自适应过滤算法在社区 E-learning 的个性化服务系统中的研究

罗 奇[1,2],谈宏华[1]

(1. 武汉工程大学 电气信息学院,湖北 武汉 430073;2. 武汉科技大学中南分校 信息工程学院,湖北 武汉 430223)

摘 要:针对学习型社区中的教育需求,在传统算法上加以改进,提出了一种基于向量空间模型的教育资源自适应过滤算法. 通过训练算法,提取特征向量和伪反馈建立初始模板,设置初始阈值. 然后通过过滤算法根据用户的反馈信息自适应地调整模板和阈值. 该算法在执行过程中,不需要大量的初始文本,同时在过滤的过程中可不断地进行自主学习来提高过滤精度. 该算法已在个性化知识服务系统中进行验证,结果表明是有效的.

关键词:自适应过滤;个性化知识服务; 相似度;终身化学习

中图分类号: TP302 **文献标识码**: A **文章编号**:1673-4785(2007)05-0091-04

Research on a personalized knowledge service system for community E-learning using an adaptive filtering algorithm

LUO Qi[1,2], TAN Hong-hua[1]

(1. School of Electrical Information, Wuhan Institute of Technology, Wuhan 430073, China; 2. Department of Information Engineering, Wuhan University of Science and Technology Zhongnan Branch, Wuhan 430223, China)

Abstract: To effectively provide personalized E-learning in a community, an adaptive filtering algorithm for identifying appropriate teaching resources was developed. It is based on a vector space model, an improvement on traditional algorithms used for this purpose. Firstly, feature selection and pseudo feedback were used to establish the initial templates and thresholds through a training algorithm. Then the user's feedback was utilized to modify the templates and thresholds adaptively for the filtering algorithm. The algorithm did not need massive quantities of initial texts to begin the process of filtering. Furthermore, filtering precision improved during the process through self learning. The algorithm proved effective as a personalized knowledge service system for community E-learning.

Keywords: adaptive filtering; personalized knowledge service; similarity; lifelong education

当今,为了实现社区居民终身学习、主动学习、全面学习的教育理念,国内外建立了很多的学习型社区. 随着拥有电脑和网络的社区居民数的增加,基于 E-learning 社区教育可突破时空的限制并降低学习成本、显著提高学习效率等优点日益受到人们的关注[1]. 因此,不少学习型社区也基于 E-learning 方式建设了社区网站,为社区成员提供一些信息,或提供一些学习课程. 但在实践应用中,这些网站却难以吸引社区居民的主动参与[2]. 经调查研究表明,这主要是由于个性化的知识服务体系还不完善,提供的信息准确度不高,有效性差,导致社区学习者兴趣度低,对 E-learning 社区教育信心不足或持怀疑态度. 基于 E-learning 的社区教育要想很好地吸引住社区居民,就要有个性化设计的思想,即为社区居民提供个性化的量身定做的知识和信息服务. 而个性化设计的关键在于如何根据用户的个性兴趣进行教学资源的过滤. 目前,国内外也有不少学者对过滤算法进行了大量的研究,例如传统的批过滤算法[3]. 他们的算法在过滤的过程中,需要大量初始训练文本,同时准确率和查全率也不高[4]. 基于此,文中在改进传统

收稿日期:2007-01-25.
基金项目:国家自然科学基金资助项目(60533080);"973"基金资助项目(2002CB312100);"863"基金资助项目(2006AA01Z303).
通讯作者:罗 奇. E-mail:ccnu_luo2008@yahoo.com.cn.

算法基础上,引入智能控制中的自适应反馈学习机制,提出了一种基于向量空间模型的教育资源自适应过滤算法.该算法在执行过程中,不需要大量的初始文本,同时在过滤的过程中可不断的进行自主学习来提高过滤精度.将该算法应用于基于社区 E-learning 的个性化知识服务系统中,能更好地支持社区教育的开展.

1 基于向量空间模型的教育资源自适应过滤算法

基于向量空间模型的教育资源过滤算法包括训练和过滤 2 个阶段.训练阶段的目的是根据给定的教育资源训练文本,生成初始的过滤模板,并决定初始的阈值.在自适应过滤阶段,对于教育资源中的每篇文本,系统判断它是否和过滤模板相关,再根据用户的反馈信息,自动调整过滤模板和阈值,以获得最佳的过滤性能.

