doi:10.3969/j.issn.1673-4785.2010.02.014

混合模型的用户兴趣漂移算法

郭新明,弋改珍
(咸阳师范学院 信息工程学院,陕西 咸阳 712000)

摘 要:针对个性化信息服务中的用户兴趣漂移问题,提出了一种新的正态分布密度曲线遗忘函数,该函数符合用户兴趣遗忘的规律.并且将用户模型定义为长期模型和短期模型相结合的混合模型,其中短期模型使用最近最久未使用的滑动窗口算法进行更新,长期模型采用正态渐进遗忘算法进行更新.实验表明,该方法能够较迅速地发现和准确地跟踪用户的兴趣变化,提高了个性化信息服务的效率.

关键词:个性化;混合模型;兴趣漂移;遗忘函数
中图分类号:TP391.3 **文献标识码**:A **文章编号**:1673-4785(2010)02-0181-04

A hybrid algorithm to track drift of user's interests

GUO Xin-ming, YI Gai-zhen
(School of Information Engineering, Xianyang Normal University, Xianyang 712000, China)

Abstract: User interests inevitably drift while using a personalized information service. A new forgetting function with normal distribution density was proposed to accommodate drift. The function conformed to a user interest forgetting law. User interest was defined in a hybrid model that contained both long and a short-term components. The short-term component was renewed by using the least recently used algorithm. The long-term component was renewed by using the normal incremental forgetting distribution algorithm. Experiments showed that the algorithm noted changes in user's interests more quickly and tracked them more accurately, greatly improving the efficiency of personalized information services.

Keywords: personalization; hybrid model; interest drift; forgetting function

个性化信息服务的质量主要依赖于用户兴趣模型的准确程度,因此构造恰当的用户兴趣模型就成了个性化信息服务的关键技术.用户在进行信息检索的时候,个人兴趣可能会随时发生变化,因此用户的兴趣模型也应随着用户兴趣的变化而调整,使其能够准确地描述用户的当前兴趣特征.

目前,有关用户兴趣漂移的研究较多,其中包括漂移概念研究[1]、兴趣变化规律研究[2-3]、兴趣漂移模型研究[5-6]、兴趣跟踪研究[5-7]等.用户兴趣漂移算法主要有2种:时间窗口法[1]和遗忘函数法[2],时间窗口法是利用滑动时间窗滤除过时的兴趣,使窗口中存放用户的最新兴趣;遗忘函数法是利用遗忘函数对用户兴趣的权重进行衰减处理,从而得到用户的真正兴趣.

本文将进一步研究用户兴趣的漂移问题,包括用户兴趣的变化规律,以及采用怎样的用户兴趣模型来快速跟踪用户的兴趣变化.

1 渐进遗忘函数

1.1 线性遗忘函数

由于人对事物的遗忘是一个渐进的过程,相同的兴趣在不同阶段对人的重要性是不同的,因此,兴趣的权重可以通过遗忘函数来计算.对用户来说,兴趣刚出现时重要性最高,随着时间的推移,重要性逐渐下降,因此遗忘函数应该是一个连续递减的函数.文献[2]提出了一种渐进的线性遗忘函数,如式(1):

$$w_i = -\frac{2k}{n-1}(i-1) + 1 + k. \qquad (1)$$

收稿日期:2009-12-05.
基金项目:陕西省科技厅自然科学基础研究计划资助项目(SJ08ZT14-8);陕西省教育厅科学研究计划资助项目(08JK481);咸阳师范学院专项科研基金资助项目(08XSYK335).
通信作者:郭新明. E-mail:guoxinming118@126.com.

该函数表示兴趣的权重随其出现的时间次序而变化,参数 n 为特征序列的长度;$i \in \{1,2,\cdots,n\}$ 是计数器,按照从最近选择的特征到第一次选择的特征的顺序,依次取 $i=1,i=2,\cdots,i=n,k \in [0,1]$,表示遗忘的快慢. 在某一时刻,有新的特征值出现时,重新计算所有特征的重要性. 对每一个特征 j,用户对其感兴趣的程度可以通过观测序列中其出现的情况计算得到,可以采用式(2)来计算.

$$c_j = \sum_{i=1}^{n} w_i a_i^j. \tag{2}$$

式中:i 为计数器,n 为用户的行为数,w_i 为遗忘函数计算的权重. 在 a_i^j 确定问题上,考虑概念之间的相关性,令 $a_i^j \in [0,1]$,表示某次观测对此特征的影响程度,若观测值就是特征 j,则 $a_i^j=1$,若观测值不是特征 j,则 $a_i^j=0$.

