一种基于协同过滤的网格门户推荐模型

方 娟 梁文灿
(北京工业大学计算机学院 北京 100124)

摘 要：随着网格技术的不断发展，作为网格资源管理接口的网格门户也迅速发展起来。访问网格门户的用户数和门户管理的资源数也越来越多。为了解决网格门户系统资源管理信息规模过载、服务器大规模查询和处理资源负载较高、用户获取所需资源的满意度较低等问题，该文通过分析网格门户的主要功能和特点，在结合现有协同过滤推荐算法并改进的基础上，提出了基于协同过滤的网格门户推荐模型。推荐模型包括协同过滤交互模型、处理模型和展现模型。在设计模型的基础上提出了二次组合协同推荐算法并且进行了和现有算法的比较工作。实验表明：该文提出的推荐模型可以较好地实现网格门户的个性化协同推荐功能，并且可以保证个性化推荐的准确度和质量。

关键词：网格计算；网格门户；门户协同；协同过滤；个性化推荐
中图分类号：TP309 **文献标识码**： A **文章编号**：1009-5896(2010)07-1585-06
DOI: 10.3724/SP.J.1146.2009.00916

A Grid Portal Recommendation Model Based on Collaborative Filtering

Fang Juan Liang Wen-can
(*College of Computer Science*, *Beijing University of Technology*, *Beijing* 100124, *China*)

Abstract: Grid portal is playing an important role in grid computing area. However, with the large-scale diverse resource need to be dispatched and coordinated, the portal shows its insufficient ability to deal with this complex situation and can not bear the overload of long-time transaction querying. What is worse, the users can not get their desired or expected resources. To solve these problems, grid portal recommendation architecture is presented. It consists of collaborative filter interaction layer, action layer and user render layer. In addition, a 2-way combo collaborative filter algorithm is put forward, and then the algorithm comparison is shown. Finally the experiment results improve that this architecture can be used to obtain the expected portal recommendation function and guarantee quality of personalized recommendation.

Key words: Grid computing; Grid portal; Portal collaboration; Collaboration filtering; Personalized recommendation

1 引言

随着网格[1]应用能力不断加强，加入的网格节点会越来越多，各网格分支站点，分支工作站数量和多种分支资源数量可能会成几何倍数增长，这时对于门户[2,3]来说需要发现、管理和监视的资源越来越多。与此同时由于网格资源的异构、透明等特点，肯定存在着大量处理作用和能力大体相同的情况，普通门户用户为了找到相应的网格资源去执行某个特定任务可能要进行大量的搜索活动。这个盲目的基于客户端的搜索肯定会对服务器构成一定的负载压力，性能会有所下降并间接影响其服务于其它用户，而且不利于基于服务器端的网格门户后台统一调度。

广大学者在网格门户的理论和应用研究也是越来越深入。国内外研究方向主要涉及到门户体系、门户单点登录[4]、处理和监视作业[5]、发现和注册网格资源、网格 FTP 和网格门户安全等方面。与此同时网格门户协同推荐的研究在国内还刚刚起步，也就是说基于协同过滤网格门户推荐模型的研究还是很少。目前的门户系统没有考虑到资源的协同推荐功能，即不能提供主动式协同推荐服务。本文在利用现有协同过滤算法和网格门户的基础上提出一种基于协同过滤网格门户推荐模型，通过该模型设计能够保证门户执行效率和效果，实现门户的协同推荐功能，并且保证个性化推荐的准确度和质量。

2009-06-23 收到，2009-12-30 改回
北京市优秀人才基金(Q0007013200801)，国家自然科学基金(60873145)和国家 973 计划项目(2007CB311100)资助课题
通信作者：方娟 fangjuan@bjut.edu.cn

2 网格门户推荐模型

根据现阶段网格门户的需求和特点，充分考虑了将来网格门户发展的趋势，为了合理利用网格资源，在结合现有的协同过滤推荐算法[6]并加以改进的基础上，本文提出了一种基于协同过滤的网格门户推荐模型，并就整体协同推荐框架结构、与门户其它重要构件的协同交互、协同过滤处理层设计、协同过滤用户展现层设计等关键技术进行了较为详细的分析。该网格门户推荐模型充分考虑到了网格资源推荐准确性、推荐实时性、推荐速度以及推荐安全性等问题。该模型的设计可以很好地提高网格门户的使用率、减轻网格门户负载压力和增强后台资源的利用率。

