网络首发时间：2019-04-22 14:59:37
网络首发地址：http://kns.cnki.net/kcms/detail/11.2223.N.20190422.0815.001.html

ISSN 1000-0054　　清华大学学报（自然科学版）
CN 11-2223/N　　J Tsinghua Univ (Sci & Technol)

服务系统中冷启动服务协作关系挖掘与预测

郝予实，　范玉顺

（清华大学 自动化系，北京 100084）

摘　要：Web服务系统中大量无使用记录的服务和不断发布的新创建服务被称为冷启动服务。为了帮助服务组合开发者了解冷启动服务的特性，提高冷启动服务的关注率与使用率，从而增强服务系统的元素多样性和系统鲁棒性，该文提出了一种冷启动服务协作关系挖掘与预测方法。该方法利用服务描述重构和功能主题分析为每个服务建立功能属性向量。对非冷启动服务，基于其历史协作关系和功能属性向量为其建立协作属性向量。通过对冷启动服务功能属性向量与非冷启动服务协作属性向量进行相似性比较，实现冷启动服务组合协作关系预测。真实数据集上的实验结果证明该方法在预测效果上显著强于当前最优方法。

关键词：服务系统；冷启动服务；服务组合；协作关系；主题模型

中图分类号：TP399　　**文献标志码**：A

DOI：10.16511/j.cnki.qhdxxb.2019.22.022

Mining and predicting of cold start service collaboration relationships in service systems

HAO Yushi, FAN Yushun

(Department of Automation, Tsinghua University, Beijing 100084, China)

Abstract: Services that have never been used and newly released services in web service systems are called cold start services. A cold start service collaboration relationship mining and predicting method is developed to help service composition developers identify the characteristics of cold start services, increase attention to and usage of cold start services, and enhance the element diversity and robustness of service systems. The method first establishes a functional vector for each service using service description reconstruction and a functional topic analysis. Next, the method builds a collaborative vector for each non-cold start service based on its historical collaboration record and functional vector. Finally, the method compares the functional vectors of cold start services with the collaborative vectors of non-cold start services to predict the collaboration relationships for cold start services. Tests on real-world data show that this method more effectively predicts the relationships than state-of-the-art methods.

Key words: service systems; cold start services; service composition; collaboration relationships; topic models

随着面向服务的体系架构（service-oriented architecture，SOA）和云计算技术的快速发展，以服务为导向的业务发展模式被广为接受。大量企业将其核心业务资源以Web服务的形式在互联网上进行发布和交付[1]，第三方用户可以在互联网上查找和使用这些服务。然而，由于用户需求往往具有复杂性和个性化的特点，而单个的服务通常仅能满足单一的需求，因此在实际操作中，用户通常将多个服务组合使用，以共同满足其个性化的复杂需求[2]。服务组合的创建拓展了服务的功能边界，实现了服务的重复利用，也缩短了针对新需求的开发周期。互联网上不断增长的Web服务和服务组合以及它们蕴含的各类信息和关联关系共同构成了Web服务系统[3]。

近年来，Web服务的数量呈现出快速增长的态势，海量的服务给用户的服务查找和筛选过程带来了严重的信息过载问题[4]。为了帮助用户快速进行服务组合开发，服务推荐问题与服务协作关系挖掘问题获得了学界的广泛关注和研究[5]。服务推荐问题旨在通过识别用户需求，为其推荐创建服务组合的备选服务；服务协作关系挖掘问题旨在挖掘服务之间的组合模式和协作关系，从而辅助用户进行服务组合的创建。

收稿日期：2018-11-29
基金项目：国家自然科学基金资助项目(61673230)；
国家高技术船舶科研项目(17GC26102.01)
作者简介：郝予实(1992—)，男，博士研究生。
通信作者：范玉顺，教授，E-mail：fanyus@tsinghua.edu.cn

针对服务推荐问题的早期研究主要依赖于对用户需求和服务描述文档进行关键词匹配[6]。然而，单纯的关键词匹配往往难以取得理想的效果，研究者尝试将协同过滤、矩阵分解等工具应用到服务推荐问题上，通过对服务历史使用记录的建模提升服务推荐效果[7]。Zhang 等[8]认为 LDA(latent Dirichlet allocation)模型能够从主题的角度理解用户需求和服务功能，并通过主题分析的手段提升了服务推荐的准确性。近年来，研究人员从服务系统演化模式分析、服务流行度分析、服务系统网络模型分析、服务功能领域分析等角度出发，分别构建了一系列复杂的服务推荐模型，获得了较好的推荐效果[9-10]。针对服务协作关系挖掘问题，早期的研究多利用 Apriori 算法挖掘服务之间的关联规则，从而实现协作关系的预测[11]。Huang 等[12]对服务系统进行网络化建模，利用链路预测工具实现对服务潜在协作关系的预测。Gao 等[5]利用 LDA 主题模型挖掘服务的潜在共现主题关系，进而预测服务未来的组合模式。