1.2 非线性遗忘函数

根据心理学的知识,人的遗忘规律不是完全线性变化的,对刚出现的兴趣在最近一段时间内关注度很高,立即遗忘的可能性极小,即使再有新的兴趣出现,该兴趣的关注度也不会马上降低. 但当用户较长时间不再使用某兴趣时,该兴趣的关注度会随时间迅速降低,当一个兴趣的关注度降低到一定程度时,即它能被用户回忆起来的可能性很小时,那么它的遗忘速度应该迅速减小[3]. 本文提出了一个基于正态分布密度函数的非线性遗忘函数,如式(3)所示(以下称正态遗忘函数):

$$w_i = \frac{1}{\alpha\sigma\sqrt{2\pi}} e^{\frac{-i^2}{2(\beta\sigma)^2}}. \tag{3}$$

式中:α、β 为调节因子,$\alpha=0.52,\beta=2,i \in \{1,2,\cdots,n\}$ 是计数器,按照从最近选择的特征到第一次选择的特征的顺序,依次取 $i=1,i=2,\cdots,i=n$(n 为序列的长度). 图 1 是线性遗忘函数与正态遗忘函数的对比图,容易看出正态遗忘函数对刚出现的新兴趣的遗忘速度要比线性遗忘函数慢,对于出现时间较长的兴趣正态遗忘函数要比线性遗忘函数遗忘得快.

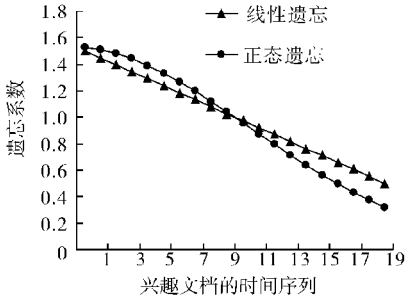

图 1 线性遗忘函数与正态遗忘函数的对比图

Fig. 1 Comparison chart of linear forgetting function and normal forgetting function

2 混合模型

心理学研究认为,人的记忆分为长期记忆和短期记忆,对于短期记忆,由于容量非常有限,因此当信息不能很快重现时,会被很快遗忘;对于长期记忆,当环境或场合的改变使得长期记忆中某些信息长期不用时,这些信息才会逐渐被遗忘[4]. 只有把与目前真正无关的信息剔除出去,才能快速准确地提取与当前环境有关的信息.

本文采用混合模型,将用户的兴趣模型看作由长期模型和短期模型共同组成,短期模型代表用户的近期兴趣,由用户的短期行为观察得到;长期模型表示用户的长期偏好,由长时间发展积累得到. 当一种短期兴趣的出现次数积累到一定程度时,则将此兴趣交给长期模型处理. 根据 2 种模型的特点,对于长期模型和短期模型采用不同的遗忘方法[5-6].

2.1 短期兴趣模型漂移算法

短期兴趣模型代表用户近期的兴趣偏好,由于其容量小,所以变化较快[7]. 对于该模型将最近最久未使用的方法与滑动窗口方法相结合,来实现对用户短期兴趣漂移的跟踪. 具体算法是:

当有新兴趣出现时:

1)若兴趣窗口未满,则将新兴趣加入到窗口的前端;

2)若兴趣窗口已满,则将兴趣窗口中最近最久未使用的兴趣滑出窗口,然后将新兴趣加入到兴趣窗口的前端.

2.2 长期兴趣模型漂移算法

长期模型对应用户的长期记忆,是用户比较固定的偏好,相对稳定. 因此当有新的兴趣发现时,需要采用一定的方法,将新加入的兴趣和以前的兴趣进行合并,得到当前的用户兴趣信息.

