第 40 卷 第 11 期 计 算 机 工 程 2014 年 11 月
Vol. 40 No. 11 Computer Engineering November 2014

· 开发研究与工程应用 · 文章编号: 1000-3428(2014)11-0318-04 文献标识码: A 中图分类号: F724.6

适应用户兴趣变化的社会化标签推荐算法研究

张艳梅,王 璐

(中央财经大学信息学院,北京 100081)

摘 要: 目前许多基于社会化标签的推荐均忽视用户的兴趣变化及反复性,影响了推荐质量。针对该问题,提出一种将指数遗忘权重和时间窗口相结合的算法,既突出了近期兴趣的重要性,又强调了反复出现的早期数据。建立基准标签集,根据指数偏移后的标签向量选出目标用户的最近邻居,通过目标用户时间窗内标记的资源计算其所有资源的推荐权重向量,结合推荐权重和资源相似度给出最近邻居标记资源的推荐分数,取分数最高的前 K 个资源做出推荐。仿真实验结果表明,改进后的算法能动态地跟踪、学习用户的兴趣变化,提高推荐精度。

关键词: 协同过滤;标签;兴趣变化;指数遗忘;时间窗;推荐

中文引用格式: 张艳梅,王 璐. 适应用户兴趣变化的社会化标签推荐算法研究[J]. 计算机工程,2014,40(11):318-321.

英文引用格式: Zhang Yanmei, Wang Lu. Research on Social Tagging Recommendation Algorithm Incorporated with User Interest Change[J]. Computer Engineering,2014,40(11):318-321.

Research on Social Tagging Recommendation Algorithm Incorporated with User Interest Change

ZHANG Yanmei, WANG Lu

(School of Information, Central University of Finance and Economics, Beijing 100081, China)

【**Abstract**】Many recommendation methods based on social tagging ignore the change and repeatability of user interests, which may lead to unsatisfactory results. In order to solve these problems, a new method which efficiently combines exponential forgetting-based data weight and time windows is proposed. The method not only highlights the importance of recent interest, but also stresses the recurring early data. Based on standard tag set of the target user, the nearest neighbour set can be gained according to exponential offset tag vectors, and then calculates weight vectors via items within time windows. Recommendation values of the nearest neighbour set are computed by weight vectors and similarity. Finally, it makes recommendation of items within the top K predicted values. Simulation experimental results show that the proposed algorithm for recommendation can dynamically track the changes in user's interest and has high quality of precision to some extent.

【**Key words**】collaborative filtering; tag; interest change; exponential forgetting; time window; recommendation

DOI: 10.3969/j.issn.1000-3428.2014.11.062

1 概述

近年来,互联网规模及 Web2.0 迅速发展,大量有用信息和垃圾数据充斥着网络,面对信息过载问题,信息检索领域迫切需要个性化推荐服务为用户提供有价值信息,完善用户体验,使用户由被动的信息浏览者转变为主动的参与者。当前主流的推荐方法包括协同过滤推荐[1]、基于内容推荐[2]、基于图结构的推荐[3]、混合推荐[4]等。其中,协同过滤技术是实际应用中使用最广泛的推荐技术[5]。联系用户和资源,标签就是其中一种,它既能表现用户的兴趣取向,同时还能体现资源的属性特征。文献[6-7]的研究显示标签可以准确显示用户对网页内容的判断,同时标签也是描述资源的一种好方法。目前,社会化标签系统越来越流行,比如 MovieLens,Flickr,del.icio.us,Citeulike,国内的豆瓣网站等。

标签反映了用户对物品的看法,是一种重要的用户行为,也是反映用户兴趣的数据来源。处理用户兴趣变化的方法主要有 2 种:时间窗口法和遗忘函数法[8],其中时间窗口法利用滑动时间窗反映用户近期兴趣及反复兴趣;遗忘函数法则利用遗忘函数对用户的兴趣权重进行衰减,从而得到真正的兴趣趋势。

基金项目:教育部人文社会科学研究基金资助项目(11YJC880163);北京市哲学社会科学规划基金资助项目(11JGC136)。
作者简介:张艳梅(1976-),女,副教授、博士,主研方向:电子商务,服务计算;王 璐,硕士研究生。
收稿日期:2013-09-24 **修回日期**:2013-12-27 **E-mail**:jlzym0309@sina.com