2.1 协同过滤交互模型

设计网格门户推荐模型的重点在于如何将协同过滤模块与现有网格门户各模块安全高效地无缝结合，并且能提供给网格用户一个较为准确的资源访问推荐。该设计体系示意图如图 1 所示。

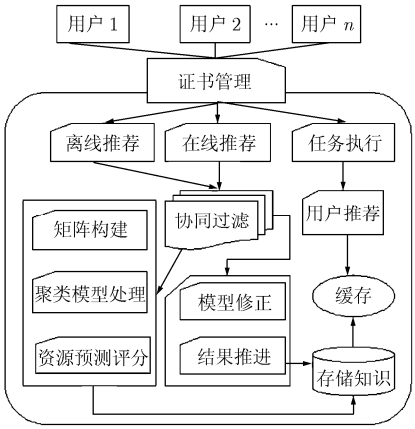

图 1 基于协同过滤的门户推荐系统模型

(1)整体交互　图 1 主要包括 3 部分：网格用户、存放 Portlet[7]的 Portal Server、资源推荐知识库。第 1 部分指网格用户。在门户中其指的是需要登录进门户系统进行实际操作的普通用户；第 2 部分和第 3 部分是重点部分，也是结构设计的核心。在第 2 部分中存放 Portlet 的 Portal Server 发挥主要功能。在网格门户中所有的操作都是通过 Portlet 来进行的，不同的 Portlet 根据用户的需要可以设计成不同的逻辑业务实体，设计出来的业务 Portlet 和系统核心 Portlet 一起完整地构成了整个门户。在该门户推荐模型中，主要相关的 Portlet 构件有：证书管理 Portlet，协同过滤 Portlet，离线定时推荐 Portlet，推荐展现 Portlet 和执行任务 Portlet 等。

证书管理 Portlet 的作用是：网格门户用户要想成功执行某种资源或任务，必须要有相应未过期的有效证书。这个证书是需要和网格门户的安全机制配合来完成的。只有当用户有相应的用户代理证书的情况下才可以进行下一步的操作。如果通过了安全的证书认证，协同过滤 Portlet 就会启动。该 Portlet 的主要作用就是协同、指挥和协调几个子处理 Portlet。他们分别是资源模型建立、聚类模型处理、资源推荐和预测和推荐形成等 4 个模块。

(2)离线交互　此外，由于协同过滤的计算量大且相对耗资源的特点，所以可以选择离线运算。作业提交 Portlet 可以很好地实现这一功能。由于作业提交 Portlet 本质上是把任务提交给后台服务器执行，并且会和后台服务器进行交互进而更新作业状态，所以在门户层次可以做到离线进行协同过滤数据运算，并且适时地返回协同过滤结果。

(3)存储交互　最外一部分就是资源推荐知识库。该知识库主要功能是保存协同过滤模块产生的推荐结果。此外考虑到了数据规模较大的问题，设计了缓存机制，用来保证产生推荐结果的高效和快速。

2.2 协同过滤处理模型

二次组合协同推荐　在结合已有的传统协同过滤算法并充分考虑网格门户的基础上提出了二次组合协同推荐的设计思想。

图 2 基本表示了二次组合协同过滤推荐的基本步骤。分为：基于资源聚类的原始模型建立、一次预测资源评分、二次资源模型修正、基于资源的二次预测、寻找相似项内最近资源邻居和产生推荐结果。二次组合协同过滤推荐要是在一次协同过滤的基础上采用的二次组合协同过滤。其主要思想是进一步减少数据稀疏度、提高推荐准确度，从而提高门户推荐的质量。在最后进行二次预测的基础上根据预测评分的高低，选取前 N 项，产生 TOP-N 推荐列表。

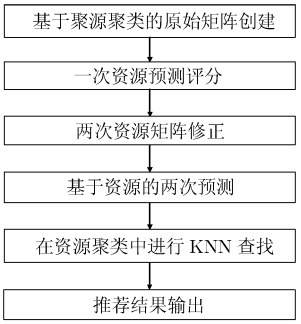

图 2　二次组合协同推荐模块

(1)基于资源聚类的原始模型建立　因为本文采用协同过滤方法需要前期建立一个较好的数据模型，原始的数据模型计算量大，没有把相关的资源进行分类，根据聚类理论和结合门户的特点我们使用基于资源聚类的模型建立方法。