然而，当前研究均局限在对拥有使用记录的服务的推荐和协作关系挖掘上，而对冷启动服务的特性鲜有研究，这制约了服务系统的健康发展。本文所述冷启动服务是指 Web 服务系统中无使用记录的已有服务和新增的新创建服务。当前，服务系统具有严重的长尾特性，而现有的服务推荐算法正加剧这种特性。以 ProgrammableWeb.com(http://www.programmableweb.com)服务系统为例，系统上共有 Web 服务 13 269 个，而被使用过的服务仅占 9%，有 91%的服务从未被使用过，体现了严重的长尾特性。在这种背景下，现有的服务推荐算法为了提升自身的推荐效果或协作关系预测效果，均着力对拥有使用记录的服务进行研究和推荐，这又反过来加剧了服务系统的长尾特性和服务使用的不平衡，形成了恶性循环。

服务系统的长尾特性大大制约了服务系统的发展。当服务系统具有严重的长尾特性时，其内部的服务发展极不平衡。一小部分服务被多次使用，而大量服务没有使用机会，这导致服务系统元素多样性下降，服务之间不能形成良性的协作、竞争关系，热门服务的排挤使得大量服务退出服务系统，这一方面导致用户的选择性变小，另一方面促使新增的服务提供者转而选择其他服务系统。此外，服务发展的不平衡会导致系统鲁棒性下降，个别服务的消亡可能导致大范围服务组合的不可用，而系统鲁棒性的持续下降极易引起系统整体的崩溃。

对冷启动服务协作关系挖掘的不足使得服务系统大量潜在资源被忽视，也不利于服务组合开发者进行最适服务筛选。因为当前研究对冷启动服务协作关系挖掘不足，用户在筛选服务时不能很好地辨识冷启动服务的特性和组合模式，所以难以对冷启动服务作出准确判断、组合和使用。事实上，冷启动服务中不乏大量能够满足用户需求的优质服务。上述现象导致服务系统中这些潜在的资源被忽视和浪费，也使服务组合开发者不能寻找到真正合适的服务。对冷启动服务协作关系研究的不足也加剧了服务系统的长尾特性。

为了解决上述问题，本文提出了一种冷启动服务协作关系挖掘与预测模型(mining and predicting model of cold start service collaboration relationship，CSCR)。CSCR 模型首先利用服务描述重构和功能主题分析为每个服务建立功能属性向量，进而基于服务历史协作关系和功能属性向量为非冷启动服务建立协作属性向量。最后，模型通过对冷启动服务功能属性向量与非冷启动服务协作属性向量进行相似度比较，完成冷启动服务协作关系的挖掘和预测。通过对冷启动服务协作关系的挖掘，CSCR 模型能够帮助用户了解冷启动服务的特性和组合模式，提高服务系统中冷启动服务的关注率和使用率，进而帮助提升服务系统的元素多样性和系统鲁棒性。本文在真实数据集上对模型预测效果进行了验证，实验结果表明 CSCR 模型在以 MAP(mean average precision)指标评估的综合预测效果上较当前最优模型提升了 5.2%。

1 问题定义

1.1 Web 服务系统

本文采用 7 元组 SS=(S，SD，ST，M，MD，MT，$\boldsymbol{R}$)来定义 Web 服务系统。其中：$S=\{s_1, s_2, \cdots, s_{\mathrm{SN}}\}$表示服务系统内服务的集合，SN=$|S|$表示服务数量；$\mathrm{SD}_i=\{w_{i1}, w_{i2}, \cdots, w_{im_i}\}$表示服务提供者为服务 i 撰写的描述，其中 w_{ik} 表示描述中的第 k 个单词；$\mathrm{ST}=\{t_{s1}, t_{s2}, \cdots, t_{s\mathrm{SN}}\}$表示各服务的发布时间；$M=\{m_1, m_2, \cdots, m_{\mathrm{MN}}\}$为服务组合集合，MN=$|M|$表示服务组合数量；$\mathrm{MD}_j=\{w_{j1}, w_{j2}, \cdots, w_{jn_j}\}$表示服务组合开发者为服务组合 j 撰写的描述，其中 w_{jk} 表示描述中的第 k 个单词；$\mathrm{MT}=\{t_{m1}, t_{m2}, \cdots, t_{m\mathrm{MN}}\}$表示各服务组合的创建时间；矩阵 $\boldsymbol{R}=(r_{ij})_{i=1,\,j=1}^{\mathrm{MN}\times\mathrm{SN}}$ 表示服务与服务组合的

历史使用关系，$r_{ij}=1$ 表示服务组合使用了服务 j，$r_{ij}=0$ 表示服务组合 i 没有使用服务 j。

1.2 冷启动服务协作关系预测问题

对时间点 t，把时间点 $t+1$ 新创建的服务和时间点 t 及以前创建的但没有使用记录的服务记为冷启动服务，表示为 $\mathrm{CS}_t=\{s_{Ct1}, s_{Ct2}, \cdots, s_{CtNCt}\}$；把时间点 t 及以前创建的并已经存在使用记录的服务记为非冷启动服务，表示为 $\mathrm{NS}_t=\{s_{Nt1}, s_{Nt2}, \cdots, s_{NtNNt}\}$。其中：$s_{Cti}$ 表示 t 时间点的第 i 个冷启动服务，$\mathrm{NC}t$ 表示 t 时间点冷启动服务的数量；s_{Nti} 表示 t 时间点的第 i 个非冷启动服务，$\mathrm{NN}t$ 表示 t 时间点非冷启动服务的数量。冷启动服务协作关系预测问题就是通过对已有信息的建模和挖掘，实现对未来冷启动服务可能的协作模式的预测，也就是预测在接下来的一段时间内可能与该服务共同形成服务组合的服务集合。

