基于描述逻辑的语义 P2P 网络构架及资源定位*)

卢正鼎　孙小林　李瑞轩　文坤梅

(华中科技大学计算机科学与技术学院　武汉 430074)

摘　要　随着网络资源的日益增长,以及人们查询要求的复杂化,如何合理地在 P2P 网络中分配和查询资源已经变得极为重要。本文计划基于描述逻辑介绍一种应用在语义 P2P 网络上的算法思想,以期实现资源的概念化分布,使基于语义的查询和检索变得简单。我们的算法采用 Chord 算法的相容散列思想,将资源的关键字和资源所在节点的 IP 地址散列为相同的数据类型来进行实例选择。除此之外,每个节点拥有自己的本体系统,并将与其他节点交换 CHG 来达到知识库的完备。由于在 CHG 中所有的概念拥有同一个根结点,所以不停地向层次分类的概念树上层节点询问,一定可以找到目标资源的信息。

关键词　语义 P2P,描述逻辑,相容散列,本体,CHG

The Construction and Resource Locating of the Semantic P2P Grid Based on Description Logics

LU Zheng-Ding　SUN Xiao-Lin　LI Rui-Xuan　WEN Kun-Mei

(Dept. of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074)

Abstract　Since the resources increase day by day and the inquiry requests become more and more complicated, how to distributing and querying resources in P2P network in reason has already been extremely important. This paper proposes an algorithm applied in semantic P2P network based on the description logics with the purpose for realizing the concepts distribution of resources, which makes the resources semantic locating easy. With the idea of the consistent hashing in the Chord, our algorithm stores the addresses and resources with the values of the same type to select instance. In addition, each peer has its own ontology, which will be completed by the knowledge distributed over the network during the exchange of CHGs. The hierarchy classification of concepts allows to find matching resource by querying to the upper level concept because the all concepts described in the CHG have the same root.

Keywords　Semantic P2P, Description logics, Consistent hashing, Ontology, CHG

1　简介

当今,资源的概念被应用到非常广泛的领域,不同领域各种各样可获得的资源呈爆炸性增长。与此同时,随着本体与语义网的出现,对资源查询的要求也变得越来越高、越来越复杂。这就意味着有些时候我们需要分析语义的元素。因此,P2P 网络中的资源定位和匹配变得越来越困难。

语义网格将注意力放在在一些集中式[1]的语义方法上,例如:网格中间件上的本体和描述逻辑,并使用本体在资源匹配的问题上取得了一些成果[2~4]。但是,上述文献使用的仅仅是本地的本体,并且假定资源查询者和资源提供者从一开始就拥有相同的本体知识库。这就表明资源的节点必须共享同一个知识库(TBox, ABox)。但是,由于知识库每时每刻都在动态地扩展,所以将完整的、统一的知识库配置到所有的资源节点来保证它们拥有及时的、一致的本体是非常困难的。所以,一些基于本体的 P2P 网格资源定位和匹配的研究提出分布式的节点应该使用分布式的知识库[5]。这意味着所有的节点可以独自地进行推理和路由搜索。

本文首先简单介绍所要涉及到的方法:描述逻辑(DL)、分布式哈希表(DHT)和 Chord 算法,它们是我们算法的基础;第 2 节详细介绍每个资源上的 DL 系统的基础设计和格式;在第 3 节描述寻找资源节点算法的核心思想;最后给出结论并展望下一步的工作。

2　相关技术

上面已经提到,我们的目标是建立一个分布式的演绎系统来实现语义资源的定位。对我们作为基础的推理系统以及分布式点对点计算的研究已经有了很多成果。接下来,我们介绍以下几种技术:描述逻辑、分布式哈希表和 Chord 算法思想。

2.1　描述逻辑(DL)

描述逻辑是一个最近才出现的名词,是知识表示(Knowledge Representation)家族的一个新的名称。它定义域内相关概念来表示应用域知识并且使用这些概念来确定事物的属性和应用域内出现的个体实例(ABox)。

概念的层次分类决定了术语集(TBox)的概念之间子概念 sub-concept(父概念 super-concept)的关系(包含与被包含),因此描述逻辑允许在术语集内定义包含的层次关系。这种层次关系为不同的概念之间提供了非常有用的信息,用来建立彼此的联系。除此之外,这种层次关系还可以提高推理服务的效率。