1.1 训 练

图 1 说明了训练算法的流程图.首先,将主题转变为向量形式,同时从正例文本和伪正例文本中抽取特征向量.而初始的模板则是正例特征向量和伪正例特征向量的加权和.于是,就可以计算初始模块向量和全部的训练样本之间的相似度,从而为每个主题选择最优的初始相似度阈值.

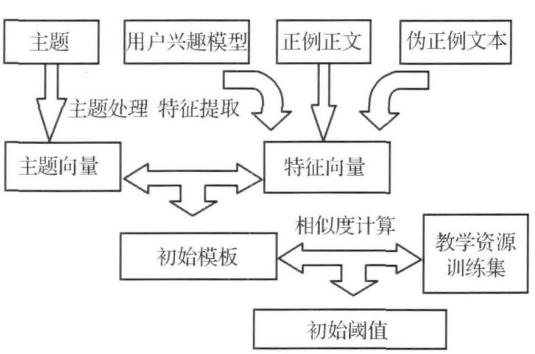


图 1 训练算法流程图

Fig.1 The flow chart of training algorithm

1.1.1 初始模板的建立

1) 对于每个主题,只能得到少的正例文本.因此需加入伪反馈的功能,从训练文本中挖掘出更多的相关文本来补充正例文本,和模板向量具有高度相似度而不是给定的正例文本的那些作为伪正例文本.

2) 获得正例文本和伪正例文本后,采用计算互信息量的方法计算每个词的权重[5].

$$\log \mathrm{MI}(w_i, t_j) = \frac{\log(P(w_i/t_j))}{P(w_i)}. \qquad (1)$$

式中:w_i 为文档中的第 i 个词,t_j 为第 j 个主题.$P(w_j/t_j)$ 和 $P(w_i)$ 采用最大似然法进行估计.

3) 在获得主题的正例和伪正例向量后,初始模板向量是正例特征向量、伪正例特征向量、用户兴趣向量和主题向量 4 个向量的加权和,权重分别为 α、β、γ、δ 即

$$pf_0(Q) = \alpha P_0(Q) + \beta P_1(Q) + \gamma P_2(Q) + \delta P_3(Q). \qquad (2)$$

式中:Q 表示主题,$Pf_0(Q)$ 是主题 Q 的初始模板向量,而 P_0、P_1、P_2、P_3 是它的 4 个分量.

1.1.2 初始模板的建立

教育资源训练集中,比该主题模板的相似度大的阈值的文档将作为该主题的相关文本而检出.计算初始模块向量和全部训练样本之间的相似度,可以为每个主题选择最优的初始相似度阈值.相似度采用余弦公式进行计算,如式 3 所示.

$$\mathrm{sim}(d, p_f) = \frac{\sum d_k pf_k}{\sqrt{\sum d_k^2 \sum pf_k^2}}. \qquad (3)$$

式中:p_f 表示初始模块向量,d 表示文本.d_k 是 d 中第 k 个词的权重[6].

1.2 过 滤

1.2.1 自适应过滤算法

初始的过滤模板建立,并且设置好初始阈值之后,过滤的过程就是自适应地修改过滤模板和阈值,使过滤性能不断提高,是一个机器学习的过程.图 2 是自适应过滤算法流程图.对于教学资源的每个文本,可计算它和某个主题的模板向量的相似度.若相似度大于阈值,就被认为是相关文本.然后由用户判断这篇文本是否真正与主题相关.根据不同的结果相应地修改模板向量或调整阈值.

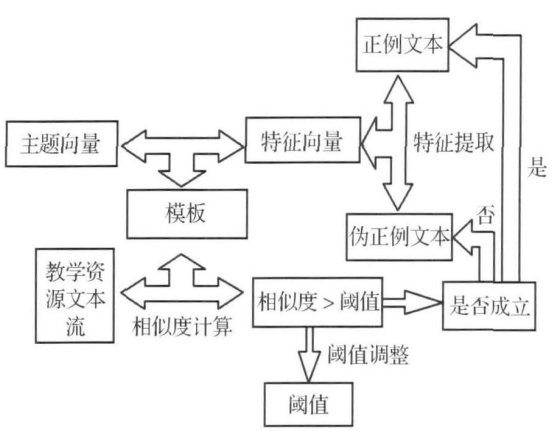


图 2 自适应算法流程图

Fig.2 The flow chart of training algorithm

1.2.2　阈值自适应调整

由于教学资源文本流中相关文本的比例是很低的,因此过滤出文档后就需要进行自适应调整阈值.提高阈值的目的是过滤出较少的正例文档,从而提高准确率;而降低阈值的目的是过滤出较多的正例文档. 文中提出采用概率分布密度的思想,如果近期过滤的正例分布超过期望的正例分布范围,则升高阈值. 否则,降低阈值,以增加正例文档数.

