基于随机游走和多样性图排序的个性化服务推荐方法

方 晨[1,2],张恒巍[1,2],王 娜[1],王晋东[1,2]
(1. 信息工程大学,河南郑州 450001;2. 数字工程与先进计算国家重点实验室,河南郑州 450001)

摘 要: 针对传统服务推荐算法由于数据稀疏性而导致推荐准确性不高,以及推荐结果缺乏多样性等缺陷,提出基于随机游走和多样性图排序的个性化服务推荐方法(PRWDR). 在分析直接相似关系稀疏性的基础上提出带权重的随机游走模型,通过在用户网络上进行随机游走来挖掘更多的相似关系;基于所有相似用户预测服务的 QoS 值,并给出服务图模型构建方法,以过滤大量性能过低的候选服务;提出最优节点集合选取策略,利用贪婪算法得到兼具推荐准确性和功能多样性的服务推荐列表. 在公开发布的数据集上进行实验,并与多个经典算法进行比较,验证了本算法的有效性.

关键词: 服务推荐; 数据稀疏性; 多样性; 随机游走模型
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2018)11-2773-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.11.027

Personalized Service Recommendation Method Based on Random Walk and Diversified Graph Ranking

FANG Chen[1,2], ZHANG Heng-wei[1,2], WANG Na[1], WANG Jin-dong[1,2]
(1. *Information Engineering University, Zhengzhou, Henan* 450001, *China*;
2. *State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, Henan* 450001, *China*)

Abstract: In view of the low recommendation accuracy due to the sparseness of data, and the lack of diversity in traditional service recommendation algorithms, personalized service recommendation method based on random walking and diversified graph ranking (PRWDR) is proposed. On the basis of analyzing the sparseness of direct similarity relationships, a weighted random walk model is proposed, which can excavate more similarity relationships by random walk on the user network. The QoS value of services is predicted based on all similar users, and then the service graph model construction method is presented to filter those services with low performance. By using the greedy algorithm, the optimal node collection selection strategy is proposed to obtain the service recommendation list with both accuracy and diversity. By testing the algorithm on the public dataset and also comparing with several classic algorithms, the validity of PRWDR is verified.

Key words: service recommendation; data sparseness; diversity; random walk model

1 引言

云计算模式的进一步推广催生了大量功能相似但服务质量(Quality of Service, QoS)各异的候选服务,用户根据自身有限的知识和经验难以从中选取出满足其需求的最佳服务. 服务推荐技术因此成为解决用户服务选择困境的重要手段,近几年来在服务计算领域受到了越来越多的关注.

服务推荐的主要依据是服务的 QoS 值,包括响应时间、成本、吞吐量、可靠性、评分等[1]. 目前广泛采用的服务推荐方法是应用协同过滤算法预测服务的 QoS 值,然后以此为依据进行推荐. 但是随着研究的深入,越来越多的实验表明云环境下用户的评分数据十分稀疏,传统算法难以从中找到准确的相似用户或相似服务用于预测. 因此基于稀疏数据的服务推荐方法近年来受到重视. 胡堰等[2]提出基于隐语义概率模型的个性化服务推荐方法,借助隐含类别建立用户指标偏好、用户及服务情境三者之间的关联关系,有效缓解了数据稀疏性带来的不良影响. Liu 等[3]同时利用用户的评分数据和服务的描述信息,采用概率矩阵分解和 LDA

收稿日期:2017-05-16;修回日期:2018-01-01;责任编辑:孙瑶
基金项目:国家自然科学基金(No. 61309013, No. 61303074);河南省科技攻关计划项目(No. 12210231003)

模型来学习用户、服务和相关主题的隐含特征用以推荐. 郭弘毅等[4]通过重叠社区发现算法挖掘用户在社交关系层面的聚类,同时设计模糊聚类算法挖掘用户在兴趣偏好层面的聚类,然后结合两种聚类信息进行推荐.

上述研究工作均聚集于推荐准确性,但仅有推荐准确性是远远不够的. 在用户个性化需求日益增长的今天,用户希望获得既具有准确性又具有多样性的推荐结果. 实际上,推荐多样性在信息检索、数据挖掘等领域已经成为了一个重要的标准. 现有的多样性排序算法主要分为以下三大类:竞争随机游走方法、排序和聚类互增强方法和边际效益最大化方法. 其中边际效益最大化方法由于实现简单得到广泛应用. 基于子话题覆盖的边际效益方法[5]利用贪心策略惩罚多样性差的节点,根据信息的覆盖程度和节点多样性的综合得分对图中的节点进行重排序,从而实现排序的多样性;基于相似度的边际效益方法[6]采用 PageRank 算法度量单个节点的中心度,采用连边权重来度量节点集合的多样性;基于扩散机制的边际效益方法[7]综合考虑节点的相关性和多样性,通过最大化子模性边际效益函数来获取最优的节点集合. 上述算法效果较好,但均不能直接用于服务推荐的场景中. 原因在于,提高服务推荐结果的多样性就意味着要损失一部分准确性. 如何权衡两者之间的关系,从而达到最优的个性化推荐效果,成为服务推荐领域的一大难题.

针对以上问题,本文提出基于随机游走和多样性图排序的个性化服务推荐方法(Personalized Service Recommendation Method Based on Random Walking and Diversified Graph Ranking,PRWDR). 该方法首先在分析直接相似关系稀疏的基础上给出带权重的随机游走模型,通过在用户网络上进行多次随机游走,为目标用户寻找更多的相似用户,从而克服数据的稀疏性;然后,基于所有相似邻居预测候选服务的 QoS 值,选取一定比例 QoS 值较优的服务作为节点,将服务之间的功能相似性作为连边,完成服务图模型的构建;最后,提出最优节点集合选取策略,通过贪婪算法在服务图模型上寻找最优节点集合,使该集合中的服务具有最优的推荐准确性和功能多样性,从而完成个性化服务的推荐.

