基于内容的发布订阅系统路由算法

薛小平[1,2], 张思东[1], 张宏科[1], 王小平[2], 葛 乐[2], 尹 琴[2]
(1. 北京交通大学电子信息工程学院, 北京 100044; 2. 同济大学电子与信息工程学院, 上海 200092)

摘 要: 本文综合评述了基于内容的 P/S 系统路由机制和算法. 根据客户的移动性和网络结构的变化对路由算法进行了分类和归纳, 分别论述了静态和变化拓扑环境下、支持客户移动情况下 P/S 系统各种路由算法的基本思想和优缺点等. 在此基础上, 针对 P/S 系统的动态、松耦合、多对多通信的特征, 分析和提出了有待解决的问题以及进一步的研究方向.

关键词: 发布订阅系统; 路由算法; 网络拓扑; 事件代理网络
中图分类号: TP393.01 文献标识码: A 文章编号: 0372-2112 (2008) 05-0953-09

Content-Based Routing Algorithms of the Publish-Subscribe Systems

XUE Xiao-ping[1,2], ZHANG Si-dong[1], ZHANG Hong-ke[1], WANG Xiao-ping[2], GE Le[2], YIN Qin[2]
(1. *School of Electronics and Information Engineering, Beijing Jiaotong University, Beijing* 100044, *China*;
2. *School of Electronics and Information Engineering, Tongji University. Shanghai* 200092, *China*)

Abstract: This paper discussed and reviewed various routing mechanisms and algorithms of the content-based publish-subscribe systems. By the client mobility and dynamically changed network topology the routing algorithms were classified, with their basic ideas, merits and shortages described in details respectively. According to the characters of dynamics, loose coupling, and many-many communication in the publish-subscribe systems, some unsolved problems were introduced and analyzed for the future research.

Key words: publish-subscribe systems; routing algorithms; network topology; event brokering network

1 引言

发布订阅系统(Publish-Subscribe Systems, 以下简称 P/S 系统) 由于具有异步、多点通信的特点, 近年来受到了研究人员的广泛关注. P/S 系统以传统网络为基础的新的重叠网(Overlay Network) 通信系统, 发布者以事件的形式将信息发送给事件代理, 订阅者向事件代理订阅感兴趣的内容, 事件在事件代理网络中传播, 最终路由到感兴趣的订阅者, 如图 1 所示.

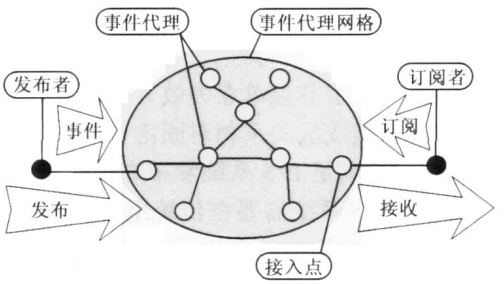


图 1 P/S 系统的通信模型

通常 P/S 系统可以大致划分成基于通道[1,2]、基于主题[3,4]以及基于内容[5,6]等三类, 已有的实验原型系统包括 Elvin, Gryphon, Siena, JEDI, Hermes 和 Narada Brokering 等. 本文主要研究基于内容的事件路由问题, 订阅者和发布者之间基于内容建立匹配或对应关系, 并将匹配后的事件正确、安全且高效地路由到订阅者. 在大规模 P/S 系统中, 由于匹配和路由计算给事件代理带来巨大的计算和通信负载, 因此, 性能卓越的匹配和路由算法是发布订阅系统的核心问题[1~6].

2 基于内容的 P/S 系统事件路由

2.1 基于内容的 P/S 系统事件路由

匹配和路由算法是 P/S 系统的核心. 为描述匹配和路由算法, 与 Siena 系统类似[10,11], 本文采用属性-值对 (a, val) 的组合来描述事件. 订阅是谓词组合, 通常由属性、运算符和属性值构成的谓词约束(a op val)组成.

定义 **1** 匹配 属性-值对(a_i, val_i)与谓词约束$(a_j$ op $val_j)$相匹配, 当且仅当属性名(a_i, a_j)相同, 且 op 对属

收稿日期: 2007-06-21; 修回日期: 2008-01-14
基金项目: 国家自然科学基金(NO. 60473001); 国家 973 重点基础研究计划(NO. 2007CB307100)

性值 val_i 和 val_j 操作(val_i op val_j)结果为True. 事件 e 与订阅 s 相匹配, 当且仅当 s 中每一个谓词在 e 中都存在与之匹配的属性-值对. P/S 系统中, 事件 e 与订阅 s 在事件路由器上的匹配通常由过滤器(Filter)实现.

定义2 **P/S** 系统的路由 在事件代理网络中根据事件的内容进行过滤, 寻找恰当的事件传播路径, 并将过滤后的事件转发给感兴趣的订阅者, 且满足开销、效率、可靠及安全的约束.

P/S 系统的路由与传统网络路由思想基本类似, 主要区别在于传统三层路由是按目的地地址寻址, 而基于内容的 P/S 系统则是以事件的内容作为寻址和路由的依据, 事件的内容隐含了路由的目的地.

2.2 基于内容的 P/S 系统路由分类

P/S 系统是由客户端(发布者和订阅者的统称)、网络拓扑组成的, 且在网络中根据事件语义进行匹配和路由. 因此, 本文将 P/S 系统的路由归结为以下类型, 如表 1 所示.

