基于动态自适应性的语义对等网

刘 晔,马 慧

许昌学院计算机科学与技术学院,河南许昌 461000

摘要　针对 P2P 网络的聚类性以及网络的逻辑拓扑与物理拓扑不匹配所造成的通信延迟增加、网络开销增大等问题,根据节点的物理位置及节点所包含信息的语义相似度,提出了一种 3 层结构的语义对等网模型。该模型基于节点物理位置与属性特征动态调整网络结构,自适应地改善自身的资源搜索性能。仿真结果显示,该 P2P 覆盖网能够以较低的查找时延和代价获得较高的查全率。

关键词　对等网络;语义;物理拓扑;自适应性

中图分类号　TP393.02　　**文献标识码**　A　　**doi**　10.3981/j.issn.1000-7857.2011.32.011

Semantic P2P Networks Based on Dynamic Self-adaptability

LIU Ye, MA Hui

College of Computer Science and Technology, Xuchang University, Xuchang 461000, Henan Province, China

Abstract　In view of the clustering of P2P networks and the issues of increased communication latency and network costs caused by mismatch between the logical topology and the physical topology, this paper proposes a mixed three-layer P2P network model. The model adjusts the network structure dynamically based on the position and the property features of nodes, so that the performance of searching resources can be improved adaptively. The simulation result shows that the P2P networks can achieve a high recall ratio with low search delay and cost.

Keywords　P2P networks; semantic; physical topology; self-adaptability

0　引言

如何提高资源搜索性能一直是非结构化 P2P 系统研究领域的一个热点,而网络模型是影响资源搜索性能的重要因素。非结构化 P2P 网络[1-2]允许节点随机接入,结构简单,然而其采用的洪泛搜索机制使带宽资源浪费严重、网络负载加重。混合型 P2P 网络 Gnutella 0.6 [3] 兼具了集中式与分布式 P2P 网络的优点,但其模型中超级节点既要负责节点管理,又要进行消息路由,负载过重。文献[4]提出,P2P 网络逻辑拓扑与物理拓扑的严重不匹配延长了节点间的通信延迟,增加了网络的开销,从而降低了资源搜索效率。文献[5]表明,P2P 网络的拓扑图不是完全随机的,而是具有明显的聚类性。

本文根据实际网络特征把所有节点按物理位置划分成若干个域,在每个域内根据节点的语义相关度划分若干个群,建立了 3 层结构的语义对等网模型。该模型既考虑了网络逻辑拓扑与物理拓扑的不匹配,同时缩小了资源查找的范围。在 Peersim 模拟器上搭建 P2P 模拟环境,对资源查找算法进行仿真分析,并与 Gnutella 0.6 进行对比。

1　属性相似度

在 P2P 网络中,节点之间的资源查找与现实世界中的社会活动具有一定的相似性。现实世界中的资源查找一般在一个特定的群体内进行,该群体的人有某种共同的爱好与兴趣,这样查找成功的概率非常大,同样在 P2P 网络中,一些节点拥有的资源具有一定的语义相关性,如果资源查找在具有语义相关性的节点间进行,不仅能提高资源查找的效率,而且可以减少网络流量。

定义 1(语义相似度)　设资源空间 R 与查询空间 S 都采用 VSM 模型[6-7]表示,即 $\forall r \in R$, $\forall s \in S$, $V_r=[W_{1r},\cdots,W_{Nr}]$, $V_s=[W_{1s},\cdots,W_{Ns}]$,其中 W_{ir}、W_{is} 分别表示 i 项在 R 与 S 中的权重,其值采用 TF-idf[7-8]计算,则节点 r_1 与 r_2 的语义相似度计算公

收稿日期:2011-08-11;修回日期:2011-11-04

基金项目:河南省教育厅科技攻关项目(2009B520026)

