面向虚拟组织资源发现的语义模糊匹配*)

唐九阳 张维明 肖卫东 宋峻峰 汤大权
(国防科技大学信息系统与管理学院 长沙 410073)

摘 要 资源发现是虚拟组织提供资源共享和协同工作的前提。本文首先提出一种支持资源动态匹配的资源发现结构;然后针对传统资源匹配技术的不足,在资源元数据本体描述的基础上,结合概念的语言特征和语义特征,提出一种语义模糊匹配算法。通过语义相似度的计算,进而对语义关联进行分类,实现了一定程度的资源模糊匹配,为资源发现提供了新的解决方法。

关键词 虚拟组织,资源发现,概念上下文,语义模糊匹配,语义相似度

Semantic Fuzzy Matching towards Resource Discovery in Virtual Organization

TANG Jiu-Yang ZHANG Wei-Ming XIAO Wei-Dong SONG Jun-feng TANG Da-Quan
(School of Information System and Management, National University of Defense Technology, Changsha 410073)

Abstract Finding appropriate resource within the virtual organization (VO) is frequently a key and difficult work. First a flexible resource discovery architecture for resource dynamic matching is presented in this paper. Based on ontological descriptions of the metadata characterizing the resources to be shared, a semantic fuzzy matching algorithm is proposed by exploiting the linguistic and contextual characteristic of concepts. Through computing the semantic affinities between ontological resource descriptions and a target resource request, and classifying the semantic association, resource fuzzy matching is supported, which gives a new resolution for resource discovery.

Keywords Virtual organization, Resource discovery, Concept context, Semantic fuzzy matching, Semantic affinity

计算机网络的普及和WWW的迅猛发展,使得可访问的信息源和异构计算资源的数量迅速增加。为提升组织核心竞争力、快速响应环境需求和变化,虚拟组织综合利用信息和网络技术将时空上分布但能力和资源互补的组织高效地组合起来,形成动态联盟,是组织、人力、技术、信息等资源在完善的网络组织结构基础上的有效集成[1,2],为企业组织、科研机构以及电子商务等领域的蓬勃发展奠定了基础。

为了分散和简化应用逻辑,提高组织单元可重用性,单个组织单元都不可能做得非常复杂。基于分布式协作的组织单元为客户提供了不同类型的资源,而这些资源可能具有不同的形式,并且它们的复杂程度不相同。面对数量庞大的语义异构资源,基于给定所需资源的描述,返回与之相匹配的资源集合,成为当前虚拟组织资源发现的关键[3]。

虚拟组织中资源匹配方式通常局限于按照预定的模式。用户表达一个查询后,系统只在查询和内部表达之间执行关键字或者子串的匹配。典型的如Condor Matchmaker[4]和PBS[5],要求资源提供者和资源需求者之间进行(属性,值)元组匹配。Redline匹配系统[6]借助现有的约束求解技术,将属性匹配问题转化到约束满足求解。这些方法停留在语法层次,没有利用任何语义信息,用户难以得到语义上相关联的、更多的数据。为了解决上述方法的不足,文[7,8]和Edamok[9]借助逻辑描述资源概念及上下文,通过逻辑推理实现语义匹配。然而逻辑表示和逻辑推理都具有很高的计算复杂性,实用性不强。文[10]基于分布式本体间的动态匹配实现资源发现,不能较好地支持基于语义约束的模糊匹配,使得资源发现执行的整个过程受到影响。

本文首先提出一种支持资源动态匹配的分布式资源发现结构。在资源元数据的本体描述的基础上,从概念的名称和概念间的语义关系两方面着手,结合概念的语言特点和语义特点,提出一种语义模糊匹配算法。通过对语义关联进行分类,辅助用户进行资源模糊查找。最后以一个实例检验了本方法的应用。

1 资源发现的结构

目前虚拟组织中的资源发现基本上采用以第三方服务实现的集中式方式[11,12],即利用一个集中式的结点负责搜集虚拟组织内各资源的信息,用户可以通过该结点查找虚拟组织范围内所有满足特定条件的资源。尽管这种集中式的方式实现简单,但随着虚拟组织规模的增大、资源数量和类型的丰富,集中式的资源发现方式将会产生可伸缩性差、单点失效等问题。