表 1 P/S 系统路由分类

Classification	Route Semantic	Topology Change	Mobility of Client
1	Subscribe Semantic	Unsupported	Unsupported
2	Publish Semantic	Unsupported	Unsupported
3	Subscribe Semantic	Unsupported	Supported
4	Publish Semantic	Unsupported	Supported
5	Subscribe Semantic	Supported	Unsupported
6	Publish Semantic	Supported	Unsupported
7	Subscribe Semantic	Supported	Supported
8	Publish Semantic	Supported	Supported

根据 P/S 系统的路由建立过程, P/S 系统的路由可分为发布方匹配路由和订阅方匹配路由两类. 这两者的基本思想是一致的, 都基于洪泛、匹配并转发的, 但在路由效率、性能及优化方法等方面存在差异.

3 静态拓扑中的路由

由于基于内容 P/S 系统中接收方的不确定性, 路由时采用遍历整个代理网络来确定通信对象, 为了达到这个目的, 通常采用洪泛、广播来遍历整个代理网络; 在确定了通信对象后, P/S 的路由机制一般采用组播技术, 在事件代理网络中传播事件.

3.1 P/S 系统基本的路由算法

P/S 系统最基本的路由方法是洪泛法及匹配优先法[12]. 这两种基本路由算法会带来大的网络负载和节点计算开销, 不适用于大规模网络的情形, 却是现有成熟 P/S 系统路由算法的基础. 现有的 P/S 系统路由算法大致有两类: 基于订阅语义和基于广告语义的路由. 基于广告语义的路由通常有三个阶段: (1) 广告广播: 在事件代理网络中广播广告, 确定通信对象; (2) 订阅传播: 广告与订阅相交的代理将订阅沿着广告传播的反向路径传送给该广告发布代理; (3) 事件传播: 发布事件的代理将事件沿着订阅的反向路径发送给感兴趣的订阅者. 基于订阅语义的路由通常有两个阶段: (1) 订阅广播; (2) 事件传播. 这两种路由机制都假设: 匹配位置要么在事件的发布方要么在事件的接收方. 两种语义下的路由尽管在具体阶段上有所不同, 本质上却是一致的.

3.2 Gossip 算法机制[24~26]

传统的洪泛和广播在大规模应用时会带来巨大的网络负载. 为此, 文献[24~26]提出了 Gossip 算法(也称为 Epidemic 算法).

Gossip 机制类似于流行病或谣言的传播, 算法可以抵抗失效. 算法思想如下: 当事件代理要发送消息(事件/订阅)时, 将消息发送给一些随机选择的代理. 代理收到消息后重复同样的过程, 再将消息转发给随机选择的其他代理. Gossip 机制可以限制消息在网络中的传播量, 同时获得与洪泛或广播类似的效果.

P/S 系统中的 Gossip 算法可以分散计算的负载. 在数据流量较小时, 对系统施加恒定的负载, 且不随系统规模的扩大而增加, 不存在链路或进程的单点失效, 健壮性好. 但增加网络的负载, 耗费网络资源, 在大数据量传输时也会成为瓶颈, 因此 Gossip 算法一般适用于小规模 P/S 系统.

为避免相同订阅或相同事件在网络中的传播, 通过调节以下三个参数, 可在可靠性和扩展性之间取得折衷[10]: (1) 每个进程的消息缓冲区大小; (2) 启动 Gossip 算法的周期; (3) 每次随机选择的代理数目. 此外, 每个代理上能重复 Gossip 消息的次数、同一 Gossip 消息可向前传播的跳数等因素对可靠性和扩展性也有一定影响.

3.3 组播机制

传统的组播可以用于 P/S 系统的事件传播, 以实现高效、低负载的事件传播. 为此, 很多 P/S 系统将事件代理网络预先组织成特定的拓扑结构[10,11,13~16], 并采用适合于这些拓扑结构的匹配和路由算法实现高效的消息转发, 如 Gryphon、Siena、JEDI、Narada Brokering 等采用的层次结构以及 Elvin 等采用的无环图结构, 并分别采用生成树转发算法[12,17,18]和基于反向路径转发算法[19,20].

传统组播技术与特定拓扑相结合, 简化了相应的路由算法, 优化了某些网络特性, 如流量、负载平衡等; 但特定的拓扑也带来了节点单点失效等问题. 传统的组播技术适用于组播组成员关系相对固定、并且接收方数量不太多的情况, 不满足 P/S 系统发布方和订阅方之间的动态变化关系. P/S 系统需要在传统组播树算法基础上进一步动态、快速地构造组播树[22,23].

文献[22]提出面向内容网络的结合广播和内容的路由机制(Combined Broadcast and Content Based, CBCB). 这种机制将事件代理网络看成两个层次: 传统的广播层和内容层. 如果某个代理需要发布消息, 首先在传统广播层上

广播以形成广播树, 然后, 对广播树进行基于内容的修剪形成所需的组播树. 广播层确保消息能通过无环或最短路径到达所有的目的地; 而内容层被认为是动态可配置的广播网络, 其中采用基于内容的地址对广播树进行动态修剪, 避免向不感兴趣的节点发送消息.