本文的主要贡献有 3 点:(1)基于相似度的传递特性提出带权重的随机游走模型,以寻找更多的相似用户,克服用户评分数据的稀疏性. (2)结合服务的 QoS 预测值和服务间的功能相似性构建服务图模型,缩小后续多样性排序算法的寻优空间. (3)提出最优节点集合选取策略来寻找兼具推荐准确性和多样性的服务推荐列表.

2 PRWDR 算法

在分析了已有服务推荐算法所存在缺陷的基础上,本节给出基于随机游走和多样性图排序的个性化服务推荐方法 PRWDR. 该方法首先利用 Pearson 相关系数计算用户相似度,并提出带权重的随机游走模型来传递用户间的相似关系,从而有效克服数据的稀疏性问题;然后,充分利用所有相似用户预测 QoS 值,并结合服务功能相似度构建服务图模型;最后,利用贪婪算法在服务图模型上寻找最优节点集合,从而得到兼具推荐准确性和功能多样性的服务推荐列表. 其中,带权重的随机游走模型(Weighted Random Walk Model,WRW)、服务图模型构建方法(Service Graph Model Construction Method,SGMC)以及最优节点集合选取策略(Optimal Node Collection Selection,ONCS)是 PRWDR 的核心,下面将对其进行详细介绍.

2.1 带权重的随机游走模型

基于协同过滤服务推荐算法的核心是相似度的计算. 目前文献中采用最多的相似度计算方法是 Pearson 相关系数,其定义如下[8].

定义 1 (Pearson 相关系数)若 $r(u,i)$ 和 $r(v,i)$ 分别为用户 u 和 v 对服务 i 的 QoS 评价值,$\bar{r}(u)$ 和 $\bar{r}(v)$ 分别为用户 u 和 v 的平均 QoS 评价值,$I_{u,v}$ 表示用户 u 和 v 共同调用的服务集合,则利用 Pearson 相关系数计算用户 u 和 v 的相似度为

$$PCC(u,v)=\frac{\sum_{i\in I_{u,v}}(r(u,i)-\bar{r}(u))(r(v,i)-\bar{r}(v))}{\sqrt{\sum_{i\in I_{u,v}}(r(u,i)-\bar{r}(u))^2}\sqrt{\sum_{i\in I_{u,v}}(r(v,i)-\bar{r}(v))^2}} \tag{1}$$

然而,在海量服务环境下,用户调用过的服务只占很小一部分,很多用户之间根本没有共同调用的服务,由式(1)可得他们之间的相似度为 0,这将导致难以为目标用户找到足够的相似用户. 实际场景中,如果两个用户共享同一相似用户,则根据相似关系的传递性,他们之间也可能相似. 因此,通过挖掘用户之间的间接相似关系,可以有效缓解直接相似关系的稀疏性问题.

近年来,复杂网络中的随机游走模型在克服数据稀疏性问题上取得了良好的效果. 它可以看作是一个描述随机游走者访问顶点序列的马尔可夫链,当游走达到稳态后,每一个节点被访问的概率即为该节点的得分[9]. 而文献[9]仅仅利用传统的随机游走算法,将节点间的转移概率设为定值,未考虑到推荐系统中不同用户间具有不同相似度的特点. 因此,本文在其基础上进行改进,将用户间的相似度作为边的权值,构造转移概率矩阵,通过随机游走来传递用户间的相似关系,

从而为目标用户找到更多的相似用户. 为方便描述,首先给出如下定义.

定义 2　(用户邻接矩阵)基于定义 1 为每个用户寻找与其具有直接相似关系且相似度大于 0 的所有用户,构建用户邻接矩阵 $\boldsymbol{S}$:

$$\boldsymbol{S}=\begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{m2} & \cdots & s_{mm} \end{pmatrix}$$

其中 m 为用户的总数;当用户 u_j 为用户 u_i 的直接相似用户且 $PCC(u_i,u_j)>0$ 时,$s_{ij}=PCC(u_i,u_j)$,否则 $s_{ij}=0$(即不考虑相似度小于 0 的用户);$s_{ii}=0(1\leqslant i\leqslant m)$ 表示不考虑用户与其自身的相似度.

定义 3　(转移概率矩阵)本文将用户的相似度作为边的权重,那么将用户邻接矩阵 $\boldsymbol{S}$ 进行行向量归一化后便得到转移概率矩阵 $\boldsymbol{T}=(t_{ij})_{m\times m}$,$t_{ij}$ 表示游走者由用户节点 u_j 移动到用户节点 u_i 的概率,其表达式为

$$t_{ij}=\frac{s_{ij}}{\sum_{k=1}^{m}s_{ik}} \tag{2}$$

定义 4　(随机游走策略)令 $\boldsymbol{r}$ 表示用户列向量,其中每个元素 $r_j(1\leqslant j\leqslant m)$ 表示用户节点 j 被访问的概率,则随机游走策略可表示为数学表达式如下

$$\boldsymbol{r}^n=c\times\boldsymbol{T}\times\boldsymbol{r}^{n-1}+(1-c)\times\boldsymbol{r}^0 \tag{3}$$

其中 c 为游走者下一步移动到与其最近的邻居的概率,$1-c$ 为游走者下一步返回到开始节点 i 的概率,$\boldsymbol{r}^n$ 表示第 n 步到达各用户节点的概率分布,$\boldsymbol{r}^0$ 表示初始概率分布,它每个元素 $r_j(1\leqslant j\leqslant m)$ 的取值为

$$r_j=\begin{cases}1, & \text{if } j=i(u_i \text{ is the target user})\\ 0, & \text{otherwise}\end{cases} \tag{4}$$

定理 1　根据定义 4 中的随机游走策略,经过多次迭代,最终用户列向量 $\boldsymbol{r}$ 会收敛到一个静态概率分布,记为 $\boldsymbol{r}^*=(1-c)\times(\boldsymbol{I}-c\times\boldsymbol{T})^{-1}\times\boldsymbol{r}^0$.

