基于覆盖网络模型的跨领域组合服务优化问题研究

张艳梅 曹怀虎
(中央财经大学信息学院 北京 100081)

摘 要 对基于覆盖网络模型的跨领域的组合服务优化问题进行了深入研究。首先考虑到跨领域策略路由的影响因素,将跨领域组合服务优化问题建模为带有功能约束和多QoS约束的多目标优化问题。然后利用层次算法和蚁群算法求解,先利用层次模型解决功能约束中的服务次序问题,再用改进的蚁群算法在层次模型中求出最优解集。仿真实验表明,随着进化代数的递增,非支配解在解集空间中呈均匀分布状态,说明求解算法的性能较好,跨领域组合服务优化策略具有可行性。

关键词 覆盖网络,服务组合,分层算法,蚁群算法,组合优化
中图法分类号 TP393 **文献标识码** A

Research on Routing Problem of Inter-domain Composed Service Based on Overlay Network

ZHANG Yan-mei CAO Huai-hu
(Information School, Central University of Finance and Economics, Beijing 100081, China)

Abstract The routing problem of inter-domain composed service based on overlay network was deeply researched. Since the strategy-routing policy affects a lot on the inter-domain routing, the multi-goals optimal model with multi-constraints was built on the inter-domain composed service routing. The layered algorithm was adopted to solve the function constraint of optimal model, and then the improved ants algorithm was employed to solve this problem in the layered model. The simulation shows the non-dominants solutions are evenly distributing, which means the algorithms perform well and the inter-domain routing method based on overlay network is feasible.

Keywords Overlay network, Service composition, Layered algorithm, Ants algorithm, Composition optimization

1 引言

近年来,越来越多的企业和组织都选择依托开放互联网实现动态业务协作,共享业务数据,实现跨组织数据动态集成,因此跨领域的服务组合日益增多,同时云计算等新兴技术的发展也为跨领域的组合服务带来了新的机遇。目前不少专家学者对组合服务优化问题进行了研究。Raman[1]分析了线性结构的组合服务优化问题;Manish[2]基于树结构研究了组合服务发现和路由过程;Tang[3]在服务云的探测研究中采用基于代价-质量折中的拓扑感知Overlay路径探测方法;Farshad[4]在文献[4]的基础上,提出根据早期发现的部分服务路径质量来决定放弃或者转发路径探测包的思想。以上研究都是通过实时探测网络状况的方式来发现组合服务的路由路径,虽简化了问题的复杂性,但探测数据包会占用较大的带宽资源。研究组合服务优化问题的另一个思路是基于网络拓扑图的全自动搜索求解,如Gu[5]对基于P2P网络的组合服务优化问题进行建模,并证明其属于完全NP问题,但没有给出求解方法。夏亚梅[7]等基于改进蚁群算法进行服务组合优化,提出一种多信息素动态更新的蚁群算法,包括局部优化算法和全局优化算法。该算法可以适应服务组合优化过程中发生的服务无效以及服务中QoS变化等情况,但它仅针对一般意义上的服务组合,并没有专门针对跨领域服务组合优化问题进行研究。跨领域服务组合优化问题的主要特点是:不但要考虑领域服务本身的代价,也要考虑不同领域服务之间的链路代价。因此本文对跨领域组合服务优化问题进行研究。

2 问题描述

如图1所示,在基于覆盖网络的跨领域组合服务优化问题模型中,每个组织被抽象为一个覆盖节点,称为领域。跨领域的业务协作在这些领域覆盖层的节点之间进行。组合服务的路由结构可以抽象为线性和并行两种。并行结构组合服务的优化问题可以转化为线性结构进行处理[6],因此本文重点研究跨领域线性结构组合服务优化问题的建模和求解。一个跨领域的组合服务可能既包括企业或者组织内部的域内服务组合,也包括跨领域的域间服务组合,为了突出跨领域组合服务问题,忽略领域内的服务组合。

如图1所示,跨领域线性结构组合服务优化问题ILCRP(Inter-domain Linear-structured Compositon Routing Prob-