由于人对事物的遗忘是一个渐进的过程,相同的兴趣在不同阶段对人的重要性是一个逐渐变化的过程,因此,兴趣的权重可以通过遗忘函数来计算. 对用户来说,兴趣出现时重要性最高,随着时间的推移,重要性逐渐下降,采用正态遗忘函数的方法来跟踪用户的长期兴趣漂移. 若用户兴趣模型没有发生明显变化时,那么用户的长期兴趣模型与短期兴趣模型的相似性保持在一个固定的范围内(比如:相似性大于或等于 0.3);当用户的长期兴趣模型发生改变时,那么用户的长期兴趣模型与短期兴趣模型的相似性会突破相似性的下限,而且会随着时间的推移不断降低,这就表明用户的长期兴趣模型已经

发生了改变. 跟踪用户长期兴趣模型的算法如下:

1)每隔一定的时间间隔,计算用户的长期兴趣模型与短期兴趣模型的相似性,如式(4):

$$\mathrm{sim}(V,W) = \frac{\sum_{i=1}^{n} v_i \times \omega_i}{\sqrt{\sum_{i=1}^{n} v_i^2} \times \sqrt{\sum_{i=1}^{n} w_i^2}} . \quad (4)$$

式中:V 和 W 分别是用向量空间模型表示的长期兴趣模型和短期兴趣模型.

2)若用户的长期兴趣模型与短期兴趣模型的相似性没有突破相似性下限(在本算法中下限统一定为 0.3),则什么也不做.

3)若用户的长期兴趣模型与短期兴趣模型的相似性突破了相似性下限,但没有迅速降低,表明这是一次突发性的兴趣震荡,也不进行长期兴趣模型的更新.

4)若用户的长期兴趣模型与短期兴趣模型的相似性突破了相似性下限,并且迅速降低,表明用户的长期兴趣模型已经发生了明显的改变,按照正态遗忘算法对用户的长期兴趣模型进行更新,直到用户的长期兴趣模型与短期兴趣模型的相似性大于等于相似性的下限为止.

5)返回 1).

3　实验分析

表 1 模拟了一个用户的兴趣变化序列,可以看出用户对兴趣 1 的关注度逐渐降低,对兴趣 2 的关注度逐渐升高,对兴趣 3 始终保持一定的关注度. 如果用户兴趣模型只能容纳 2 个兴趣,那么表 1 中描述的用户兴趣变化过程应该是从(兴趣 1,兴趣 3)变为(兴趣 2,兴趣 3),转折点是序列 9. 对表中的数据分别按照线性遗忘函数式(1)和正态遗忘函数式(3)进行计算,得到 2 个用户兴趣权重变化图如图 2 ~3 所示. 从图 2 可以看出采用线性遗忘函数在序列 16 处兴趣发生漂移,而在采用正态遗忘函数的图 3 中在序列 14 处兴趣发生了漂移,可见正态遗忘函数更接近实际情况.

表 1　各兴趣在用户兴趣序列(1-20)中出现的概率

Table 1　Probability of each interest in user's interest sequence

兴趣编号	P																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0.7	0.7	0.7	0.6	0.5	0.4	0.4	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.4	0.4	0.5	0.6	0.5	0.6	0.6	0.5	0.7	0.6	0.6	0.5
3	0.3	0.3	0.3	0.4	0.4	0.5	0.5	0.4	0.4	0.6	0.5	0.4	0.5	0.4	0.4	0.5	0.3	0.4	0.4	0.5

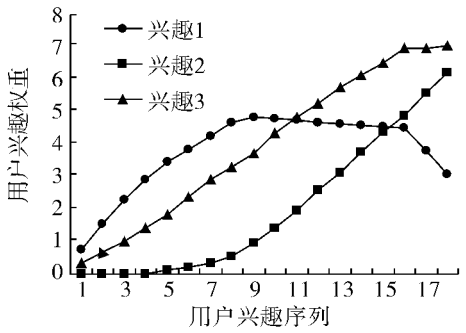

图 2　线性遗忘函数

Fig. 2　Linear forgetting function

使用网络爬虫 Heritrix 从 Internet 上收集到主题蕴含“搜索引擎”、“计算机网络”和“网络游戏”的 500 个页面作为测试数据,采用上述的混合模型和单纯使用正态渐进模型对用户浏览页面的兴趣进行追踪. 图 4 是 2 种模型的精确度对比图,可以明显看出采用混合模型比单纯采用正态渐进模型的精确度高.

