基于树状向量空间模型的用户兴趣建模

费洪晓,蒋 翀,徐丽娟
(中南大学 信息科学与工程学院,湖南 长沙 410075)

摘 要:提出了一种基于树状向量空间模型的用户兴趣建模和更新方法,以满足网络消费者个性化的服务需求。根据用户在注册信息中提供的兴趣喜好建立兴趣模型,利用用户的反馈自适应地调整主题特征值向量和阈值,更新用户模型。通过加入时间向量区别短期兴趣和长期兴趣,及时准确地反映用户兴趣变化,提高个性化服务性能。

关键词:个性化;混合兴趣模型;树状向量空间模型
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2009)05-0079-03

User Profile Based on Dendriform Vector Space Model

FEI Hong-xiao, JIANG Chong, XU Li-juan
(School of Information Science & Engineering, Central South University, Changsha 410075, China)

Abstract: Proposes a new approach for user modeling and updating based on dendriform vector space model, in order to meet the personalized needs of customers in e-shopping. This profile is initially founded according to user's interest and modifies the profiles and threshold by user's feedback. This profile distinguishes between short-term interest and long-term interest by time vector, which can reflect user's interest change in time and offer personal service for better.

Key words: personalization; hybrid interest model; dendriform vector space model

0 引 言

互联网的飞速发展正逐步改变着人们的生产生活方式,并且成为人们获取和交流信息的重要途径。因为互联网的开放性,网络信息量正以惊人的速度增长,WWW庞大且无序的海量信息与用户狭窄专一需求之间的矛盾日益凸显。特别是在电子商务领域,网络消费者往往会被林林总总的商品所淹没,通过个性化服务可以有针对性地推荐满足用户兴趣的商品,将浏览者转变为购买者[1]。

用户兴趣建模是个性化服务技术的关键,决定了个性化服务系统的性能优劣。只有当用户的兴趣、偏好和访问模式等可以被系统很好地"理解"时,才可能实现理想的个性化服务。用户兴趣建模是指根据用户提供的信息(如用户浏览内容、浏览行为、基本信息、显式反馈和订单信息等)归纳出用户兴趣模型[2]。因为在个性化服务系统中的重要地位,如何构建、更新和进一步优化用户兴趣模型已经成为近年来的研究重点,也是个性化应用系统亟待解决的热点问题[3]。

收稿日期:2008-09-03
基金项目:湖南省科技计划基金资助项目(2006JT1040)
作者简介:费洪晓(1967-),男,副教授,硕士,研究方向为网络安全、数据挖掘、个性化服务。

用户兴趣建模的方法很多,包括向量空间模型、Navie Bayes、神经网路、遗传算法等[4]。文中提出一种混合用户兴趣模型,采用树状向量空间模型来表示,加入时间因素来区别用户的短期兴趣和长期兴趣,通过用户的兴趣喜好创建兴趣模型,根据用户的反馈来自适应地修改兴趣模型。与利用个性化服务技术对文本信息进行过滤不同,文中主要研究该模型在网络购物平台的应用,为电子商务平台提供个性化技术支持。

1 混合用户兴趣模型的表示

用户兴趣的改变是一种遗忘现象,和人的记忆相对应,也分短期兴趣和长期兴趣,长期兴趣比较广泛,遗忘速度慢,短期兴趣变化快,遗忘速度快[5]。混合用户兴趣模型将用户兴趣模型分为短期兴趣模型和长期兴趣模型,短期兴趣模型中存储用户的近期兴趣,长期兴趣模型中存储用户的长期偏好。用户兴趣模型处理流程如图1所示。

通过用户填写的注册信息得到感兴趣的信息集合,采用向量的方式来表示,通过聚类分析和兴趣度调整,得到用户短期兴趣模型,采用树状向量空间模型表示。除此之外,还可以通过对用户浏览历史和模式的捕捉,对用户行为进行分析,得到反映用户兴趣的参考

数据，这是一种隐式获取用户兴趣的方式。当用户的某一短期兴趣主题加入时间间隔达到一定程度，则认为该兴趣为用户的长期兴趣，把此兴趣加入用户的长期兴趣模型中，并在短期兴趣模型中删除此兴趣。