在具体操作上，本文算法为各个冷启动服务 $s_{Cti}\in\mathrm{CS}_t$ 给出一串排序服务列表 RL_{Cti}，RL_{Cti} 中位置越靠前的服务被认为在接下来的时间段 $[t+1, t+t']$ 内越倾向于和冷启动服务 s_{Cti} 构成服务组合。

2 CSCR 模型及其实现

图 1 为 CSCR 模型整体架构图。冷启动服务协作关系挖掘与预测模型(CSCR 模型)的实现过程可分为 3 个步骤：服务功能属性向量构建、服务协作属性向量构建、冷启动服务协作关系预测。

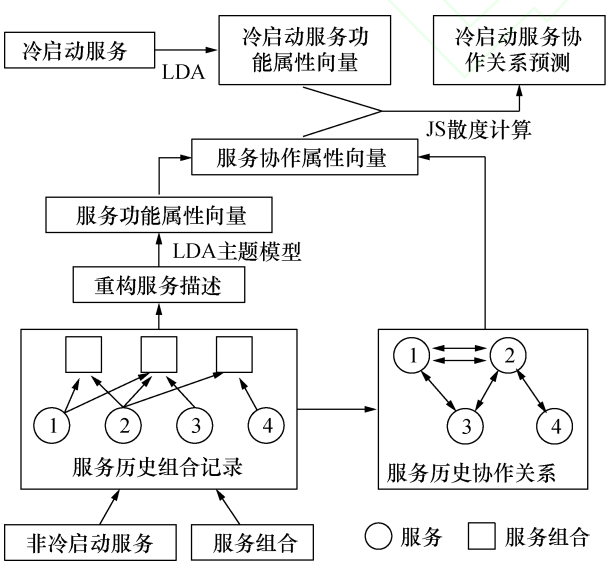

图 1 CSCR 模型整体架构

2.1 服务功能属性向量构建

为了实现对服务的建模和特征提取，本文算法首先为全部非冷启动服务建立功能属性向量，以实现服务特征的向量化。服务功能属性向量在功能属性层面描绘服务的特征，包含服务提供功能、服务适用应用场景等信息。

服务提供者在发布服务时会为服务撰写一份描述文档，该文档描述了服务所提供的功能，然而该原始服务描述并不能全面描述服务潜在应用场景信息[13]。在服务系统中，每一个服务都可以应用到多个应用场景之中，每一个服务组合对应于其所使用服务的一个应用场景，而服务组合的描述则详细描述了该应用场景的特征和信息。因此，通过对原始服务描述进行重构，引入服务组合描述，可以补充原始服务描述中缺失的应用场景信息，使得重构后的服务描述能够更全面地描述服务的特征，为服务功能属性向量的构建奠定基础。

利用服务系统中已有的服务历史组合记录和服务组合描述文档，对于每一个服务 $s_i\in\mathrm{NS}_t$，其服务描述重构过程如下：

$$\boldsymbol{H}_i=\sum_{j=1}^{\mathrm{MN}} r_{ji}\boldsymbol{P}_i+\sum_{j=1}^{\mathrm{MN}} r_{ji}\boldsymbol{Q}_j,\quad s_i\in\mathrm{NS}_t. \tag{1}$$

式中：$\boldsymbol{P}_i$ 和 $\boldsymbol{Q}_j$ 为标准化的词向量，分别由服务 s_i 的原始描述 SD_i 和服务组合 m_j 的描述 MD_j 转化而来。$\sum_{j=1}^{\mathrm{MN}} r_{ji}\boldsymbol{Q}_j$ 为补充应用场景信息的过程，只有调用了此服务的服务组合才会被引入进来。为了避免引入的服务组合描述过多而稀释了原始描述中蕴含的信息，本文算法为原始服务描述乘以了一个权重因子 $\sum_{j=1}^{\mathrm{MN}} r_{ji}$，即调用服务 s_i 的服务组合的个数。$\boldsymbol{H}_i$ 为得到的重构服务描述词向量。

LDA 主题模型[14]被认为是提取文本主题信息、实现文本向量化的有力工具。LDA 主题模型认为文本的生成过程是由 T 个不同权重的主题所指导的，不同的主题能够表达出不同的单词，最终形成文本。因此，文本的 T 个主题及其对应的权重能够表示文本更为本质的信息。为了实现重构服务描述文本的向量化和服务功能属性向量的构建，本文算法采用 LDA 主题模型对重构服务描述文本进行建模，并使用 Gibbs 采样[15]工具推断相关参数的取值。LDA 主题模型[14]和 Gibbs 采样[15]的具体实现过程为标准化过程。