为了建立知识库,推理问题的内容并且处理它们,基于描述逻辑的知识表示系统提供了许多工具和模块。这样一个知识库包含两个部分:TBox 和 ABox。其中 Tbox 陈述例如术语和公理等等应用域的词汇表。同时,ABox 根据词汇表提供命名的个体的断言。词汇表由表示个体集和角色集的概念组成,其中角色集表示的是两个个体之间的二元关系。一个

*)基金项目:国家自然科学基金(No:6040327)。孙小林　博士研究生,研究方向:数据挖掘,描述逻辑,语义 Web。

描述逻辑系统不仅仅存储术语和断言，还提供推理它们的各种服务[6]。我们的描述逻辑系统将把重点放在查询方法，所以在这里暂时不讨论推理算法。

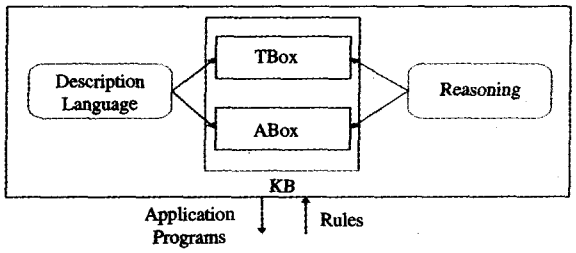


图 1　基于描述逻辑的知识表示系统的架构

2.2　分布式哈希表(DHT)

分布式哈希表意味着 P2P 网络中每一个对等节点拥有自己的哈希表。这样通过增加术语集来扩展一个 P2P 网络就变得非常容易。第二代 P2P 网络都是基于分布式哈希表(DHT)算法的，这种算法根据包含信息的对等节点所拥有确定的关键字来定位节点在 P2P 网络中的位置。现在存在许多不同的 DHT 算法，我们现在使用相容散列算法 Chord[7]，这个算法在大小为 n 的 P2P 网络中(n 表示网络中节点个数)找到目标节点的算法的复杂度为 O(log(n))。

Chord 是一种根据不同的唯一标识将资源和节点分布在 P2P 网络上的 DHT 算法，它采用变种的计算相容哈希值的方法。相容散列有几个很好的特点，首先是散列函数可以做到负载平衡，也即所有的节点都可以接收到基本相同数量的关键字。另外，当第 n 个节点加入或者离开系统时，只有 $1/n$ 的关键字需要移动到另外的位置。相容散列可以根据哈希算法生成的唯一的标识符分布每个资源和节点的关键字。节点的标识符可以通过散列节点的 IP 地址产生，而关键字的标识符可以直接散列此关键字。例如 IP 地址为 120. 10. 10. 1 的节点经过散列之后得到的标识符为 54，而关键字“LifAtGo”散列之后的关键字为 30。

3　P2P 网络中的描述逻辑系统

我们的算法是基于构造 P2P 网的每个节点上 DL 系统。分类层次图是分布式 DL 系统中保证其组织一致很重要的部分。

3.1　分类层次图 (CHG)

每个对等节点拥有本地的描述逻辑系统并且使用这个描述逻辑系统来存储它自己的资源列表。网络中某一确定的资源可以通过 ABox 中的实例断言来描述出来。每一个资源可以描述为一个被声明为某个或多个概念的成员的个体实例。这些概念描述资源所有的特征。由于 P2P 网络是一个开放的系统，所以强迫每个对等节点使用相同的 TBox 词汇表示很困难的。

当不同 TBox 分布在各个节点上的时候，我们使用概念的分类层次图 (CHG)来让不同的对等节点协同工作。就像前面提到的，DL 提供概念之间的包含关系，我们称之为子概念和父概念。根据这种包含关系我们计算分类图。分类图中以包含关系为边，以概念为顶点。DL 系统中存在一个通用概念用来包含所有 DL 系统中所有声明的概念。我们指定概念“Resource”作为整个 CHG 中分类树的根节点。图 2 展示了这种非循环的概念分类层次图。