作者简介:刘晔,讲师,研究方向为计算机网络和数字媒体,电子信箱:eduliuye@126.com

式可表示为

$$similarity(r_1,r_2)=\frac{\sqrt{\sum_{k=1}^{N}(w_{1rk}-w_{2rk})^2}}{\sqrt{\sum_{k=1}^{N}w_{1rk}^2}\cdot\sqrt{\sum_{k=1}^{N}w_{2rk}^2}} \quad (1)$$

查询请求与节点间的语义相似度计算方法同式(1)。

2 节点性能评价

在本文中将节点的物理性能(CPU,存储能力与带宽)与稳定值作为评价节点性能的主要参数。

定义 2(物理性能) 设语义群内所有节点的 CPU 频率总和为 C,内存总和为 M,带宽总和为 B,语义群内某节点 P_i 的 CPU 频率为 c_i,内存为 m_i,带宽为 b_i,为这 3 个参数分配的权重系数分别为 λ_1、λ_2、λ_3,则节点 P_i 的物理性能为

$$Capability(P_i)=\sqrt{\left(\lambda_1\frac{c_i}{C}\right)^2+\left(\lambda_2\frac{m_i}{M}\right)^2+\left(\lambda_3\frac{b_i}{B}\right)^2} \quad (2)$$

$Capability(P_i)$反映了节点 P_i 的物理性能,该值越大,节点 P_i 的物理性能就越高。

定义 3(稳定值) 设语义群内某节点 P_i 某段时间的登陆次数为 $LandTotalNum_{P_i}$, 某次登陆在线时间为 $LandTime_{P_i}$,在线时间总和为 $TotalOnTime_{P_i}$,则节点 P_i 的稳定值为

$$Stablity(P_i)=\frac{\text{在线时间大于或等于 } AverOnTime_{P_i} \text{ 的登陆次数}}{LandTotalNum_{P_i}} \quad (3)$$

$$AverOnTime_{P_i}=TotalOnTime_{P_i}/LandTotalNum_{P_i} \quad (4)$$

定义 3(综合性能) 假设某节点 P_i 的物理性能为 $Capability(P_i)$,稳定性为 $Stability(P_i)$,节点性能的权重系数为 λ_P,稳定性权重系数为 λ_S,λ 依实际情况动态调整,且 $\sum\lambda=1$。则节点 P_i 的综合性能为

$$W(P_i)=Capability(P_i)\times\lambda_P+Stability(P_i)\times\lambda_S \quad (5)$$

式中,λ_P、λ_S 表示节点性能与稳定性 2 项指标对节点综合性能的影响程度,且系数越大,影响越大。

3 基于语义与物理拓扑的对等网

基于语义与物理拓扑的对等网由物理位置邻近节点所构成的若干个域组成,在每个域内,具有相似语义特征的节点构成群,网络内节点数为 n 的对等网的形式化描述为

$$P2P\ networks=\{<CN,DCN,CND,\delta,RN>\} \quad (6)$$

式中,CN (Character Nodes)$=\{m_i,i=1,2,\cdots,n\}$;DCN(Domain Character Nodes)表示经过节点间距离计算,满足节点间邻近阈值 δ 的构成域 (Domain) 的属性节点集;CND(Character Node Distance)表示节点间距离计算函数,用于计算属性节点间的距离;δ 表示检验属性节点是否属于同一域的邻近阈值;RN(Routing Node)表示域中的路由节点,负责查找请求在域间的路由。其中

$$Domain=\{<DCN,GCN,CSF,\eta,IN>\} \quad (7)$$

式中,GCN(Group Character Nodes)表示经过节点语义相似度计算后,满足语义相似度阈值 η 的构成群(Group)的属性节点集; CSF (Character Similarity Function)$=<F:x,y\rightarrow z>(x,y\in GCN,z\in(0,1))$,表示语义相似度计算函数,用于计算属性节点的语义相似度;$\eta\in\{0,1\}$ 表示检验属性节点是否属于同一群的语义相似度阈值;IN(Intelligent Node)表示群中的智能节点,负责管理群中所有终端节点(Terminal Node,TN)以及群内的资源查找。基于语义与物理拓扑的对等网结构如图 1 所示。

