一种基于内容过滤的科技文献推荐算法

王嫣然,陈 梅,王翰虎,张 鑫
(贵州大学 计算机科学与信息学院,贵州 贵阳 550025)

摘 要:个性化推荐技术能够帮助用户快速方便地从大量的电子文献中获得感兴趣的文献,但传统的基于内容过滤的推荐算法不能反映用户对文献需求的兴趣变化,难以区分文献质量的高低。针对上述问题,提出了基于用户访问时间的数据权重和文献重要度,以便更好反应用户的兴趣以及文献质量的优劣,并将这两种度量引入到基于内容过滤的科技文献推荐算法中。实验和分析表明,改进后的算法比传统的内容过滤推荐算法在对文献推荐的准确度上有所提高。

关键词:科技文献;内容过滤;个性化推荐;时间权重;文献重要度
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2011)02-0066-04

A Content-Based Filtering Algorithm for Scientific Literature Recommendation

WANG Yan-ran, CHEN Mei, WANG Han-hu, ZHANG Xin
(College of Computer Science and Information, Guizhou University, Guiyang 550025, China)

Abstract: Personalized recommendation techniques can help users to retrieval documents quickly and effectively from mass of electronic documents. However, existing content-based filtering algorithm don't reflect that the users information demands change over time, it is difficult to distinguish the literature quality difference. According to these questions, time-based data weight and literature importance are proposed, in order to response user's interest as well as literature quality fit and unfit quality better. On this basis, a content-based filtering of scientific literature recommendation algorithm is presented in this paper. Experimental evaluations and theoretical analysis demonstrate that the proposed algorithm can improve the accuracy of recommendations.

Key words: *scientific literature; content-based filtering; personalized recommendation; time weight; literature importance*

0 引 言

随着互联网技术飞速发展,电子文献数量越来越庞大,如何帮助用户尤其是科研工作者从海量的电子科技文献中快速有效地找到所需的相关文献就成为急需解决的问题。因此,需要研究不同用户的兴趣,主动为用户推荐最需要的资源,克服用户信息获取的困难。

近年来,应用到各种个性化推荐系统的推荐算法主要有基于内容过滤和协同过滤[1]。协同过滤推荐算法的基本思想是根据用户兴趣的相似性来推荐资源,用户要找到他感兴趣的文献首先要找到与他兴趣相似的用户,然后把这些最相似的用户所感兴趣的内容推荐给该用户[2],如 SiteSeer[3],GroupLens[4]。基于内容的推荐算法的基本思想是根据资源与用户感兴趣信息的相似性推荐内容,如 Personal Web Watcher[5],CiteSeer[6]等,其关键问题是相似性计算。

由于科技文献资源可以获得其完整的内容描述,基于内容过滤更能从内容本质上推荐给用户真正感兴趣的文献,所以采用基于内容过滤的算法来推荐科技文献。另外,针对传统的基于内容过滤算法存在的弊端:不能及时反映用户兴趣的动态变化和难以区分资源内容的品质和风格。鉴于此,我们对传统的算法进行了改进,从而提高了推荐的精度。

1 相关工作

基于内容过滤的基本问题包括用户兴趣的建模与更新以及相似性计算方法。本节首先给出获取用户兴趣信息的方法,建立用户兴趣特征描述的模型;然后使用向量空间模型(Vector Space Model,即 VSM)来表示用户兴趣模型和文献资源模型;最后给出相似度计算方法。

收稿日期:2010-06-23;**修回日期**:2010-09-29
基金项目:贵州省2008年省级信息化专项基金项目(0830);贵州省科技计划工业攻关基金项目(黔科合 GY 字[2008]3035)
作者简介:王嫣然(1986-),女,山东济宁人,硕士生,CCF 会员,研究方向为数据库技术与软件工程;陈 梅,副教授,硕士生导师,研究方向为数据库技术与软件工程;王翰虎,教授,硕士生导师,研究方向为数据库系统、分布式系统、面向对象方法。

1.1 用户兴趣

文中根据用户感兴趣的文档选取合适的主题词来表达用户兴趣。对于用户的兴趣,采用隐式方式确定,即收集用户的访问模式,用户访问日志中记录了用户的操作类型(如浏览或保存)、操作发出的时间、浏览操作持续的时间长度等[7]。给出用户兴趣度的以下定义:

设 u 是一个用户,用户 u 对文档 i 的偏好程度 P_{ui} 定义为

$$P_{ui}=\begin{cases}1, D_{ui}=1\\ 0, D_{ui}=0 \wedge t_{ui}<\delta_1\\ \dfrac{\dfrac{t_{ui}}{D_{\mathrm{Len}(i)}}}{1+\dfrac{t_{ui}}{D_{\mathrm{Len}}(i)}}, D_{ui}=0 \wedge \delta_1<t_{ui}<\delta_2\end{cases} \tag{1}$$

其中,$D_{ui}=1$ 表示用户 u 保存、打印或收藏了文档 i,$D_{ui}=0$ 表示用户 u 没有保存、打印或收藏文档 i,只是浏览了文档 i。t_{ui} 表示用户 u 浏览文档 i 的时间长度,δ_1、δ_2 为浏览时间长度的阈值,当用户阅读文档的时间超过 δ_1,但是小于 δ_2 时,认为用户对此文档感兴趣;如果时间超过 δ_2,则认为是一种无效操作,如用户打开页面后,出去接电话等;如果时间小于 δ_1 则认为用户只是从此页面经过。$D_{\mathrm{Len}}(i)$ 表示文档 i 的长度。

1.2 用户兴趣模型及文献资源模型的表达

使用向量空间模型(Vector Space Model,即 VSM)来表示用户兴趣模型和文献资源模型[8]。向量空间模型用特征项及其相应权值代表文档信息。用户兴趣特征项主要以特征文档中提取出的关键词来表示。关键词的提取非常重要,关系到能否准确代表文献内容。

关键词的权重可以用 TFIDF[8] 方法来表示,TFIDF 法是以关键词在文档 d 中出现的次数与包含该关键词的文档数之比作为该词的权重,即

$$w_i=\frac{TF_i(t,d)\log(N/DF(t)+0.01)}{\sqrt{\sum_k TF_i^2(t,d)\log^2(N/DF(t)+0.01)}} \tag{2}$$

其中,w_i 表示第 i 个特征词的权重,$TF(t,d)$ 表示词 t 在文档 d 中的出现频率,N 表示总的文档数,$DF(t)$ 表示包含 t 的文档数。用 TFIDF 算法来计算特征词的权重值是表示当一个词在这篇文档中出现的频率越高,同时在其他文档中出现的次数越少,则表明该词对于表示这篇文档的区分能力越强,所以其权重值就应该越大。

每个用户的兴趣表示成一个向量:$V_u=(\langle t_1,w_1\rangle,\langle t_2,w_2\rangle,\cdots,\langle t_n,w_n\rangle)$,$V_u$ 记录了能反映用户兴趣的关键词 t 以及关键词的权重 w。

每篇文献表示成一个向量:$V_d=(\langle t_1,f_1\rangle,\langle t_2,f_2\rangle,\cdots,\langle t_n,f_n\rangle)$,$V_d$ 记录了文献中每个关键词 t 和关键词出现的频率 f。

1.3 相似度计算方法

相似度是衡量用户兴趣特征与文献内容特征相似程度的变量。对于向量空间模型来说,相似度的计算传统方法是计算向量间的余弦相似度。文中即采用这种方法并借鉴了文献[9]给出的形式。

文献资源 i 的 VSM 模型 V_d 的关键词向量为 $d_i=(t_1,t_2,\cdots,t_m)$,其中 t 为关键词,总共 m 个。用户兴趣的 VSM 模型 V_u 的关键词向量为 $u=(t_1,t_2,\cdots,t_n)$,其中 t 为关键词,总共 n 个。将两个特征词向量合并,得到本次相似度计算的合并向量 D,$D=d_i\cup u=(t_1,t_2,\cdots,t_k)$,其中 t 为合并后的关键词,总共 k 个。因为 d_i 和 u 中可能存在重复的关键词,所以合并后的关键词向量中关键词个数 k,满足 $\max(m,n)\leqslant k\leqslant m+n$。