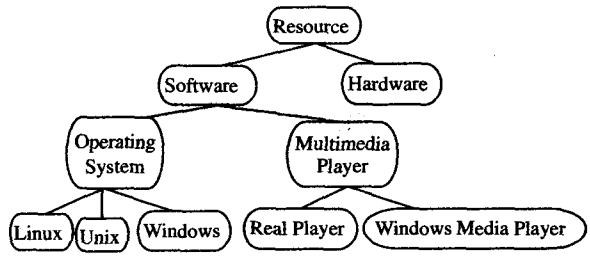


图 2　分类层次图 (CHG)

当节点加入 P2P 网络的时候会将自己的本地 CHG 发送给 P2P 网中离自己最近的节点。由于 P2P 网中不同的对等节点拥有相同的 CHG，新加入的节点收到一张唯一的 CHG 后会进行 CHG 的交换。在 CCG 的交换过程中，如果收到新加入节点的 CCG 与本地的 CCG 有冲突，则 P2P 网络的节点会根据本地的 DL 系统检查 CCG 的正确性。如果本地 CCG 无误，则需要将正确的 CCG 发还给申请加入的节点，否则将向 P2P 网络中所有节点发送错误报告和 CCG 修改信息。

这样就可以保证 P2P 网络中的节点的概念层次的一致性，有助于资源的搜索和定位。这里我们要注意的是，同一个概念可能会被分配给不同的节点，所以概念层次图虽然是不循环的树状结构，但是节点之间的关系却要复杂得多。

3.2　分布式的知识库

同样，分布式的 DL 系统中也有 TBox 和 ABox 两个部分。我们的目标是为一个允许 ABox 和 TBox 都分布在各个节点上的 P2P 网络提供分布式资源搜索、实例检测以及公理推理等服务。

在我们的模型里，每一个对等节点拥有自己的知识库，如图 1 所示，TBox 和 ABox 都有着自己的格式。我们在这里定义公理集中两种格式中的包含关系来表示节点与节点之间的关系作为分布式 TBox 中的主要内容。其格式如下：

$$ConceptX \subseteq Super\text{-}conceptY$$

$$Sub\text{-}conceptA_1 \subseteq ConceptX$$

$$Sub\text{-}conceptA_2 \subseteq ConceptX$$

……

$$Sub\text{-}conceptA_n \subseteq ConceptX$$

其中 X 代表该节点所表示的概念，Y 表示该节点的 super-concept。我们更倾向于使用“super-concept”和“sub-concept”来表示那些与目标节点有着直接包含或被包含关系的节点。如果不是直接包含或者被包含的关系，而是间接的可推理(k_{m+1}，k_{m+2}，k_{m+n})出来的包含关系，我们可以用“super-super-concept”和“sub-sub-concept”来表示。举个例子，如果概念 A 是概念 B 的 super-concept，而 B 又是 C 的 super-concept，我们就称概念 A 是概念 C 的 super-super-concept。就像我们提到的，在 CCG 中我们定义的概念最多只能有一个 super-concept，所以我们可以从一个概念节点的 TBox 和 ABox 中找到 super-concept 的名字和 sub-soncept 的列表。

在每个节点的 ABox 中会存放本地节点所拥有的资源列表：除此之外，节点应该保存其 sub-concept (k_1，k_2，…，k_m)和 super-concept(k_{m+n+1}) 节点的 IP 地址的哈希散列值。这些声明的格式如下：

$$X(k_1), X(k_2), \cdots, X(k_m)$$

$$A_1(k_{m+1}), A_2(k_{m+2}), \cdots, A_n(k_{m+n}), Y(k_{m+n+1})$$

此外，我们利用 DL 的思想给每个知识库赋予一定的基

本规则。例如资源实例声明 X(k)中的 k 值不允许在该节点其他的资源实例声明中出现，或者 TBox 中关于 super-concept 的定义必须有且仅有一条等等。

3.3 语义 P2P 系统的构建

本节我们将会通过 peer 加入和离开的例子来展现基于分布式 DL 系统的 P2P 平台的构建过程。

3.3.1 加入 P2P 网络的过程

当一个节点想要成为 P2P 网络中的对等节点，首先要做的是由底层的本地 DL 系统决定其概念名称。图 3 显示了一个名为“Windows”的概念节点(节点 1)的 TBox 和 ABox 的细节。其中值 22 作为 DHT 算法如 Chord 根据资源关键字算出的哈希值。值 17 是描述概念“Windows”的节点 IP 地址的相容散列值。接着，名为“Windows”将它的 CHG 发送给 P2P 网格中最近的节点，并与之交流信息确定正确的 CHG。正如我们在 2.1 中讨论的一样会有两种交换过程，但最后会达到 P2P 网络 CHG 的统一。