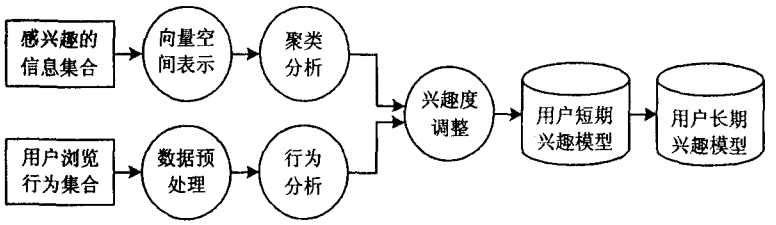

图 1 用户兴趣模型建立数据流图

2 混合用户兴趣模型的表示

混合用户兴趣模型的表示包括作为个性化过滤对象的商品和反映用户喜好的兴趣模型树的表示。

2.1 商品的向量表示

和传统的以向量空间模型表示文档[6]类似，每个具体的商品也可以设定一个向量 $M(<t_1,v_1>,<t_2,v_2>,\cdots,<t_i,v_i>,\cdots,<t_n,v_n>)$ 来表示，其中 t_i 表示特征值，v_i 表示对应的权重，通过以下公式计算得出：

$$v_i = N(v_i)/\mathrm{NUM} \tag{1}$$

其中 $N(v_i)$ 表示在所有特征项中 v_i 出现的次数，NUM 表示的是所有特征值的总数。

2.2 兴趣模型树的表示

兴趣模型树表示成二层结构，上层父概念类是对下层所有子类的共同属性的概括，而下层子概念类则是从不同角度对上层父概念类加以细化，所有子节点之间形成平等的兄弟关系，这满足本系统能区分不同兴趣类别的要求。

考虑到以上因素，通过基于时间向量的二层树状结构来表示用户兴趣模型，第一层节点表示用户的兴趣主题，一个主题可以有很多主题特征项，第二层节点表示用户某个兴趣主题下的特征项。

用户兴趣模型的表示方法如图 2 所示。

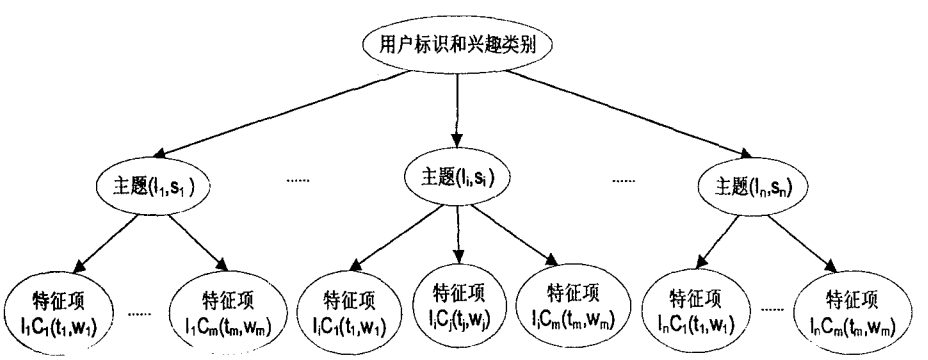

图 2 二层树状向量空间模型表示

用户标识是每个用户的唯一 ID，对每个用户有两棵树来表示用户兴趣模型，分别是短期兴趣树(STIT，Short - Term Interest Tree)和长期兴趣树(LTIT，Long - Term Interest Tree)，通过兴趣类别来区分。

