数字图书馆个性化资源推荐系统算法研究

张克柱
(宿州职业技术学院 计算机信息系,安徽 宿州 234000)

【摘 要】针对当前数字图书馆存在的问题,提出了基于用户聚类技术的数字图书馆个性化资源推荐算法,使得图书馆资源利用率得到了较大的提高,并为管理者提供了决策支持。

【关键词】个性化资源推荐;数字图书馆;聚类
【中图分类号】TP274 【文献标识码】A 【文章编号】1673-1891(2013)01-0065-03

引言

随着计算机网络技术、通讯技术的快速发展和应用,人们的学习、生活方式等发生了翻天覆地的变化,人类获取知识的途径也变得越来越多、越来越方便,网络资源正变得越来越丰富,以致很多人习惯于到网上查找资源,不愿意到图书馆查找资源,这就使得传统图书馆资源必须数字化、网络化。图书馆本身就是为广大读者服务的,为读者提供学习资源的宝库,但目前有不少图书馆的资源利用率并不是很高,分析其主要原因:一方面图书馆资源过多,读者很难找到自己喜欢的图书;另一方面图书馆新进的图书资源,读者不能在第一时间知道,以致于很多读者错过了最佳阅读时间,所以图书馆数字化、个性化推荐是今后图书馆发展的趋势。

本文在协同过滤算法的基础上,提出了基于聚类技术的数字图书馆个性化资源推荐算法。

1 协同过滤技术

1.1 协同过滤的方法

采用协同过滤技术的方法很多,主要有以下几种方法:

(1)基于用户的协同过滤

是指根据某些用户的访问记录的评分计算结果,对另外一些与此用户相关属性比较接近的目标用户进行个性化推荐的一种算法,该算法在电子商务等领域得到了广泛的应用。

(2)基于项目的协同过滤

是指通过对大量用户评价的一些项目的计算,得出同类用户目标项目的推荐。

(3)基于模型的协同过滤

是指首先在用户原来提供或产生的资料基础上,构建出评价等级模型,再采用相关技术把用户的相关信息项目与此模型进行比较,从而预测用户的兴趣爱好级别。

1.2 协同过滤中的相似度计算

在个性化资源推荐算法中,最近邻协同过滤是当前用的最多的算法之一,该算法的基本原理是:为了给某个特定用户进行个性化推荐,首先根据该用户所提供的相关信息,计算出与该用户相似度较高的最近邻居,再对最近邻居的活动记录等数据进行分析,预测出该用户可能对资源比较感兴趣,从而实现对该用户进行个性化推荐。

从该算法的原理可以看出,该算法的核心技术就是用户相似度计算,相似度计算的准确性直接决定个性化推荐的质量高低。常见的相似性计算方法有以下几种。

(1)余弦相似性

把每个用户的评分当作一个向量(n维的),若用户对某个项目没有评分,那么该用户此项目的向量分量值记为0,两个向量夹角之间的余弦值称之为两个用户之间的相似性。假设有两个用户A和B,分别用α和β表示两个用户评分向量,则用户A与B之间的相似度计算可用以下公式表示。

$$sim(\alpha,\beta)=\cos(\alpha,\beta)\frac{\alpha\cdot\beta}{|\alpha|\times|\beta|}$$

从公式可以看出,两个向量的夹角越小越好。

(2)相关相似性

如果两个用户A和B对某些项目(该项目集记为$I_{\alpha\beta}$)都评过分,那么这两个用户的相似性可通过以下公式计算。

$$sim(\alpha,\beta)=\frac{\sum_{\gamma\in I_{\alpha\beta}}R_{\alpha,\gamma}\cdot R_{\beta,\gamma}}{\sqrt{\sum_{\lambda\in I_{\alpha\beta}}R_{\alpha,\gamma}^{2}}\sqrt{\sum_{\gamma\in I_{\alpha\beta}}R_{\beta,\gamma}^{2}}}$$

其中$R_{\alpha,\beta}$表示A对γ的评分,$R_{\alpha,\gamma}$与$R_{\beta,\gamma}$分别表示A与B对γ的平均评分。