到稿日期:2013-07-29 返修日期:2013-09-01 本文受国家自然科学基金(61103198),教育部人文社会科学研究青年基金(11YJC880163),北京市哲学社会科学规划项目(11JGC136)资助。

张艳梅(1976—),女,博士,副教授,主要研究方向为服务计算、智能优化算法,E-mail:jlzym0309@sina.com;曹怀虎(1977—),男,博士,副教授,主要研究方向为网络计算、社会网络。

lem)是指根据功能图描述的用户需求(包括功能需求和 QoS 属性需求),选择一条跨越不同组织的路径,该条路径依次经过功能图要求的各功能服务组件,且满足 QoS 属性(如带宽、延迟、可靠性和价格等),如图 1 所示,从 Start 到 End 是一条跨领域线性结构组合路由。通常既满足功能需求,也满足 QoS 属性等非功能需求的路径不止一个,此时需要根据某种优化目标选择最佳路由。

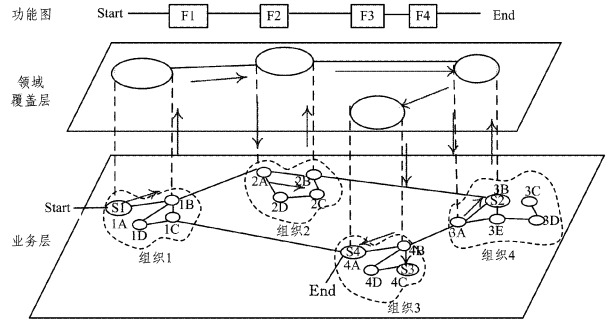


图 1　基于覆盖网络的跨领域线性结构组合服务路由示意图

跨领域组合服务路由和域内组合服务路由是有区别的,基于覆盖网络的跨领域优化问题需要考虑组织间跨领域路由的实际情况。跨领域路由由于域间"互相不信任"使得网络资源状态信息更新不及时,而且策略路由优先于最优路由,难以实现真正意义上的负载均衡。基于以上因素可知,对于跨领域组合服务而言,组合路径涉及的领域越多,涉及的域间的策略路由就越多,就越有可能降低组合服务的性能。因此,基于覆盖网络的跨领域组合服务路由应重点考虑领域间链路的选择问题,使得组合服务跨越领域数量尽可能少,涉及的领域尽量邻近,同时考虑不同领域间的负载均衡问题。跨领域的组合服务路由同样受到用户需求的约束,包括功能约束和 QoS 约束。

3　跨领域线性结构组合服务优化问题模型

在描述跨领域优化问题时,为了便于描述问题,将企业或者组织抽象为一个点,忽略域内路由代价。因此跨领域优化问题可以形式化表述如下:设无向图 $G_{HSON}=(V_{AS},E_{AS})$ 表示核心覆盖网络,V_{AS} 表示领域的集合,E_{AS} 表示领域之间边的集合,即跨领域链路的集合,边权表示节点之间的权值,可解释为延迟、带宽等性能度量指标。对于任意领域,其处理能力用 $cpu(v_{AS}):E_{AS}\rightarrow R^{+}$ 表示,因为领域可能已经有了一定的负载,其剩余负载能力表示其实际的处理能力,我们用函数 $cpu^{available}(v_{AS}):E_{AS}\rightarrow R^{+}$ 表示。其提供的服务花费表示为 $price(v_{AS}):V_{AS}\rightarrow R^{+}$。对任意覆盖链路 e_{AS},分别采用如下 4 种不同的度量指标:$bandwidth^{available}(e_{AS}):E_{AS}\rightarrow R^{+}$ 表示链路的实际可用带宽函数;$delay(e_{AS}):E_{AS}\rightarrow R^{+}$ 表示链路的延迟函数;$reliabitlity(e_{AS}):E_{AS}\rightarrow R^{+}$ 表示链路的可靠性函数;为了表达自治跨领域负载均衡这一指标,定义带宽比率(简称 BR,Bandwidth Ratio)为所需要的链路带宽和跨领域覆盖链路剩余带宽的比值,对于组合路由上的任意跨领域链路 e_{AS},有 $BR(e_{AS})=\dfrac{bandwidth^{required}(e_{AS})}{bandwidth^{available}(e_{AS})}$。在跨领域线性结构组合服务的候选路由图中,存在多条潜在候选路由,其中每一条候选路由都来自多个领域,表示为 P_{AS}。每一条候选路由 P_{AS} 就是指从源端开始,依次经过各功能组件($F_1,F_2,\cdots,F_i$)的一条组合路由,跨领域链路不同代价指标的计算方法如下:

$$(1)\ price(P_{AS})=\sum_{v_{AS}\in P_{AS}} price(v_{AS})$$

$$(2)\ delay(P_{AS})=\sum_{e_{AS}(u,v)\in P_{AS}} delay(e_{AS}(u,v))$$

$$(3)\ reliability(P_{AS})=\prod_{e_{AS}(u,v)\in P_{AS}} reliability(e_{AS}(u,v))$$

$$(4)\ BR(P_{AS})=\sum_{e_{AS}(u,v)\in P_{AS}}\frac{bandwidth^{required}(e_{AS}(u,v))}{bandwidth^{available}(e_{AS}(u,v))}$$

在为跨领域线性组合服务路由建模时,需要考虑对跨领域组合服务性能影响的主要因素:(1)路由路径上包含的领域数量尽可能少,即跨领域跳数最小;(2)跨领域链路的带宽比率尽可能小,即能够承受的负载最大,所以选择该路径能够实现含有相同服务组件的不同领域的负载均衡。跨领域线性结构组合路由选择的约束条件是:用户对组合服务的功能需求和延迟、可靠性和价格等 QoS 需求。综上所述,跨领域线性结构组合服务优化问题可以建模为:

$$\begin{cases}\text{Min}hop(P_{AS}) & (a)\\ \text{Min}BR(P_{AS}) & (b)\\ \text{s.t. } Function(F_1\rightarrow F_2\rightarrow F_3) & (c)\\ delay(P_{AS})\leqslant D & (d)\\ reliability(P_{AS})\geqslant R & (e)\\ price(P_{AS})\leqslant C & (f)\end{cases}\quad(1)$$

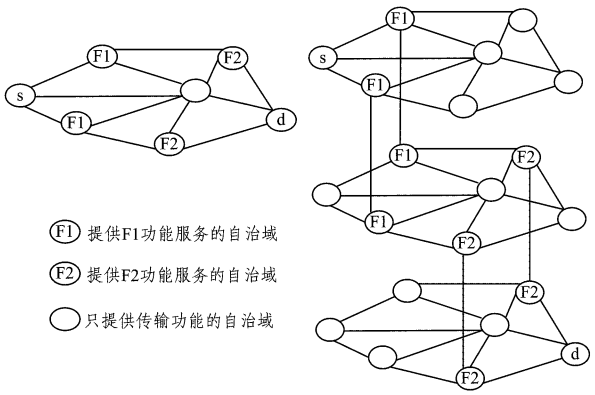
其中,式(a)表示跨领域跳数最小,hop 是跳数函数。式(b)表示尽可能实现跨领域链路的负载均衡。式(c)表示组合服务路由的功能约束,即路由需要依次经过功能图中的各个功能的服务组件。式(d)、式(e)和式(f)分别表示组合服务路由的延迟、可靠性和花费约束,其中 $D\in R^{+}$、$R\in R^{+}$ 和 $C\in R^{+}$ 分别表示请求服务对路由的最大延迟、最小可靠性和最大花费约束。从式(1)可以看出,该问题是多目标多约束的数学规划问题,属于完全 NP 问题。

4　跨领域线性结构组合服务路由算法

跨领域线性结构组合服务优化问题的求解难点是功能约束,因此需要对该问题采用分步求解法。首先引入层次算法,建立层次模型,在层次模型中实现功能约束;然后采用多目标蚁群算法在层次模型中求解带有多个 QoS 约束的线性结构组合服务的优化问题。