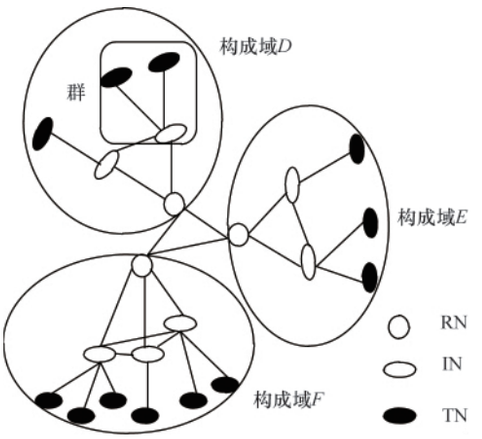

图 1 基于语义与物理拓扑的对等网模型

Fig. 1 P2P network model based on semantic and physical topology

3.1 对等网的构建

在传统混合 P2P 网络中,某节点加入网络时只需任意选择 1 个域并在该域的超级节点注册,当要加入的域节点数超过某个阈值时,就选择另一个域加入。在该 P2P 网络中,节点加入时不仅要考虑节点的物理位置,还要考虑节点的语义特征,不能随意加入某个域内的某个语义群,要选择与物理位置最近且语义相似的语义群的智能节点相连。

对等网生成算法如下所示。

输入:节点 Nt 的加入请求

输出:对等网

```
Procedure from_network (Node Nt)
Let D represent the set of domain, Di represent the domain in D
While(true) do {
Listening to adding request of nodes;
If(listened adding request of Nt ) {
  If(CND(Nt,RNi))≤CND(Nt,RNk)) RNk∈D
    Getconn(Nt,RNi);
  If(CSF(Nt,groupij)≥CSF(Nt,groupik)) and CSF(Nt,groupij) ≥η and (Capacity_IPij+1≤MAX)  groupik∈Di
    /* groupik 表示域 Di 中的群 k*/
    {
```

```
        Add (INij,Nt,Nt_Charcater);
        Send (INij,Nt,acceptmessage);
        Join (Nt,INij);
      }
    }
  }
```

3.2 对等网的动态自适应性

研究表明,每个域的性能与该域内所含的节点数量满足关系式 $K \leqslant Capa(M) \leqslant 3K-1$,其中 K 由域的路由节点的负载能力和带宽确定。当域内所含节点数量超过 $3K-1$ 时,智能节点与路由节点的负担加重,系统性能会下降;当域内所含节点数量少于 K 时,域的使用效率会较低。采用域分裂和域合并可以控制网络的规模,从而达到提高系统性能的目的。

3.2.1 域的分裂

为域设定 1 个节点数量的最大阈值 6,当域内节点数量超过常数 6 时,从域内选择 1 个节点作为新的路由节点,并与域内原来的路由节点相连,然后将域中部分智能节点及其管辖的终端节点从原来域中移出,通过与新的路由节点相连组成 1 个新域。域分裂算法如下所示。

```
Procedure Domain_Split (Node OldRN)
  If(NodeNumofDomain>6){
  NewRN=getNode(max(W(Pi));
  Conect(NewRN,OldRN);
  q=NumofIN/2+1;
  m[q]=ChooseINofShortestDist(NewRN,q);
    for(i=0;i<q;i++){
      disconnect(m[i],OldRN);
      Connect(m[i],NewRN);
       For(j=q; j<NumofIN; j++){
         Disconnect(m[i],INj);
        }
      }
    }
```