证明　根据式(3)可得 $\boldsymbol{r}^{n-1}=c\times\boldsymbol{T}\times\boldsymbol{r}^{n-2}+(1-c)\times\boldsymbol{r}^0$,令 $\Delta_n=\boldsymbol{r}^n-\boldsymbol{r}^{n-1}$,带入公式可得

$$\Delta_n=c\times\boldsymbol{T}\times\Delta_{n-1};n=2,3,\cdots \tag{5}$$

根据递推式(5)可得

$$\Delta_n=c\times\boldsymbol{T}\times\Delta_{n-1}=(c\times\boldsymbol{T})^{n-1}\times\Delta_1 \tag{6}$$

根据式(3)可得 $\Delta_1=c\times(\boldsymbol{T}-\boldsymbol{I})\times\boldsymbol{r}^0$,将 Δ_1 代入式(6)得

$$\Delta_n=c^n\times(\boldsymbol{T})^{n-1}\times(\boldsymbol{T}-\boldsymbol{I})\times\boldsymbol{r}^0 \tag{7}$$

由于转移概率 $c\in(0,1)$,有 $\lim_{n\to\infty}c^n=0$,故当 $n\to\infty$ 时,$\Delta_n=0$,即 $\boldsymbol{r}^n=\boldsymbol{r}^{n-1}=\boldsymbol{r}^*$,代入式(3)得

$$\boldsymbol{r}^*=(1-c)\times(\boldsymbol{I}-c\times\boldsymbol{T})^{-1}\times\boldsymbol{r}^0 \tag{8}$$

因为 $\|c\times\boldsymbol{T}\|_1<1$,根据矩阵条件数中相关定理知 $\boldsymbol{I}-c\times\boldsymbol{T}$ 的逆一定存在. 证毕.

定义 5　(用户相邻度)当随机游走模型达到稳态后,用户列向量 $\boldsymbol{r}$ 中每个元素 $r_j(1\leqslant j\leqslant m)$ 即代表从目标用户节点 i 随机游走到用户节点 j 的概率,将其定义为目标用户 u_i 与用户 u_j 的相邻度

$$pro(u_i,u_j)=\begin{cases}r_j, & \text{if } j\neq i\\ 0, & \text{if } j=i\end{cases} \tag{9}$$

其中 $pro(u_i,u_i)=0$ 即不考虑目标用户与它自身的相邻度. 但是,用户相邻度仅仅代表用户之间的相近程度,并不是实际的用户相似度. 因此需要对用户相邻度进行变换,得到修正用户相似度.

定义 6　(修正用户相似度)设共有 m 个用户构成集合 U,基于定义 1 为目标用户 u_i 构建直接相似邻居集合 $N_{u_i}=\{u_{i_l}\mid PCC(u_i,u_{i_l})>0,u_{i_l}\in U\}$. 则对于任一用户 $u_j\in U$,其与目标用户 u_i 的修正用户相似度为

$$sim(u_i,u_j)=\frac{1}{|N_{u_i}|}\times\sum_{l=1}^{|N_{u_i}|}\frac{PCC(u_i,u_{i_l})}{pro(u_i,u_{i_l})}\times pro(u_i,u_j);\ 1\leqslant j\leqslant m \tag{10}$$

在上述定义的基础上,给出带权重的随机游走模型的算法伪代码(如算法 1 所示).

算法 1　带权重的随机游走模型 WRW

输入　全体用户集合 U,目标用户 u_i,概率 c
输出　全体用户与目标用户 u_i 的修正相似度 $sim(u_i,u_j)$
1)基于定义 2,构建用户邻接矩阵 $\boldsymbol{S}=(s_{ij})_{m\times m}$//选取所有相似度大于 0 的用户对
2)用户初始列向量 $\boldsymbol{r}^0\leftarrow[0,\ldots,1,0,\ldots,0]^{\mathrm{T}}$　//其中只有第 i 个元素为 1,代表目标用户 u_i
3)for $j\leftarrow$1 to m do
4)　$rowsum\leftarrow\sum_{k=1}^{m}s_{jk}$
5)　for $k\leftarrow$1 to m do
6)　$t_{jk}\leftarrow s_{jk}/rowsum$
7)　end for
8)end for　　//构建转移概率矩阵 $\boldsymbol{T}$
9)$\boldsymbol{r}\leftarrow(1-c)\times(\boldsymbol{I}-c\times\boldsymbol{T})^{-1}\times\boldsymbol{r}^0$
10)利用式(9)得到每一个用户 $u_j\in U$ 与目标用户 u_i 的相邻度 $pro(u_i,u_j)$
11)$sum\leftarrow0$,$N_{u_i}\leftarrow\varnothing$
12)for each user $u_j\in U$
13)　if $PCC(u_i,u_j)>0$,then
14)　$N_{u_i}\leftarrow N_{u_i}+\{u_j\}$
15)　$sum\leftarrow sum+PCC(u_i,u_j)/pro(u_i,u_j)$
16)　end if
17)end for
18)for each user $u_j\in U$
19)　$sim(u_i,u_j)\leftarrow1/|N_{u_i}|\times sum\times pro(u_i,u_j)$
20)end for

2.2 服务图模型构建方法

通过 WRW 模型得到修正用户相似度后,便可以基于相似用户对服务的 QoS 值进行预测,进而做出推荐. 但是,仅仅推荐 QoS 值较优的服务难以满足用户的个性化需求,还需提高服务推荐结果的多样性. 文献[7]利用多样性图排序算法来提高推荐多样性,但是其侧重于通过节点间的相似度来度量多样性,容易造成局部收敛. 本文在其基础上,从单个节点的角度转向节点集合,通过节点集合对网络的覆盖程度来度量多样性,充分考虑到了网络的拓扑结构.