1)求解问题的层次算法

在进行跨领域路由时,覆盖网络中的每个领域抽象为一个点。图 2(a)给出的是一个具有两类服务的覆盖网络的拓扑结构,其中标注为 F1 的节点代表可以提供 F1 功能服务的领域,标注为 F2 的代表可以提供 F2 功能服务的领域,白色节点代表只具有传输功能的领域;s 和 d 所指节点分别为源点和终点所在的领域。图 2(b)是利用分层思想求解该问题的示意图。下面两层的图是上面图层的"拷贝"(拓扑结构和边权完全相同);将第 1 层和第 2 层中可以提供 F1 功能服务的相应领域用新增加的边连接起来,边权就是相应领域提供 F1 功能服务的代价;将第 2 层和第 3 层中可以提供 F2 功能服务的相应领域用新增加的边连接起来,边权就是相应领域提供 F2 功能服务的代价;然后将所有图层中的其它领域都看成具有转发功能的点;同时源点 s(终点 d)是使用第 1 层(第 3 层)的。在层次模型中 ILCRP 问题不需要考虑功能约束,就等价于在层次模型中求解传统的多约束最短路问题。



(a) 覆盖网络拓扑结构图　　(b) 覆盖网络拓扑结构图的分层模型

图 2　跨领域线性结构组合服务路由分层求解示意图

在建立分层模型之前，需要根据功能约束所定义的偏序关系进行拓扑排序，使得功能约束中的 k 种功能的服务组件的顺序关系可以由从 1 到 k 的一个全序关系所替代。然后根据服务类别集合$\{1,2,\cdots,k\}$上的全序，在原图的基础上拷贝图层，建立分层模型。该算法描述如下：

Algorithm　LA_ILCRP (Layered Algorithm for ILCRP)

Input: AS network topology graph $G_{ASON}=(V_{AS},E_{AS})$; and composed service request $CS_{req}=(F_{req},QoS_{req})$, where $F_{req}=\{F_1\rightarrow F_2,\cdots,\rightarrow F_n\}$ is the function constraint which includes the function components and the function order. $QoS_{req}=\{q_1,q_2,\cdots,q_n\}$ denotes QoS constraint

Output: composed service routing path$=(s,v_1,v_2,\cdots,v_m,d)$. s and d denote start node and end node

```
Begin
  toplogysort(L)
  for l=1 to k
    begin
      //copy gengranl edges
        V_AS={v_1,v_2,…,v_n}, V_AS^(l)={v_1^(l),v_2^(l),…,v_n^(l)} ;//copy nodes
        E_AS^(l)=Φ, ∀⟨u,v⟩∈E,⟨u^(l),v^(l)⟩∈E^(l)
        c(⟨u^(l),v^(l)⟩)=c(⟨u,v⟩)
    End
    //append vertical service edge in G'_ASON
      V'_AS=V_AS∪V_AS^(1)∪…∪V_AS^(k)
      E'_AS=E_AS∪E_AS^(1)∪…∪E_AS^(k)
      G'_ASON'=(V'_AS,E'_AS)
    Insertedge(G'_ASON,F_req)
    pathset=MAA_ILCRP(G'_ASON,s,d)
    path=Mapping(pathset,G_ASON)
    Output (pathset);
End
```

分层思想不仅是确保获得所需服务的有效手段，而且是寻找满足服务次序约束路由的有效手段，可以看作是动态规划思想的实现方式。通过以上分层算法的思想，解决了功能约束这一难题，但其因为有多个 QoS 约束条件，所以仍为完全 NP 问题，下面采用多目标蚁群算法继续进行求解。