3.2.2 域的合并

同时为域设定 1 个节点数量的最小阈值§,当域内节点数量少于常数§时,首先选择 1 个与该域相距最近的域,然后比较 2 个域路由节点的性能,选择性能较优的节点作为合并后新域的路由节点;其次,如果 2 个域有相同的群,合并后共享 1 个性能较优的智能节点,并将该智能节点与选择出来的较优路由节点相连,否则智能节点将直接与较优的路由节点相连。

```
Procedure Domain_Combine(Node RNi)
  If(NodeNumofDomain<§){
    RNj=min(CND(RNi,RNk));
      /* k=i,…,T,T 为 P2P 网络的路由节点数/
     NewRN=max(W(RNi,RNj))
    If(NewRN=RNi) {
    For(q=1;q<M; q++) {
      /*M 为域 J 中的智能节点数 */
     If CSF(INq,INi)<η{
      Disconnect(INq,RNj);
      Connect(INq,RNi);
        }
    Else{
      NewIN=max(INq,INi)
      For(s=1; s< W; s++){
      Disconnect(OldIN,TNs)
      Connect(NewIN,TNs)
      /*W 为群 P 中的终端结点数 */
        }
      }
    }
  }
}
```

3.3 资源查找算法

在此覆盖网中,资源查找分为域内查找与域间查找。其中,域内根据查找请求与节点的语义相似度进行路由选择;域间查找除了考虑语义相似度外,还要考虑域间距离。

当终端节点需要查找资源时,首先向本群内智能节点发送查找请求,智能节点查询自己的本地资源列表,若命中,则将拥有所需资源的节点信息发送给请求节点,否则向域内一定跳数范围内属性相似度大于等于 θ 的智能节点转发请求,当有多个智能节点满足要求时,选择相似度最大的智能节点转发请求,然后按群内资源查找的方法进行查找。

如果在请求节点所在域内没有找到满足要求的资源,则采用域间查找,由域内智能节点把查找请求发送给域内的路由节点。在每个路由节点上,都维护有一定跳数范围内相邻域的域路由表,路由节点根据 2 个域的相似度及距离将查找请求发送给最合适的路由节点进行资源查找。

在资源查找过程中,为了避免后续节点向已经发过查找请求的节点发送相同的查找请求,每 1 条转发的资源请求中都含有请求转发所经过的节点路径信息,当找到所需资源时,节点将结果按路径的逆路径返回给请求者。同时为了避免请求在网络中转发时出现回路,在每 1 条查找请求中规定请求转发的最大跳数,当转发跳数达到规定的最大值时,转发节点抛弃该查找请求,同时向源点返回查找失败消息。

4 仿真与分析

度量语义对等网构建方法性能采用的主要指标为查找时延、查全率与查找跳数。查找时延指单位时间内成功查找到资源所需时间的平均值。查全率指在一段预定的时间内查找到满足服务请求的资源数占整个系统内此类资源总数的

比率。查找跳数指单位时间内所有搜索请求的查询跳数的平均值。查找时延与查全率主要衡量该网络构建方法在资源搜索方面的性能,而查找跳数主要衡量资源搜索的代价。

本文在 Peersim 模拟器上搭建 P2P 模拟环境，通过仿真对该对等网的综合性能进行分析，并与 Gnutella 0.6 进行对比。网络初始化结点为 2000 个,分为 10 个域,每个域根据节点语义相似性分为 4 个群。整个网络包含 5000 个资源,设置查询时的语义相似度的阈值为 0.8,节点间按相似度阈值 0.6 生成群。

在仿真中只要某节点拥有被查找的资源就认定搜索成功。仿真结果如图 2—图 4 所示。

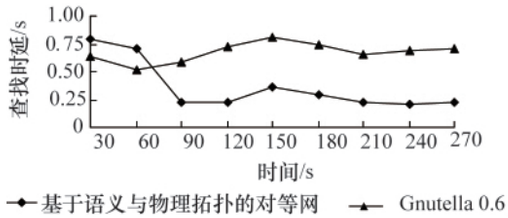

图 2　基于语义与物理拓扑的对等网与 Gnutella 0.6 查找时延的对比

Fig. 2　Search delay comparisons between P2P network based on semantic and physical topology and Gnutella 0.6