文献[9]首先使用线性遗忘函数并引入了渐进遗忘的思想。文献[10]用幂函数曲线跟踪模拟用户兴趣,将基于时间窗与基于资源相似度的数据权重结合起来反映用户的兴趣度,但没有考虑标签这一联系用户和资源的主要因素。文献[11]通过指数遗忘函数衰减用户兴趣模型,以捕获用户兴趣的变化,拟合记忆遗忘机制,但忽视了兴趣的反复性与稳定性,未能准确跟踪用户兴趣变化。针对现有的基于标签协同过滤算法不能快速、准确地反映用户兴趣变化趋势的问题[12],本文在社会化标签推荐的基础上,综合处理用户兴趣变化的 2 种方法,提出一种适应用户兴趣变化的标签推荐算法(User Interest Change-tagging Collaborative Filtering,UIC-TCF)。

2　UIC-TCF 算法描述

UIC-TCF 算法大致分为 2 个阶段:首先根据偏移后的用户标签矩阵选出相似度最高的 N 个邻居用户,然后将这 N 个用户标记过的资源与目标用户最近一段时间内标记的资源一一计算相似度,取相似度最高的 K 个资源推荐给目标用户,两阶段中相似度均通过基准标签集来计算。

2.1　最近邻居选取

给定目标用户 u_0 后,u_0 使用过的标签集合构成基准标签集 $T=(t_1,t_2,\cdots,t_n)$。使用过基准标签集中至少一个标签 t_i 的用户集合称为候选用户集。用户的标签权重向量中每个标签 t_i 的权重由 TF-IDF 加权方案得出。一般来说,用户近期标记资源对生成未来可能感兴趣的资源起相对重要的作用,而早期的标记记录对推荐影响相对较小。因此,为拟合人脑的遗忘规律,本文采用指数遗忘曲线以提高近期标记资源的重要性。

定义 1(用户标签向量)　设基准标签集 T 中有 n 个标签。用户 u_j 的标签权重向量:

$$\boldsymbol{ut}_j = (Tagw_{1,j},Tagw_{2,j},\cdots,Tagw_{n,j})$$

其中,$Tagw_{i,j}$ 为用户 u_j 在第 i 个标签上的权重;$Tagw_{i,j}=TF_{i,j}\times IDF_i$,$TF_{i,j}=\dfrac{num_{i,j}}{\sum_{i=1}^{n} num_{i,j}}$,$IDF_i=\log\dfrac{m}{n_i}$;$num_{i,j}$ 为用户 u_j 使用标签 t_i 标注资源的数量;$\sum_{i=1}^{n} num_{i,j}$ 是用户 u_j 使用标签集 T 中所有标签对所有资源标注的总次数;m 为用户数量;n_i 为使用过标签 t_i 的用户数量。

定义 2(指数偏移权重)　用户 u_j 在标签 t_i 上的偏移权重为:

$$OS(u_j,t_i) = e^{-\partial\times\frac{Date_{\text{last}}-Date_i}{Date_{\text{last}}-Date_{\text{begin}}}}$$

其中,∂ 为衰减系数,$0<\partial<1$;$Date_{\text{begin}}$ 为用户 u_j 开始使用标签系统的时间;$Date_{\text{last}}$ 为用户 u_j 最后一次使用标记系统的时间;$Date_i$ 为用户 u_j 最后一次使用标签 t_i 的时间。

定义 3(偏移后的用户标签向量)　偏移后用户 u_j 的标签向量为:

$$\begin{aligned}\boldsymbol{ut}'_j &= (Tagw'_{1,j},Tagw'_{2,j},\cdots,Tagw'_{n,j}) = \\ &(Tagw_{1,j},Tagw_{2,j},\cdots,Tagw_{n,j})\times \\ &(OS(u_j,t_1),OS(u_j,t_2),\cdots,OS(u_j,t_n))\end{aligned}$$

定义 4(用户相似度)　目标用户 u_0 与比对用户 u_j 之间的余弦相似度公式为:

$$sim(u_0,u_j)=\frac{\sum_{r=1}^{n}(Tagw_{r,u_0}\times Tagw_{r,u_j})}{\sqrt{\sum_{r=1}^{n}(Tagw_{r,u_0})^2}\times\sqrt{\sum_{r=1}^{n}(Tagw_{r,u_j})^2}}$$