文献[23]在匹配优先的基础上, 为适应 P/S 系统的动态性和基于内容 P/S 通信的多样性, 提出了 MEDYM 算法(Matchr Early with Dynamic Multicast). 将发布订阅服务分成两个功能模块: 在网络边缘应用特定的复杂匹配模块和核心网络中通用的简单组播路由模块. 算法中, 发布的事件首先与来自远程服务器的订阅进行匹配, 以获得成功匹配服务器的目的地列表(DL, Destination List). 然后, 根据事件消息头中携带的目的地清单, 计算并构建瞬时的、无状态组播树. 这种方法允许对单个事件的提交进行细粒度的优化, 最小化 P/S 服务器上的通信负载. 采用配置后的重叠网. MEDYM 也很容易配置, 并且具有高度的灵活性, 以支持各种匹配和路由策略. 组播树构建采用了以延时为代价的 SPMST(Short-path MST) 算法. 如图 2 所示, 以快速和分布式的方式, 在目的地服务器之间计算一棵近似最小生成树. 算法中, 先离线计算下一跳的最短路径, 然后在线构造组播树. 在离线状态下, 维护称为 ShadowBitVectors 的数组, 帮助快速地确定时延代价最短(可用距离尺度)的下一跳服务器; 可由 ShadowBitVector 和 DLBitVector 的位矢量的交集快速确定. 在选择好下一跳以后, 其他目的地被分配到邻近的下一跳服务器目的地列表, 从而动态地实现构建组播树. 这种算法在事件匹配上的计算量较小, 且在事件路由具有高效灵活性. 为满足系统可扩展性要求, 进一步设计了基于分层的 MEDYM(H-MEDYM).

```
computeShadowBitVectors_s(){          //offline
  foreach server s_i{
    foreach server s_j
      if (DistanceMatrix[i][j] < DistanceMatrix[s][i]) &&
        (DistanceMatrix[s][j] < DistanceMatrix[s][i])
          Set_jth_bit_in_ShadowBitVector[i];}}
SPMSTRouting_s(DL){                   //online
  Nexthops = DL;
  foreach server s_i in DL
    if (ShadowBitVector{i} & DLBitVector != 0)
      Nexthops_remove{s_i};
  if(|Nexthops| > maxNextHops)
    Nexthops = closer_nexthops(maxNextHops)
  foreach server s_j in (DL-Nexthops){
    n_i = closest_nexthop_to(s_j)
    DL_i += {s_j};}
  return(⟨n_i, DL_i⟩);}
```

图 2　SPMST(Short-path MST) 算法

3.4　优化网络负载的路由机制

为了减轻 P/S 系统的负载, 减少不必要的事件或订阅在网络中的传播, 使系统具有更好的扩展性, 研究者们提出了许多限制和约束网络中的事件传递数量的方法, 如广告(Advertisement)、覆盖和合并(Merge)等.

3.4.1　广告机制[11,18]

广告的目的是通知事件代理网络发布者将要发布事件的类型, 可以更为合理地引导订阅的传播. 订阅定义感兴趣的事件集合, 广告则定义了发布方将要发布的事件集, 通过订阅与广告的相关性, 引导并限制订阅在事件代理网络的传播路径. 广告谓词与订阅谓词一样, 由谓词集组成, 但采用离散的结构.

定义 3　事件与广告关系　当且仅当 e 中的每个属性-值对都包含在广告谓词集中, 则广告 a 覆盖事件 e. 用 $\overset{N}{A}$ 表示由广告覆盖的事件集, 事件 e 与广告 a 关系描述为: $e \overset{N}{A} A \Leftrightarrow \forall \alpha \in e: \exists \varphi \in a: a \varphi$, 式中: φ 为广告谓词, α 为事件的属性-值对, A 为广告集合.

根据定义 3, 当且仅当广告覆盖了事件中的每个属性, 才能说广告覆盖了事件. 与订阅相对照, 当广告采用过滤器时, 同一属性多个约束仅需要满足其中一个约束即可.

定义 4　订阅与广告的相关性　当且仅当广告与订阅所定义的集合有非空交集时, 广告与订阅相关.

每个事件代理通过维护广告路由表来转发新增和取消的广告, 为订阅构建从订阅者到发布者的路由路径. 订阅路由表用来转发新增和取消的订阅, 为事件构建从发布者到订阅者的路由路径. 这里, 事件源代理可采用"基于源转发"广播算法, 将广告消息转发到其他各代理; 订阅消息根据广告消息的逆向路径, 到达各个可能发布相匹配事件的代理, 事件再根据订阅的逆向路径到达订阅者.

SIENA 和 Hermes 等系统采用了基于广告的优化路由, 减少订阅请求中的消息转发数量, 但这种方法额外增加了广告的转发数量. 因此, 该方法适用于系统中发布事件的客户较少、客户发布的事件满足关系不常变化的情形.

3.4.2　覆盖及合并机制[10,11,27,28]

覆盖和合并机制[10,11,27,28]是在过滤器的基础上针对订阅而提出的, 目的是有效地减少订阅消息的传播以及组播树建立的数量. 利用订阅中过滤条件之间的覆盖关系来优化订阅消息在网络中的传播数量, 这种路由机制可减少路由表中路由实体的数量, 也可以减少必须要转发的控制消息数量. 著名的 Siena、Rebeca、JEDI 和 Hermes 等都采用了支持覆盖关系的路由.

定义 5　覆盖关系　给定两个订阅 s_1 和 s_2, 当且仅当所有与 s_2 相匹配的事件都与 s_1 相匹配时, 则称 s_1 覆

盖 s_2, 表示为 $s_1 \overset{S}{\supseteq} s_2$.

在定义 5 的基础上, 通常将过滤器组织成偏序集(POSET), 图 3 给出了简单订阅偏序关系的例子[10, 11].

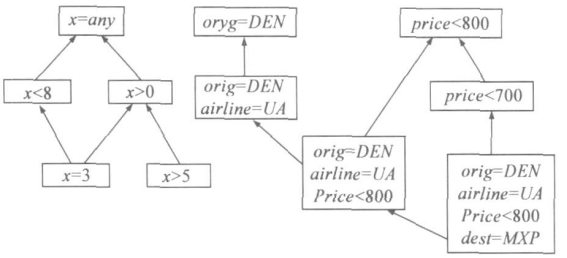


图 3 简单订阅偏序集的例子

过滤器的偏序集(POSET)适用于事件代理的不同算法和拓扑结构. 用 $\overset{S}{\supseteq}$ 表示订阅过滤器的覆盖关系. 基于偏序关系, 在代理订阅表中形成层次式订阅关系, 当且仅当新增或删除的订阅为最上层订阅(没有任何订阅可以覆盖该订阅)时, 才进行订阅信息的传播.