LDA 主题建模过程为每一个重构服务描述建立了一个 $1\times T$ 的向量 $\boldsymbol{\theta}_i=(\theta_i^1, \theta_i^2, \cdots, \theta_i^T)$，表示服务 s_i 的主题分布，$\boldsymbol{\theta}_i$ 满足以下条件：

$$\sum_{t=1}^{T}\theta_i^t = 1, \quad \theta_i^t \in [0, 1]. \tag{2}$$

将 $\boldsymbol{\theta}_i$ 记为服务 s_i 的功能属性向量，表示服务 s_i 在功能属性层面的特征。

2.2 服务协作属性向量构建

在进行服务协作关系预测时，除在服务个体层面展开分析外，还需对服务与服务之间的协作模式进行挖掘。本节利用服务历史协作关系和服务功能属性向量，为非冷启动服务建立协作属性向量，从而实现对服务协作属性信息的挖掘和向量化描述。服务功能属性向量和服务协作属性向量分别从两个层面展示服务的信息，功能属性向量描述了服务的功能特性、应用场景信息等个体层面信息，而协作属性向量则是从服务之间的协作模式出发，描述了服务的协作、组合模式特征。这两个向量都描述服务的特征信息，但在描述层面和实际意义上却有根本性的差异。

服务系统中服务与服务组合的历史组合关系用矩阵 $\boldsymbol{R}=(r_{ij})_{i=1,\,j=1}^{\mathrm{MN}\times\mathrm{SN}}$ 表示，$\boldsymbol{R}$ 包含了服务的历史组合记录信息，也隐含了服务之间的历史协作关系信息。例如，当某两个服务共同构成了某一服务组合时，表示这两个服务构成了一次协作。通过对矩阵 $\boldsymbol{R}$ 进行变体操作，可以获得服务历史协作关系矩阵 $\boldsymbol{C}=(c_{ij})_{i=1,\,j=1}^{\mathrm{NN}t\times\mathrm{NN}t}$。$\boldsymbol{C}$ 为一对称矩阵，表明了各非冷启动服务 $s_i\in\mathrm{NS}_t=\{s_{\mathrm{N}t1}, s_{\mathrm{N}t2}, \cdots, s_{\mathrm{N}t\mathrm{NN}t}\}$ 之间的历史协作关系，$c_{ij}=c_{ji}=n$ 表示服务 s_i 与服务 s_j 在历史上曾协作过 n 次，即共同创建过 n 个服务组合。不考虑服务自身与自身的协作，即对任意 $i\in[1, \mathrm{NN}t]$，$c_{ii}\equiv 0$。

应用历史协作关系矩阵 $\boldsymbol{C}$ 和服务功能属性向量 $\boldsymbol{\theta}_i$，能够为各服务 $s_i\in\mathrm{NS}_t$ 构建服务协作属性向量，

$$\boldsymbol{\varphi}_i = \sum_{j=1}^{\mathrm{NN}t} c_{ij}\boldsymbol{\theta}_j \Big/ \sum_{j=1}^{\mathrm{NN}t} c_{ij}. \tag{3}$$

由式(3)可知，服务 s_i 协作属性向量的构建过程为将全部与之有过协作关系的服务的功能属性向量 $\boldsymbol{\theta}_j$ 进行加总和归一化，其中 $\boldsymbol{\theta}_j$ 在加总时采用加权方式，权重为 s_i 和 s_j 的协作次数 c_{ij}。

至此，本文算法为每个服务 $s_i\in\mathrm{NS}_t$ 构建了协作属性向量，记为 $\boldsymbol{\varphi}_i$，$\boldsymbol{\varphi}_i$ 也满足式(2)的约束。从服务协作属性向量的构建过程可以看出，$\boldsymbol{\varphi}_i$ 并不包含服务自身的功能信息，而其信息来源为与该服务拥有协作关系的服务的功能属性向量，$\boldsymbol{\varphi}_i$ 从服务协作属性的层面刻画了服务的特征。因而，当某一服务的功能属性向量与另一服务的协作属性向量相似度较高时，即当一个服务的功能属性与另一服务的历史协作服务的功能属性相似时，则这两个服务在以后更有可能产生协作、建立组合。

2.3 冷启动服务协作关系预测

CSCR 模型的输入为各冷启动服务的信息，输出为各冷启动服务协作关系预测结果，即可能与该冷启动服务产生协作的服务列表。为了实现这一过程，算法首先为各冷启动服务构建功能属性向量。

对于每一个冷启动服务 $s_i\in\mathrm{CS}_t$，其功能属性向量的构建过程与 2.1 节中 CSCR 模型的向量建构过程一致，即通过 LDA 模型对冷启动服务描述信息进行建模，并利用 Gibbs 采样工具推断相关参数取值，从而实现服务描述文档的主题特征提取和向量化。通过 LDA 模型得到的 $1\times T$ 主题分布向量 $\boldsymbol{\theta}_i$ 表示冷启动服务 s_i 在功能属性层面的特征，称 $\boldsymbol{\theta}_i$ 为冷启动服务 $s_i\in\mathrm{CS}_t$ 的功能属性向量。服务系统中存在少量特别的服务，它们也曾创建过服务组合，但其创建的服务组合均仅使用了该服务自己这一个服务。由于这些服务在历史上并未与其他服务形成协作，这些服务也被认定为冷启动服务，但是在进行冷启动服务功能属性向量构建前，与 2.1 节一致，本文算法也对它们进行服务描述重构，即为它们补充服务组合所包含的应用场景描述信息，之后的操作与其他普通冷启动服务没有区别。