因此,应对虚拟组织向更大规模发展的趋势,需要一种不依赖集中控制的、分布式、可扩展、能适应资源动态变化并且定位性能好的资源发现方式。借鉴对等网的研究成果,一个自然的解决办法是允许每个组织单元设置一个或多个提供本地资源信息存储和访问的信息结点,控制对其本地共享资源信息的访问。信息结点接收到虚拟组织网络上的任一资源请求,都可进行本地动态匹配。系统的总体结构如图1所示,包

*)国家自然科学基金资助(60172012)、湖南省自然科学基金重点项目资助(03JJY3110)项目。唐九阳 博士研究生,主要从事信息系统集成、对等网、智能决策支持技术等领域研究。

含以下几个模块。

用户接口：接受用户的资源查询请求，同时对用户的输入做预处理，完成概念的标准化，并负责显示查询结果；

匹配模块：计算资源请求中的本体概念和本地组织单元本体概念的匹配程度。

通信管理器：负责本地组织单元与网络中其它组织单元的通信，包括接受资源请求、返回匹配结果等。

查询管理器：执行查询处理和结果整合。当查询管理器接收到其它组织单元的查询请求时，首先进行解析，抽取出目标概念的本体描述，交给匹配模块，根据相应的匹配算法，得到满足用户需要的本地组织单元匹配概念，返回相应的数据，流程如图中标注的数字次序。图中的虚箭头则描述了本地提交查询请求的处理过程。用户通过用户接口提交查询请求，查询管理器分两步处理：一方面在本体管理器的元数据库中直接查找，另一方面通过通信管理器向网络中的其它组织单元发出查询请求，最后对返回的所有结果进行整合。

本体管理器：对本地组织单元的资源元数据提供本体描述，并提供本体编辑工具。

虚拟组织中的组织单元提供不同类型的资源，可以是数据、服务和物理资源等。为解决资源异构问题，本文中组织单元的共享资源应用本体描述，简称组织单元本体。一个组织单元本体 UO 定义为三元组：

$$UO=(C,P,SR) \tag{1}$$

其中 C 为组织单元中的概念集合；P 为概念集 C 的属性集合。对任一属性 $p\in P$，存在 $c\in C$ 使得 $p(c)$ 成立；SR 为概念间的语义关系，有：

$$SR=\{equals,similar\text{-}to,part\text{-}of,includes,overlays\} \tag{2}$$

其中 *equals* 关系描述现实世界中语义相同的两个概念，如 *equals*(Father,Dad)；*similar-to* 关系描述现实世界中语义相似的概念，如 *similar-to* (Book,Volume)；*part-of* 关系表示语义上概念前者是后者的组成部分，如 *part-of* (Tyre,Car)；*includes* 关系表示概念前者的语义包含后者，如 *includes* (Document,Article)；*overlays* 关系表示有别于上述的概念间有语义重叠的关系，如 *overlays* (Journal,Book)。

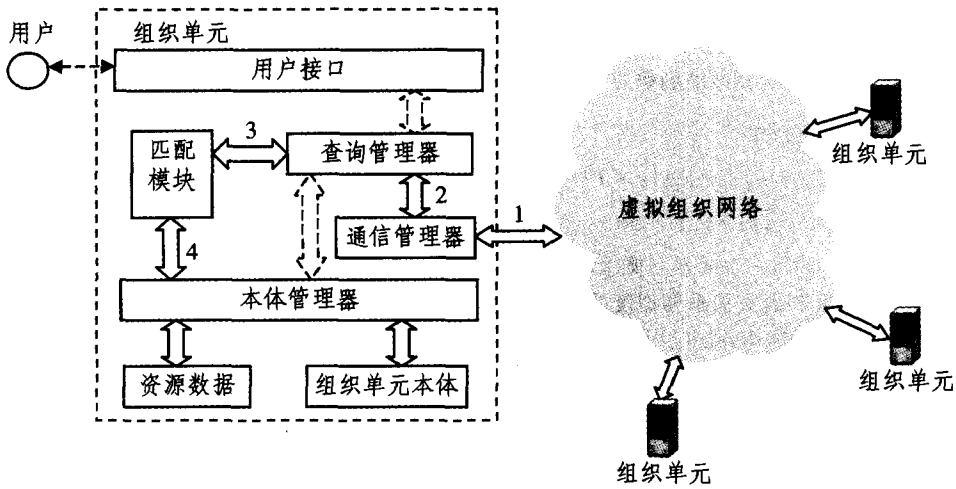

图 1　资源发现的结构

2　语义模糊匹配

本文中虚拟组织的资源匹配是针对资源请求中的目标概念，计算它与组织单元本体中各个概念的语义关联类别。从语义的角度出发，概念的内涵不仅取决于概念名和概念属性，还依赖它和其它概念的语义关系。因此，从两个方面综合考虑概念的名称以及概念间的语义关系对资源匹配的影响。