(3)修正的余弦相似性

由于每个用户对项目的评分标准可能不一样,造成利用余弦相似性公式计算出来的结果可能有

收稿日期:2012-11-19
作者简介:张克柱(1979-),男,安徽庐江人,讲师,硕士,研究方向:计算机网络技术、数据挖掘。

一定的误差，通过此方法可以避免此误差。假设用户A与B共同对某项目集（记为：$I_{\alpha\beta}$）参与评分，A与B分别评分的项目集分别记为I_α与I_β，R表示平均评分[1]，具体公式如下：

$$sim(\alpha,\beta)=\frac{\sum_{\gamma\in I_{\alpha\beta}}(R_{\alpha,\gamma}-\overline{R}_\alpha)(R_{\beta,\gamma}-\overline{R}_\beta)}{\sqrt{\sum_{\gamma\in I_\alpha}(R_{\alpha,\gamma}-\overline{R}_\alpha)^2}\sqrt{\sum_{\gamma\in I_\beta}(R_{\beta,\gamma}-\overline{R}_\beta)^2}}$$

随着时间的推移，数字图书馆用户的数量将会不断增加，另外，为了知识更新，图书馆每年都要购进一批新的资源，日积月累，用户数量及项目数量急剧上升，通过调查发现，很多用户访问资源后并没有评价，这将造成相似度计算存在弊端，计算出来的结果不够准确，因此，本系统在此算法的基础上采用了聚类技术。

2 聚类技术

2.1 聚类的定义

聚类就是在所有数据项集合中，根据每个数据项之间的相似度，把相似度较高的数据项分别集合起来形成多个分组集合，这个分组的过程我们就称之为聚类。每个分组内的数据项高度相似，不同分组中的数据项相关很大。

例如：假设有一个数据集合$I=\{x_1,x_2,x_3,\cdots\cdots,x_n\}$，根据$x_1,x_2,x_3,\cdots\cdots,x_n$之间的相似度，把相似度接近的数据项组成一个新的分组集合M_1、$M_2\cdots\cdots$，其中$I=M_1\cup M_2\cup\cdots\cdots$，且$M_1\cap M_2\cap\cdots\cdots=\Phi$。

聚类与我们通常所说的分类有点相似，但它们又是不同的，分类是根据原先指定的要求进行分类，而聚类则是经过动态计算的，聚类是数据挖掘的一种。

2.2 聚类的主要方法

（1）分层法

利用相关算法规则，把原来的数据集合划分成一个树状层次结构的数据分组，每个分组是一个子集合，具体划分方法有以下两种：一、凝聚法，即：先任意选择一个数据项为集合，然后把与此相似的数据项合并过来成为新集合，以此类推，直至所有数据项全部划分到不同的分组集合。二、分解法，此方法刚好与凝聚法相反，它是先把所有数据项看成一个大的集合，后根据各数据项的相似度差异，逐一分裂成新的集合的过程。

（2）划分法

根据原先指定划分的集合数，把原来数据集合中所有数据划分到相应集合中去。典型的有：K-means、K-medoids等划分方法。

其中K-means算法的过程：第一步：在原先给定的数据集合中随机选择k个数据项成为聚类的中心，组成k个分组集合。第二步：把其余每个数据项全部划分到离它距离最短的聚类中心点。第三步：再次计算每个分组的平均值，得出新的聚类中心，当最新得出的聚类中心点与原来中心点有变化时，需回到第二步，继续执行，否则整个流程结束[2]。

（3）基于密度算法

本算法主要是每个聚类内数据项有个规定值，当某个聚类的数据项密度超过规定值时，需要放到其它近似的集合中。与前面所说的划分法是不同的，划分法主要是根据数据项之间的距离来划分的。

（4）基于网格算法

把原来数据集合划分成不同单元，建立成一个个网格，再利用STING等算法把集合进行聚类。

（5）基于模型算法

在数据没有被聚类之前，先为每个聚类建立一个模型，根据模型要求，把符合模型要求的数据放入相应聚类中去。

3 基于聚类技术的数字图书馆个性化资源推荐系统算法研究

关于个性化资源推荐系统的算法很多，不同的算法适应的环境有所不同，本文主要采用的算法是基于用户模式聚类的个性化资源推荐算法。本算法的主要思想是：先根据所有用户的注册信息、访问记录、个性化资源定制等信息，产生用户聚类，再与每个用户特有信息相匹配，形成个性化推荐资源集。