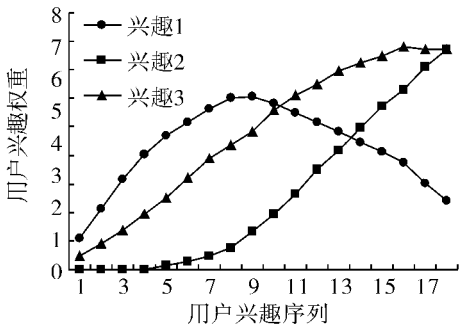

图 3　正态遗忘函数

Fig. 3　Normal forgetting function

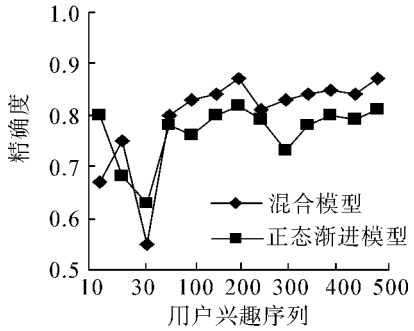

图4　2种算法精确度比较

Fig.4　Accuracy comparison of two algorithms

4　结束语

本文提出了基于混合模型的用户兴趣漂移算法,将用户兴趣分为长期模型和短期模型,短期模型使用最近最久未使用的兴趣淘汰方法来更新用户兴趣,长期模型采用基于正态遗忘函数的渐进遗忘方法进行更新.该模型比较准确地跟踪了用户的兴趣变化,具有较高的效率.在以后的工作中,还可以对算法进行其他方面的尝试,如根据时间等因素预测用户的兴趣变化规律等.

参考文献:

[1]KLINKENBERG R. Learning drifting concepts: example selection vs example weighting[J]. Intelligent Data Analysis, 2004,8(3):281-300.

[2]KOYCHEV I, SCHWAB I. Adaptation to drifting user's intersects [C]//Proceedings of ECML. Barcelona, Spain: IEEE Press,2000:39-45.

[3]郑先荣,汤泽滢.适应用户兴趣变化的非线性逐步遗忘协同过滤算法[J].计算机辅助工程,2007,16(2):69-73.
ZHENG Xianrong, TANG Zeying. Non-lineal gradual forgetting collaborative filtering algorithm capable of adapting to users drifting interest[J]. Computer Aided Engineering, 2007,16(2):69-73.

[4]杨炳钧.认知心理学[M].北京:中国轻工业出版社,2006:137-164.

[5]宋丽哲,牛振东.一种基于混合模型的用户兴趣漂移方法[J].计算机工程,2006,32(1):4-6.
SONG Lizhe, NIU Zhendong. A method of drifting user's interests based on hybrid model[J]. Computer Engineering, 2006,32(1):4-6.

[6]曹　毅,贺卫红.基于用户兴趣的混合推荐模型[J].系统工程,2009,27(6):68-72.
CAO Yi, HE Weihong. Mixed recommender model based on user's interest[J]. Systems Engineering, 2009,27(6):68-72.

[7]费洪晓,戴　弋.基于优化时间窗的用户兴趣模型漂移方法[J].计算机工程,2008,34(16):210-214.
FEI Hongxiao, DAI Yi. Method of drifting user's interests based on time window optimization[J]. Computer Engineering, 2008,34(16):210-214.

作者简介:

郭新明,男,1979年生,讲师,主要研究方向为信息检索与网络安全技术,参与省级科研项目2项,主持厅局级科研项目1项,发表学术论文7篇.

弋改珍,女,1969年生,副教授,主要研究方向为无线网络、网络仿真.参与省级科研项目2项,主持省级科研项目1项,主持厅局级科研项目1项,发表学术论文20余篇,其中被EI检索1篇.