基于合并机制的路由算法也是基于过滤器的, 可以在基于覆盖关系的路由上实现. 不同之处在于, 它不仅依赖于客户发布的过滤器, 而且还对已有路由表项的过滤器进行合并, 并将合并后的过滤器转发给邻居子集.

定义 **6** 合并 过滤器 F 是 $\{F_1, \cdots, F_N\}$ 的过滤器集的合并(或覆盖), 表示为:

$$F \propto \{F_1, \cdots, F_N\}, \text{iff} N(F) \supseteq (\bigcup N(F_i))$$

文献[27]根据内容在最短路径上分发的思想, 提出了基于链路聚合(Link-Based Aggregation, Content Routing Approach, LACRA)和基于代理聚合(Broker Based Aggregation Content Routing Approach, BARCA)的思想, 两者区别在于代理中完成订阅信息的聚合层次不同.

订阅覆盖和合并通过减少订阅传播, 减轻网络负载, 有效地减少路由表的容量以及交换的控制信息数量, 这种覆盖关系也可用在广告机制中. 但偏序集的建立与维护需要相当的开销, 覆盖和合并关系的应用使算法复杂化.

4 拓扑动态变化环境下的路由

由于受到可控因素(如系统管理员为平衡负载添加新代理)或不可控因素(如物理主机的加入或离开、连接失败、移动等)的影响, P/S 系统的网络拓扑会发生变化. 在动态 P/S 系统中很难保持动态变化的发布订阅关系. 文献[9]针对这种情况, 提出了弱有效路由配置的概念, 保证仅向更新过程已经结束的订阅者发送通告. 本文将网络拓扑变化的路由分为两类: 拓扑结构相对固定的网络路由及拓扑结构频繁变化的网络路由.

4.1 拓扑结构变化相对固定的路由[10, 29~33]

当事件代理网络拓扑变化时需要维护网络拓扑结构, 修复或重构路由, 并维护相关事件代理的发布和订阅信息, 且不影响系统其他代理的正常工作, 以保持P/S系统的性能. 整个过程主要涉及: (1)节点或路径失效的发现; (2)拓扑和生成树的重构; (3)订阅、广告信息的维护和一致性; (4)事件恢复等问题. 主要的约束条件包括重构期间的事件丢失率、拓扑维护时间、路由重构时间等.

Scribe[33]在基于主题分层网络中, 采用"心跳"算法以发现失效节点. 某些基于内容的 P/S 系统也采用这一算法. 但采用这种算法需在失效发现和额外负载之间找到平衡点.

文献[32]提出了一种在动态环境下维持事件发送的树型拓扑结构. 利用分布式哈希表(DHT)技术, 对每个主机分配一组键值(Key), 并建立特定的参考树拓扑(Reference Tree Topology), 通过广度优先搜索算法搜索最高键值(Topmost Key), 将相应的主机映射到参考树中, 然后利用父节点、子节点寻找算法及邻居管理算法, 维护参考树的网络拓扑结构. 其优点在于对节点的实际拓扑无要求, 但仅实现了网络拓扑的重构.

文献[10, 29~31]中基于无环图, 提出了 Strawman 方法, 解决代理网中子树在合并和分裂时对订阅信息的恢复问题. 以基于逆向路径转发算法为例, 将节点移动引起的变化看成是节点的插入或离开, 而节点的插入可看作建立了新的连接, 节点离开看作与原有连接断开. 当链路断开时, 断开链路的两端分别向所在的子树发送来自链路另一端订阅的取消订阅信息; 当某条新链路出现时, 如当事件代理有新邻居 n 时, 将其订阅表中的所有订阅通过新链路发送给 n, 但新链路的出现不应造成拓扑无环图性质的变化.

值得注意的是, Strawman 方法只适用于某一时间段内单一路径变化的情况, 对如多链路断开、链路断开与新链路建立同时发生、或者断开的链接很快又得到恢复等情况的处理存在问题. 文献[31]对 Strawman 方法进行了改进, 考虑了订阅者的汇聚性和重构订阅时的影响范围等问题. 如在链路断开时, 有新的链路代替断开的链路连接两个子树, 则可延迟一段时间再执行订阅的取消操作, 减少不必要的订阅取消信息的传播; 此外, 重构订阅时受影响的只是从断开链路的一端经过新链路的重构路径到另一端路径上的代理, 因此仅需重配置重构路径上的订阅信息等等. 但这些改进增加了算法的复杂度.

4.2 拓扑频繁变化情况下的 P/S 系统路由

网络拓扑频繁变化情况下, P/S 系统的路由机制虽然可借鉴 MANET(Mobile Ad-hoc Network)网中的路由机制[33~35], 但代理所保存的发布/订阅信息也会随网络拓扑结构变化而变化, 因此, 现有的 MANET 的路由算法并不适合于 P/S 系统.

文献[35]结合扩展的按需组播路由协议(ODMRP)和基于内容的订阅机制提出了 P/S 系统路由. ODMRP 支持优化的具有上下文感知能力的数据发布机制, 感知的上下文包括物理地址、网络拓扑结构、网络能力(例如带宽和稳定性)及移动性. 代理节点聚合基于内容的订阅后, 送入 Bloom 滤波器, 事件源代理通过检测已经传播的订阅来确定组播组. 算法实现了高效的动态事件路由机制, 在吞吐量和控制数据报的开销方面表现均十分出色. 但其路由思想近似于基于主题的机制, 也没有有效减少消息传播.

文献[36]研究了将基于 MANET 的动态代理网络划分多个虚拟层次, 每个虚拟层次都为传播特定事件而构造有向无环图(DAG). 如果有大量的数据分发时, 网络中每个结点都要维护所有层次上的邻居信息, 这需要大量的存储资源, 造成巨大的数据处理负载, 因此不适合大规模网络的情形.