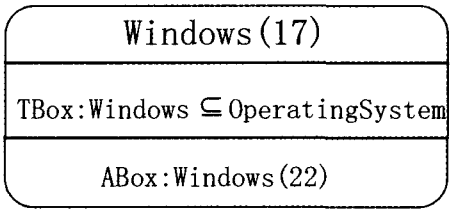


图 3 表示“Windows”概念的节点

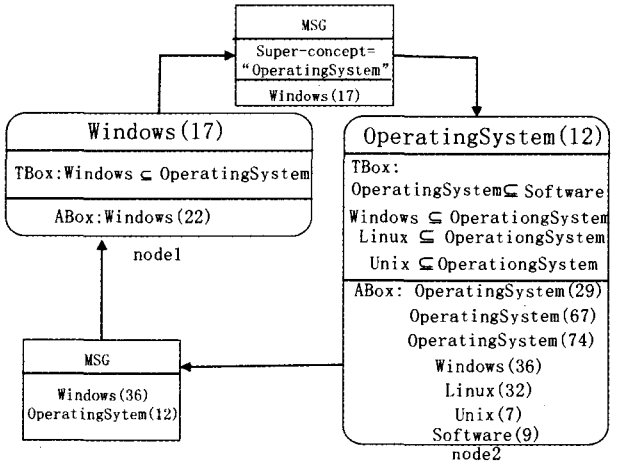


图 4 发送消息第一步

节点 1 为了找到它在网络中的位置，它必须向 P2P 网络中的节点发送两种消息，如图 4、5 所示。第一种消息包括其 super-concept 的概念名。凡是可以表示节点 1 所标示的概念“Windows”的 super-concept——“Operating System”的节点(节点 2)都会收到这个消息。如果该网络中已经存在名为“Windows”的节点，则节点 2 返回一个包含这些已存在的同名节点信息的回应消息。如果节点 1 在目标网络中找到相同的概念节点，它将向最靠近它的同名节点发送第二种消息，其中包括节点 2 的本地资源列表。接着，节点 2 会收到 P2P 网络中“Windows”节点(节点 3)的回应信息包括“Windows”概念的 sub-concept 列表。节点 3 会检查收到的资源实例列表并与本地的资源列表进行合并然后将这张资源列表和 sub-concept 列表一同发还给节点 1 来实现同概念的节点拥有完全相同的知识库。

如果概念“Windows”在网络中是一个新的没有定义过的概念，那么节点 1 在修改了 P2P 网中的 CCG 后并且发送了第一种 message 后就不必再寻找已经存在的同名概念的资源列表和 sub-concept 列表。

节点加入完成情况后见图 6。我们可以发现节点 1 和节点 3 有同样的 TBox 和 Abox，节点 2 Abox 中关于 Windows 的断言也变成了 Windows(17)。