根据 D 可以得到本次相似度计算时文献的相似度计算向量为 $D_i=(f_1,f_2,\cdots f_k)$,D_i 中的 f 表示 D 中对应的关键词在文献中的频率,如果这个词不存在,则取 0。同样得到用户的相似度计算向量为 $D_u=(w_1,w_2,\cdots,w_k)$,D_u 中的 w 表示 D 中对应的关键词在用户模型中的权重,如果这个词不存在,则取 0。

采用向量的夹角余弦值来计算 D_u 与 D_i 之间的相似度:

$$\mathrm{Sim}(D_u,D_i)=\left(\sum_{j=1}^{k}w_j\times f_j\right)\Big/\sqrt{\left(\sum_{j=1}^{k}w_j^2\right)\left(\sum_{j=1}^{k}f_j^2\right)}$$

2 基于内容过滤的科技文献推荐算法

2.1 现有算法的不足

虽然使用基于内容的过滤方法,可以依据用户过去的偏好,推荐出符合用户兴趣的项目,但是现有的基于内容过滤的推荐算法用于科技文献推荐仍存在以下不足之处:一是没有考虑用户兴趣的动态变化,将用户访问过的每个资源同等对待;二是只考虑资源间的相似性,没有考虑资源质量的高低,推荐的结果难以区分资源内容的品质和风格。为解决上述两个问题给出了基于时间的权重函数和文献重要度的定义,并将其引入到基于内容过滤的推荐算法中。

2.2 基于时间的权重函数

传统的内容过滤方法不能将用户的兴趣变化表现出来,对于科技用户来说,其研究领域大致不变,研究的问题是随着时间的推移不断变化的,较短的一段时间内用户的兴趣相对固定。用户近期关注的文献对推荐该用户未来可能感兴趣的文献起比较重要的作用,而早期的访问记录对生成推荐影响相对较小,一个用户感兴趣的资源最可能和他近期访问过的资源相似。因此,文中引入基于用户访问时间的权重函数到兴趣

预测中,以提高最近访问数据在推荐生成过程中的重要性,充分考虑到"时间效应"的影响,得到更准确的推荐。

设 x_{ui} 表示用户 u 访问资源 i 的时间与用户 u 最近一次访问某资源的时间间隔,间隔越小,说明资源 i 越近时间被访问。定义基于时间的权重函数 $f(x_{ui})$ 表示资源 i 对用户 u 的权重,它是一个和 x_{ui} 相关的函数值,权重函数应该设计成关于 x_{ui} 的递减函数。文中将基于时间的权重函数作如下定义:

$$f(x_{ui}) = (1 - \alpha) + \alpha e^{-x_{ui}} \tag{3}$$

其中,$\alpha \in (0,1)$ 称为权重增长指数,改变 α 的值可以调整权重随访问时间的变化的速度。α 越大权重增长速度越快,α 的大小可以影响到算法的性能。

2.3 文献重要度

基于内容过滤的推荐系统在碰到相同主题的内容时,很难区分质量的高低[10]。例如在对论文的推荐中,同一研究方向的论文往往内容相近而水平相差很大,因此应具有不同的推荐度,但是基于内容的推荐系统不能识别其质量差异,这也从一定程度上影响了其推荐质量[11]。决定一篇文献质量高低的因素有很多,比如文献所发表的期刊或会议的级别、发表时间、被引次数、作者的权威性等。因此,可以在推荐过程中把文献的价值考虑进来,定义一篇文献 i 的重要度 $R(i)$ 的计算公式如下:

$$R(i) = IF(i) * TW(i) \tag{4}$$

其中 $IF(i)$ 表示文献 i 所在的期刊影响因子,$TW(i)$ 表示时间因子,$TW(i) = \dfrac{T(i) - T_e + 1}{T_r(i) - T_e + 1}$,这里,$T(i)$ 表示文献 i 发表的年份,T_e 表示最早发表文献的年份,$T_r(i)$ 表示用户检索文献 i 的时间,可以看出,较晚发表论文的时间因子较大。

