文章编号：2095－0411（2014）02 - 0022 - 04

云环境下基于资源类别聚合的算法*

徐守坤，庄福宝
（常州大学 信息科学与工程学院，江苏 常州 213164）

摘要：随着云服务类型和数量不断增长，用户很难从中选择有效的云服务。为解决云环境下海量服务的个性化推荐问题，提出了一种基于类别聚合的个性化推荐算法。首先对数据存储节点上的资源进行分类；然后计算类别之间的相关性；其次寻找资源的最近邻；最后产生推荐集。通过实验数据进行验证，提出的云环境下的协同过滤算法与传统协同过滤算法相比，推荐质量和系统性能都有很大提高。

关键词：云环境；资源分类；类别聚合；个性化推荐

中图分类号：TP 301.6　　**文献标识码**：A　　**doi**：10.3969/j.issn.2095－0411.2014.02.007

Research of Resource Aggregation Algorithm Based on Cloud Environment

XU Shou-kun，ZHUANG Fu-bao
（School of Information Science and Engineering，Changzhou University，Changzhou 213164，China）

Abstract：With the growing of numbers and types of cloud services，user are faced with the issue of how to choose the best cloud service. In order to solve the problem of personalized recommendation of cloud environment，this paper presents a cloud service recommendation algorithm based on category aggregation. Firstly，the resources on the data storage node is classified；secondly，the correlation between the categories is calculated；thirdly，a search of the nearest neighbor of the resources is made；finally，the user recommendation sets are created. Validated by the experiment data，the collaborative filtering algorithm based on cloud computing in this paper，compared with the traditional collaborative filtering algorithm，there has been great improvement in the recommendation quality and system performance.

Key words：cloud environment；resource classification；category aggregation；personalized recommendation

随着网络的普及和信息的爆炸式增长，人们对海量数据的计算和存储的要求越来越高。为了能够有效地存储和计算海量的数据，在网格计算、并行计算和分布式计算等相关技术的基础上产生了云计算。随着云中服务类型和数量不断增长，用户很难从海量的数据中快速的选择有效地数据。即云计算中存在着这样的问题，面对海量数据用户不能有效选择最佳云服务[1-3]。

云计算环境下服务推荐涉及的对象与传统的服务推荐研究涉及的对象相比较，除了包括用户终端、服务和服务供应商 3 个部分，还包括基础设施供应商部分[4]。由于云环境下的服务推荐系统结构的特点，使传统服务推荐算法的研究成果并不能直接应用于云计算环境。而随着云端服务类型和数量的持续增长，如何有效的选择和推荐服务成为个性

* **收稿日期**：2013 - 08 - 30。
作者简介：徐守坤（1972－），男，吉林蛟河人，教授，博士，主要从事计算机应用、软件开发、智能空间等研究。

化推荐领域面临的重要问题[5]。

为解决个性化推荐问题，有很多学者提出了不同的改进方法。例如，文献［6］增加了对用户评分矩阵的预处理；文献［7－8］对属性进行加权，提高了属性相似度的精确；文献［9］引入用户对项目关注因子来修正原始相似性计算。文献［10］通过对用户进行聚类处理来实现个性化推荐。

文献［6－9］对传统的个性化推荐算法中的步骤进行了不同的改善。其优点是可以不同程度地改善个性化推荐的质量，但是，由于它没有对用户和资源信息进行筛选，所以仍存在推荐效率低等问题。文献［10］是通过用户的聚类处理实现了对用户信息的筛选，减少了要处理的用户和资源的信息，提高了个性化推荐的效率。但是，该算法是以用户进行分类，而用户是灵活多变的，这就给用户聚类处理带来了困难，从而影响个性化推荐的质量和效率。针对以上问题，本文提出了一种基于资源类别聚合的个性化推荐算法。实验结果证明：在云计算服务推荐系统中，本文所提算法能够有效而快速的为用户提供个性化推荐。

1 云环境下个性化推荐算法

1.1 资源的分类

在云计算环境下，数据存储节点存储着海量数据，而用户评价信息却很稀少且大部分都集中在几个类别，那么如何有效的利用这些少量且分布不均的用户评价信息来实现个性化推荐就成了研究的重点。为实现云环境下的个性化推荐，现将每个数据节点上的资源都按照相同的标准划分到各个类型之中[11]。本文基于这样一种假设：如果用户对此类别中的某些资源感兴趣，那么他也对此类别下的其他资源更有可能感兴趣。这样就可以以用户感兴趣的资源是否属于此类来判断用户是否对此类中的其他资源感兴趣。如果用户对某类别感兴趣，就在此类别下处理与目标用户有相同爱好的用户信息。如果用户对某类别不感兴趣，则不处理。因此，当对目标用户进行个性化推荐时，就可以降低处理资源和用户的数量，实现有效而迅速的个性化推荐。