2)求解问题的多目标蚁群算法

在求解多目标优化问题时，由于各个目标之间往往是相互冲突的，往往不存在能满足所有约束条件，且使所有目标函数都能达到全局最优的解，而是存在一组 Pareto 最优解。近几年来，如何应用蚁群等智能优化算法来求解多目标优化问题成为该领域研究的热点。这些基于种群的智能优化方法具有较高的并行性，尤其在求解多目标问题时，一次运行可以求得多个 Pareto 最优解，具有单目标优化方法不可比拟的优势[6]。本文对传统蚁群算法进行改进后，把它应用到上述层次模型的求解中。

假设蚂蚁所释放信息素的量与其所表示的解的优劣成正比。多目标优化问题由于其解的多样性，不存在绝对的最优解，不容易比较解的改进程度。通过比较解之间的 Pareto 支配关系来决定蚂蚁所释放的信息素的多少。同时，蚂蚁在觅食过程中不断更新自己的位置，即便发现了一个较优解，也会在下次移动中将此解丢失，因此将当前发现的所有非支配解保存起来，进而用这些解来指导蚂蚁整体寻优。具体改进有如下几点：

(1)信息素浓度

在多目标优化问题中没有绝对的最优解，解的优劣是相对的。比较蚁群中各只蚂蚁所表示的解的 Pareto 支配关系。这种 Pareto 支配关系决定了在它们所经路径上释放的信息素的浓度。对于规模为 N 的蚁群，如果蚂蚁 $k(k=1,2,\cdots,N)$表示的解 x_k 为非可行解，则说明蚂蚁 k 所经路径对于寻优帮助不大，故释放很少的信息素。如果 x_k 为可行解且支配其他的解 $x_{k'}$，则说明选 x_k 是一条较优路径，有利于算法朝着 Pareto 前沿或者可行解的方向进化。因此，蚂蚁 k 在其所经路径上大量释放信息素，按照这一想法，蚁群中第 k 蚂蚁在其所经路径上释放的信息素浓度 θ_k 定义如下：

$$\theta_k=\begin{cases}\lambda_1, & x_k \text{ 为非可行解}\\ \lambda_2, & x_k \text{ 为可行解，但不是非支配解}\\ \lambda_3, & x_k \text{ 为可行解，且为非支配解}\end{cases}\qquad(2)$$

其中，λ_1、λ_2 和 λ_3 为 3 个参数，且 $\lambda_3>\lambda_2>\lambda_1$。第 k 只蚂蚁在节点 i 的寻优方向与其到节点集 $allowed_k$ 的链路上的信息量有关。信息量越大，蚂蚁向该方向寻优的概率就越大。

(2)保存非支配解

仅依靠蚂蚁留下的信息素来进行寻优，算法搜索需要的时间长，且群体的多样性不易保持。这里在全局最优经验指导下进行寻优。在算法中，设立一个外部集合 P_{ndom}，用来保存整个蚁群当前所发现的所有非支配解。具体策略如下：

①通过支配关系选择出当前代种群中的非支配个体集 X。

②将 X 与外部集合 P_{ndom} 中的个体放在一起进行比较：若 X 被外部集合 P_{ndom} 中的个体所支配，则 X 不能进入外部集合 P_{ndom}；若 X 与外部集合 P_{ndom} 中的个体无支配关系，则 X 进入外部集合 P_{ndom}；若 X 支配外部集合 P_{ndom} 中的某些个体，则 X 进入外部集合 P_{ndom}，并剔除那些被支配个体。外部集合 P_{ndom} 中的个体在进化过程中保持非支配地位。

③当准则终止时，外部集合 P_{ndom} 中的解集即为所要求的 Pareto 最优集的近似解集。

(3)算法伪代码

Algorithm　MAA_ILCRP (G'_{ASON}, s, d)

Input: Layered gragh G'_{ASON} which is transformed from network graph $G_{ASON}=(V,E)$; and composed service path QoS constraints $QoS_{req}=\{q_1,q_2,\cdots,q_n\}$, and each q_i denotes a type of QoS constraint, such a delay, reliability and so on

Output: the best paths set P_{ndom}

```
Begin
  Step 1  paramters initialization. t=0, τ_ij(t)=0, Δτ_ij=0; P_ndom=ϕ;//Inintialize the parameters
```