2.4 改进的算法描述

把上述指标引入到传统的内容过滤算法中,提出一种改进的基于内容过滤的推荐算法。首先根据式(1)得出用户兴趣度,然后将用户兴趣度 P_{ui} 与基于时间的权重函数 $f(x_{ui})$ 结合起来,计算 $P_{ui} * f(x_{ui})$ 得出值最大的 N 篇文档作为能够反应与目前用户感兴趣的资源最接近的文档,即为特征文档。进而按照 1.2 节的说明使用向量空间模型来表示用户兴趣模型和文献资源模型。最后根据 1.3 节给出的方法计算用户兴趣特征与文献内容特征相似度,然后与文献的重要度结合,得出推荐度 $\mathrm{rec}(u,i) = (1 - d) * R(i) + d * \mathrm{Sim}(D_u, D_i)$,即为文献资源 i 推荐给用户 u 的推荐度,这里 d 是一个比例系数,它决定了一篇文献固有的重要度和与用户兴趣特征的相似度各自所占的比重,我们取 0.85。接着计算每篇文献对用户的推荐度得到有序排列,把推荐度最大的 N 篇文献作为用户的 Top - N 推荐集。算法描述如下:

算法 1. 基于时间加权的内容过滤推荐算法

输入:用户 u、与之对应的访问日志、M 篇文献资源的 VSM 模型 $V_{di}(i = 1,2,\cdots,M)$

输出:用户 u 的 top - N 推荐集

过程:

Step1. 根据式(1)计算用户的兴趣度 P_{ui},根据式(3)计算时间权重 $f(x_{ui})$;

Step2. 计算 $P_{ui} * f(x_{ui})$ 取值最大的 N 篇作为特征文档,并给出用户的 VSM 模型;

Step3. 从用户的 VSM 模型 v_u 获得它的关键词向量 u:

$u = (t_1, t_2, \cdots, t_m)$;

Step4. 从文献 VSM 模型 v_{di} 获得它的关键词向量 d_i:

$d_i = (t_1, t_2, \cdots, t_n)$;

Step5. 合并两个关键词向量 u 和 d_i,获得本次相似度计算的合并向量 D:

$D = (t_1, t_2, \cdots, t_k)$, $\max(m,n) \leqslant k \leqslant m + n$;

Step6. 依据 D,计算用户和文献资源在这次相似度计算中各自的向量值 D_u 和 D_i:

$D_u = (w_1, w_2, \cdots, w_k)$ $D_i = (f_1, f_2, \cdots, f_k)$;

Step7. 计算文献资源 i 推荐给用户的推荐度:

$\mathrm{rec}(u,i) = (1 - d) * R(i) + d * \mathrm{Sim}(D_u, D_i)$

Step8. 重复 Step4,计算每一篇文献对与该用户的推荐度;

Step9. 将所有文献资源按推荐度大小排列,其中推荐度最大的 N 个资源将作为用户的推荐集。

3 实验结果及分析

实验是基于我们自行开发的科技文献异构数据库共享检索平台进行,实验使用的数据来自网络爬虫从万方数据资源系统上下载的部分计算机类科技文献,共 5825 篇。实验针对 20 个用户进行,收集用户的 web 访问日志并进行预处理。

用信息检索领域中评估系统效果的准确率(Precision)标准作为对比传统内容过滤推荐算法和改进后的算法推荐精度的标准:

$$\text{准确率} = \frac{\text{正确推荐数}}{\text{生成的推荐总数}} \quad [12]$$

根据实验需要,把用户访问日志分成两部分,用户最近一段时间的访问记录隐藏起来作为测试数据(占 20%),用来判断推荐的准确性,其余的访问记录作为训练数据(占 80%),用来产生推荐结果。实验过程中根据用户在训练集中的访问记录为其计算推荐集,如

果推荐集中的某个资源 i 出现在该用户测试集中的访问记录里,则认为生成了一个正确的推荐。

在实验时,权重增长指数 α 及推荐文本个数的不同都会对推荐的准确率产生影响,这里权重增长指数 α 分别取0.3,0.5,0.7,推荐度中的比例系数 $d=0.85$。我们观察推荐文本的数目从 5 到 30 每次增加 5 不同情况下推荐算法的准确率。图 1 给出了传统内容过滤推荐算法与本文改进算法的准确率对比图,实验中取 20 个用户数据的平均值。