在实施多样性图排序算法之前,需要构建服务图模型. 下面首先给出 QoS 预测值和服务功能相似度的计算方法.

定义 7 (QoS 预测值)将修正用户相似度大于 0 的用户作为目标用户 u_i 的相似邻居集合 $R_{u_i}=\{u_j \mid sim(u_i,u_j)>0, u_j\in U\}$,则目标用户 u_i 对服务 k 的 QoS 预测值为

$$q_{ik}=\bar{q}_i+\frac{\sum_{u_j\in R_{u_i}} sim(u_i,u_j)(q_{jk}-\bar{q}_j)}{\sum_{u_j\in R_{u_i}} sim(u_i,u_j)} \tag{11}$$

其中 $\bar{q}_i$ 表示目标用户 u_i 的平均 QoS 评分值,q_{jk} 表示用户 u_j 对服务 k 的 QoS 评分值.

为了方便统一比较,本文将所有 QoS 值进行归一化至区间[0,1]内. 对于效益型 QoS,采用式(12)进行归一化;对于成本型 QoS,采用式(13)进行归一化.

$$q'_{ik}=\begin{cases}\dfrac{q_{ik}-q_{\min}(k)}{q_{\max}(k)-q_{\min}(k)}, & \text{if } q_{\max}(k)\neq q_{\min}(k)\\ 1, & \text{if } q_{\max}(k)=q_{\min}(k)\end{cases} \tag{12}$$

$$q'_{ik}=\begin{cases}\dfrac{q_{\max}(k)-q_{ik}}{q_{\max}(k)-q_{\min}(k)}, & \text{if } q_{\max}(k)\neq q_{\min}(k)\\ 1, & \text{if } q_{\max}(k)=q_{\min}(k)\end{cases} \tag{13}$$

其中 $q_{\max}(k)$ 代表所有候选服务中最大的 QoS 预测值,$q_{\min}(k)$ 代表所有候选服务中最小的 QoS 预测值.

基于内容的推荐算法通常利用词频-逆文档频(Term Frequency-Inverse Document Frequency, TF-IDF)算法来比较两个产品的相似度. 而在服务推荐领域,每个 Web 服务对应一个 WSDL(Web Service Description Language)文档和相应的标签集合,用于描述该 Web 服务的相关功能及特点. 基于此,本文给出服务功能相似度的计算方法:根据每个 Web 服务的 WSDL 文档和标签集合,提取出一系列关键词 k(具体的关键词提取技术见文献[10],本文不作探讨). 关键词 k_i 对于服务 j 的重要性权重 w_{ij} 可以用下式计算得到

$$\begin{aligned}w_{ij}&=tf(k_i,WSDL_j)\times idf(k_i,WSDL_j)\\&=\frac{freq(k_i,WSDL_j)}{|WSDL_j|}\times\log_2\frac{|WSDL|}{|\{WSDL_j:k_i\in WSDL_j\}|}\end{aligned} \tag{14}$$

其中 $tf(k_i,WSDL_j)$ 表示词频,记录关键词 k_i 在 $WSDL_j$ 文档中出现的频率,出现次数(记为 $freq(k_i,WSDL_j)$)越多,表示关键词 k_i 越重要;$idf(k_i,WSDL_j)$ 表示逆文档频率,包含关键词 k_i 的文档数(记为 $|\{WSDL_j:k_i\in WSDL_j\}|$)越少,表示关键词 k_i 越能唯一体现出 Web 服务的功能特征;$|WSDL_j|$ 表示 $WSDL_j$ 文档中关键词总数;$|WSDL|$ 为 WSDL 文档的总数,即候选服务总数.

由此可以将服务 j 表示为关键词权重向量 $\boldsymbol{w}_j=(w_{1j},w_{2j},\cdots,w_{lj})$,其中 l 为关键词总数. 因此,两个服务的功能相似度可以利用对应的关键词权重向量的相似度来表示. 为此给出如下定义.

定义 8 (功能相似度)服务 i 用向量 $\boldsymbol{w}_i$ 表示,服务 j 用向量 $\boldsymbol{w}_j$ 表示,则服务 i 和服务 j 的功能相似度为

$$funsim(i,j)=\cos(\boldsymbol{w}_i,\boldsymbol{w}_j)=\frac{\boldsymbol{w}_i\cdot\boldsymbol{w}_j}{|\boldsymbol{w}_i|\times|\boldsymbol{w}_j|} \tag{15}$$

在给出 QoS 预测值和服务功能相似度计算方法后,现将服务图模型构建方法 SGMC 描述如算法 2 所示.