Step 2 locate m ants in the source node s;

Step 3 For each ant does

Begin

Step 3.1 Search the next node j on the path to destination node d; while j≠d, each ant k(k=1,2,…,m) select it's next node j∈allowed$_k$ with the probability p_{ij}^k;

Step 3.2 Compute every chromsome fitness value, where $f_1=hop(x_k(t))$, $f_2=BR(x_k(t))$;

Step 3.3 Select the non-dominant solution, and the archive the non-dominant solutions to out set P_{ndom};// archive operation

Step 3.4 Update the information parmeaters θ_k, and $\tau_{ij}(t+n)$.

End;

Step 4 Compute the condition expression. If $t\geqslant t_{max}$ or the Pareto best solutions set P_{ndom} is satisfied, exit the genetic operation and output the solutions of P_{ndom}, otherwise t=t+1, return to step2 to the next genetic operation

End

5 仿真模拟及分析

使用 BRITE 工具来生成网络的拓扑图，其代表跨领域覆盖网络，网络的拓扑结构基于 Waxman 的拓扑生成算法 $P_e(u,v)=\beta\exp\frac{-l(u,v)}{L\alpha}$ 生成。各参数取值如下：(1)节点表示领域，个数为 50，权值在[5,20]之间均匀分布，表示领域处理能力。有 10 种不同功能的服务组件，每种组件 5 个副本随机分布在这些节点中，服务的价格在[1,5]上均匀分布。(2)边的代价有 3 种：$e_B\in[0,5]$表示可用带宽，$e_D\in[1,10]$表示延迟，$e_R\in[0,1]$表示可靠性。(3)平均每个节点连接边数 $m=3$，$\alpha=0.15$，$\beta=0.2$。

1)算法求解

该算法每次迭代运行都产生一些非支配解，并保存于外部的非支配解集中，通过观察非支配解集中的 Pareto 最优解的情况，可以了解各指标的最优解随迭代次数的变化情况，也反映了算法进化和收敛过程。由于使用外部集保留每次迭代非支配解策略，算法的渐近收敛速度得到了保证。实验中测试对象是包含 5 个功能服务组件的跨领域线性组合服务。优化目标的函数值随迭代次数的变化情况如图 3 所示，图 3 中给出了第 0 次到第 8 次迭代后的非支配解集情况。水平坐标轴分别表示该解的两项目标函数值，即跨领域跳数和链路可用带宽；垂直坐标轴表示迭代次数。结果显示：每次迭代后，非支配解的数量和质量都可能有变化，由于每次迭代后需要更新非支配解集，因此解数量可能增加，也可能减少。但解的质量随着迭代次数的增加会越来越好，逐渐接近 Pareto 最优解。

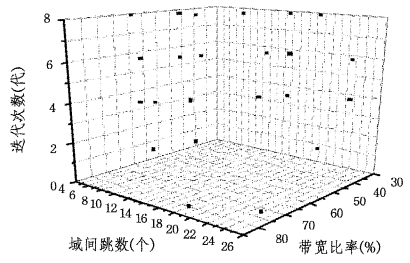


图 3 非支配解集随迭代次数的变化情况

图 4 是第 10 次迭代后的 Pareto 非支配解集。从图 4 可以看出，解的多样性好，分布比较均匀。

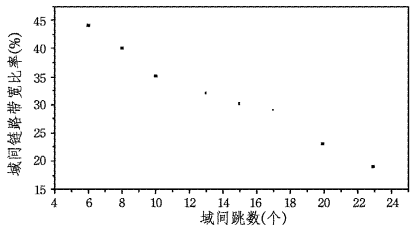


图 4 经 10 次迭代后的非支配解集

最终解结果从单个的优化目标角度看可能并不是最优的，但同时考虑两个子目标时则是最优的。最终的这些最优解提供给用户后，用户可根据通信业务需要从中选取偏好的解，例如在实时视频传输时更关注的是延迟，则可将领域跳数作为一种硬约束条件，选择符合该约束或者优化目标最小的路由，这也是多目标求解的优势，单目标路由优化只能产生一个最优解，不具备这样的能力。