定义 1　设 n 为过滤文档在教育资源中的顺序编号, D 为期望的正例分布密度.

定义 2　$S(n)$ 为截至文档 n 时过滤的文档总数.

定义 3　$S_R(n)$ 为截至文档 n 时过滤得到的正例文档.

定义 4　$O(n)$ 为截至文档 n 时过滤阈值.

定义 5　$D_R(n_t, n_{t+1})$ 为上次阈值调整后得到的正例文档的概率分布密度.

$$D_R(n_t, n_{t+1}) = \frac{S_R(t+1) - S_R(t)}{S(t+1) - S(t)}. \qquad (4)$$

如下的阈值调整算法:

1) 若 $D_R(n_t, n_{t+1}) > \max(D, 0.2)$ 且 $S_R(n) < 0.2S(n)$,则 $O(n+1) = O(n) \times 1.2$,即如准确率过低,过滤出的文档数量又不太少,则迅速提高阈值.

2) 若 $D_R(n_t, n_{t+1}) > D$,则 $O(n+1) = O(n) \times 1.1$,即如过滤出的文档数多于必需的,则提高阈值.

3) 若 $D_R(n_t, n_{t+1}) < D$,则 $O(n+1) = O(n) \times 0.9$,如果过滤出的文档数少于必需的,则降低阈值.

如果检出的文本被用户判断为相关文本,将它加入到正例文本集合中,否则加入到伪正例文本集合中. 在调整模板向量时,从正例文本和伪正例文本中抽取出特征向量. 于是新的模板向量就是正例特征向量、伪正例特征向量、用户兴趣向量和主题向量 4 个向量的加权和,权重分别为 α、β、γ、δ 即

$$pf_0(Q) = \alpha P_0(Q) + \beta P_1(Q) + \gamma P_2(Q) + \delta P_3(Q). \qquad (5)$$

2　基于社区 E-learning 的个性化知识服务系统

该文应用个性化、数据挖掘、自适应过滤等技术提出了一种基于社区 E-learning 的个性化知识服务系统模型(ECPKSS)[7],如图 3 所示.

ECPKSS 模型的主要功能是在社区网络教学资源中学习和跟踪用户的个性化兴趣,并根据用户的个性化兴趣特征对教育资源进行过滤,帮助用户在海量的社区网络教学资源中快速而准确得到或者推荐用户感兴趣的教学资源.

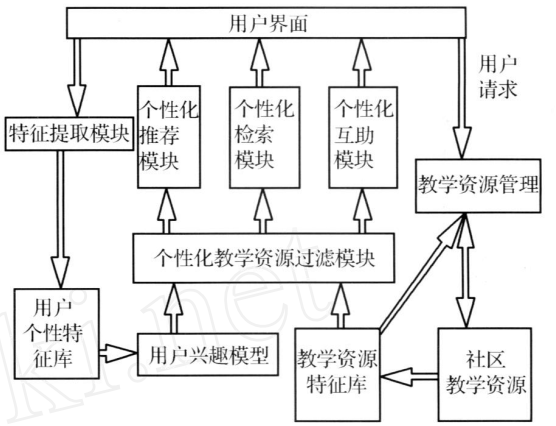


图 3　ECPKSS 模型

Fig. 3　ECPKSS model

特征提取模块实现根据用户对社区教育资源游览和相关反馈,来提取用户感兴趣的教学资源的特征信息,然后把这些特征信息保存在用户个性特征中并及时跟踪和更新.

用户个性特征库记录用户个性化信息. 它动态跟踪用户的兴趣,提取并记录关键词作为用户的个性特征,并作为用户模型构建模块提供用户特征.

用户兴趣模型构建模块从用户个性特征库提取关键词构成个体用户模型.

个性化社区教育资源过滤模块可以根据用户模型分别对社区教育资源进行过滤.

个性化推荐模块实现教育资源自动推荐和用户请求推荐 2 种个性化推荐功能. 用户请求推荐通过对教学资源库的管理模块和教学资源过滤模块调用来实现.

个性化检索模块是接受用户的检索请求,由过滤模块根据用户对社区教育资源的过滤形成个性化的检索结果.

个性化的互助模块使用户在学习过程中出现问题而得到及时、准确地指导与帮助.