对数据节点上资源的分类，可以根据资源的属性资源所处的节点、发布时间、是否更新等信息为标准进行分类。如果某资源所属类别与现有类别都不相同，则将其划入新的类别。如果某资源所属类别与多个类别相同，则将其划入到各个类别。设对资源进行分类的类别集合为：$C=\{C_1, C_2 \cdots C_{|C|}\}$，其中 C_i 代表一个类别；资源在类别集合中的概率向量为：$\boldsymbol{p}=(p(d_i \mid C_1), p(d_i \mid C_2) \cdots p(d_i \mid C_{|C|}))$，其中 d_i 代表资源 i 的加权值，$p(d_i \mid C_j)$ 表示资源 i 在类别 C_j 中的概率，用 p_{ij} 表示。那么云端资源在类别集合 C 中的概率矩阵为：

$$\boldsymbol{P}=\begin{bmatrix} p_{11} & \cdots & p_{1j} & \cdots & p_{i|C|} \\ \vdots & & \vdots & & \vdots \\ p_{i1} & \cdots & p_{ij} & \cdots & p_{i|C|} \\ \vdots & & \vdots & & \vdots \\ p_{n1} & \cdots & p_{nj} & \cdots & p_{n|C|} \end{bmatrix}$$

1.2 类别的相关性

1.2.1 资源类别的关联度

$$P_{\mathrm{CoA}}(C_i, C_j)=\frac{|U(C_i) \cap U(C_j)|}{|U(C_i) \cup U(C_j)|}$$

上式表示 2 个类别 C_i 和 C_j 间的关联度，其值越大则类别关联度越大。其中 $|U(C_i) \cap U(C_j)|$ 是 2 个类别共有的资源数，$|U(C_i) \cup U(C_j)|$ 是 2 个类别总的资源数。

1.2.2 资源类别的相似度

除了考虑资源类别的关联度，还要计算资源类别的相似度。本文使用类别中心的皮尔逊积矩相关系数（PPMCC）进行度量，则两个类别间的相似性 $\mathrm{sim}(C_i, C_j)$ 为：

$$\mathrm{sim}(C_i, C_j)=\frac{\sum_{s \in \Omega}(C_i(s)-\overline{C_i}) \cdot (C_j(s)-\overline{C_j})}{\sqrt{\sum_{s \in \Omega}(C_i(s)-\overline{C_i})^2} \cdot \sqrt{\sum_{s \in \Omega}(C_j(s)-\overline{C_j})^2}}$$

令 $\Omega=S(C_i) \cap S(C_j)$，

$\Psi=S(C_i) \cup S(C_j)$

式中：$S(C_i)$ －类别 C_i 中已评用户集合，$C_i(s)$ －类别 C_i 中心对用户 s 的评分，$\overline{C_i}$－类别 C_i 中心对所有用户评分。

因为 PPMCC 仅考虑两个类别 C_i 和 C_j 中共同评价服务的相似性，因此，当两个类别仅有少量共同评价资源时，它通常高估两者的相似性。此处对上式进行如下修正：

$$\mathrm{sim}'(C_i, C_j)=\frac{|\Omega|}{|\Psi|} \cdot \mathrm{sim}(C_i, C_j)$$

式中：$|\Omega|$－2 个类别 C_i 和 C_j 共同评价的用户数量，$|\Psi|$－2 个类别 C_i 和 C_j 评价的非重复的总用户数量。

1.2.3 资源类别的相关性

类别相关性概率 $P_{i,j}$ 是由侧重于类别中已评用户的类别相似度和侧重于类别资源的类别关联度的乘积，为：

$$P_{i,j}=\mathrm{sim}'(C_i, C_j)\cdot P_{\mathrm{CoA}}(C_i, C_j)$$

1.3 生成资源的最近邻

设资源 i 所属类别集合为 C_i，资源 j 所属类别集合为 C_j，则资源 i 和资源 j 的相似性为：

$$\mathrm{sim}(i, j)=\sum_{m\in C_i}\sum_{n\in C_j}\frac{P_{m,n}\cdot p_{im}\cdot p_{jn}\cdot (r_{u,i}-\overline{r_i})(r_{u,j}-\overline{r_j})}{\sqrt{\sum_{u\in\Omega}p_{im}^2(r_{u,i}-\overline{r_i})^2}\cdot\sqrt{\sum_{u\in\Omega}p_{jn}^2(r_{u,j}-\overline{r_j})^2}}$$