概念名和属性名由词汇表示，词汇的含义取决于它和其它词汇的词汇关系，这种词汇关系往往通过语言学本体定义。语言学本体指关于语言、词汇等的本体，典型的如 WordNet[13] 采用语义网络作为其词汇本体的基本表示形式，其中节点词汇之间的关系分为同义关系(Synonymy)，反义关系(Antonymy)、上义/下义关系(Hypernym/Hyponym)、继承关系(Hyponymy)、部分/整体关系(Meronymy)等。为简化对问题的讨论，参考统一词典 WordNet，我们采用具有代表意义的同义关系、上义/下义关系和部分/整体关系作为词汇关系，并定义相应的关联权值(表 1)，量化地表征两个节点词汇的关联程度。

在语义网络中，如果词汇 t 和 t' 之间存在 k 条可达路径，$t\rightarrow_i^m t'$ 代表第 i 条路径，路径长度为 $m(m\in Z$ 且 $m\geqslant 1)$，σ_{ij} 为该路径上词汇 t_i 和 t_j 的关联权值，则第 i 条路径的关联强度为该路径上词汇之间的关联权值的乘积：

$$\sigma(t\rightarrow_i^m t')=\sigma_{12}\cdot\sigma_{23}\cdot\cdots\sigma_{(m-1)m} \tag{3}$$

那么词汇的名称相似度定义为它们之间存在的所有可达路径的最大关联强度：

$$LA(t,t')=\begin{cases}\max\limits_{i=1\cdots k}\{\sigma(t\rightarrow_i^m t')\} & \text{if}\quad k\geqslant 1\\ 0 & \text{otherwise}\end{cases} \tag{4}$$

例如 Hypernym(Publication, Book)，Meronymy (Book, Title)，则 Publication 到 Title 之间存在一条关联路径，其关联强度是 $\sigma_{12}\cdot\sigma_{23}=0.8\cdot 0.5=0.4$。

从准确获取语义信息的角度出发，引入概念上下文。

记 $P(c)$ 为概念 c 的属性集合，有 $P(c)=\{p_i\mid p_i(c)\}$。与 c 有语义关系的邻接概念集合为 $SR(c)=\{c_j\mid sr_j(c,c_j),sr_j\in SR\}$。概念上下文 $Ctx(c)$ 表示 c 的属性以及和 c 有语义关系的概念与相应语义关系的笛卡儿乘积的集合(概念的属性也看成一种特殊的语义关系)：

$$Ctx(c)=\left\{(f,r)\mid(f,r)\in\begin{Bmatrix}(p_i(c),property)\cup\\(c_j,sr_j(c,c_j)),p_i(c)\in P(c),\\sr_j(c,c_j)\in SR(c)\end{Bmatrix}\right\} \tag{5}$$

为建立概念上下文的度量，首先定义语义关系相似度

语义关系相似度是对两个语义关系的相似度或者语义关系和属性的相似度进行度量，表示为：

$$RA(r,r')=1-|\sigma_r-\sigma_{r'}| \tag{6}$$

其中 σ_r 和 $\sigma_{r'}$ 分别是语义关系 r 和 r' 的权值，由专家设定，见表1。

例如，part-of 和 property 的语义关系相似度

$RA(part\text{-}of,property)=1-|0.7-1|=0.7$

表1 词汇关系和语义关系的权值

	关系	权值
语言特点	Synonymy	1.0
	Hypernym/Hyponym	0.8
	Meronymy	0.5
语义特点	property	1.0
	equals	1.0
	similar-to	0.8
	part-of	0.7
	includes	0.5
	overlays	0.3

上下文相似度是对两个概念的上下文语义的相似度进行度量。

记 $|Ctx(c)|$ 和 $|Ctx(c')|$ 分别为概念 c 和 c' 的上下文集合中元素的个数，概念 c 和 c' 的上下文相似度函数为：

$$CA(c,c')=\frac{1}{|Ctx(c)|+|Ctx(c')|}\cdot\{\sum_{(f,r)\in Ctx(c)}\max_{(f',r')\in Ctx(c')}LA(f,f')\cdot RA(r,r')+\sum_{(f',r')\in Ctx(c')}\max_{(f,r)\in Ctx(c)}LA(f',f)\cdot RA(r',r)\} \tag{7}$$