2)与平面图蚁群求解算法性能比较

将本文提出的层次图蚁群求解算法与传统的平面图蚁群求解算法的时间开销进行对比发现：当服务数目较小(小于 4)或者较大(大于 7)时，平面图蚁群算法开销较小；服务数在 4～6时，层次图蚁群算法开销小于平面图蚁群算法。因为在层次图中，每一层的查找开销虽然小，但是层次图的拓扑排序和构造是需要一定的时间开销的。算法时间性能对比分析如图 5 所示。

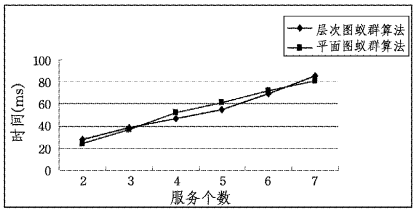


图 5 算法时间性能对比分析

结束语 研究了基于覆盖网络模型的跨领域线性结构组合服务优化问题。由于网络层的跨领域路由是在一种“互相不信任”的环境中的策略路由，因此覆盖网络组合服务的跨领域路由受网络层策略路由的影响和制约。在考虑以上各种因素的基础上，将跨领域线性结构组合服务优化问题建模为带功能约束和多 QoS 约束的多目标优化问题，并提出全自动服务搜索与优化算法。仿真实验表明，提出的跨领域线性组合服务优化策略具有可行性，选用的求解算法的性能较好。本文算法的性能可以从下面两个角度进一步提升：第一，通过对组合服务中的每组候选服务进行凸包构建[8]来减少搜索空间，并通过对初始解向量的多次升级和一次降级操作来达到全局优化的目标，从而提升算法的时间和空间效率；第二，引入 PROMETHEE[9]方法进一步区分 Pareto 最优解集合的优劣程度，从而选出 Top-k PROMETHEE 最优方案作为优化目标。

参 考 文 献

[1] Raman B, Katz R H. Load balancing and stability issues in algorithms for service composition[C]// Proceedings IEEE INFOCOM 2003. San Francisco, CA, 2003:1477-1487

[2] Jai M, Sharma P, Banerjee S. QoS-Guaranteed Path Selection Algorithm for Service Composition [C] // IEEE IWQoS, 2006. 2006: 288-289

[3] Tang C, McKinley P K. On the cost-quality tradeoff in topology-aware overlay path probing[C] // Proceedings of the 11th IEEE International Conference on Network Protocols (ICNP), (Atlanta, Georgia). November 2003: 268-279

[4] Samimi F A, McKinley P K. Dynamis: Dynamic Overlay Service Composition for Distributed Stream Processing [C] // SEKE 2008. 2008: 881-886

[5] Gu Xiao-hui. Spidernet: a Quality-Aware Service Composition Middleware [D]. University of Illinois at Urbana-Champaign, 2004

[6] Deb K. Multi-objective evolutionary algorithms: Introducingbias among Pareto-optimal solutions[C] // Ghosh A, Tsutsui S, eds. Advances in Evolutionary Computing: Theory and Applications. London: Springer-Verlag, 2003: 263-292

[7] 夏亚梅，孟祥武，陈俊亮，等. 基于改进蚁群算法的服务组合优化[J]. 计算机学报，2012，35(2)：207-281

[8] 赵欣，沈立炜，彭鑫，等. P MOEA：一种多目标决策辅助遗传算法用于服务组合 QoS 优化[J]. 中国科学：信息科学，2013，43(1)：73-89

[9] 李俊，郑小林，陈松涛，等. 一种高效的服务组合优化算法[J]. 中国科学：信息科学，2012，42：280-289

(上接第 64 页)

图 6 给出了 $k=15$ 时特征个数 m 与 OOB 的曲线变化。由图可见，随着 m 值增加，OOB 误差值先迅速地下降，当 $m=3$ 时，处于最低值，然后逐步回升。因而，当 $m=3$ 时，OOB 误差最小，故 $m=3$ 为最优。最终我们选定随机森林 $k=18$，$m=3$。