本文算法对各冷启动服务 $s_i\in\mathrm{CS}_t$ 的功能属性向量 $\boldsymbol{\theta}_i$ 与所有非冷启动服务 $s_j\in\mathrm{NS}_t$ 的协作属性向量 $\boldsymbol{\varphi}_j$ 进行相似度比较，应用的相似度比较工具为 JS 散度(Jensen-Shannon divergence)[16]：

$$\mathrm{KLD}(\boldsymbol{\theta}_i \,\|\, \boldsymbol{\varphi}_j) = \sum_{t=1}^{T}\theta_i^t \lg(\theta_i^t/\varphi_j^t), \tag{4}$$

$$\boldsymbol{\theta}_m = (\boldsymbol{\theta}_i + \boldsymbol{\varphi}_j)/2, \tag{5}$$

$$\mathrm{JSD}(\boldsymbol{\theta}_i \,\|\, \boldsymbol{\varphi}_j) = (\mathrm{KLD}(\boldsymbol{\theta}_i \,\|\, \boldsymbol{\theta}_m) + \mathrm{KLD}(\boldsymbol{\varphi}_j \,\|\, \boldsymbol{\theta}_m))/2. \tag{6}$$

其中：$\mathrm{JSD}(\boldsymbol{\theta}_i \,\|\, \boldsymbol{\varphi}_j)$表示 $\boldsymbol{\theta}_i$ 与 $\boldsymbol{\varphi}_j$ 的 JS 散度值，$\mathrm{JSD}(\boldsymbol{\theta}_i \,\|\, \boldsymbol{\varphi}_j)$的值越小表示 $\boldsymbol{\theta}_i$ 与 $\boldsymbol{\varphi}_j$ 的相似度越高。

若冷启动服务的功能属性向量与非冷启动服务的协作属性向量相似度较高，则表明这两个服务在未来有更大的可能性形成组合服务。因而，对于各冷启动服务 $s_i\in\mathrm{CS}_t$，按照 $\mathrm{JSD}(\boldsymbol{\theta}_i \,\|\, \boldsymbol{\varphi}_j)$的值由低到高(即相似度由高到低)对非冷启动服务 $s_i\in\mathrm{NS}_t$ 进行排序，可以得到一列带有顺序的服务列表 $\mathrm{RL}_{\mathrm{C}ti}$，$\mathrm{RL}_{\mathrm{C}ti}$中位置越靠前的服务被认为在之后一段时间窗口内越有可能和冷启动服务 s_i 构成协作，即形成

服务组合。

3 实验设计与结果分析

3.1 实验数据集

本文选取 ProgrammableWeb.com 数据集对算法进行实验验证。ProgrammableWeb.com 数据集是当前最大的服务和服务组合在线资源库。本文作者爬取了其上自 2005 年 9 月至 2016 年 6 月的全部服务和服务组合数据，共计包含服务 13 269 个、服务组合 5 840 个，各类描述文本总词汇量为 21 891。

爬取到的服务数据包含服务名称、发布日期和服务描述等信息，服务组合数据包含服务组合名称、创建日期、调用服务列表和服务组合描述等信息。针对服务与服务组合的描述文本预先进行了分词、停止词去除、词干提取等预处理和数据清洗工作[17]，并将处理得到的描述文本存储为词向量的形式作为算法的输入，该词向量被视为原始描述。

3.2 评估指标与实验设置

实验采用在推荐及预测领域被广泛接受的评估指标 MAP@N(mean average precision @ top N)[13]来评估 CSCR 模型及各对比方法的预测效果，即评估输出的服务列表的准确性。MAP@N 的取值范围为[0，1]，其值越接近 1 表示算法的预测效果越好。MAP@N 指标中的 N 代表在验证时只对输出的服务列表的前 N 项的准确性进行评估，N 值取整个输出服务列表的长度时的 MAP 值记为 MAP@J。

将测试时间窗口的移动步长设置为 1 个月，即令测试时间窗口按月移动。当时间窗口移到某个月 t 时，记 $t+1$ 月新创建的服务以及 t 月及以前创建的但没有协作记录的服务为冷启动服务，作为对应的时间窗口的测试数据集；令 t 月及以前创建的已经存在协作记录的服务和各服务组合为训练数据集。通过对训练数据集的挖掘，实现对测试数据集中各冷启动服务未来协作服务的预测。在评估实验结果时，针对时间段$[t+1, t+12]$，即当前时间窗口 t 往后的 1 年以内，把这 1 年以内与测试数据集中的各冷启动服务真实建立了协作和组合关系的非冷启动服务的列表视为预测结果的真实值，并将其与算法输出的预测服务列表作比较分析，最终得出月份 t 的 MAP@N 评估指标。针对在时间段$[t+1, t+12]$中仍没有和其他服务建立协作关系的冷启动服务，在当次实验中不进行计算和考虑，但此类服务能够随着 t 的前进被之后的实验所检验。

在实验中，将时间点 t 从 2010 年 7 月按月前进到 2015 年 6 月，一共得到 60 个 MAP@N 评估结果。而后，依据各测试集冷启动服务数量对这 60 个 MAP@N 结果进行加权平均，并将该加权平均值作为各算法的 MAP@N 评估结果。实验过程实现了对近 6 年数据的按月评估，充分评估了 CSCR 模型及各对比方法的预测效果。