该兴趣模型树共有两级非根节点：第一级节点代表用户的兴趣类别，用一组兴趣主题词 $(I_1,I_2,\cdots,I_n)$ 来代表用户的 n 个兴趣类别，每一兴趣类 I_i 对应的 S_i 表示该兴趣主题持续的时间长度，初始值为零，在兴趣模型更新过程中可通过以下公式计算得出：

$$S_i = \log(NT_i - OT) \tag{2}$$

$NT_i - OT$ 表示当前时间与特定兴趣主题加入时间相差的秒数，以此来区别长期兴趣主题和短期兴趣主题。

第二级节点，即叶子节点，它代表用户某一兴趣类别下的特征项，每个主题 $I_i(i=1,2,\cdots,n)$ 都有若干个特征项 $I_iC_j(j=1,2,\cdots,m)$，t_j 表示特征项的描述，w_j 表示对应特征项的权重，通过以下公式计算得出：

$$w_j = \log(\alpha C(o) + \beta C(b) + \gamma C(d)) \tag{3}$$

$C(o)$ 表示在用户基本信息中所出现的该特征项的次数，$C(b)$ 表示在用户所购买的商品中该特征项出现的次数，$C(d)$ 表示用户所浏览的商品中该特征项出现的次数，α,β,γ 分别表示不同的权值，$0\leqslant\alpha,\beta,\gamma<1$，且 $\alpha+\beta+\gamma=1$。

3 用户兴趣模型的建立和更新

3.1 相似度的计算

某商品是否符合用户的兴趣喜好由商品特征值向量与用户兴趣模型的主题向量来决定，实际上也就是计算两个向量之间的夹角，文中采用常用的计算 cosine[7]来实现，具体计算公式如下：

$$\mathrm{sim}(M,C) = \sum_{i=1}^{n} w_i * v_i \Big/ \sqrt{\left(\sum_{i=1}^{n} w_i^2\right)\left(\sum_{i=1}^{n} v_i^2\right)} \tag{4}$$

M 和 C 为特征值向量，n 为向量的维数，w_i 和 v_i 分别为两个向量的第 i 维的权重。

3.2 用户兴趣模型的建立算法

在建立初始用户兴趣模型时，主要考虑建立 STIT 和确定初始阈值，具体算法描述如下：

Step1　从用户注册信息中提取兴趣主题 I_i 和特征项 t_j,用公式(3)计算对应权值 w_j,得到兴趣主题向量组$(I_1, I_2, \cdots, I_n)$。

Step2　根据用户注册 ID 和步骤 1 得到的兴趣主题向量组$(I_1, I_2, \cdots, I_n)$建立 STIT。

Step3　从候选商品训练集中选出 N 种商品,表示成向量的形式,用公式(1)计算特征项的权值,得到商品向量组$(M_1, M_2, \cdots, M_N)$。

Step4　用公式(4)计算商品向量与用户兴趣主题向量的相似度,根据计算结果和需要产生的推荐商品数确定初始阈值。

Step5　算法结束。

初始状态下并不产生 LTIT,把所有用户的兴趣喜好都看做是短期兴趣,只有当兴趣主题持续一段时候后才会被当作长期兴趣主题,加入到 LTIT,并在 STIT 中删除该兴趣主题。商品特征向量与用户兴趣主题向量相似度大于初始阈值则确定是用户可能感兴趣的商品,推荐给用户,否则不予推荐。

3.3　用户兴趣模型的更新算法

随着用户活跃度的增加和兴趣喜好的变化,长期兴趣和短期兴趣逐渐显露,该模型有一个自适应的过程,主要包括 STIT、LTIT 和阈值的修改。

兴趣模型的自适应的过程主要是依据用户对推荐商品的判定,具体如表 1 所示[8]。

表 1　用户对过滤结果的判定

	感兴趣	不感兴趣
推荐	R^+	R^-
未推荐	N^+	N^-

在这里,N^+ 和 N^- 很难直接获取,主要考虑 R^+ 和 R^- 对兴趣模型树和阈值的影响,这主要来自于用户对过滤结果的显式反馈,更新算法描述如下:

Step1　用公式(4)计算待过滤商品特征值向量$(M_1, M_2, \cdots, M_k)$与兴趣主题特征值向量$(I_1, I_2, \cdots, I_n)$的相似度。

