毅¹, 徐志明²

(1. 哈尔滨华德学院 电子与信息工程学院 计算机应用技术系, 黑龙江 哈尔滨 150025;
2. 哈尔滨工业大学 计算机学院, 黑龙江 哈尔滨 150025)

摘要:提出了一种自适应的用户兴趣模型更新学习及评价方法。为了给用户提供更精准的查询结果,将用户兴趣模型加入自适应调整算法后进行验证,研究通过分析用户短期兴趣、长期兴趣规律,成为该系统建立用户的兴趣模型可能。随着时间等的变化,用户兴趣也会发生相应变化。通过自适应学习过程,为了更好地识别用户感兴趣的信息,通过研究规律进行总结分析。对兴趣学习技术进行研究,同时对该算法进行了评价。主要计算了查准率等参数,为此通过评价得出该用户兴趣挖掘精准率较好,对于现代计算机网络购物,以及网络应用过程挖掘用户行为和兴趣提供了良好的方案,也为个性化推荐应用提供了帮助。

关键词:搜索;兴趣;数据挖掘;学习;评价

中图分类号:TP302

文献标识码:A

文章编号:1673-629X(2018)06-0064-03

doi:10.3969/j.issn.1673-629X.2018.06.014

Research on Personalized Search User Interest Updating Learning and Evaluation

SONG Yi¹, XU Zhi-ming²

(1. Department of Computer Application and Technology, School of Electronic Information Engineering, Harbin Huade University, Harbin 150025, China;
2. School of Computer, Harbin Institute of Technology, Harbin 150025, China)

Abstract: An adaptive updating learning and evaluating method for user interest model is proposed. In order to provide users with more accurate search results, the user interest model is verified after adding adaptive adjustment algorithm. Through the analysis of user short-term interest and long-term interest in law, it becomes interested in model of users of the system. With the change of time, the user's interests will change accordingly. We analyze the algorithm of user's interest by the adaptive learning process, in which the rules change, so as to obtain the user's interest points. We also research on interest learning technology and evaluate it. Main parameters like precision is calculated, and the evaluation shows the user interest mining precision rate is better, providing a well solution for the modern computer network shopping and network application and process of mining user behavior and interest, with aid to personalized recommendation application.

Key words: search; interest; data mining; learning; evaluation

0 引言

每个用户总体兴趣是个恒定常数。人的精力是有限的,用户兴趣类别偏好也是有限的,如果对某些类兴趣度高,对其他类兴趣度必然降低。文中关注用户感兴趣的类别,用户整体兴趣满足固定常数,也就是随着更新学习,用户某些兴趣可能由高到低递减变化,而有些类别兴趣由低到高递增变化,但是用户在整个类别

偏好体系中兴趣度总和个恒定常数用户兴趣能够反映用户主题偏好^[1]。然而现有大部分个性化搜索引擎没有识别用户长期兴趣和短期兴趣,因此提出基于短期兴趣来学习用户长期兴趣^[2]。

用户兴趣随时间变化符合一定规律,基本规律是先快后慢,先多后少,逐渐遗忘。面对兴趣遗忘过程,如果兴趣模型不进行更新,将会出现用户兴趣漂移现

收稿日期:2017-05-22

修回日期:2017-08-26

网络出版时间:2018-02-24

基金项目:国家自然科学基金(61672185);黑龙江省自然科学基金资助项目(F2015046)

作者简介:宋毅(1981-),女,讲师,研究方向为自然语言处理、数据挖掘、计算机网络安全;徐志明,博士,教授,研究方向为信息检索、社会计算、移动电商。

网络出版地址: <http://cnki.net/kcms/detail/61.1450.TP.20180224.1510.004.html>

象;也就是随着时间变化,用户对某类兴趣可能增加,对另一类兴趣可能减小,也会有短期兴趣积累一定时间,将会向长期兴趣演变,用户兴趣需要定期更新,可使模型自动发现用户的新兴趣,并能适应用户兴趣的变化,从而能更好、更准确地反映用户的真实兴趣。具体更新需要对增量数据进行处理,因为如果用户对某类兴趣增加,相关文档会增加,对新增数据的大量数据计算需要本文高效处理^[3]。