在实验参数设置上，对于 CSCR 模型，根据经验和参数寻优结果，设置 LDA 模型 Dirichlet 分布超参数 $\alpha=50/T$，$\beta=0.01$，LDA 模型主题数 $T=40$，Gibbs 采样迭代次数 $N_{\text{iter}}=1\,000$。对比方法的实验参数均设为各自的最优参数。

3.3 对比方法

本文共选取 5 种对比方法用以和所提出的 CSCR 模型进行比较分析。

3.3.1 服务流行度分析(SUF)

服务流行度分析(SUF)[18]方法认为服务历史组合频次信息代表服务的热度，而热度越高的服务越有可能与冷启动服务构建服务组合。SUF 方法的协作关系预测结果即为按照服务历史组合频次由高到低排序的结果。各非冷启动服务预测得分为

$$\mathrm{SUF}(s_{\mathrm{C}ti}, s_{\mathrm{N}ti}) = \frac{\sum_{j=1}^{\mathrm{NN}t} c_{ij}}{\sum_{i=1}^{\mathrm{NN}t}\sum_{j=1}^{\mathrm{NN}t} c_{ij}}. \tag{7}$$

3.3.2 服务端主题匹配(SDCM)

服务端主题匹配(SDCM)方法由文[19]提出，利用 LDA 模型对冷启动服务 $s_{\mathrm{C}ti}\in \mathrm{CS}_t$ 及非冷启动服务 $s_{\mathrm{N}ti}\in \mathrm{NS}_t$ 进行建模，并对服务描述信息进行主题分析，预测服务-主题概率分布 $p(t\mid s)$ 与主题-单词概率分布 $p(w\mid t)$。SDCM 方法认为当两个服务具有更高的相似性时，它们之间更有可能产生协作，因此算法通过比较服务间的主题相似性实现冷启动服务协作组合关系的预测。各非冷启动服务预测得分为

$$\mathrm{SDCM}(s_{\mathrm{C}ti}, s_{\mathrm{N}ti}) = \prod_{w\in \mathrm{SD}_{\mathrm{C}ti}} \sum_{t=1}^{T} p(w\mid t)\,p(t\mid s_{\mathrm{N}ti}). \tag{8}$$

3.3.3 服务组合端协同过滤(MDCF)

服务组合端协同过滤(MDCF)方法立足于原始的协同过滤模型，在服务推荐与预测领域被广泛认

可和使用[20]。MDCF 方法认为当服务的主题向量与服务组合的主题向量相似度较高时，该服务更有可能与该服务组合所使用的服务产生组合关系。MDCF 算法利用 LDA 模型对各冷启动服务和服务组合进行建模，获取相应描述文本的主题向量，并依据各冷启动服务与服务组合的主题向量相似度匹配，计算各非冷启动服务预测得分，

$$\mathrm{MDCF}(s_{Cti},\ s_{Nti})=\frac{\sum\limits_{m_j\in U(N,\ s_{Cti})}\mathrm{sim}(s_{Cti},\ m_j)r_{jNti}}{\sum\limits_{m_j\in U(N,\ s_{Cti})}\mathrm{sim}(s_{Cti},\ m_j)}. \tag{9}$$

其中：$U(N,\ s_{Cti})$为与冷启动服务 $s_{Cti}\in \mathrm{CS}_t$ 相似性最高的 N 个服务组合集，$\mathrm{sim}(s_{Cti},\ m_j)$计算冷启动服务 s_{Cti} 与服务组合 m_j 间的主题向量的余弦相似度。

3.3.4 基于服务使用记录与类别协作记录的协同过滤(CCF)

基于服务使用记录与类别协作记录的协同过滤(CCF)方法由文[21]提出。CCF 方法一方面对服务系统中的服务历史使用记录进行建模，并认为服务更有可能与和自身相似的服务组合所包含的服务形成组合；另一方面，通过对服务历史协作关系开展挖掘，构建服务类别协作关系矩阵，并认为当两个服务的所属类别在历史上曾多次协作时，这两个服务将有更大的可能性构建服务组合。CCF 方法基于协同过滤思想从以上两个层面预测冷启动服务的协作关系，并将两个层面的预测结果进行合理统一。CCF 方法在进行相似性比较等相关操作时，除利用文本描述外，也将标签信息纳入考虑。CCF 方法的详细介绍、计算过程和参数设置等可参阅文[21]。CCF 方法在实际应用中取得了很好的预测效果，被广泛视为当前最先进的冷启动服务协作关系预测方法之一。

3.3.5 省略服务描述重构的 CSCR 模型(N-CSCR)

省略服务描述重构的 CSCR 模型(N-CSCR)是本文提出的 CSCR 模型的变体。与 CSCR 模型相比，N-CSCR 模型省略了服务功能属性向量构建前的服务描述重构过程，直接使用原始服务描述文档进行 LDA 主题建模和功能属性向量构建，N-CSCR 模型的其他环节和细节均与 CSCR 模型一致。

通过与 N-CSCR 模型进行比较，可以验证对原始服务描述进行重构，即补充服务组合描述中所含的场景信息，是否能够改善服务描述的质量和信息含量，从而使得服务功能属性向量获得更好的提取。