Step2　将相似度大于阈值 THR 的商品推荐给用户。

Step3　用户对推荐的商品做出显示反馈,得到 R^+ 和 R^-。

Step4　用公式(5)修改用户兴趣模型:

$$VSM(up') = \eta V(R^+) + \omega VSM(up) \qquad (5)$$

式中 $VSM(up')$ 表示修改后用户兴趣模型,$V(R^+)$ 表示正例向量,也就是满足用户兴趣的商品向量,$VSM(up)$ 表示原用户兴趣模型,η 和 ω 是表示权值的参数。

Step5　调整阈值:

(1) 若 $Count(R^+)/[Count(R^+) + Count(R^-)] \geqslant 60\%$ 且 $Count(R^+)$ 小于用户需求,则 THR = THR * 0.9,降低阈值 THR。

(2) 若 $Count(R^+)/[Count(R^+) + Count(R^-)] < 60\%$ 且 $Count(R^+)$ 大于用户需求,则 THR = THR * 1.1,提高阈值 THR,若修改后新阈值大于 1 则回退该修改,阈值 THR 不变。

(3) 其他情况阈值不变。

Step6　算法结束。

STIT 表示用户的短期兴趣主题,LTIT 表示用户的长期兴趣主题,当用户 STIT 中兴趣主题持续一段时间后会将它加入到 LTIT 中,并在 STIT 中删除。图 2 中的主题 I_i 被访问时,对应的 S_i 会依据公式(2)进行修改,当 $S_i > MT$,主题 I_i 被加入到 LTIT 中,并在 STIT 中删除,其中 MT 表示设定的持续时间长度。所以,LTIT 表示用户的长期兴趣主题,相对稳定,STIT 表示用户的短期兴趣主题,相对比较活跃,变化较快。

4　结束语

用户兴趣模型的建立和更新是个性化服务的基础和关键,直接关系到个性化推荐的效果。文中以目前发展势头迅猛的网络购物平台为背景,提出采用基于树状向量空间模型的表示和更新机制,根据用户主动提供的兴趣喜好建立兴趣模型,构建了用户兴趣模型更新的自适应机制,区分用户的短期兴趣和长期兴趣,能够准确表达和动态更新用户兴趣模型,对个性化服务在网络购物平台的实现具有较大意义。在接下来的工作中,主要通过实验确定兴趣模型建立和更新中的最优参数,优化该模型结构,同时研究如何隐式获取用户浏览行为和浏览模式,为建模提供更丰富数据。

参考文献:

[1]　雷　莹,冯玉强.基于软约束满足理论的用户偏好建模方法[J].华南理工大学学报:自然科学版,2008,36(4):115-121.

[2]　应晓敏,刘　明,窦文华.一种面向个性化服务的无需反例集的用户建模方法[J].国防科技大学学报,2002,24(3):67-71.

[3]　Ardissono L, Gena C, Torasso P, et al. Personalized recommendation of TV programs[C]// Proc of AI * IA 2003: Advances in Artificial Intelligence. Berlin: Springer,2003:474-486.

[4]　张东礼,汪东升,郑纬民.基于 VSM 的中文文本分类系统的设计与实现[J].清华大学学报:自然科学版,2003,43(9):1288-1291.

(下转第 85 页)

所示的模板图像在图 4 所示的待匹配图像中的最佳匹配。使用表 1 所述的算法进行匹配消耗了 250ms,而使用表 2 所述的改进算法进行此匹配过程只消耗了 150ms,与改进前的算法相比,减少了 40% 的匹配时间,提高了匹配效率。