该模型的工作过程是首先由特征提取模块提取用户感兴趣的社区教学资源的特征信息,并把这些信息保存在用户个性特征库中并及时跟踪和更新,其次由用户模型构建模块根据用户的个性化特征信息构成用户模型. 然后由个性化教学资源过滤模块根据用户模型实现对社区教育资源的过滤,最后由个性化检索、个性化推荐、个性化互助等模块根据过滤结果分别实现个性化检索服务、个性化推荐、个性化互助服务等.

3 实 验

在上述研究的基础上,结合与某社区的合作课题"个性化知识服务系统"的研究,为某社区建设了一个提供个性化知识和信息服务系统网站(供小区局域网接入,外网不可访问),为了得到实验的对比结果,文中在个性化教学资源过滤模块中分别采用传统批过滤算法和自适应过滤算法.实验数据来自社区教育中的法律知识,所有的文档为 XML 格式,分为训练集(83 650个文档)和测试集(723 420个有序文档).测试共使用 63 个不同的主题,实验结果如图所示,横轴是 63 个主题,按批过滤的准确率从大到小的顺序排列.纵轴则出了每个主题自适应过滤和批过滤的准确率数值.批过滤每个主题平均提供了 10 篇相关文本,而自适应过滤提供了 2 篇.此外,不进行自适应的情况下,每个主题只提供 2 篇相关文本.

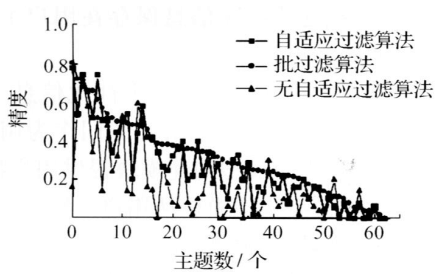


图 4 算法性能比较

Fig. 4 Performance comparison of batch algorithm and adaptive filtering algorithm

从图 4 中可以发现,与传统批过滤相比,自适应过滤的性能下降得并不是很大,2 条曲线非常接近.事实上,两者的平均数值分别是 31.7 %和 26.5 %,下降幅度仅为 16.4 %.相比之下,在不进行自适应的情况下,大多数主题的准确率均有很大幅度的下降,且平均准确率仅为 17.5 %,下降了 45.8 %.

4 结束语

综上所述,文中提出了一种基于向量空间模型的教育资源自适应过滤算法.将该算法应用于基于社区 E-learning 的个性化知识服务系统中,能更好地支持社区教育的开展.该算法已在实验中得到验证,结果表明是有效的.希望本文的工作能给相关人员有所参考.

参考文献:

[1]LUO Qi, XUE Qiang. Research on application of association rule mining algorithm in learning community [C]// Proceedings of CAAI-11. Wuhan, 2005.

[2]WU Yanwen, WU Zhonghong. Knowledge adaptive presentation strategy in E-learning [C]// Proceedings of Second International Conference on Knowledge Economy and Development of Science and Technology. Beijing, 2004.

[3]HU Tian, XIA Yingju, HUANG Xuanjing. A web-based Chinese information filtering system base on VSM [J]. Computer Engineering, 2003, 29(3):25-27.

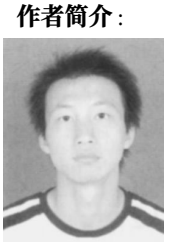
[4]LI Dun, CAO Yuanda. A new weighted text filtering method [C]// Proceedings of International Conference on Natural Language Processing and Knowledge Engineering. Wuhan, 2005.

[5]ROBERTSON S, HULL D A. The TREC-9 filtering track final report[C]// Proceedings of the 9th Text Retrieval Conference. Gaithersburg, Maryland, USA, 2001.

[6] HUANG X J, WU L D, ISHIZAKI H, et al. Language independent text categorization [J]. Journal of Chinese Information Processing, 2000, 14(2):1-7.

[7]LAWRENCE R D, ALMASI G S. Personalization of supermarket product recommendations [J]. Special Issue of the International Journal of Data Mining and Knowledge Discovery, 2001, 15(5):11-32.

作者简介:



罗 奇,男,1982 生,讲师,主要研究方向为情感计算、智能计算| 体育工程等,发表论文多篇.

谈宏华,男,1963 生,博士,教授,武汉工程大学电气信息学院院长,主要研究方向为机电一体化、智能控制等,曾主持项目 40 余项,发表文章 30 余篇.