设用户集合 $\mathrm{UC}=\{u_1,u_2,\cdots,u_m\}$，资源集合 $\mathrm{RC}=\{r_1,r_2,\cdots,r_n\}$，Portlet 集合 $\mathrm{PC}=\{p_1,p_2,\cdots,p_k\}$，集合 RC 和集合 PC 中的元素一一对应。设矩阵 $\boldsymbol{R}=(r_{ab})$ 是用户的兴趣模型，其中 $a=1,2,\cdots,m$，$b=1,2,\cdots,n$。r_{ab} 表示用户 a 和资源 b 所对应的规范化的评分值。

在获取初始的用户-资源矩阵之后，需要对最初的矩阵模型进行聚类处理[8]。本推荐模型采用目前应用最为广泛的 K-means[9]算法。聚类的计算依据是根据资源之间的相似性[10]来进行。其中资源相似性的计算方法采用 PCC 相关[11]。

$$\frac{\sum_{r\in R_{pq}}\left(R_{r,p}-\overline{R_p}\right).\left(R_{r,q}-\overline{R_q}\right)}{\sqrt{\sum_{r\in R_{pq}}\left(R_{r,p}-\overline{R_p}\right)^2}\sqrt{\sum_{r\in R_{pq}}\left(R_{r,q}-\overline{R_q}\right)^2}} \tag{1}$$

R_{pq} 表示项目 p 和 q 在项目集中所有共有评价分的用户子集。其中 $\overline{R_p}$ $\overline{R_q}$ 分别为属于 R_{pq} 集合中的用户 r 对项 p 以及 q 的平均评分。

由聚类的性质得知，目标资源的前 M 个最近邻居都在其最相似的聚类中，因此不需要在整个空间上进行查找推荐，故而提高了在线查找最近邻居的时间，满足了模型的实时性要求。K-means 算法主要过程如下：

(a)首先在 M 个项目中选取 K 个项目初始化，初始化时每个聚类中只有一个元素，并将其初始向量作为初始的聚类中心。

(b)依据每个聚类对象中心计算其均值，计算每个项目与这个聚类中心的距离即相似性，并根据最大相似性原则重新划分聚类对象，重新调整聚类中心。

(c)重新计算每个聚类中心的中心对象，重复步骤 2，比较聚类中心是否发生变化，即比较聚类中心对象是否发生变化，如果无变化，聚类结束，否则，重新聚类。

(2)一次资源预测评分　在形成初始聚类模型的基础上，该模块需要考虑大部分的用户对资源的评分为 0 的情况。这时需要对这部分的资源进行评分预测，并为形成二次资源模型做准备。

其算法过程如下：

输入：初始聚类矩阵(含大量空缺数据)，最相邻资源个数为 K。

输出：预测后聚类矩阵。

实例化初始矩阵：对于空缺评分资源，即对于用户-资源空缺评分的，暂以 2.5 代替。

(a)利用不同的相似度计算公式，查找每个资源的总数为 K 个的邻居集。

(b) 在获得每个资源的邻居和相似度后，根据评分预测公式计算未评分资源的评分，并形成预测评分集合。

(c) 对于每一个空缺评分资源，形成 n 个预测评分集合。将这 n 个预测评分集合与初始化矩阵进行并集运算，最终形成一次预测的用户-资源评分矩阵。

(3)二次资源模型修正　在使用一次协同过滤方法的基础上数据集已经有了初步规模，但是推荐质量可能还是会有所欠缺。为了获得更好的推荐质量和更准确的推荐资源，必须根据现有的信息在形成矩阵资源的基础之上进行用户资源模型修正。

算法主要过程如下：

输入：一次评分后的资源矩阵，邻居资源个数 NC，推荐资源个数 RC。

输出：修正后的聚类模型

(a)对于每一个资源 R，如果资源对应的用户有原始评分，则转入步骤(c)。否则转入步骤(b)。

(b)对于每一个目标资源来说，计算该目标资源和其它资源的相似度，并且根据要求把最相似的邻居资源作为该目标资源的邻居。需要注意是：在计算该相似度时需要把第一次预测评分值进行填充处理以降低数据稀疏度的影响。