通过此试验能够定量的看出改进算法与改进前的算法相比,在不降低匹配精度的前提下在执行效率上所具有的优势。

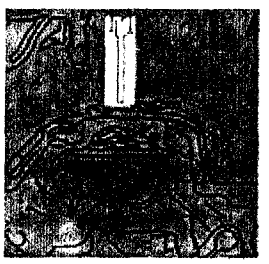

图 5　使用改进前算法进行模板匹配的结果图

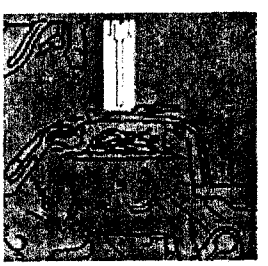

图 6　使用改进后算法进行模板匹配的结果图

4　结束语

Hausdorff 距离用于测量模板边缘图像点集合与待匹配图像点集合的相似程度,当两个集合中点的数目比较小时,基于 Hausdorff 距离的模板匹配算法具有良好的效果和匹配效率,但是当集合的数目比较庞大时,此算法的匹配效率会迅速降低。文中就该算法的此缺点提出了一种改进算法。此改进算法能够在上述情况下,在不降低匹配效果的前提下,提高匹配效率,并且对视频序列图像也具有很好的匹配效果和较高的匹配效率,对于目标图像出现遮挡情况时,此算法也具有较好的鲁棒性。

致谢　在此,笔者向对本文的工作给予极大的支持和帮助的同行表示真切的感谢和致敬!

参考文献:

[1] Clifton C, Zhang Zhijia, Huang Shabai, et al. A Fast Strategy for Image Matching Using Hausdorff Distance[C]// Proceedings of the 2003 IEEE International Conference on Robotics, Intelligent Systems and Signal Processing. Changsha, China:[s.n.], 2003:915-919.

[2] Huttenlocher D P, Klanderman G A, Rucklidge W J. Comparing Images Using the Hausdorff Distance[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1993,26(13):850-863.

[3] Zhang Jian, Zhao Baojun. A Quantitative Evaluating Method for Camouflage Effect Based on Wavelet Transform and Improved Hausdorff Distance[C]// The Seventh International Conference on Electronic Measurement and Instruments. Yantai, China:[s.n.], 2005:556-561.

[4] Wang Jing, He Peilun, Zhu Mengyu, et al. A Similarity Measure between the Target and Its Decoy Based on the Improved Hausdorff Distance[C]// The Seventh International Conference on Electronic Measurement and Instruments. Nanchang, China:[s.n.], 2005:204-210.

[5] Borgefors G. Distance Transforms in Digital Images[J]. IEEE Trans. Patt. Anal. Intell, 1986,8(14):334-343.

[6] Huttenlocher D P, Kedem K, Kleingerg J M. On Dynamic Woronoi Diagrams and the Minimum Hausdorff Distance for Point Sets under Euclidean Motion in the Plane[C]// Eighth ACM Symp. Barcelona, Spain:[s.n.], 2003:336-340.

[7] Alt H, Behrends B, Blomer J. Measuring the Resemblance of Polygonal Shapes[C]// Seventh ACM Symp. Comput. Geometry. New York:ACM Press,1991:204-208.

[8] Sim Dong-Gyu, Kwon Oh-Kyu, Park Rae-Hong. Pyramidal Robust Hausdorff Distance for Object Matching[C]// Proceedings of the International Conference on Image Processing. Piscataway, NJ, USA:[s.n.], 1999:88-92.

[9] 刘福新,杜世培,陈益强. 基于改进 Hausdorff 距离的人脸匹配方法[J]. 计算机工程与应用,2007,43(35):169-171.

[10] 孟　飞,王仕成,杨小冈,等. 基于 Hausdorff 距离和免疫遗传算法在图像匹配的应用研究[J]. 兵工自动化,2008,27(2):79-81.

[11] 蒋新士,吕　岳. 基于改进的加权 Hausdorff 距离的图像匹配[J]. 计算机应用研究,2007,24(4):182-184.

[12] 邱志敏,李　军,葛　军,等. 基于 Hausdorff 距离的自动目标识别算法的研究[J]. 红外技术,2004,28(4):199-202.

(上接第 81 页)

[5] Best J B. 认知心理学[M]. 北京:北京轻工业出版社,2000.

[6] 庞剑锋,卜东波,白　硕. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究,2001,9(9):23-26.

[7] Debole F, Sebastiani f. Supervised term weighting for automated text categorization[C]// Proc of the 2003 ACM symposium on Applied Computing. New York: ACM Press, 2003:784-788.

[8] 黄萱菁,夏迎炬,吴立德. 基于向量空间模型的文本过滤系统[J]. 软件学报, 2003,14(3):1538-1543.