1 用户兴趣更新学习方法

1.1 时间窗原理

时间窗通过时间的阈值来设定,有很多研究均采用此方案。基于优化时间窗的用户兴趣漂移算法^[4],利用分类错误率的变化跟踪用户兴趣的漂移,当用户兴趣发生变化时,通过优化时间窗算法自动调节时间窗的大小^[5],用户模型根据该值来进行改进。该算法主要通过客观的时间来设定,因此对于用户遗忘比较公正。目前有学者讨论了个性化技术兼顾时间窗算法的模型^[6]。在此,考虑长期因素,也包括短期因素,两者兼顾观察用户兴趣的趋势。该机制效率良好。

1.2 相关反馈原理

为了改进用户兴趣模型的精准率,加入相关反馈知识^[7]。该算法是根据原来的文本时间,当有更新文本值时,加入新的文本,同时原来文本相同的不进行更新,只更新不同的差值,这样对于更新时间明显减少,更新效率大大提高,对于发现用户最新的兴趣节省了时间。

1.3 遗忘规律

有研究学者根据遗忘规律进行衰减^[8],通过不同的年龄来标识样本信息,时间增长,标识信息的日期也增长,如果时间超出设定数值,忽略该样本信息。改进用户模型仅用没有被筛选掉的数据,被筛选留下的数据可以反映用户随时间变化的兴趣规律。

1.4 更新学习思想

第一是用户短期兴趣更新学习,采用遗忘因子进行更新;第二是短期兴趣向长期兴趣变化更新,由于短期兴趣经过一定时期累加^[9],随着兴趣度累加到一定时期^[10],短期兴趣会演变为长期兴趣,面对增大的数据量,文中考虑增量学习方法,所以采用改进的 Rocchio 定期自动调整学习模型^[11];最后是长期兴趣学习,由于长期兴趣具有变化缓慢、稳定的特点,如果长时间内长期兴趣的兴趣度仍然较小,可以判断用户对该类兴趣不感兴趣,可以对该类兴趣进行淘汰。由此启发,联想到操作系统中的最近最少使用算法(LRU)^[12],对应最近一段时间内最久没有使用的兴趣类别进行淘汰,也就是对最近一段时间内长时间兴趣

度低的兴趣类别进行淘汰,将新加入的兴趣度高的兴趣类别更新进行替换,进行长期用户兴趣更新学习。

2 短期用户兴趣更新学习

用户兴趣更新学习包括加入用户的最新兴趣和对旧兴趣的遗忘^[12]。实验证明,人们在学习中的遗忘是有规律的,遗忘的进程很快,并且先快后慢。观察曲线会发现,学得的知识在一天后如不抓紧复习,就只剩下原来的 25%。随着时间的推移,遗忘的速度减慢,遗忘的数量也就减少。有人做过一个实验^[13],两组学生学习一段课文,甲组在学习后不复习,一天后记忆率 36%,一周后只剩 13%。乙组按艾宾浩斯记忆规律复习,一天后保持记忆率 98%,一周后保持 86%,乙组的记忆率明显高于甲组。遗忘因子^[7] $F(x)$ 如式 1 所示:

$$F(x) = e^{-\frac{\log 2}{hl}(cur-est)} \quad (1)$$

其中,cur 为当前日期;est 为兴趣词在用户兴趣库中出现的最近日期;hl 为减弱值。

经过弱化,用户兴趣遗忘一半,但并不是线性遗忘,遗忘速度是先快后慢。hl 可以根据大量实验测试确定,也可以人为确定,确保在短期兴趣中历史兴趣遗忘快些,长期兴趣中遗忘速度适当慢些。hl 短期=2, hl 长期=7,随着时间的流逝,用户兴趣也会有对应一些规律性变更,因此挖掘用户模型也对应参数调整。随着个性化信息推荐的发展,研究人员进行了时间参数更新的模型研究,对于存在的差异也就是兴趣的偏移解决策略提出了对应方案:时间窗方法、遗忘函数方法、混合用户模型等^[14]。以上思想基本是 FIFO 算法原理,缺乏考虑用户长期和短期结合的思想^[14]。