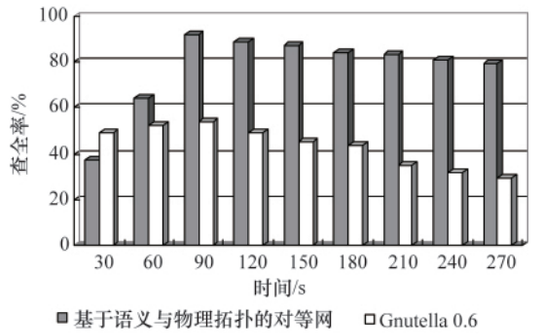

图 3　基于语义与物理拓扑的对等网与 Gnutella 0.6 查全率的对比

Fig. 3　Recall ratio comparisons between P2P network based on semantic and physical topology and Gnutella 0.6

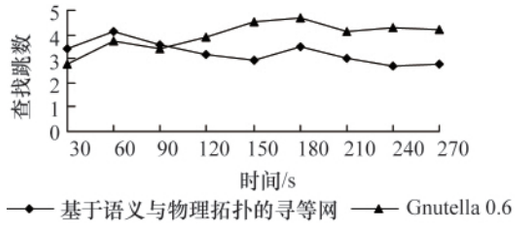

图 4　基于语义与物理拓扑的对等网与 Gnutella 0.6 查找跳数的对比

Fig. 4　Hop comparisons between P2P network based on semantic and physical topology and Gnutella 0.6

仿真结果显示，在试验初始阶段,2 个对等网的查找时延、查全率与查找跳数都有较大幅度的波动;随着时间的推移,基于语义与物理拓扑的对等网的查找时延比 Gnutella 0.6 缩短了约 60%,查全率较 Gnutella 0.6 提高了约 50%,查找跳数比 Gnutella 0.6 减少了约 40%,这是由于本文提出的基于语义与物理拓扑的对等网中节点根据所处物理位置以及语义相似性进行自聚，网络的物理结构与逻辑结构更加匹配,节点在群内或域内就能查找到自己所需资源,从而缩短了查找时延,提高了查全率,大幅提高了网络性能。

5　结论

本文根据 P2P 网络的聚类性,同时考虑 P2P 网络的物理拓扑构造了基于动态自适应的三层语义对等网,详细描述了节点间语义相似度的计算与节点性能评价方法、P2P 对等网的形成以及对等网的动态自适应性。仿真结果显示,基于语义与物理拓扑的对等网能够有效地提高资源查找效率,缩短查找时延,提高资源查全率,具有良好的实用性。

参考文献(References)

[1] Clip2 Distributed Search Services. The Gnutella Protocol Specification v0.4 [EB/OL]. [2007-12-20]. http://wenku.baidu.com/view/1635bad97f1922791688e82a.html.

[2] Zhuang L, Pan C J, Guo Y Q, *et al*. Connection management based on gnutella network[J]. *Journal of Software*, 2005, 16(1): 158-164.

[3] Ralf S, Klaus W. P2P 系统及其应用[M]. 王玲芳, 陈焱, 译. 北京: 机械工业出版社, 2008.
Ralf S, Klaus W. Peer-to-peer systems and applications[M]. Wang Lingfang, Chen Yan, trans. Beijing: China Machine Press, 2008.

[4] Xiao L, Liu Y H, Ni L M. Improving unstructured peer-to-peer system by adaptive connection establishment [J]. *IEEE Transactions on Computers (TC)*, 2005, 54(9): 1091-1103.

[5] Fast A, Jensen D, Levine B N. Creating social networks to improve peer-to-peer networking [C]. Proc of 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. Chicago: ACM, 2005: 568-573.

[6] Schutze H, Silverstein C. A comparison of projections for efficient document clustering [C]. Proc of ACM SIGIR. New York: ACM, 1997: 74-81.

[7] Wikipedia. Vector space model [EB/OL]. [2008-06-20]. http://en.wikipedia.org/wiki/vector_space_model.

[8] Salton G, Buckley C. Term weighting approaches in automatic text retrieval[R]. New York: Cornell University, 1987.

(责任编辑　孙秀云,代丽)