算法 2 服务图模型构建方法 SGMC

输入 算法 1 的输出,候选服务集合 I,阈值 θ、τ
输出 服务图模型 $G=(V,E)$
1) $V=\varnothing, E=\varnothing$
2) for each service $k\in I$, do
3) 利用式(11)计算服务 k 的 QoS 预测值 q_{ik}
4) end for
5) for each service $k\in I$, do
6) 利用式(12)(13)对 QoS 预测值进行归一化,得 q'_{ik}
7) if $q'_{ik}>\theta$, then
8) $V\leftarrow V+\{k\}$ //将服务 k 作为节点加入到图模型中
9) end if
10) end for
11) for each pair of service nodes k、$k'\in V$, do
12) 利用式(15)计算服务 k 和服务 k' 的功能相似度 $funsim(k,k')$
13) if $funsim(k,k')>\tau$, then
14) $E\leftarrow E+edge(k,k')$ //将服务节点 k 和 k' 进行连边
15) end if
16) end for

综上,SGMC 算法首先从候选服务集合中选取出 QoS 预测值较优的服务作为节点加入到图模型中,这样可以过滤掉大量性能过低的服务,缩小候选服务空间;然后,如果两个服务的功能相似,则将该对服务节点进行连边,以利于后续多样性图排序算法寻找最优节点集合.

2.3 最优节点集合选取策略

构建出服务图模型后,便可以利用多样性图排序算法来寻找既具有较优 QoS 值也具有较好的功能多样

性的服务节点集合. 在给出最优节点集合选取策略之前,进行如下定义.

定义 9　(扩展集合)假设 S 为服务图模型 $G=(V,E)$ 中节点集合 V 的子集,则 S 的扩展集合 $N(S)$ 定义为

$$N(S)=S\cup\{v\in(V-S)\mid\exists u\in S,(u,v)\in E\}$$

定义 10　(扩展比)已知节点集合 S 的扩展集合为 $N(S)$,其中 K 为服务图模型中节点总数,$|N(S)|$ 为 $N(S)$ 中节点的个数,则节点集合 S 的扩展比定义为

$$\sigma=\frac{|N(S)|}{K}\tag{16}$$

根据定义 9 和定义 10 可知,扩展比与服务图模型的拓扑结构相关. 显而易见,对于节点集合 S,如果其扩展比越大,就意味着该集合中的节点在服务图模型中越分散. 根据服务图模型的构建过程可知,只有功能相似的服务节点之间有连边. 那么集合中的节点越分散,就表示它们之间的功能越不相似,即多样性越好. 由此可以得出结论:节点集合的扩展比可以衡量该集合中服务功能的多样性.

基于上述结论,向用户推荐 k 个最优服务的问题可转化为在服务图模型 $G=(V,E)$ 上选取一个包含 k 个服务节点的最优集合 S,使其具有最大的 QoS 值和最大的扩展比,从而同时确保推荐准确性和功能多样性,可表示为数学表达式

$$\arg\max_{S\subseteq V}F(S)=(1-\lambda)\sum_{v\in S}q'_{iv}+\lambda\frac{|N(S)|}{K}$$
$$\text{s.t. }|S|=k\tag{17}$$

其中 V 为图模型 $G=(V,E)$ 中全体服务节点集合,q'_{iv} 表示目标用户 u_i 对于服务 v 的归一化 QoS 值;$\sum_{v\in S}q'_{iv}$ 为集合 S 中所有服务的归一化 QoS 值之和,其代表推荐准确性;$\frac{|N(S)|}{K}$ 为集合 S 的扩展比,其代表推荐多样性;参数 λ 用于权衡推荐准确性和功能多样性. 当 $\lambda=1$ 时,最优节点集合选取的问题就变成最大化扩展比问题,而在文献[11]中该问题被证明是 NP-hard. 由此可见推断,最优节点集合选取问题也是 NP-hard 的.

文献[11]中指出,对于像函数 $F(S)$ 这样的非减性子模函数,采用贪心算法得到的集合 S 不会比最优的集合 S^* 的 $1-1/e$ 差,即 $f(S)\geq(1-1/e)f(S^*)$,且没有其它算法能够在多项式时间内得到一个更加近似的结果. 因此,最优节点集合选取问题可以采用贪婪算法进行近似求解. 现给出最优节点集合选取策略的算法伪代码,如算法 3 所示.

算法 3　最优节点集合选取策略 ONCS

输入　服务图模型 $G=(V,E)$,参数 λ、k
输出　最优节点集合 S

Begin
1) $S=\varnothing$
2) while $|S|\leq k$, do
3)　find
$$v_{\max}=\arg\max_{v\in(V-S)}(1-\lambda)q'_{iv}+\lambda\frac{|N(S\cup\{v\})|-|N(S)|}{K}$$
4)　$S\leftarrow S+\{v_{\max}\}$
5) end while
End

2.4　PRWDR 算法描述

经过上述分析,PRWDR 算法的具体过程如下:

步骤 1　利用 WRW 模型在用户网络上进行随机游走,得到全体用户与目标用户 u_i 的修正相似度 $sim(u_i,u_j)$;

步骤 2　基于步骤 1 得到的结果,利用 SGMC 算法构建服务图模型,过滤大量性能过低的服务,缩小候选服务规模;

步骤 3　在构建的服务图模型上运行 ONCS 算法,得到最优节点集合,即为向用户推荐的兼具推荐准确性和功能多样性的 k 个最优服务.

2.5　算法时间复杂度分析

设推荐系统中共有 m 个用户,n 个服务,每对用户平均共同调用服务数为 n_1,服务图模型中节点数为 $K(K\leq n)$,连边数为 $|E|$. 对于带权重的随机游走模型 WRW,其时间复杂度为 $O(m^3+m^2n_1)$,其中构建用户邻接矩阵的复杂度为 $O(m^2n_1)$,计算矩阵逆的复杂度为 $O(m^3)$;对于服务图模型构建方法 SGMC,其时间复杂度为 $O(mn+|E|)$,其中预测所有服务 QoS 值的复杂度为 $O(mn)$,构建图模型的复杂度为 $O(n+|E|)$;对于最优节点集合选取策略 ONCS,其时间复杂度为 $O(K|E|)$.