实验中,每天用户兴趣度更新都有所变化,或增大,或减小,以第 7 天为例,用户兴趣更新遗忘规律如图 1 所示。通过图 1 可以看出,用户在第 7 天时,在各类兴趣度都有所减小,在体育和军事类兴趣减小均等,在旅游类兴趣减小幅度大,可以推测用户在一周后对旅游领域兴趣明显降低,相对不感兴趣了,而对汽车和军事类别还是比较感兴趣。

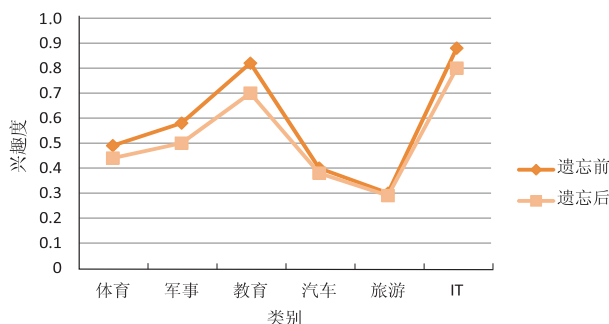


图1 用户兴趣遗忘结果

文中将 10 天设定为短期用户兴趣,具体更新结果如图 2 所示。可以看出,整体衰减速度是先快后慢,先

多后少的趋势符合人们的遗忘规律。

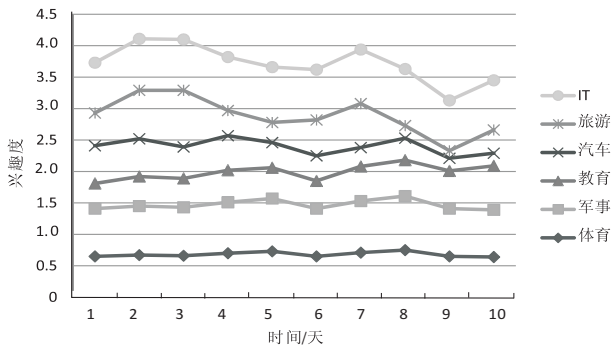


图 2 短期兴趣更新

3 长期用户兴趣更新学习

3.1 长期兴趣更新学习方法

个性化原理是按照用户所感兴趣的数据,根据时间的变化以及兴趣的热点来获取用户兴趣点,用户实际需要的数据也会根据模型而输出具体数值^[15]。该算法通过最近最久未用方法改进用户模型,设定阈值的尺寸为 L ,当有多于 L 个兴趣出现时,利用“访问的局部问题”,按照“到目前为止最少使用的兴趣,很可能也就是将来最少使用的兴趣”的原则,把兴趣点最低的值淘汰。

3.2 长期兴趣更新学习实验

根据原理,被移除的兴趣应该是那些在近期内被再次访问的可能性最低的兴趣对象^[16]。该算法优于时间窗机制进行淘汰的方法,优点是命中率较高。根据用户在半个月内在体育、军事、教育、汽车、旅游和IT六类的兴趣度淘汰表,可以计算出命中率,就是新加入兴趣已在原用户兴趣序列中的命中次数与新加入兴趣的总数之比。长期兴趣更新结果如图3所示。

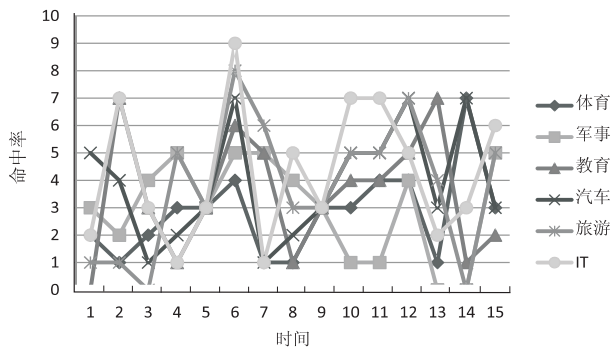


图 3 长期兴趣更新

4 实验结果及结论

4.1 兴趣度的相对误差

兴趣计算的准确程度需要衡量,所以采用传统的相对误差方法,如式2所示。

$$E = \frac{|V - V'|}{\text{万条数据}}$$

(2)