3.4 实验结果分析

图 2 所示为 CSCR 模型及各对比方法的 MAP@N 实验结果，横坐标表示 N 的不同取值，即对输出的预测服务列表的前 N 项的准确性进行评估，纵坐标表示各算法的 MAP@N 实验结果。

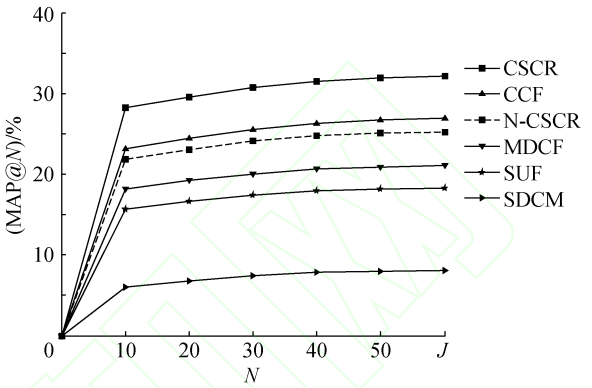

图 2 各预测模型 MAP@N 结果

SDCM 模型从语义匹配的角度出发，认为当两个服务主题相似性较高时，它们倾向于形成协作关系。然而在实际中，构成协作关系的服务往往不是相似的服务，而是能够形成功能互补的服务，而两个功能互补的服务在主题特征上可能不具有相似性，甚至完全相反。另外，SDCM 模型仅考虑了服务描述相关信息，而对于服务系统中的组合记录、协作记录等并未加以考虑，使得模型不能充分利用服务系统的信息。因此，SDCM 模型的预测效果在各方法中列于末位。

SUF 方法认为服务历史组合频次信息代表服务的热度，而热度越高的服务越有可能与冷启动服务构建新的协作关系。实验中证明 SUF 方法能够取得一定的效果。但是，SUF 方法在本质上仍然是利用服务分布的不均衡性展开预测，而这种模式会进一步加剧服务系统的长尾特性，这与本文进行冷启动服务特性挖掘的目标不符。并且，SUF 方法在预测结果上不具有多样性，实际操作意义不大。

MDCF 模型是典型的基于协同过滤思想构建的模型，能够利用服务系统中的服务历史组合记录信息和服务组合端的描述信息。在实验中，MDCF 模型取得了强于 SDCM 模型和 SUF 方法的预测效果，这证明了协同过滤思想在解决该问题时的有效性。然而，MDCF 模型没有对服务的协作、组合关系进行深入挖掘，这导致其预测效果仍落后于其他算法。

CCF 方法分别对服务历史使用记录和服务类别协作关系进行建模，综合利用服务的描述信息和标签信息，并应用协同过滤手段对冷启动服务协作

关系展开预测。然而，CCF 方法仍依赖于传统的协同过滤思想，对于服务的协作、组合属性挖掘不够深入和具体，因此其预测效果虽显著优于其他对比方法，但仍落后于 CSCR 模型。CCF 方法在实际应用中取得了很好的预测效果[21]，并被视为当前最优的冷启动服务协作关系预测算法之一，而本文提出的 CSCR 模型的预测效果显著优于 CCF 方法，证明 CSCR 模型较当前最佳模型在预测效果上有显著提升。

N-CSCR 模型是 CSCR 模型的变体，它省略了服务功能属性向量构建前的服务描述重构过程，直接应用原始服务描述作为后续流程的输入。实验结果表明，其预测效果差于 CCF 方法与 CSCR 模型。通过与 N-CSCR 模型的对比，能够证明对原始服务描述进行重构，即补充服务描述中缺失的场景信息，能够增加服务描述的信息含量，从而使得服务功能属性向量获得更好的提取。

CSCR 模型克服了当前研究对冷启动服务协作组合关系探索的不足，显著提升了预测效果。实验结果表明，CSCR 模型的 MAP@N 结果明显高于所有对比方法，这证明了 CSCR 模型的有效性。

表 1 中列出了 CSCR 模型及各对比方法的具体实验结果。比较 CSCR 算法和 CCF 算法结果可知，CSCR 算法预测效果比当前最优方法提升了 5.2%，这得益于对服务协作属性关系的挖掘。比较 CSCR 算法和 N-CSCR 算法结果可知，为原始服务描述补充场景信息能够将预测效果增强约 6.9%。

表 1 CSCR 模型及各对比方法 MAP@N 实验结果

预测方法	实验结果/%		
	MAP@10	MAP@30	MAP@J
CSCR	28.21	30.73	32.10
CCF	23.13	25.56	26.90
N-CSCR	21.82	24.09	25.17
MDCF	18.19	20.02	21.03
SUF	15.69	17.37	18.23
SDCM	5.99	7.44	8.11

4 结 论

针对 Web 服务系统中各服务发展不平衡、冷启动服务协作关系挖掘不足等问题，本文提出了一种冷启动服务协作关系挖掘与预测(CSCR)模型。CSCR 模型对服务系统内服务历史组合记录和服务历史协作关系进行挖掘建模，利用服务描述重构和功能主题分析等方法，从服务的功能性特征和协作性特征两个角度出发，为服务建立功能属性向量和协作属性向量，进而通过对冷启动服务功能属性向量与非冷启动服务协作属性向量进行相似性比较，实现冷启动服务协作关系的挖掘与预测。真实数据集上的实验结果证明本文方法的预测效果显著强于当前最优方法。