2.2　资源推荐

使用指数衰减后,用户近期标记资源的权重总是大于早期标记记录的权重,从而突出了近期数据的重要性。但一般情况下,不同用户兴趣变化速度和趋势不同。此外,用户兴趣还会经常表现出反复性,所以用户早期的标记数据往往对于生成推荐也很重要,单纯使用指数函数衰减权重,削弱所有早期资源在推荐生成中的作用,可能会降低推荐效果。

设用户 u_0 标记过的资源集合为 I_u。设定一个时间窗 win,得到目标用户最近一段时间 win 内标记资源集合 I_{win},该集合内的资源一定程度上反映了用户近期的兴趣偏好。无论资源标记时间早晚,只要和 I_{win} 的相似度较高,说明该资源和用户当前的兴趣很相关,即有较高的推荐权重,依次算出 I_u 中每一资源对 u_0 的推荐权重。在上一步最近邻居用户选定后,获取邻居用户访问资源,构成候选推荐资源集合,结合推荐权重和资源相似度算出推荐分数,给出 TOP-K 推荐。

定义 5(资源的标签向量)　资源 $\boldsymbol{R}_x$ 的标签权重向量为:

$$\boldsymbol{R}_x = (RTagw_{1,x},RTagw_{2,x},\cdots,RTagw_{n,x})$$

其中,$RTagw_{n,x}=RTagTF_{n,x}\times RTagIDF_n$;$RTagTF_{n,x}=\dfrac{RTagNum_{n,x}}{\sum_{r=1}^{n}RTagNum_{n,x}}$;$RTagIDF_n=\log\dfrac{N_n}{n_n}$;$RTagNum_{n,x}$ 为资源 x 被标签 n 标记的次数(包括系统中所有用户);$\sum_{r=1}^{n}RTagNum_{n,x}$ 为资源 x 被所有标签集内标签标记的次数和;N_n 为用户 u_j 标记资源的数量;n_n 为标签 t_i 出现在多少个用户 u_j 标记过的资源中。

定义 6(资源相似度)　资源 q 与资源 k 间的余弦相似度为:

$$sim(q,k)=\frac{\sum_{r=1}^{n}(RTagw_{r,q}\times RTagw_{r,k})}{\sqrt{\sum_{r=1}^{n}(RTagw_{r,q})^2}\times\sqrt{\sum_{r=1}^{n}(RTagw_{r,k})^2}}$$

定义 7(资源的推荐权重) 对于用户 u_0 标记过的每一个资源 q, q 对 u_0 的推荐权重公式为:

$$wq(u_0,x)=sim(q,I_{win})=\frac{\sum_{k\in I_{win}}sim(q,k)}{size(I_{win})}$$

定义 8(资源推荐分数) 邻居用户标记过的资源 X 对目标用户 u_0 的推荐分数为:

$$rec(u_0,x)=\sum_{i\in I_{win}}wq(u_0,i)\times sim(i,x)$$

2.3 推荐算法

推荐算法描述如下:

输入 用户标记项的矩阵;m:候选用户数;n:社区用户数;win:时间窗;u_0:用户

输出 推荐给 u_0 的分数最高的 K 个资源列表

Step1 标准标签集初始化 $T=(t_1,t_2,\cdots,t_n)$;

Step2 为 u_0 创建标签矢量 $\mathbf{ut}_0$,乘以指数抵消标签矢量,then 得到 $\mathbf{ut}'_0$;

Step3 For j = 1 to m − 1
 为 u_j 创建标签矢量 $\mathbf{ut}_j$;
 乘以指数抵消标签矢量,得到 $\mathbf{ut}'_j$;
 计算 $sim(u_0,u_j)$;
 End For

Step4 对 $sim(u_0,u_j)$降序排列,then 得到分数最高的 N 个邻居;

Step5 $I_{win}=GetLatestItemSet(u_0,win)$;

Step6 For q = 1 to Length($Item_{u_0}$)
//$Item_{u_0}$代表 u_0 标签的物品
 计算推荐量 $wq(u_0,q)$;
 End For