其中, E 为相对误差; V 为真实兴趣度; V' 为测量兴趣度。

表1是用户在体育、军事、汽车、教育、旅游和IT六类中兴趣度相对误差实验结果,相对误差率越小,表明兴趣度越准确,用户兴趣模型性能越好。表中显示了用户的兴趣误差:误差范围在0.011之内,兴趣度计算相对误差率较低,表明用户兴趣度的计算相对准确率较高。

表 1 误差分析

体育	军事	汽车	教育	旅游	IT
0.12	0.34	0.40	0.80	0.40	0.89

4.2 查询分类的准确率

采用传统的两个参数评价分类性能,即查准率及召回率。具体定义如式3所示。

$$P = \frac{Q_T}{Q_A}$$

(3)

其中, Q_T 为查询分类正确数量; Q_A 为所有查询数量。

查询串有相应类别,文中模型的本质是将查询分类,以查询分类的准确率来评价分类准确性。输入查询串320个,分别属于体育、军事、汽车、教育、旅游、IT六类,分类准确率平均值为0.86,每类分类性能如表2所示。

表 2 查询分类准确率

体育	军事	汽车	教育	旅游	IT
0.92	0.87	0.81	0.88	0.82	0.89

5 结束语

阐述了用户兴趣更新学习意义和现有方法,基本的用户兴趣更新学习方法包括时间窗机制、遗忘因子更新学习和最近最少使用算法等。分为短期用户兴趣更新学习和长期用户兴趣更新学习。短期兴趣学习方法采用遗忘因子进行更新学习,长期兴趣学习方法采用最近最少使用算法。通过更新学习,能够动态识别用户兴趣。评价方法包括相对误差分析方法、传统的准确率方法。相对误差值越小,查询串分类准确率越高,说明用户兴趣模型识别用户兴趣类别越准确。相应地给出了实验分析,并且具体评价了用户兴趣模型的性能。

参考文献:

[1] 邢春晓,高风荣,战思南,等. 适应用户兴趣变化的协同过滤推荐算法[J]. 计算机研究与发展,2007,44(2):296-301.

[2] 费洪晓,戴 弋,穆 珺,等. 基于优化时间窗的用户兴趣

(下转第 72 页)

的绘制时间的增量略有波动。

综合图 5 和图 6 的数据可知,硬件加速后,在矩形填充这个基本图形操作上,硬件加速获得了较好的效果,加速比基本维持在 2 以上,实现了预期的目的。

5 结束语

通过对 Qt/Embedded 体系结构和 Qt/Embedded 图形引擎架构的分析,利用嵌入式 Linux 下的帧缓冲体系结构,提出一种 Qt/Embedded GUI 环境下图形硬件加速实现方法和图形加速实现架构。实现了帧缓冲体系对 Qt/Embedded GUI 体系的底层硬件支持,依赖于 Linux 帧缓冲体系的加速接口,使用硬件 GPU 内部加速功能对 Qt/Embedded GUI 环境下的基本图形绘制进行了硬件加速,取得了较好的加速效果。

参考文献:

- [1] 蒋 飞. 基于嵌入式 Linux 系统的数字电视 GUI 图形加速设计[D]. 北京:北京邮电大学,2010.
- [2] 李升亮,徐剑锋,李峻林. 嵌入式系统中的多窗口 GUI 系统的研究[J]. 计算机与数字工程,2008,36(10):126-128.
- [3] 周 鸿. 基于嵌入式系统的智能缝制设备研究[D]. 西安:西安电子科技大学,2010.
- [4] 李军民,祝红军. 基于 ARMLinux 平台的 QT/E 键盘实现[J]. 微计算机信息,2008,24(26):27-29.
- [5] 周 开,倪 伟. 基于 Qt/E 的嵌入式 Linux GUI 研究与实现[J]. 淮阴工学院学报,2015,24(3):10-13.
- [6] 严吉国. 基于嵌入式 Linux 的 200MHz 数字存储示波器的

(上接第 66 页)