式中：$P_{m,n}$－类别 m、n 的相关性，p_{im}－项目 i 在其所属类别 C_m 中的概率，$\overline{r_i}$－项目的平均评分。

通过上面公式，求出资源 i 的最近邻 $S_{i,\mathrm{n}}$。

1.4 产生推荐集

1.4.1 基于项目的评分预测

对文献［12］中的用户 u 对未评项目 i 的预测评分公式进行改进，引入了用户 u 对项目 i 的权值 w_{ui}，如下所示：

$$P_{ui}=\overline{r_i}+\frac{\sum_{i\in S_{i,\mathrm{n}}}w_{ui}\cdot\mathrm{sim}(i, j)\cdot(r_{u,j}-\overline{r_j})}{\sum_{v\in S_{i,\mathrm{n}}}w_{ui}\cdot\mathrm{sim}(i, j)}$$

1.4.2 top - N 推荐集的产生

预测用户 u 对未评项目 i 总的评分，选择偏爱度比较高的的前 N 个项目作为 top - N 推荐集。

2 实验结果与分析

2.1 实验条件

实验采用 MovieLens 站点提供的数据集（http：//mvielens. umn. edu/），选取包含 500 位用户和对 1000 部电影的评分数据作为数据集。其中，数据集中的 400 名用户作为训练集，剩余 100 名作为测试集。根据用户信息对训练集中的用户进行分类，同时根据电影信息按类别对电影进行分类。

如图 1 所示，云平台的架构是由 4 台机器构建而成，其中每台机器都安装了 Linux 环境，让其中 1 台做 Master（名称）节点，另外 3 台做 Slaver（数据）节点。用路由器连接这 4 台机器使其彼此相通，并实现相互通信和传输数据。另外，还可以通过路由器来访问 Internet 来采集网页文档。

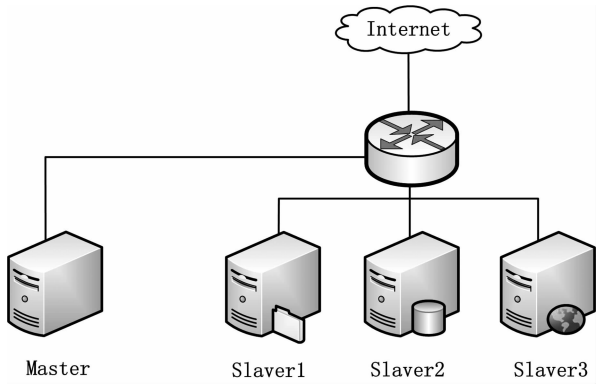

图 1 云平台的架构

Fig. 1 The architecture of cloud platform

2.2 评价标准

本文用 MAE（Mean Absolute Error）来衡量推荐的质量。MAE 是所有用户的预测评分与实际评分差的绝对值的和的平均。与平均误差相比，MAE 是不会出现正负相抵消的情况，因此它可以直观地反映个性化推荐的质量。由此可知，MAE 值越小，表明预测误差越小，即推荐质量越好。设用户预测评分集 $\{p_1, p_2\cdots p_N\}$ 和实际评分集 $\{q_1, q_2\cdots q_N\}$，则平均绝对偏差为[16]：

$$\overline{E}_{\mathrm{A}}=\frac{\sum_{i=1}^{N}|p_i-q_i|}{N}$$

2.3 实验结果和分析

实验 1：本实验随机的移除训练矩阵中一定数量的评价值，使当前用户所提供的评价条目数为 20，而被移除的这些评价值则被用来作为期望值以对预测性能进行研究。实验方案如表 1 所示，其中，Name 为方案名称，Density 为用户评分矩阵的密度，Users 为训练集用户数。

表 1 实验方案

Table 1 The experimental scheme

Name	a1	a2	a3	a4
Density/%	15	15	30	30
Users	300	400	300	400

4 次测试的结果如图 2 所示。从这 4 次测试结果的两两比较可以看出，用户评分矩阵密度大或训练集用户数越多的，其个性化推荐的质量越好，即其值越小。同时也可以看出，在每一次测试中，本文所提算法比传统的协同过滤算法和都有更好的推荐质量，即其值更低。其中有两个原因：①本文所提算法对云中数据进行类别的聚合并根据类别之间

的相关性对所求资源的最近邻进行加权。这样就减少了个性化推荐过程中要处理的用户和资源的数量，为推荐效率的提高作了铺垫。②通过对资源所属类别的详细讨论，并求出基于资源的相似度的最近邻，为最后 top - N 推荐集的产生作了准备并提高了推荐的精度。