(c)根据步骤(b)计算出来的相似度和原有模型中的原始评分和领居计算，将评分预测范围限制在该 RC 的资源以内，即只对资源的最近邻居做处理。

(d)形成资源 R 的评分推荐集合。

(e)再次循环处理步骤(a)-步骤(c)，至此形成所有资源的评分推荐集合。

(f)二次资源修正后的模型形成。

(4)基于资源的二次预测　在获得目标资源的历史信息以及最近资源邻居集合之后，可以通过预测的方法对目标资源进行预测评分，获取预测评分值。评分方法[12]如下：

$$PR_{u_a,r_t}=\frac{\sum_{r\in \mathrm{KNN}(r_t)}\mathrm{sim}(r,r_t)\cdot(R_{u_a,r}-\overline{R_r})}{\sum_{r\in \mathrm{KNN}(r_t)}\left|\mathrm{sim}(r,r_t)\right|}+R_{r_t} \tag{2}$$

通过最近邻居查找后，需要获得用户 u_a 对目标资源 r_t 的评分，需要计算 r_t 资源平均分值，$\mathrm{KNN}(r_t)$ 表示 r_t 的最相邻邻居集合，$\mathrm{sim}(r,r_t)$ 表示属于集合

KNN(r_t) 中的某个资源 r 相对于目标评分资源 r_t 的相似度。$\overline{R_r}$，R_{r_t} 分别表示 r 和 r_t 资源平均分。

(5)寻找相似项内最近资源邻居　最近邻居查找是指 KNN[13]的查找，通过此查找过程得到目标资源的邻居。两个对象之间的相似性通过 $\mathrm{sim}(j,k)$ 来度量。对象的邻居关系集合是指相对于每一个特定的对象而言都有其对应的邻居集 $\mathrm{NC}=\{nc_1,nc_2,\cdots,nc_j\}$，使得 $\mathrm{sim}(o,nc_1)>\mathrm{sim}(o,nc_2)>\cdots>\mathrm{sim}(o,nc_k)$。每一特定的对象获得了最近相邻的邻居集合即 KNN(M_T)。通过推荐算法运行获得了最近邻居之后即可以将其写入推荐知识库。推荐知识库的协同过滤数据信息可以供用户展现 Portlet 调用。

二次组合协同推荐方法可以最大程度地提高门户协同推荐效果和推荐质量。首先，该算法运用将数据邻居搜索限制在最相近的聚类中；其次，只对项目的最近邻居集合进行处理。以上改进的二次组合协同经实验论证，推荐效果和质量的确有所提高。

2.3 协同过滤展现模型

上面主要介绍门户协同处理框架来说明推荐模型的内部处理过程。而图 3 主要说明了门户模型在展现层次推荐用户资源的情况。

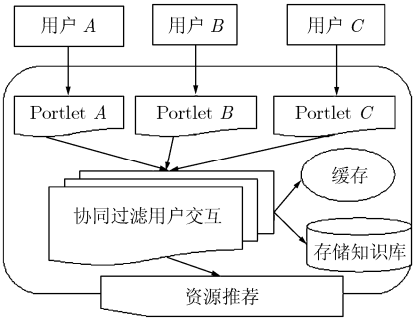

图 3 用户展现模块设计

图 3 由 3 部分组成。依次为：普通网格用户、资源推荐模块以及资源推荐知识库。其中普通网格用户、资源推荐知识库以及 CACHE 功能同上节所述。资源推荐具体展现方式有隐式展现和显式展现。

(1)显式展现　初始用户-资源经过协同过滤处理层之后，产生的协同过滤推荐结果已经保存在相应的存储知识库中。相应的用户当进入门户时可以获得推荐资源。在有相应权限条件下用户可以选择不同的算法查看资源推荐的结果。只要选择不同推荐算法点击后就可以进入到展现模块。展现模块会从系统中去读取相应的推荐资源，从而得到与目标用户最为相似的一些资源返回给用户。

(2)隐式展现　隐式展现是另外一种门户主动式服务的方法。当用户登录进门户系统时，系统在个性化页面上展现用户所感兴趣的内容。其工作流程如下：以门户用户 A 工作为例，用户 A 在门户中调用 Portlet A 执行相应的资源，在 Portlet 获取某个特定资源运算任务的同时推荐系统会同时工作，门户系统会根据用户的当前行为和历史信息数据根据协同过滤 Portlet 工作结果自动进行协同过滤计算，并且协同过滤 Portlet 会通知资源推荐模块去推荐知识库获取当前这个资源的多个推荐资源形成一个推荐集合。该集合中资源的相似度和用户将来所需要的资源相似度很高。这些资源也就是用户将来需要实际使用的资源，最后通过单独 Portlet 处理的方式以 Web 形式将结果返回给用户。具体消息格式如下：资源号*网络号*节点号*站点号*主机号*Portlet 号* 资源拥有者*。