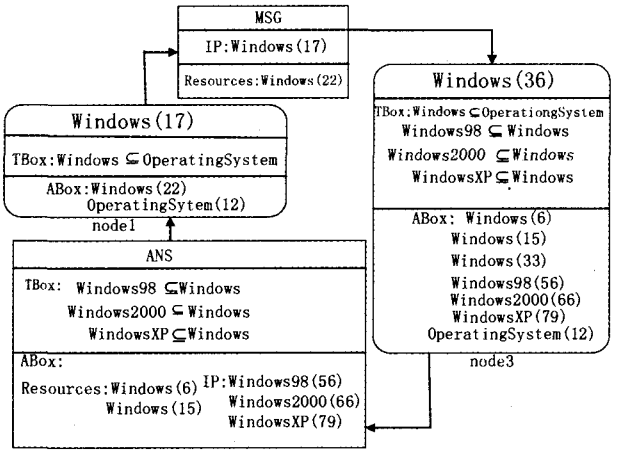


图 5 发送消息第二步

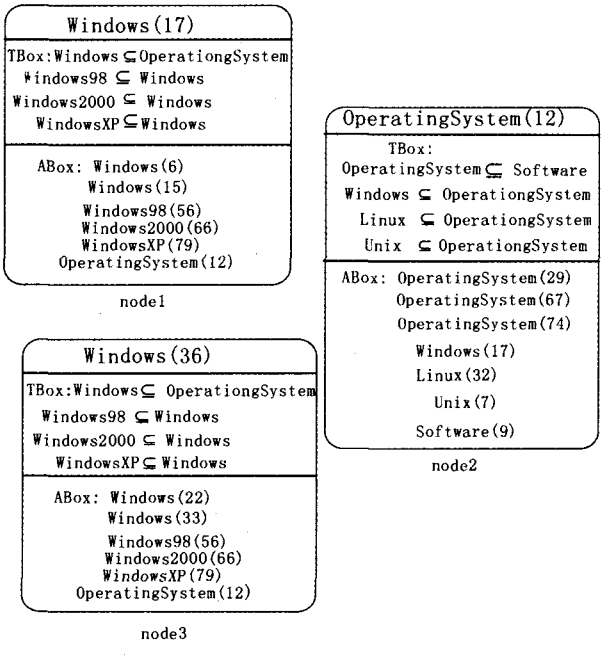
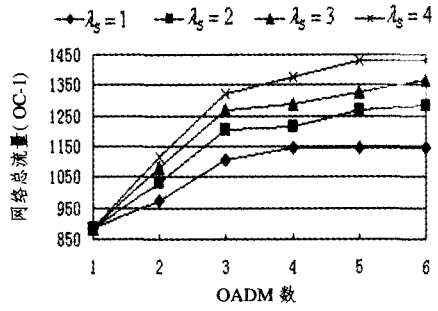


图 6 消息传递后 3 个节点各自的信息

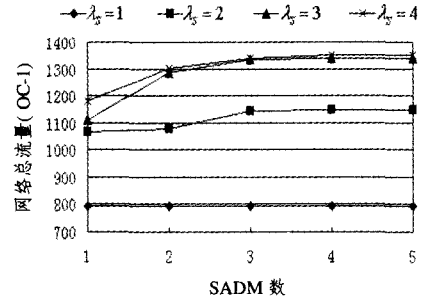
3.3.2 peer 离开 P2P 网络的过程

我们这里只考虑 peer 正常离开的情况，不对意外退出的情况进行研究。peer 离开 P2Pnetwork 的过程与加入很相似。首先也是向其 super-concept 节点发送消息，然后 super-concept 节点根据退出节点是否最后一个表示该概念的节点作出不同的回应。如果还有表示该概念的节点存在于网络中，super-concept 节点只需向其剩下的所有 sub-concept 节点发送资源列表修改消息并修改自己的 ABox。否则，super-concept 节点只用修改自己的 ABox 来除去离开的节点，但是不需要 CCG 和 TBox 进行修改，因为公理是永远正确，概念是永远存在的，即使没有实例进行解释。

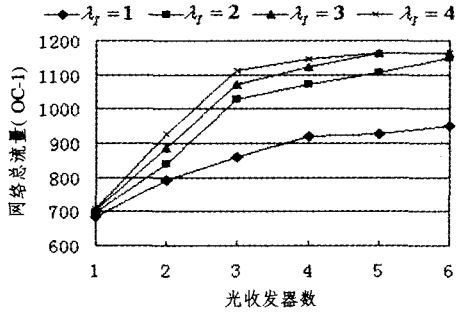
(下转第 39 页)



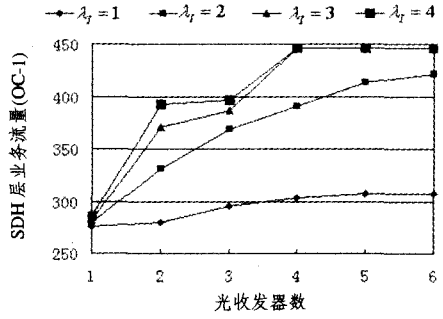
(a)SDH/WDM 层间配置对网络流量的影响



(b)IP/SDH/WDM 层间配置对网络流量与影响



(c)IP/WDM 层间配置对网络流量的影响



(d)IP/WDM 层间配置对 CDH 层业务流量的影响

图 3　不同层间资源配置下的网络流量

(下转第 64 页)

(上接第 35 页)