针对 MANET 的 P/S 路由问题, 文献[37]中提出了无结构的 P/S 系统路由, 即消息传送时不需要维护事件代理网络在移动环境中的拓扑结构, 接收到事件的代理根据其到目的地距离估计, 采用距离降序路由(Decreasing Distance Route)方法(也称 Hint 表驱动路由)决定事件是否及何时转发.

假设每个代理 B_i 都有订阅表及索引表(Hint Table), 订阅表存放了本客户的订阅, 索引表中含代理的标识、订阅摘要以及基于时间的距离估计的索引. 它定期地广播带有本地订阅摘要的 Beacon 消息. B_j 在收到 B_i 的 Beacon 消息后, 计算 B_j 到 B_i 之间用时间估计的距离 h_{ij}, 并用 Beacon 中的订阅摘要、接收时间更新索引表.

事件源代理收到客户发布的事件 m 与其索引表中的某个订阅相匹配时, 向邻节点广播 m, m 中含有关对 m 感兴趣的订阅者目的列表以及目前所知道的到目的地的最小距离估计值.

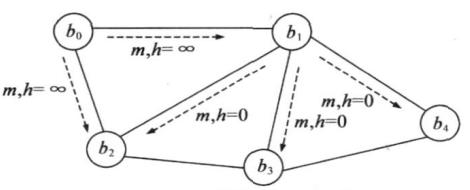


图 4　基本降矩路由机制

如图 4 所示, 代理 b_0 发布了与代理 b_4 订阅相匹配的消息. 代理 b_0 广播该消息, b_1 和 b_2 接收到该消息. 假设 b_2 到 b_4 的 $h_{24}=5$, 代理 b_2 企图在 5 倍比例延时的时间后转发. 由于 b_1 是 b_4 的邻居(即 $h_{14}=0$), b_1 立即转发消息. 代理 b_2 由于接收到了来自 b_1 的消息, 不再转发且丢弃 m. 此外, 由于 m 中到 b_4 的 Hint 值为 0, 所以 b_3 也不再转发.

无结构路由的方法不需要探测连接情况, 能很好地适应节点移动的情况, 由于采用了延时机制以及距离降序的路由, 有效地减少网络消息的传播.

5　支持客户方移动的路由

由于客户的移动, 代理之间的路由关系发生了变化, 需要重构组播树, 因此订阅或事件可能会丢失, 称之为假性失真(False Negative, FN). P/S 系统中, 支持客户方移动的路由算法应满足在客户移动过程中重建路由关系的时间段内不引起任何假性失真的约束.

5.1　支持移动的标准路由[14,38,39]

JEDI 系统首次提出了订阅者移动性支持协议, 引入了"Movein"和"Moveout"操作, 使 P/S 系统支持断开连接以及重新连接操作[14]. 文献[38]中进一步给出了发布者移动性支持协议. 这两个协议统称为移动性支持的标准路由协议. 支持移动的 P/S 系统协议涉及组播树重构和发布/订阅消息转移两个问题.

如图 5 所示, 以订阅者移动的情况为例说明客户移动的操作过程. 假设在时间段 t_1, 用户连接到代理 A 并接收事件; 时间段 t_1 的末尾, 用户与代理 A 断开连接; 在时间段 t_3 后, 用户重新连接到新代理 B. 假设 t_2 时间段被用来做下文所述的优化操作; 时间段 t_4 完成整个重新连接阶段, 包括新代理取回和重发相关订阅, 获取在用户断开连接时所遗失的事件以及旧代理取消相关订阅; 时间段 t_5, 用户像在 t_1 时间段接收到了新发布的事件.

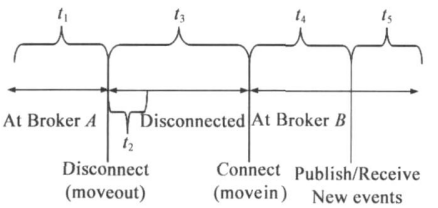


图 5　标准协议订阅者和发布者移动操作时间轴

与订阅者移动相似, 发布者移动标准协议中也可以在与旧代理断开连接后, 组播树的撤销(t_3)以及连接到新代理后的组播树重建(t_4), 但不涉及广告转移等问题, 仅需要关注组播树的重构问题.

标准协议虽然解决了移动性支持的基本问题, 但高开销、高延时等问题[39]使之几乎不能实际应用. 为此, 研究人员提出了很多优化方案, 其中, 发布者移动的主要问题在于组播重构引起的高延时, 其优化方案[37]主要包括减少树重构时间, 如预取(Prefetching)协议; 降低树重构范围, 如延迟(Delayed)协议; 减少树重构概率, 如 Proxy 协议等. 此外, 也可同时采用两种优化方案, 如延迟预取(Prefetching-Delayed)协议. 而订阅者移动除延时外, 还需解决订阅和事件单播传送的高开销问题. 在其优化方案[3]中, 通过预取(Prefetching)、家乡代理(Home-

Broker) 协议来解决高延时问题, 通过日志(Logging, 减少需单播的事件数目)、将订阅保存在设备上(Subscription-or-Device, 减少需单播的订阅数目) 协议来解决高延时问题.

5.2 支持移动性的通用路由[40, 41]

原有的移动性支持是针对某个具体的 P/S 系统(如 JEDI) 实现的. 文献[40, 41] 在 Movein 和 Moveout 的基础上提出了 Ping/Pong 协议, 通过对 Siena 系统的扩展实现通用的移动性支持. 由于 Siena 系统本身可直接支持发布者移动, 因此, 针对客户移动的情况只要考虑订阅者移动.