3 实验结果与分析

3.1 网格门户下改进的协同过滤算法测试结果与分析

测试算法改进的原始数据来自于 MovieLens[14]。为了能将该数据应用到网格门户中服务于网格门户推荐模型的研究与设计，通过编写特定的源数据处理程序，将源数据处理成便于在网格门户中进行算法推荐的可识别模型。其中包括 943 个用户和 1682 个资源，每个资源都有相应 Portlet 进行封装从而完成任务执行的功能。用于网格门户测试的数据集的数据稀疏度约为 93%。初始评分数据为 1 至 5 分的整型评分数据。0 分为空缺评分。评分指标通过准确率 MAE(平均绝对误差)来计算。平均误差公式[15]为

$$\mathrm{MAE}=\frac{\sum_{i=1}^{k}|P_i-R_i|}{k} \tag{3}$$

其中 P_i 是预测值，R_i 是实际评分值，共有 k 个预测项目。

本实验结果是在同一数据集和门户推荐平台条件下进行的对比实验和测试，比较了基于聚类的用户协同过滤、基于聚类的资源协同过滤、基于资源的二次组合协同过滤的推荐质量，并在此基础之上分析了实验结果。实验结果如图 4 所示。

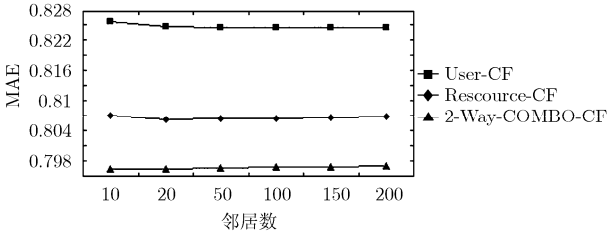

图 4 协同过滤算法比较

在图 4 中，Resource-CF 代表基于聚类的资源协同过滤，User-CF 代表基于聚类的用户协同过滤，2-way-COMBO-CF 代表基于资源的二次组合协同过滤。上述实验结果充分说明了基于资源的二次组合协同过滤的算法推荐质量较基于聚类的用户协同过滤、基于聚类的资源协同过滤都有较好的改进。因为二次过滤使得数据稀疏度降低，进而间接地改善了推荐质量。与此同时也验证了基于聚类的资源协同过滤推荐相对基于聚类的用户协同过滤推荐效果较好。

3.2 网格门户推荐的比较和分析

下面将结合具体的网格门户原型对该门户推荐体系结构进行具体比较和分析。

在当前网格门户发展中，发展较好的网格门户平台主要有：OGCE[16]Portal，Gridsphere[17]Portal，Jetspeed[18]Portal 等。但是这些现有门户平台都没有考虑到资源的推荐功能，所以当系统负载过高时会产生性能低下的情况。此外也不能提供门户主动式服务功能。在本论文中，改进的门户系统原型充分利用了协同过滤推荐的特点，能够实现主动式服务、积极推荐、减轻负载和资源利用最大化等功能。一旦门户的用户登录系统中系统可以准确实时地给用户推荐一些资源，该资源以 Portlet 形式封装，用户可以直接利用该 Portlet 去执行相应的资源。具体的比较分析如表 1 所示：

表 1 门户功能比较

相关功能	当前门户	改进的模型
协同过滤	无	有效应用
在线推荐	无	有效应用
离线推荐	无	有效应用
协同展现	无	有效应用

4　结束语

基于协同过滤的网格门户推荐模型对于协同过滤如何在网格门户中发挥作用做出了具体的分析和设计。此外，详细讨论了协同过滤方法在网格门户中应用的步骤，并且在此基础之上结合门户特点提出了改进的二次协同过滤算法，然后给出了具体的算法实验结果并对改进系统和原有系统进行了比较和分析。通过该系统的研究和设计，门户能够给出不同推荐算法下不同资源的过滤推荐列表。实验结果证明，这种方法是可行的，为个性化推荐系统在网格门户环境下的应用进行了有益的探索。在以后的工作中还需要做更深层次的研究。比如：在已经详细研究了利用协同过滤推荐模型可以为用户推荐出不同网络和站点资源的基础之上，对这些跨节点的资源向 Portlet 进行传播和捕获的研究还需要更进一步。对这些资源进行更加合理地分布式分配、传播和捕获并将其以队列形式发送给任务执行 Portlet 可以作为下一步研究工作的重点。

参 考 文 献

[1] Jiang Cong-feng. Grid computing based large scale distributed cooperative virtual environment simulation. Proceedings of the 2008 12th International Conference on Computer Supported Cooperative Work in Design, Xi'an, 2008: 507–512.