CSCR 模型实现了对 Web 服务系统中冷启动服务协作关系的挖掘和预测，从而能够帮助用户了解冷启动服务的特性和组合模式，提高服务系统中冷启动服务的关注率和使用率，进而帮助增强服务系统的元素多样性和系统鲁棒性。

在下一步研究中，将对 Web 服务系统随时间的演化特性加以建模和分析，将服务功能属性特征和协作属性特征随时间推移而可能发生演化的特性纳入考虑，从而揭示冷启动服务协作关系在不同时期的动态特征和演化特性，进而提升冷启动服务协作关系预测的时效性和准确性。

参考文献 (References)

[1] 辛乐，范玉顺. 考虑协作方的服务组合分析[J]. 清华大学学报(自然科学版)，2015，55(5)：538－542，549.
XIN L，FAN Y S. Service composition analysis with collaboration [J]. Journal of Tsinghua University (Science and Technology)，2015，55(5)：538－542，549. (in Chinese)

[2] XU W，CAO J，HU L，et al. A social-aware service recommendation approach for mashup creation [J]. International Journal of Web Services Research，2013，10(1)：53－72.

[3] YAO L，SHENG Q Z，NGU A H H，et al. Unified collaborative and content-based web service recommendation [J]. IEEE Transactions on Services Computing，2015，8(3)：453－466.

[4] CAO B，LIU J，TANG M，et al. Mashup service recommendation based on user interest and social network [C]// Proceedings of the 20th IEEE International Conference on Web Services. Santa Clara，USA，2013：99－106.

[5] GAO Z，FAN Y，WU C，et al. SeCo-LDA：Mining service co-occurrence topics for recommendation [C]// Proceedings of the 23th IEEE International Conference on Web Services. San Francisco，USA，2016：25－32.

[6] DONG X，HALEVY A，MADHAVAN J，et al. Similarity search for web services [C]// Proceedings of the 13th International Conference on Very Large Data Bases. Toronto，Canada，2004，30：372－383.

[7] MENG S, DOU W, ZHANG X, et al. KASR: A keyword-aware service recommendation method on MapReduce for big data applications [J]. IEEE Transactions on Parallel and Distributed Systems, 2014, 25(12): 3221-3231.

[8] ZHANG Y, LEI T, WANG Y. A service recommendation algorithm based on modeling of implicit demands [C]// Proceedings of the 23th IEEE International Conference on Web Services. San Francisco, USA, 2016: 17-24.

[9] LEI Y, ZHOU J, ZHANG J, et al. Time-aware semantic web service recommendation [C]// Proceedings of the 12th IEEE International Conference on Services Computing. New York, USA, 2015: 664-671.

[10] CAO B, LIU X, RAHMAN M, et al. Integrated content and network-based service clustering and web APIs recommendation for mashup development [J/OL]. IEEE Transactions on Services Computing, 2017. DOI: 10.1109/TSC.2017.2686390.

[11] LIANG Q A, CHUNG J Y, MILLER S, et al. Service pattern discovery of web service mining in web service registry-repository [C]// Proceedings of the IEEE International Conference on E-Business Engineering. Shanghai, 2006: 286-293.

[12] HUANG K, FAN Y, TAN W. Recommendation in an evolving service ecosystem based on network prediction [J]. IEEE Transactions on Automation Science and Engineering, 2014, 11(3): 906-920.

[13] HAO Y, FAN Y, TAN W, et al. Service recommendation based on targeted reconstruction of service descriptions [C]// Proceedings of the 24th IEEE International Conference on Web Services. Honolulu, USA, 2017: 285-292.

[14] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022.

[15] PORTEOUS I, NEWMAN D, IHLER A, et al. Fast collapsed Gibbs sampling for latent Dirichlet allocation [C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, 2008: 569-577.

[16] LIN J. Divergence measures based on the Shannon entropy [J]. IEEE Transactions on Information Theory, 2002, 37(1): 145-151.

[17] HUANG K, YAO J, FAN Y, et al. Mirror, mirror, on the web, which is the most reputable service of them all? [C]// Proceedings of the 11th International Conference on Service-Oriented Computing. Berlin, Germany: Springer, 2013: 343-357.

[18] CREMONESI P, PICOZZI M, MATERA M. A comparison of recommender systems for mashup composition [C]// Proceedings of the 3rd International Workshop on Recommendation Systems for Software Engineering. Zurich, Switzerland, 2012: 54-58.

[19] LI C, ZHANG R, HUAI J, et al. A probabilistic approach for web service discovery [C]// Proceedings of the 10th IEEE International Conference on Services Computing. Santa Clara, USA, 2013: 49-56.

[20] ZHENG Z, MA H, LYU M R, et al. WSRec: A collaborative filtering based web service recommender system [C]// Proceedings of the 16th IEEE International Conference on Web Services. Los Angeles, USA, 2009: 437-444.

[21] ZHANG J, FAN Y, TAN W, et al. Recommendation for newborn services by divide-and-conquer [C]// Proceedings of the 24th IEEE International Conference on Web Services. Honolulu, USA, 2017: 57-64.

(责任编辑 李丽)