在该网络架构中, 在每个代理上运行独立、固定的支持移动性服务的 Proxy, 且每个客户都保存自身的订阅信息. 在正常的通信阶段, Proxy 不起任何作用. 当客户与代理断开连接之前, 客户向该代理的 Proxy(即 moveout Proxy) 转发订阅, 由 Proxy 执行订阅操作, 接收事件并将事件保存在队列中. 当客户移动到目的地时, 与新的本地代理的 Proxy(即 Movein Proxy) 建立连接, 转发 Moveout Proxy 的地址, 并执行 Ping/Pong 协议. Ping/Pong 操作的作用在于确保旧代理取消订阅之前, 新代理的订阅已经生效.

如图6 所示, *A* (原代理) 和 *B* (新代理) 为代理, 分别运行 Moveout Proxy 和 Movein Proxy, 客户移动后接入到新代理的 Movein Proxy, 协议通过 Ping/Pong 操作 2, 3(图 6) 对两个 Proxy 进行同步, 随后如标准协议一样实现相关事件的转移 4、5、6, 只是转移操作的主体由代理变为 Proxy.

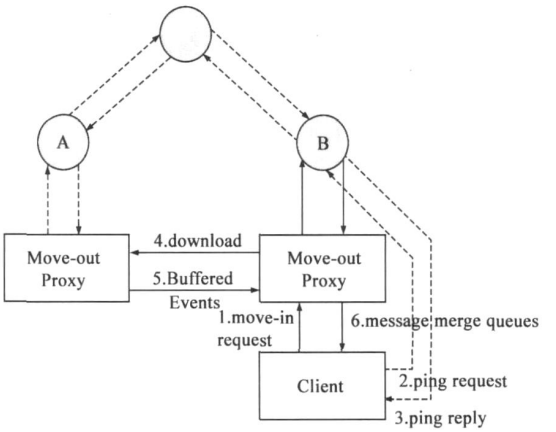


图 6 Ping/Pong 协议下的订阅者移动支持

利用 Proxy 屏蔽了下层 API 的差异, 可以工作在不同的发布/订阅系统上, 且不需要改变原系统的 API. 利用源和目的代理之间 Ping/Pong 操作对订阅进行同步. 但难以实现优化的移动性支持.

6 基于内容的 P/S 系统路由及进一步的研究方向

本文对著名的 P/S 系统原型系统的路由算法进行了研究和比较, 如表 2 所示. 现有的 P/S 系统对路由及算法进行了大量的研究, 采用了多种优化技术对 P/S 系统的路由进行了优化, 其核心思想是降低网络负载、实现高效的 P/S 系统. 但以下问题还有待于进一步研究.

表 2 典型 P/S 系统的比较

System	Type	Matching Algorithm	Multicasting Algorithm	Optimization	Topology	Client Mobility
Gryphon	Hybrid	Matching Tree	Link Matching algorithm	–	Static	–
Scribe	Subject-based	Lookup table	Multicast tree	–	Static	–
Bayuex	Subject-based	Lookup table	Hashed-suffix mesh algorithm	Covering	–	–
Rebeca	Content-based	Matching Tree, Filtering	–	Covering, merging	–	Support
JEDI	Content-based	Matching Tree	Multicast tree	Covering	Dynamic	Support
Herms	Hybrid	Matching Tree Lookup table Filtering	Core-based trees with the reverse path forwarding	Covering, Advertisement	Dynamic	Support
Sinea	Content-based	Filtering Algorithm (Binary Decision Diagrams)	Event nonfiction Service	Covering, Advertisement	–	Support
Narada	Content-based	Matching table and SQL Queries	Link Matching algorithm	–	Dynamic	–
XMessage	Hybrid	Lookup and SQL like Queries	Event nonfiction Service	–	–	–
Echo	Hybrid	Lookup and Filter function	Event nonfiction Service	–	Support	–

(1) 性能优化的 **P/S** 路由 现有的 P/S 系统的路由基于传统组播和 Gossip 的思想, 不能很好地解决动态、松耦合的多对多通信系统的问题, 传统的组播假设与 P/S 系统动态的假设存在一定的冲突, 且组播树建立和维护的开销大, 而大规模、高度动态 P/S 系统, 要求具有动态、低开销和延时地构造组播树. 在路由的研究方

面,通常将路由问题与特定网络拓扑结构(如无环图)相结合,需要进一步研究更为通用的 P/S 系统的路由理论和方法.

目前的研究通常假设匹配在发布方或订阅方的极端情况下.由于发布方和订阅方的不确定性,为了寻找特定对象需要将匹配信息遍历整个网络,给网络带来额外的压力.研究表明:合理的匹配位置能提高 P/S 系统路由的效率,减轻网络负载.此外,为了减少网络负载及代理处理开销,需要进一步研究压缩转发和匹配的理论和方法.

(2) 变化网络中的路由研究　在变化网络中,现有的算法主要基于拓扑维护及路由修补的方法,其目标是尽可能地抑制路由修复而带来的发布订阅关系的影响.无结构的路由虽然能解决拓扑变化频繁的问题,但基于 MANET 的代理本身受到计算能力、电源、通信带宽以及拓扑变化频繁等约束,这方面的研究还有待于进一步研究.

(3) 移动性支持研究　基于 Movein 和 Moveout 的机制可以解决在静态网络拓扑中支持客户移动的问题,但是,网络拓扑变化情况下发布方移动和订阅方移动还没有得到深入的研究.此外,基于 Movein 和 Moveout 的移动性支持机制,仍然没有解决静态拓扑环境下的支持发布/订阅方相关时的移动问题.

(4) 基于内容的 **P/S** 系统安全　基于内容的 P/S 系统中安全问题包括认证、机密性、完整性和审计性方面,虽然可以采用一些已有的方法(如用数字签名来解决信息完整性),但还有很多问题有待进一步地研究,如,路由需要基于内容,但又要保持内容的机密性;审计需要基于订阅,但同时又不能泄漏订阅的条件等.