3.1 用户聚类的产生

用户聚类的产生一般分为以下几步：

第一步：用户预处理。每个用户访问服务器都有个记录，通过服务器可以得知每个用户访问服务器的频度，根据此频度可以计算出每个用户的支持度，过滤掉小于原先设定的门限值的用户，剩余用户集合$X=\{x_1,x_2,\cdots\cdots,x_m\}$参与下一步操作，其中$x_i$为具体每个用户，M为用户数。

第二步：计算各用户之间的相似度。假设Y为服务器端提供给用户访问的所有网页URL地址数，即：$Y=\{u_1,u_2,u_3,\cdots\cdots u_Y\}$，第i个用户访问URL共有$m_i$次，此用户访问$u_j$的频次为$n^j_i$，则第i个用户$u_j$对应的权值为：

$$S^i_j=\begin{cases}\dfrac{n^j_i}{m_i}, & u_j\in x_i\\ 0, & Otherwise\end{cases}$$

x_i与x_j两用户的相似度为:

$$similar(x_i, x_j) = \frac{\sum_{k=1}^{Y} S_k^i \bullet S_k^j}{\sqrt{\sum_{k=1}^{Y} (S_k^i)^2} \bullet \sqrt{\sum_{k=1}^{Y} (S_k^j)^2}}$$

第三步:对用户进行聚类。本系统采用动态层次索引树聚类算法对用户进行聚类,聚类集为C。

3.2 数字图书馆个性化推荐算法

本系统主要采用的用户聚类模式算法[3],具体算法如下:

```
Recommendation_set(y)←Φ   //设置推荐集合为空
x=1;
while(Pi∈P)   //P为用户聚类
{
  if(similar(y,Pi)≥η)   //其中η为相似度门限值
    for(j=1;j<=n;j=j+1)   //n为聚类的维数
    {  dis(u^i_j,y);   //计算链接距离因子,u^i_j∈Pi
       Rec(y,u^i_j);   //Rec(y,u^i_j)为计算推荐因子
       if(Rec(y,u^i_j)≥ω)   //ω为最小推荐因子
        把u^i_j添加到集合Recommendation_set(y);
    }
  ++i;
}
```

4 小结

本文主要对数字图书馆个性化资源推荐系统的相关算法进行了研究,并在经典的协同过滤算法的基础上,提出了基于聚类技术的数字图书馆个性化资源推荐算法,并予以实现。

注释及参考文献:

[1]罗泽碧,谢庆生.基于web数据挖掘的协同过滤推荐算法[J].贵州大学学报(自然科学版)2009(2):40-43.
[2]梁伍七,江克勤.数据挖掘中的模糊聚类分析及其应用[J].安庆师范学院学报(自然科学版)2004(5):65-67.
[3]何波,杨武,张建勋,等.基于用户模式聚类的智能信息推荐算法[J].计算机工程与设计2006(7):2360-2361.

Research on Personalized Resources Recommendation System Algorithm of Digital Library

ZHANG Ke-zhu
(Department of Computer Information Systems, Suzhou Vocational Technical College, Suzhou, Anhui 234000)

Abstract: Aiming at the existing problems of digital library, digital library personalized resources recommendation algorithm is presented based on user clustering technology, so that the utilization rate of library resources is improved greatly, and decision support is provided for managers.

Key words: Personalized resources recommendation; Digital library; Clusterin

(上接64页)

Abstract: In this essay, several transition technology of IPv4 to IPv6 was introduced, tunnel technology was deeply analysed. According to the actual situation of the campus network, IPv4 network's accessing to IPv6 network was completed based on ISATAP tunneling technology, the specific implementation method was given, which has certain reference value for the campus network's transition to IPv6 network.

Key words: Transitional technology; IPv6 tunnel; Campus network