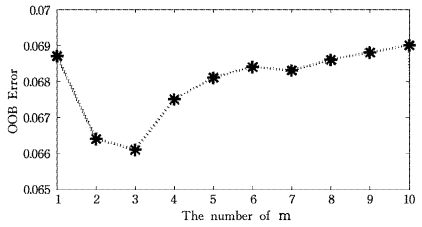


图 6 随机森林特征选择 m 与 OOB 误差

5.2 算法比较

针对 RFMR 算法，本实验将与逻辑回归(LR)、决策树(DT)、Adaboost(Ada)、朴素贝叶斯(NB)和多层感知器(MP)等经典分类算法进行比较。

由表 1 可见，在正例预测中，不同算法效果相差明显。在 Precision 中，RFMR 算法达到了 92.9%，相比于 DT 的 71.4%，提高了 21.5%，其他算法表现更加一般，NB 算法效果很差，不到 36%。总体来说，各个算法的 Recall 值较低，其原因是正负例的比例不均衡，大量被错误分类的负例降低了 Recall 值。但是 RFMR 算法的 Recall 值最大，达到了 68%，依然比 DT 算法高 4.1%，而其他算法表现很差，不到 30%。另外从 F-Measure、ROC 等指标来看，在正例预测中，RFMR 算法明显优于其他算法。

表 1 微博正例算法性能比较

	Precision	Recall	F-Measure	ROC
LR	0.645	0.133	0.221	0.79
NB	0.359	0.282	0.316	0.74
DT	0.714	0.629	0.669	0.852
MP	0.709	0.254	0.374	0.816
Ada	0.615	0.142	0.231	0.794
RFMR	0.929	0.68	0.785	0.917

表 2 给出了负例预测中各个算法的结果。相比于正例，各个算法的性能差距较小，同时各个算法的 Precision 和 Recall 值都比较高，说明这些算法预测负例的效果都比较好。但是 RFMR 算法在各项指标中最高，这说明在负例预测中它优于其他算法。

表 2 微博负例算法性能比较

	Precision	Recall	F-Measure	ROC
LR	0.883	0.989	0.933	0.79
NB	0.895	0.924	0.909	0.74
DT	0.945	0.962	0.953	0.852
MP	0.897	0.984	0.939	0.816
Ada	0.883	0.986	0.932	0.794
RFMR	0.953	0.992	0.977	0.917

综合比较正负例中各个算法的性能，可以看出 FMR 算法是最优的。

结束语 在微博网络中，用户间的微网络结构、权重比率等特征，对微博的转发行为有着显著的作用。为了研究用户的微博转发行为，本文首先分析提取相关特征，然后基于这些特征，提出了一个 RFMR 算法。实验结果表明，相比于其他分类方法，RFMR 算法性能最优。

参 考 文 献

[1] Granovetter M. The strength of weak ties[J]. The American Journal of Sociology, 1973, 78(6): 1360-1380

[2] Kam H T. Random Decision Forest[C] // Proceedings of the 3rd International Conference on Document Analysis and Recognition. 1995: 278-28

[3] Romero D M, Meeder B, Kleinberg J. Differencesin the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter[C] // Proceedings of the 20th International Conference on World Wide Web. 2011: 695-704

[4] Luo Z, Wu X, Cai W, et al. Examining Multi-factor Interactions in Microblogging based on Log-linear Modeling [C] // Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Istanbul, Turkey, August 2012: 6

[5] Yang Z, Guo J, Cai K, et al. Understanding Retweeting Behaviors in Social Networks [C] // Proceedings of the Nineteenth Conference on Information and Knowledge Management. 2010: 1633-1636

[6] Kwak H, Lee C, Park H, et al. What is twitter, a social network or a news media? [C] // Proceedings of the 19th International Conference on World Wide Web. ACM, 2010: 591-600

[7] Leo B. Random Forests[J]. Machine Learning, 2001, 45(1): 5-32