4　寻找定位资源节点

介绍了本体驱动的 P2P 网格之后，我们来讨论一下 P2P 网络中的 peers 是如何通过语义查询寻找并定位它们所需要的资源的。

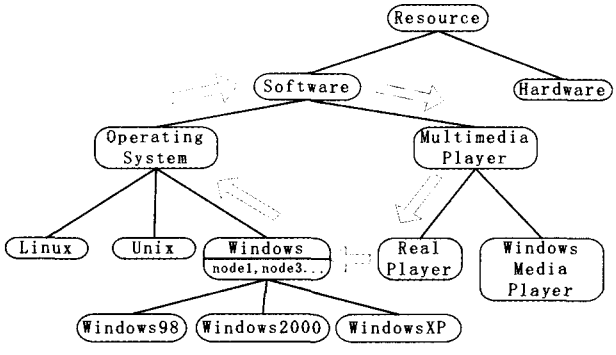


图 7　查询过程

我们仍然使用以上的例子节点定位的过程。如果节点 1 希望找到一种名字为“RealPlayer”的多媒体播放软件资源，它首先要做的就是检查自己的 CHG。由于 CHG 是一棵根节点为“Resource”的树，所以很容易找到一个概念节点具有如下特征：以该节点为根的子树必然包含目标概念“RealPlayer”和源概念“Windows”。接着选择能够表示具有这个特征的概念的最近的节点。在我们的例子里面，节点 1 找到概念“Software”并且可以不停地向上(父节点)询问确定最近表示该概念的节点的位置。被节点 1 定位的节点会收到一个类似 IP 报的信息，其中有一个包含目标概念名的栈和源节点(节点 1)的 IP 地址信息。栈内信息依次为：“Software”，“multimedia player”以及 “Real Player”。由于这个被定位的节点是标示概念“Software”的，则将与之同名的元素弹出并丢弃然后将这个被修改过的栈发送给自己的所有 sub-concept 节点。当 sub-concept 中又能够与“multimedia player”同名的节点则重复以上动作直到找到目标节点。目标节点发现栈为空的时候就知道自己就是这次查询的终点，此时按照该信息包内的 IP 地址将自己的资源列表发送过去。以上过程如图 7 所示。

结论　本文给出了一种新颖的 P2P 网格分布式资源定位的方法。我们的系统使用分布式的基于描述逻辑的分布式本体系统来描述不同的资源。通过哈希表算法将信息分布在整个点对点网络上。我们大致给出算法的核心思想并通过整合分布式的 TBox 来保证系统的正确性。通过分布式的 CHG 就可以合并所有节点的知识库就可以解决分布式 DL 系统带来的问题。我们的算法确保不同的 peer 上保存的信息量大致相同，这样可以达到负载平衡并使每个节点更为平等。

但是，文中几乎没有对推理算法进行讨论，尽管推理算法是 DL 系统中最重要的一个部分。我们将来的工作就是在这个分布式 DL 系统的基础上实现分布式推理系统。除此之外，基于本体的系统的完备性，查询的可表示性以及意外检测都将是我们研究的下一步工作。

参 考 文 献

1　Goble C, Roure D D. The Grid: An Application of the Semantic Web. SIGMOD Rec, 2002, 31(4): 65～70

2　Brooke J, Fellows D, Garwood K, et al. Semantic matching of Grid Resource Descriptions. In: 2nd European Across-Grids Conference, 2004

3　Tangmunarunkit H, Decker S, Kesselman C. Ontology-based Resource Matching in the Grid - The Grid Meets the SemanticWeb. In: Proceedings of SemPGRID '03, 2003

4　Gonz'alez-Castillo J, Trastour D, Bartolini C. Description Logics for Matchmaking of Services. In: Görz G, Haarslev V, Lutz C, et al., ed. Proceedings of the KI-2001 Workshop on Applications of Description Logics, vol 44, 2001

5　Heine F, Hovestadt M. Towards Ontology-driven P2P Grid Resource Discovery. In: Proceedings of the Fifth IEEE/ACM International Workshop on Grid Computing, 2004

6　Baader F, Calvanese D, McGuinness D, et al. The Description Logic Handbook. Cambridge University Press, 2003

7　Stoica I, Morris R, Liben-Nowell D, et al. Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications. IEEE Transactions on Networking, 2003(11)