因此综合来看,PRWDR 算法的时间复杂度为 $O(m^3+m^2n_1+mn+K|E|)$. 可见,构建用户邻接矩阵和计算矩阵逆 $(\boldsymbol{I}-c\times\boldsymbol{T})^{-1}$ 的时间复杂度相对较高,但是,利用本算法为任一目标用户进行服务推荐时都需要用到这两个计算结果,因此这两个式子可以离线进行预计算. 在实时为某一特定目标用户进行服务推荐时,PRWDR 算法的在线时间复杂度就相当于 $O(mn+K|E|)$.

3　实验结果与分析

本文使用公开的真实数据集 MovieLens[12] 作为实验数据集,此数据集中包括 6400 个用户对 3900 部电影的 1000000 个评分记录,数据稀疏度为 95.75%. 此外,由于 MovieLens 数据集中缺少电影的描述文档,我们根据电影名到 IMDB 中获取其相关信息,作为描述文档. 实验过程中,将 80% 的数据作为训练集,20% 的数据作为测试集. 随机选取 50 个用户作为目标用户,实验结果取 5 次测试的平均值.

算法最终向用户提供一个包含 k 个服务的推荐列

表,因此本文基于这 k 个服务,从以下三个方面来评价算法的有效性. ①推荐准确性:由式(17)可知, $\sum_{v\in S} q'_{iv}$ 是这 k 个服务的归一化 QoS 预测值之和, $\sum_{v\in S} q'_{iv}$ 值越大,代表所推荐的服务总体性能越好,即推荐准确性越高. ②功能多样性: $|N(S)|/K$ 是服务推荐列表的扩展比,扩展比越大,代表所推荐的 k 个服务的功能多样性越好. ③总体质量:函数值 $F(S)$ 综合考虑了推荐准确性和功能多样性,代表服务推荐列表的总体质量.

算法的主要参数设置如下. 转移概率 $c=0.85$(随机游走模型的常用取值[9]),阈值 $\theta=0.5$、参数 $\lambda=0.7$(关于 λ、θ 的取值实验中有说明),阈值 τ 取图模型中服务功能相似度的平均值,如式(18)所示

$$\tau = \frac{2 * \sum_{i,j\in V} funsim(i,j)}{K(K-1)} \tag{18}$$

3.1 参数对于算法性能的影响

首先,通过实验来考察 PRWDR 涉及到的两个主要参数 λ、θ 对于算法性能的影响.

在服务图模型 $G=(V,E)$ 的构建方法中,θ 是归一化 QoS 的阈值,它用于控制加入节点集合的候选服务比例. 直观地说,若 θ 太大,则大量候选服务被过滤掉,导致后续多样性排序算法的寻优空间缺乏功能多样性;相反,若 θ 太小,推荐结果中可能会加入一些性能过低的服务,导致算法的推荐准确性不高. 为了研究阈值 θ 的合理取值,首先令 $\lambda=0.7$,$k=20$,记录 θ 取不同值时算法推荐结果的总体质量. 由图 1 所示,当 $\theta=0.5$ 时,PRWDR 算法的总体质量最高,主要原因在于它同时保证了功能多样性和推荐准确性都维持在一个较高的水平. 在后面的实验中,均设 $\theta=0.5$.

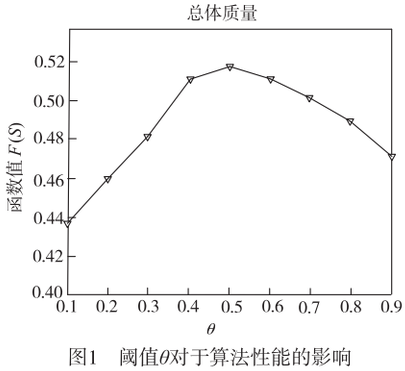

图1 阈值θ对于算法性能的影响

根据式(17),参数 λ 用于控制推荐准确性和功能多样性各自所占的权重. λ 的变化对算法推荐结果总体质量的影响如图 2 所示. 可以看出,当 $\lambda=0.7$ 时,PRWDR 的总体质量最高,说明此时得到的服务推荐列表拥有最佳的推荐准确性和功能多样性. 在后面的实验中,均设 $\lambda=0.7$.

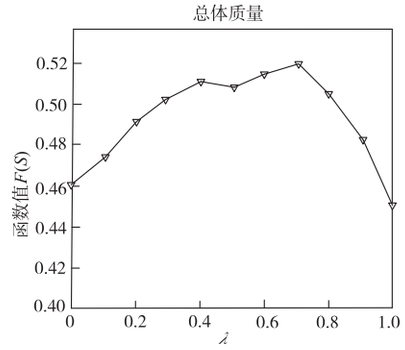

图2 参数λ对于算法性能的影响

3.2 与其它服务推荐算法比较

为了衡量本算法的推荐效果,将 PRWDR 与以下两种经典的服务推荐算法做比较,比较结果如图 3 所示.

(1)PMF-TM[3]:结合概率矩阵分解模型和概率主题模型,挖掘用户、服务和相关主题的隐含特征变量,通过向量内积来预测 QoS 评分值.

(2)HACR[13]:提出非对称相关正则化矩阵分解模型,利用用户间和服务间的非对称相关关系来缓解数据的稀疏性,提高服务推荐的准确性.

由图 3 可以得出以下结论.