- 漂移方法[J]. 计算机工程,2008,34(16):210-211.
- [3] 战守义,井 新. 加入时间因素的个性化信息过滤技术[J]. 北京理工大学学报,2005,25(9):782-785.
- [4] 蒋 萍. 基于用户兴趣挖掘的个性化模型研究与设计[D]. 苏州:苏州大学,2005.
- [5] 史朝辉,王晓丹,杨建勋. 一种 SVM 增量训练淘汰算法[J]. 计算机工程与应用,2005,41(23):187-189.
- [6] 李 娜. 基于垂直搜索引擎的农业信息推荐关键技术研究[D]. 沈阳:沈阳农业大学,2016.
- [7] 韩春晓. 中文期刊个性化搜索引擎的设计与实现[D]. 哈尔滨:哈尔滨工业大学,2014.
- [8] 张梅芳. 基于改进 PageRank 算法和用户兴趣的个性化搜索研究[D]. 天津:河北工业大学,2014.
- [9] 王 哲. 一种基于位置服务的个性化美食搜索算法研究与实现[D]. 长沙:湖南大学,2013.
- [10] 黄华东. 基于用户模型的个性化搜索研究[D]. 上海:华东理工大学,2013.
- [11] 邓晓嘉. 一种基于 RSS 用户兴趣的个性化搜索系统[D].

设计与实现[D]. 南京:东南大学,2009.

- [7] CHEN F, FAN X. Embedded system's performance analysis with RTC and QT[C]//Proceedings of the 7th international conference on advanced parallel processing technologies. Berlin:Springer-Verlag,2007:569-579.
- [8] 郭小梅. Linux 下的帧缓冲设备驱动研究与应用[J]. 工业控制计算机,2012,25(6):3-4.
- [9] YANG L, SANDER P V, LAWRENCE J. Geometry-aware framebuffer level of detail[J]. Computer Graphics Forum, 2008,27(4):1183-1188.
- [10] MAO C, JOHNSON K M. Fast-switching liquid-crystal-on-silicon microdisplay with framebuffer pixels and surface-mode optically compensated birefringence[J]. Optical Engineering, 2006,45(12):1269-1278.
- [11] 赵 洁,龚 威. 嵌入式 Linux 帧缓冲设备驱动程序[J]. 计算机系统应用,2010,19(12):208-211.
- [12] CHANG C Y, HUANG C H, CHU Y S. Efficient memory access methods for framebuffer-less video processing applications[C]//IEEE international symposium on circuits and systems. [s. l.], IEEE,2013:3026-3029.
- [13] 宋方伟,刘 勇,聂诗良,等. SM502 移动多媒体协处理器在嵌入式系统中的应用[J]. 兵工自动化,2010,29(2):91-92.
- [14] 苏哲欣,刘鸿飞,薛 晓. 基于嵌入式 Linux 的 LCD 驱动分析与实现[J]. 工业控制计算机,2009,22(2):29-30.
- [15] 黄相平,余水宝,夏 灿. 基于 S3C6410 平台的嵌入式 Linux 系统 LCD 驱动模块[J]. 微型机与应用,2013,32(13):9-12.

北京:北京工业大学,2010.

- [12] 石志伟,刘 涛,吴功宜. 一种快速高效的文本分类方法[J]. 计算机工程与应用,2005,41(29):180-183.
- [13] QIU Feng, CHO J. Automatic identification of user interest for personalized search[C]//Proceedings of the 15th international conference on world wide web. Edinburgh, Scotland, UK: ACM,2006:23-26.
- [14] KOUTRIKA G, IOANNIDISY. Personalized queries under a generalized preference model[C]//Proceedings of the 21st international conference on data engineering. Tokyo, Japan: IEEE,2005.
- [15] CLAYPOOL M, LE P, WASEDA M, et al. Implicit interest indicators[C]//Proceedings of the 6th international conference on intelligent user interfaces. Santa Fe, New Mexico, USA: ACM,2001:33-40.
- [16] SHEN Xuehua, TAN Bin, ZHAI Chengxiang. Implicit user modeling for personalized search [C]//Proceedings of the 14th ACM international conference on information and knowledge management. Bremen, Germany: ACM,2015:824-831.