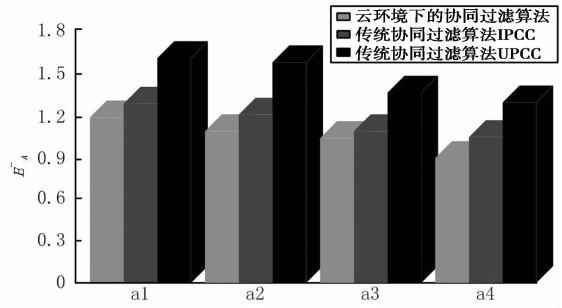

图 2　不同算法的推荐的精度

Fig. 2　The precision of different algorithms

实验 2：本实验是将云计算环境下协同过滤系统与非云计算环境下传统的协同过滤系统进行推荐效率的比较。为了将两者系统进行比较，实验样本采用实验一中 a1～a4 方案中的数据。具体实验方法是分别在云计算环境下和非云计算环境下对 4 个样本进行个性化推荐，并记录下个性化推荐所耗时间。

实验结果如表 2 所示，通过节省时间的比例可以发现，本文所提的云环境下的协同过滤系统在个性化推荐效率上要优于传统的协同过滤系统，主要原因为：在本实验中传统的协同过滤系统利用 1 台电脑来处理用户和资源信息，即单进程的处理用户和资源信息，而云计算环境下的协同过滤系统却是由 1 个 Master 节点和 3 个 Slaver 节点组成的分布式集群——云平台来处理用户和资源信息的，即并行的处理用户和资源信息。除此之外，云计算这种系统架构就是用来处理海量数据信息的，通过实验结果可知，随着用户和资源的信息的增长，云环境下的协同过滤推荐系统的推荐效率优势就越明显，即节省时间比例在不端增加。

表 2　不同计算量下的运算耗时

Fig. 2　Different calculation amount of computational time

项目名称	a1	a2	a3	a4
云环境计算下	1.850	2.216	3.105	4.216
非云环境计算下	4.526	6.035	9.025	13.085
节省时间比例	0.591	0.632	0.656	0.678

3　结　论

面对着云中海量的数据，用户面临着如何有效地选择最佳服务的问题。为解决云环境下的个性化推荐的问题，本文首先对存储节点中的资源进行类别聚合，然后求出基于项目的用户最近邻，最后产生 top - N 推荐集。通过资源分类和类别的相关性等一系列的处理，提高了个性化推荐的精度。同时，由于云计算的并行计算的特性，个性化推荐的效率也得到了很大的提升。试验结果证明，该算法能够提高个性化推荐的精度和效率。为了简化研究的复杂性，在不失一般性的条件下，本文将服务和商业服务供应商假定为一个对象。未来的工作则主要是深入探索更优的资源分类和类别聚合方法以及构建大规模的云环境下的协同过滤推荐系统中存在的问题。

参考文献：

[1] Schafer J B, Konstan J A, Riedl J. E - commerce recommendation applications [J]. Data Mining and Knowledge Discovery, 2001, 5 (1 - 2): 115 - 153.

[2] [2] Keunho Choi, Yongmoo Suh. A new similarity function for selecting neighbors for each target item in collaborative filtering [J]. Knowledge - based systems, 2013, 37 (1): 146 - 153.

[3] Gediminas Adomavicius, Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the State -of-the-art and possible extensions [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17 (6): 734 - 749.

[4] Zhang L J, Zhang J, Cai H. Service Computing [M]. Beijing: Tsinghua University Press, 2007: 7 - 8.

[5] 汪静，印鉴. 一种优化的 Item - Based 协同过滤推荐算法 [J]. 小型微型计算机系统，2010，31 (12)：2337 - 2342.

[6] 潘托宇，朱珍民. 一种改进的基于协同过滤的个性化推荐算法 [J]. 微计算机信息，2010，26 (123)：228 - 229.

[7] 陈志敏，姜艺. 综合项目评分和属性的个性化推荐算法 [J]. 微电子学与计算机，2011，28 (9)：186 - 189.

[8] 朱丽中，徐秀娟，刘宇. 基于项目和信任的协同过滤推荐算法 [J]. 计算机工程，2013，39 (1)：58 - 66.

[9] 张忠平，郭献丽. 一种优化的基于项目评分预测的协同过滤推荐算法 [J]. 计算机应用研究，2008，25 (9)：2658 - 2660.

[10] 赵玉艳，谷胜伟. 一种面向云计算环境的服务推荐算法 [J]. 巢湖学院学报，2012，14 (3)：42 - 47.

[11] 徐义峰，陈春明，徐云青. 一种基于分类的协同过滤算法 [J]. 计算机系统应用，2007 (1)：47 - 50.

[12] Resnick P, Iacovou N, Suchak M, et al. GroupLens: An open architecture for collaborative filtering of netnews [C] // Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work. New York: ACM Press, 1994: 175 - 186.