(5) QoS 研究　P/S 系统建立在尽力而为的传送机制上的,针对实时的发布和订阅关系,保障 P/S 系统发布方和订阅之间的 QoS 路由问题,也将会受到研究人员的高度关注.

7　结论

本文对基于内容的 P/S 系统的路由机制和算法进行了综合研究.归纳和分类了 P/S 系统的路由机制,根据客户移动性和网络结构的变化将路由问题分成了四种情况,并分别进行了研究和讨论;在此基础上,针对 P/S 系统的动态、松耦合、多对多通信的特征,比较并总结了现有路由机制和算法的特点,提出了进一步的研究方向.

参考文献:

[1] Duncan McCaffery, Joe Finney. Low latency optimization of content based publish subscribe for real-time mobile gaming applications[A]. Proceedings of the 25th IEEE International Conference on Distributed Computing Systems Workshops[C]. New York: ACM Press, 2005. 438- 443.

[2] Oki B, Pfluegl M, Siegel A, Skeen D. The information bus:? an architecture for extensible distributed systems[J]. ACM SIGOPS Operating Systems Review, 1993, 27(5): 8- 68.

[3] Yan TW, Garcia-Molina H. The SIFT information dissemination system[J]. ACM Transactions on Database Systems, 1999, 24(4): 529- 565.

[4] IBM RedBook. Internet application development with MQSeries and Java[EB/OL]. http://www.redbooks.ibm.com/redbooks/SG244896.html, 1997-02.

[5] Aguilera M K, Strom R E, Sturman D C, Astley M, and Chandra T D. Matching events in a content-based subscription system[A]. Proceedings of the 18th ACM Symposium on Principles of Distributed Computing[C]. New York: ACM Press, 1999. 53- 61.

[6] Cao F, Singh J P. Efficient event routing in content-based publish/subscribe service network[A]. 23rd AnnualJoint Conference of the IEEE Computer and Communications Societies[C]. USA: IEEE INFOCOM, 2004. 929- 940.

[7] Terpstra W W, Behnel S, Fiege L. A peer-to-peer approach to content-based publish/subscribe[A]. Proceedings of the 2nd International Workshop on Distributed event-based Systems[C]. New York: ACM Press, 2003. 1- 8.

[8] Srivatsa M, Liu L. Securing publish subscribe overlay services with event guard[A]. Proceedings of the 12th ACM Conference on Computer and Communications Security[C]. New York: ACM Press, 2005. 289- 298.

[9] Mühl G. Large-scale content-based publish/subscribe systems[D]. Darmstadt University of Technology, 2002.

[10] Carzaniga A, Wolf A L. Content-based networking: a new communication infrastructure[A]. Developing an Infrastructure for Mobile and Wireless Systems[C]. Berlin: LNCS, 2002. 59- 68.

[11] Carzaniga A, Rosenblum D, Wolf A L. Design and evaluation of a wide-area event notification service[J]. ACM Transactions on Computer Systems, 2001, 19(3): 332- 383.

[12] Banavar G, Chandra T, Mukherjee B. An efficient multicast protocol for content-based publish-subscribe system[A]. Proceedings of the 19th IEEE International Conference on Distributed Computing Systems[C]. New York: ACM Press, 1999. 262- 272.

[13] Robert Strom, Guruduth Banavar. Gryphon: an information flow based approach to message brokering[DB/OL]. http://researchweb.watson.ibm.com/distributedmessaging/papers/ext-abstract.htm, 1998- 10.

[14] Cugola G, Nitto E D, Fuggetta A. The JEDI event-based infrastructure and its application to the development of the OPSS

WFMS[J]. IEEE Transactions on Software Engineering, 2001, 27(9): 827-850.

[15] Indiana University. The NaradaBrokering Project[EB/OL]. http://www.naradabrokering.org, 2006-09.

[16] Fitzpatrick G, Kaplan S, Mansfield T. Supporting public availability and accessibility with Elvin: experiences and reflections [J]. Computer Supported Cooperative Work, 2002, 11(3): 447-474.

[17] 薛涛, 冯博琴. 内容发布订阅系统路由算法和自配置策略研究[J]. 软件学报, 2005, 16(2): 251-259.
Xue Tao, Feng Boqin. Research on routing algorithm and self-configuration in content-based publish-subscribe system[J]. Journal of Software, 2005, 16(2): 251-259. (in Chinese)

[18] 马建刚, 黄涛, 汪锦岭, 等. 面向大规模分布式计算发布订阅系统核心技术[J]. 软件学报, 2006, 17(1): 134-147.
Ma Jian-Gang, Huang Tao, Wang Jinling, et al. Underlying techniques for large-scale distributed computing oriented publish-subscribe system[J]. Journal of Software, 2006, 17(1): 134-147. (in Chinese)

[19] Dalal Y K, Metcalfe R. Reverse path forwarding of broadcast packet[J]. Communication of the ACM, 1978, 21(12): 1040-1048.

[20] Carzaniga A, Wolf A L. Forwarding in a content-based network[A]. Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications[C]. New York: ACM Press, 2003. 163-174.

[21] Riabov A, Liu Z, Wolf J, Yu P, Zhang L. Clustering algorithms for content-based publication-subscription systems[A]. Proceedings of the 22nd International Conference on Distributed Computing Systems[C]. USA: IEEE Computer Society, 2002. 133.

[22] Carzaniga A, Rutherford M J, Wolf A L. A routing scheme for content-based networking[A]. Proceedings of the 23rd AnnualJoint Conference of the IEEE Computer and Communications Societies[C]. USA: IEEE INFOCOM, 2004. 918-928.

[23] Fengyun Cao, Jaswinder Pal Singh, MEDYM: match-early with dynamic multicast for content-based publish-subscribe networks[A]. The 6th ACM/IFIP/USENIX International Middleware Conference[C]. Berlin: Springer, 2005. 292-313.