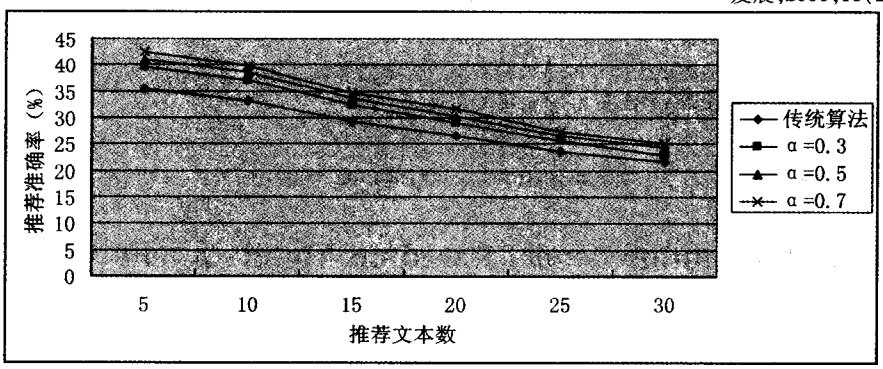

图 1　传统算法和改进算法准确率对比图

从图 1 可以看到改进后的算法推荐精度有所提高,尤其当推荐文献的数目比较少时,准确率提高更加明显一些。推荐的文献数量越少,准确率越高些。权重增长指数 α 的设置对算法准确率也有影响,对于不同的用户,用户兴趣的变化速度和变化规律是不同的。由此可见,引入基于时间的数据权重函数和文献质量度,能够及时捕捉到用户最近的兴趣,有效地突出了文献的质量高低对生成推荐的重要性,从而使算法生成的推荐文献更好地满足用户的当前需要。

4　结束语

提出了一种基于内容过滤的科技文献推荐算法,考虑到科研用户对文献需求随着时间的推移有所改变以及具有相同主题的文献质量存在差异的特点,在传统基于内容过滤的算法上进行了两方面的改进,提出了基于用户访问时间的权重函数和文献质量度的概念,将其引入到推荐算法中。实验结果表明,新算法提高了推荐的准确性。我们可以利用该方法,根据不同用户的兴趣,推荐符合其兴趣的科技文献,从而提高科研人员检索文献的效率。

下一步的工作包括两方面:不同用户的兴趣变化规律是不同的,如何针对不同的用户设置不同的参数已达到最优的推荐效果还有待于进一步研究;在评价文献质量时考虑更多的因素,如引用情况等。

参考文献:

[1] 曾　春,邢春晓,周立柱. 个性化服务技术综述[J]. 软件学报,2002,13(10):1952-1961.

[2] 吴辉娟,袁　方. 个性化服务技术研究[J]. 计算机技术与发展,2006,16(2):32-34.

[3] Rucker J, Polanco M J. Siteseer: personalized navigation for the web[J]. Communications of the ACM, 1997,40(3):73-75.

[4] Konstan J, Miller B, Maltz D, et al. GroupLens: applying collaborative filtering to usenet news[J]. Communications of the ACM, 1997,40(3):77-87.

[5] Mladenic D. Machine learning for better Web browsing[C]//In: Rogers S, Iba W. AAAI 2000 Spring Symposium Technical Reports on Adaptive User Interfaces. Menlo Park, CA: AAAI Press,2000:82-84.

[6] Bollacker K D, Lawrence S, Giles C L. Discovering relevant scientific literature on the Web[J]. IEEE Intelligent Systems, 2000,15(2):42-47.

[7] 曹　毅,贺卫红. 基于内容过滤的电子商务推荐系统研究[J]. 计算机技术与发展,2009,19(6):182-185.

[8] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Information Retrieval and Language Processing, 1975,18(11):613-620.

[9] 李　永,徐德志,张　勇,等. VRE 中基于内容过滤的论文推荐算法[J]. 计算机应用研究,2007,24(9):58-60.

[10] 赵　亮,胡乃静,张守志. 个性化推荐算法设计[J]. 计算机研究与发展,2002,39(8):986-991.

[11] Matthew R, Mclaughl N, Jonathan L. Content-based filtering & collaborative filtering: A collaborative filtering algorithm and evaluation metric that accurately model the user experience[C]//Proceeding of the 27th ACM SIGIR Conf. Sheffield: ACM Press, 2004:329-336.

[12] 秦春秀, 赵捧未, 窦永香. 基于用户兴趣的个性化检索[J]. 情报学报,2005,24(4):449-452.