(1)在推荐准确性方面,三种算法的优劣排序为 PMF-TM > PRWDR > HACR. 原因在于,HACR 仅仅利用了用户的评分数据,而 PMF-TM 和 PRWDR 同时利用评分数据和服务描述信息来预测用户偏好,使它们能够在较稀疏的数据上表现更好. 此外,PRWDR 的推荐准确性比 PMF-TM 略低,这说明提高推荐结果的多样性必然会损失一部分准确性.

(2)在功能多样性方面,PRWDR 与另外两种算法相比有较大优势. 因为 PRWDR 在推荐服务时考虑了服务集合的扩展比 $|N(S)|/K$,因此其推荐结果具有较好的功能多样性;而另外两种算法仅仅考虑服务的评分值,忽略了推荐结果的多样性.

(3)在总体质量方面,由于 PRWDR 综合考虑了推荐结果的功能多样性和推荐准确性,因此其表现最优.

为了进一步验证本算法在推荐多样性方面的表现,我们统计了当 $k=20$ 时,各方法对于 50 个目标用户的推荐结果中不同电影的数目,并计算最经典(即评分最高)的 50 部电影的出现次数占总数的比例,如表 1 所示.

表 1 各方法推荐结果统计

方法	不同电影的数目(部)	最经典 50 部电影所占比例(%)
PMF-TM	185	62.5
HACR	236	48.9
PRWDR	344	21.3

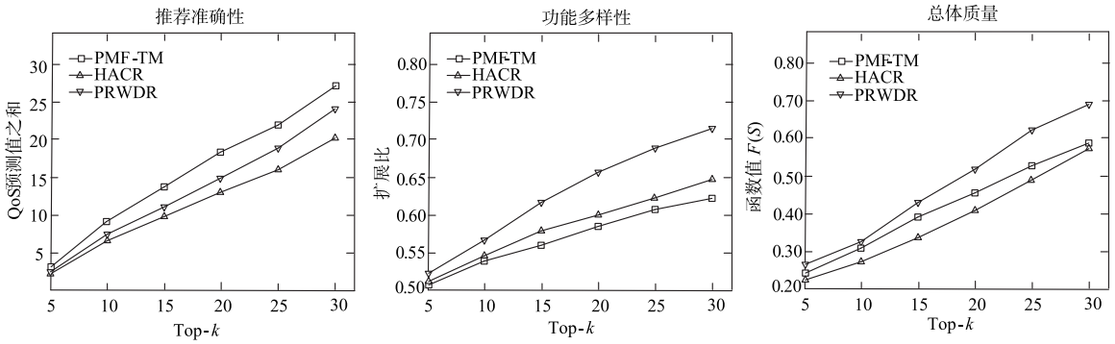

图3　PRWDR与其它服务推荐算法的实验对比

由表 1 可以看出,相比于其它两种算法,PRWDR 能够向用户推荐更多不同类型的电影. 具体地,在 PMF-TM 的推荐结果中,最经典的 50 部电影的出现次数占总数的 62.5%,这说明 PMF-TM 倾向于推荐经典电影,这些电影往往评分较高,但很有可能被大多数用户所看过. 而 PRWDR 的推荐结果分布更加均匀,最经典的 50 部电影的出现次数仅占总数的 21.3%,可以满足用户的多样化需求.

3.3　与其它多样性排序算法比较

由于传统服务推荐算法均未考虑多样性,为了进一步验证本算法的推荐效果,将 PRWDR 与以下两种经典的多样性排序算法做比较,比较结果如图 4 所示.

(1) ClusDiv[14]:通过改进多样性度量公式,利用 k-means 算法将物品聚类,然后从不同类中选取出物品构成推荐列表.

(2) CBRD[15]:提出覆盖度标准来统一概括相关性和多样性,通过贪婪算法求出 top-k 推荐结果.

由图 4 可以得出以下结论.

(1) 在推荐准确性方面,ClusDiv 和 PRWDR 表现较好,两者相比于 CBRD 具有较大优势. 原因在于,ClusDiv 和 PRWDR 分别通过聚类算法和随机游走机制来缓解 MovieLens 数据集的稀疏性,而 CBRD 并未对原始数据集进行任何处理,因此表现较差.

(2) 在功能多样性方面,PRWDR 相比另外两种算法表现较好. 原因在于,ClusDiv 和 CBRD 均是根据评分数据来度量推荐结果的多样性,但是评分数据相差大并不一定代表多样性好. PRWDR 则从语义分析角度,根据电影的描述信息来度量电影的多样性,更加合理.

(3) 在总体质量方面,由于 PRWDR 通过随机游走机制缓解数据的稀疏性,且利用语义分析模型来准确度量多样性,因此其相比另外两种算法表现较优.