[24] Eugster P T, Guerraoui R, Handurukande S B. Lightweight probabilistic broadcast[J]. ACM Transactions on Computer Systems, 2003, 21(4): 341-374.

[25] Costa P, Migliavacca M, Picco G P, Cugola G. Epidemic algorithms for reliable content-based publish-subscribe: an evaluationp[A]. Proceedings of the 24th International Conference on Distributed Computing Systems[C]. Washington: IEEE Computer Society Press, 2004. 552-561.

[26] Kermarec A M, Massoul' e L, and Ganesh A J. Probabilistic reliable dissemination in large-scale systems[J]. IEEE Transaction on Parallel and Distributed Systems, 2003, 14(3): 248-258.

[27] Gero Mühl, Ludger Fiege, Alejandro Buchmann. Filter similarities in content-based publish/subscribe systems[A]. Proceedings of the IEEE International Conference on Architecture of Computing Systems[C]. USA: ACM Press, 2002. 224-238.

[28] Mühl G. Generic constraints for content-based publish/subscribe[A]. Proceedings of the 6th International Conference on Cooperative Information Systems[C]. Trento: LNCS, 2001. 211-225.

[29] Gian Pietro Picco, GianPaolo Cugola. Efficient content-based event dispatching in the presence of topological reconfiguration[A]. Proceedings of the 23rd International Conference on Distributed Computing Systems[C]. Washington: IEEE Computer Society, 2003. 234-243.

[30] GianPaolo Cugola, Davide Frey. Minimizing the reconfiguration overhead in content-based publish-subscribe[A]. Proceedings of the 2004 ACM Symposium on Applied Computing [C]. New York: ACM Press, 2004. 1134-1140.

[31] Gianpaolo Cugola, Davide Frey. Content-based routing for publish-subscribe on a dynamic topology concepts, protocols, and evaluation[EB/OL]. http://dit.unitn.it/~picco/papers/psReconf.pdf, 2005.

[32] Paolo Costa, Davide Frey. Publish-subscribe tree maintenance over a DHT[A]. Proceedings of the 25th IEEE International Conference on Distributed Computing Systems Workshops [C]. Washington: IEEE Computer Society, 2005. 414-420.

[33] Rowstron A, Kermarrec A M. SCRIBE: the design of a large-scale event notification infrastructure[A]. Proceedings of the Third International COST264 Workshop on Networked Group Communication[C]. London: Springer-Verlag, 2001. 30-43.

[34] Virgillito A, Beraldi R, Baldoni R. On event routing in content-based publish/subscribe through dynamic networks[A]. Proceedings of the 9th IEEE Workshop[C]. USA: FTDCS, 2003. 322-329.

[35] Eiko Yoneki, Jean Bacon. An adaptive approach to content-based subscription in mobile Ad Hoc networks[A]. Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops[C]. USA: IEEE Computer Society, 2004. 92-97.

[36] Emmanuelle Anceaume, Ajoy K. Datta. Publish/Subscribe scheme for mobile networks[A]. Proceedings of the Second ACM International Workshop on Principles of Mobile Computing[C]. New York: ACM Press, 2002. 74-81.

[37] Baldoni R, Beraldi R, Cugola G, Migliavacca M, Querzoni, L. Structure-less content-based routing in mobile Ad Hoc networks[A]. Proceedings of the IEEE International Conference

on Pervasive Services[C]. USA: IEEE Computer Society, 2005. 37- 46.

[38] Vinod Muthusamy, Milenko Petrovic, HansArno Jacobsen. Effects of routing computations in content-based routing networks with mobile data sources[A]. Proceedings of the 11th Annual International Conference on Mobile Computing and Networking[C]. New York: ACM Press, 2005. 103- 116.

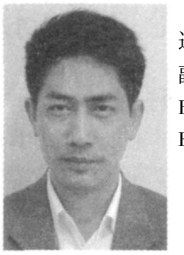
[39] Ioana Burcea, Hans Arno Jacobsen, Eyal de Lara, Vinod Muthusamy. Disconnected operation in publish/subscribe middleware[A]. Proceedings of 2004 IEEE International Conference on Mobile Data Management[C]. Berlin: LNCS, 2004. 39- 50.

[40] Caporuscio M, Carzaniga A, Wolf A L. Design and evaluation of a support service for mobile, wireless publish/subscribe applications[J]. IEEE Transactions on Software Engineering, 2003, 29(12): 1059- 1071.

[41] Caporuscio M, Carzaniga A, Wolf A L. An experience in evaluating publish/subscribe services in a wireless network [A]. Proceedings of the 3rd International Workshop on Software and Performance[C]. New York: ACM Press, 2002. 128 - 133.

[42] Ying Liu. Survey of publish subscribe event systems[EB/OL]. http://64.233.179.104/scholar?hl=zh-CN&lr=&newwindow=1&q=cache:BhBDONzf4bwJ:people.na.infn.it/~tortone/IMPORTANTTR574.pdf, 2003.

作者简介:



薛小平　男, 1963 年生于江苏金坛, 北京交通大学博士研究生, 同济大学信息与通信工程系副教授. 研究兴趣包括: 移动通信、路由理论、RFID 网络体系结构、安全苛求系统理论等.
E-mail: xuexp@mail.tongji.edu.cn

张思东　男, 1945 年生于山东寿光, 教授, 博士生导师. 主要研究领域为下一代互联网与无线传感器网络路由、资源分配与管理.
E-mail: sdzhang@center.njtu.edu.cn

张宏科　男, 1957 年生于山西大同, 教授, 博士生导师. 主要研究领域为下一代互联网与无线传感器网络路由、安全、服务质量.
E-mail: hkzhang@center.njtu.edu.cn