Step7 GetTaggedItemSetOf($u_0,u_1,\cdots,u_N$)
 For j = 1 to N //所有分数最高的 N 个邻居
 For x = 1 to Length($Item_{uj}$)
//$Item_{uj}$代表 u_j 标签的物品
 $rec(u_0,x)=\sum_{i\in I_{win}}wq(u_0,i)\times sim(i,x)$
 End For
 End For

Step8 降序排列 $rec(u_0,x)$,产生推荐给 u_0 的分数最好的 K 个资源列表。

3 实验结果与分析

3.1 实验环境

本文用 Citeulike 提供的公开数据集作为实验数据对 UIC-TCF 算法与传统的基于用户的协同过滤算法(记为 CF)作比较。Citeulike 是一个专门为学术研究人员提供分享、存储、组织学术文章的社会化书签网站,每条记录包括 <用户,论文,标签,时间>4 个字段。对获取的数据进行预处理,删除使用用户数量小于 20 的标签以及使用标签种类小于 20 的用户,同时由于用户的兴趣在较短的时间内是相对稳定的,兴趣变化只能在一段时间内体现,因此删除那些只在短期内($Date_{last}-Date_{begin}<30$ 天)使用过标签系统的用户数据。最终从以上数据集中选取 847 个用户对 3 597 个项目的评分记录,以及相应的 133 个频繁标签。其中,把每个用户最近 30% 的数据作为测试集,剩下的 70% 作为训练集。实验运行环境为:Windows XP 操作系统,CPU 主频 2.26 GHz,2 GB 内存,Eclipse 开发平台,SQL Server 2000 数据库。

3.2 评价指标

准确率(*Precision*)是广泛用于信息检索领域的评价标准之一。准确率表示用户对一个被推荐资源感兴趣的可能性。其定义为系统的推荐列表中用户喜欢的产品占所有推荐产品的比率:$Precision=\frac{Hits}{N}$。其中,其中,$Hits$ 表示正确推荐数;N 表示算法生成的推荐总数。

3.3 实验结果

为了检验 UIC-TCF 算法效果,将 UIC-TCF 算法与传统基于用户的协同过滤算法作比较,对比在推荐不同数目时,2 种方法准确率的变化情况。设置了 3 组实验,实验过程中推荐个数由 5 逐步增长到 20,间隔为 5。以下所有实验中,最近邻居数目均取为 10。

图 1 显示了衰减系数∂不同取值时 UIC-TCF 算法与传统协同过滤算法推荐准确率的对比。其中,时间窗 $win=10$,∂分别为 0.3,0.5,0.7。

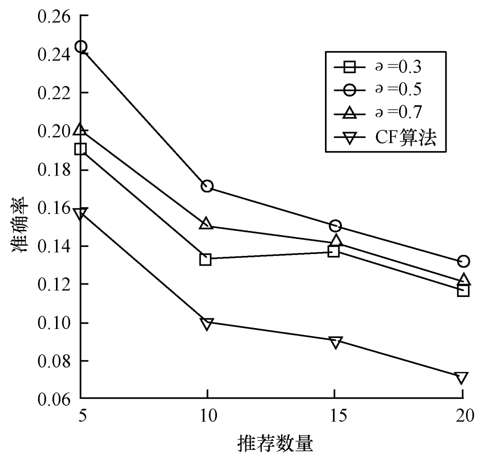

图 1 不同∂取值下算法准确率的对比

通过图 1 可以看出,本文算法明显优于传统协同过滤算法。在准确率指标下,推荐数目越少,本文算法的优势越明显。同时,衰减系数∂的取值对实验结果也有较大影响。由于存在不同种类的标签、资源,同时用户兴趣的变化速度和变化规律也不同,权重增长过快或过慢都会降低推荐精度。在∂ = 0.5 时,算法性能达到最优。如在推荐数目为 5,∂ = 0.5 时,推荐准确度提升了约 54.8% 。同时,随着推荐数量的增加,UIC-TCF 算法的性能并没有迅速下降,而是与传统协同过滤算法保持稳定的推荐精度差值。

图 2 显示了时间窗 win 的取值对推荐效果的影响。实验依次取 $win=5$ 天,10 天,15 天,其中,∂ = 0.5。