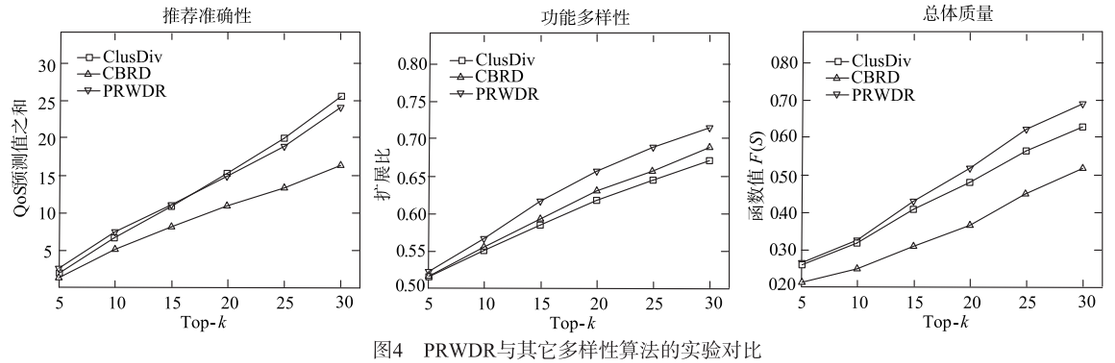

图4　PRWDR与其它多样性算法的实验对比

4　结束语

本文针对现有服务推荐算法克服数据稀疏性能力不强以及缺乏推荐多样性问题,提出基于随机游走和多样性图排序的个性化服务推荐方法 PRWDR. 该方法首先在分析直接相似关系稀疏的基础上,提出带权重的随机游走模型 WRW,通过多次传递相似关系来克服数据稀疏性问题;然后,基于所有相似用户预测服务的 QoS 值,并给出服务图模型构建方法 SGMC,过滤大量性能过低的候选服务,缩小算法寻优空间;最后,利用最优节点选取策略 ONCS 得到兼具推荐准确性和功能多样性的服务推荐列表,至此完成个性化服务推荐. 实验证明,PRWDR 具有较高的推荐准确性和较好的功能多样性,能够在保证服务推荐质量的同时还最大程度地满足用户潜在的功能需求. 下一步工作将研究引入 QoS 的动态性特征、用户上下文信息如时间、地理位置

等来进一步提高服务推荐的准确性.

参考文献

[1] 龙军,袁鑫攀,桂卫华. 基于环境感知的可信 QoS 评价与服务选取策略[J]. 电子学报,2012,40(6):1133 – 1140.
LONG Jun, YUAN Xin-pan, GUI Wei-hua. A policy for the trusted QoS evaluation and service selection with environment aware[J]. Acta Electronica Sinica,2012,40(6):1133 – 1140. (in Chinese)

[2] 胡堰,彭启民,胡晓惠. 一种基于隐语义概率模型的个性化 Web 服务推荐方法[J]. 计算机研究与发展,2014,51(8):1781 – 1793.
HU Yan,PENG Qi-min,HU Xiao-hui. A personalized web service recommendation method based on latent semantic probabilistic model[J]. Journal of Computer Research and Development,2014,51(8):1781 – 1793. (in Chinese)

[3] Liu X,Fulia I. Incorporating user,topic,and service related latent factors into web service recommendation[A]. Proceedings of the IEEE International Conference on Web Services[C]. New Work:IEEE,2015. 185 – 192.

[4] 郭弘毅,刘功申,等. 融合社区结构和兴趣聚类的协同过滤推荐算法[J]. 计算机研究与发展,2016,53(8):1664 – 1672.
GUO Hong-yi,LIU Gong-shen,et al. Collaborative filtering recommendation algorithm combining community structure and interest clusters[J]. Journal of Computer Research and Development,2016,53(8):1664 – 1672. (in Chinese)

[5] Zhang B, Li H, et al. Improving web search results using affinity graph[A]. Proceedings of the ACM SIGIR International Conference on Research & Development in Information Retrieval[C]. New Work:ACM,2005. 504 – 511.

[6] Tong H, He J, Wen Z, et al. Diversified rankingon large graphs:An optimization viewpoint[A]. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New Work:ACM,2011. 1028 – 1036.

[7] Li R H, Yu J X. Scalablediversified ranking on large graphs[J]. IEEE Transactions on Knowledge & Data Engineering,2013,25(9):2133 – 2146.

[8] 俞春花,刘学军,李斌,等. 基于上下文相似度和社会网络的移动服务推荐方法[J]. 电子学报,2017,45(6):1530 – 1536.
YU Chun-hua,LIU Xue-jun,LI Bin,et al. Mobile service recommendation based on context similarity and social network[J]. Acta Electronica Sinica,2017,45(6):1530 – 1536. (in Chinese)

[9] Bagci H, Karagoz P. Context-aware location recommendation by using a random walk-based approach[J]. Knowledge and Information Systems,2016,47(2):241 – 260.

[10] Kang G,Liu J,Tang M,et al. AWSR:Active web service recommendation based on usage history[A]. Proceedings of the International Conference on Web Services[C]. Washington:IEEE Computer Society,2012. 186 – 193.

[11] Maiya A S, Berger-Wolf T Y. Expansion and search in networks[A]. Proceedings of the 19th ACM International Conference on Information and Knowledge Management[C]. New York:ACM,2010. 239 – 248.

[12] Grouplens. MovieLens[DB/OL]. https://grouplens.org/datasets/movielens/. 2017 – 12 – 31.

[13] Xie Q,Zhao S,Zheng Z,et al. Asymmetriccorrelation regularized matrix factorization for web service recommendation[A]. Proceedings of the IEEE International Conference on Web Services[C]. New Work:IEEE,2016. 204 – 211.

[14] Aytekin T,Karakaya M Ö. Clustering-based diversity improvement in top-N recommendation[J]. Journal of Intelligent Information Systems,2014,42(1):1 – 18.

[15] Usunier N,Usunier N,Grandvalet Y. Acoverage-based approach to recommendation diversity on similarity graph[A]. Proceedings of the 10th ACM Conference on Recommender Systems[C]. New York:ACM,2016. 15 – 22.

作者简介

方 晨 男,1993 年出生于安徽宿松. 现为信息工程大学研究生. 主要研究方向为服务推荐、数据挖掘等.
E-mail:17756230629@163.com

张恒巍(通信作者) 男,1978 年出生于河南洛阳. 现为信息工程大学讲师. 主要研究方向为云资源管理、网络安全等.

王 娜 女,1970 年出生于河南郑州. 现为信息工程大学副教授. 主要研究方向为服务计算、信息安全等.

王晋东 男,1966 年出生于山西洪洞. 现为信息工程大学教授. 主要研究方向为云计算、网络攻防等.