[2] Chen Xiao-wu. A multilayer portal model for information grid. China Grid Annual Conference, China Grid '08, Gansu, 2008: 78–85.

[3] Zhang Feng-juan. Research and design of grid portal based on hibernate. *Computer Engineering and Design*, 2009, 30(1): 71–79.

[4] 罗辛, 吴晶, 熊璋等. 轻量级门户单点登录服务机制. 北京航空航天大学学报, 2008, 34(6): 721–724.
Luo Xin, Wu Jing, and Xiong Zhang, *et al.*. Lightweight single sign-on service mechanism for portal. *Journal of Beijing University of Aeronautics and Astronautics*, 2008, 34(6): 721–724.

[5] Li Ming-biao. Optimization of grid resource allocation combining fuzzy theory with generalized assignment problem. Sixth International Conference on Grid and Cooperative Computing (GCC 2007), Los Alamitos, 2007: 142–146.

[6] Hu Rong. Smoothing based approach for hybrid collaborative filtering. *Journal of Harbin Institute of Technology* (*New Series*), 2007, 14(2): 38–41.

[7] Cai Yan li. Portlet-based portal design for grid systems. Proceedings Fifth International Conference on Grid and Cooperative Computing. GCC 2006 – Workshops, Hunan, 2006: 571–575.

[8] 高凤荣, 等. 基于矩阵聚类的协作过滤算法. 华中科技大学学报(自然科学版), 2005, 33(增刊): 257–260.
Gao Feng-rong, *et al.*. Matrix clustering-based collaborative filtering algorithm. *Journal of Huazhong University of Science and Technology* (*Natural Science Edition*), 2005, 33(Suppl.): 257–260.

[9] Kim Dong-moon. A music recommendation system with a dynamic K-means clustering algorithm. Proceedings - 6th International Conference on Machine Learning and Applications, Cincinnati, 2007: 399–403.

[10] Dave T. High-dimensional similarity retrieval using dimensional choice. Proceedings of the 2008-IEEE 24th International Conference on Data Engineering Workshop,

Cancun, 2008: 330–337.

[11] George T and Merugu S. A scalable collaborative filtering framework based on co-clustering. Fifth IEEE International Conference on Data Mining, Texas, 2005: 625–628.

[12] Jin Rong and Chai Luo-si. An automatic weighting scheme for collaborative filtering. Proceedings of Sheffield SIGIR - Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, 2004: 337–344.

[13] Pouwelse J. Distributed collaborative filtering for peer-to-peer file Sharing Systems. Applied Computing 2006. 21st Annual ACM Symposium on Applied Computing, Dijon, 2006: 1026–1030.

[14] Hu Rong. A hybrid user and item-based collaborative filtering with smoothing on sparse data . 2006 6th International Conference on Artificial Reality and Telexistence, Hangzhou, 2006: 184–189.

[15] Xue Rui-rong, Lin Chen-xi, and Yang Qiang. Scalable collaborative filtering using cluster-based smoothing Proceedings of the Twenty-Eighth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, 2005: 114–121.

[16] Sun Yu-mei and Fang Yu. GridGIS portal development based on OGCE. Proceedings of the SPIE-The International Society for Optical Engineering, Guangzhou, 2009: 71462–71470.

[17] Novotny J, Russel M, and Wehrens O. GridSphere: A portal framework for building collaborations. *Middleware for Grip Computing*, 2004, 16(5): 503–513.

[18] 田虹. 基于 Jetspeed 实现企业信息门户构建. 武汉理工大学学报, 2005, 27(4): 89–92.
Tian Hong. Research of the enterprise information portal based on Jetspeed. *Journal of Wuhan University of Technology*, 2005, 27(4): 89–92.

方 娟： 女，1973 年生，博士，副教授，研究方向为计算机系统结构、分布式计算.

梁文灿： 男，1982 年生，硕士生，研究方向为计算机系统结构.