计算概念 c 上下文集合中的每一个元素与概念 c' 上下文集合中所有元素的名称相似度和语义关系相似度的乘积，取其中的最大值累计求和。同样的计算方法应用于概念上下文集合中的每个元素。最后对所有结果求平均。

有了名称相似度和上下文相似度，接下来定义语义相似度函数：

$$SA(c,c')=W_{LA}\cdot LA(t,t')+(1-W_{LA})\cdot CA(c,c') \tag{8}$$

其中 t 和 t' 分别为概念 c 和 c' 的名称，W_{LA} 权值的设置是用来调整概念名称和上下文语义对语义相似度的影响程度，本文中默认 $W_{LA}=0.5$。

由于每个人对于客观世界的认识不同，对于同一概念的理解也会有所差异，因此很难进行概念间的严格匹配。为此引入模糊集的理论，根据语义相似度对语义关联进行模糊分类，辅助用户理解语义相似度计算反映的语义内涵。

按由弱到强的次序将语义关联分为“弱关联、次关联、强关联”3类，它们分别对应于模糊集 $\widetilde{A}_1,\widetilde{A}_2,\widetilde{A}_3$。设论域 $X=[0,1]$，而且

$$\widetilde{A}_1(x)=\begin{cases}1 & 0<x\leqslant 0.2\\ 1-2\left(\frac{x-0.2}{0.2}\right)^2 & 0.2<x\leqslant 0.3\\ 2\left(\frac{x-0.4}{0.2}\right)^2 & 0.3<x\leqslant 0.4\\ 0 & 0.4<x\leqslant 1\end{cases} \tag{9}$$

$$\widetilde{A}_3(x)=\begin{cases}0 & 0<x\leqslant 0.5\\ 2\left(\frac{x-0.5}{0.2}\right)^2 & 0.5<x\leqslant 0.6\\ 1-2\left(\frac{x-0.7}{0.2}\right)^2 & 0.6<x\leqslant 0.7\\ 1 & 0.7<x\leqslant 1\end{cases} \tag{10}$$

$$\widetilde{A}_2(x)=1-\widetilde{A}(x)-\widetilde{A}_3(x) \tag{11}$$

根据最大隶属原则，如果 $\widetilde{A}_1,\widetilde{A}_2,\cdots,\widetilde{A}_n\in F(X)$，$x_0\in X$，存在 $i:1\leqslant i\leqslant n$，使得

$$\widetilde{A}_i(x_0)=\bigvee_{1\leqslant j\leqslant n}(\widetilde{A}_j(x_0)) \tag{12}$$

则认为 x_0 相对隶属于模糊集 $\widetilde{A}_i$。

因此，只要给出一个语义度量 x，由(9～11)式建立的隶属函数即可分别计算隶属于3个模糊集的隶属度，再根据最大隶属原则(式12)，即可将 x 进行模糊分类。

最后给出语义模糊匹配算法 SFM。

输入：两个组织单元的概念 c 和 c'，$W_{LA}=0.5$；

输出：概念的语义关联类别。

算法描述：

①t 和 t' 分别为概念 c 和 c' 的名称，$x=0,y=0,z=0,f=0$；

②分别对概念 c 和 c' 的属性和具有语义关系的相邻概念遍历生成上下文序列 $Ctx(c)$ 和 $Ctx(c')$；

③$x=LA(t,t')$；

④$y=CA(c,c')$；

⑤$z=W_{LA}\cdot x+(1-W_{LA})\cdot y$；

⑥$f=\max\{\widetilde{A}_1(z),\widetilde{A}_2(z),\widetilde{A}_3(z)\}$；

⑦返回 f 对应的 $\widetilde{A}_i(z)$ 的语义关联类别集合。

3 应用实例

作为资源匹配的一个例子，考虑组织单元本体1和组织单元本体2(图2，灰色的椭圆表示概念的属性)。组织单元1的用户提出基于概念 Article 的资源请求。组织单元2接收到请求后，应用语义模糊匹配算法 SFM 将资源请求中 Article 的本体描述与自身本体概念 Book，Library，Chapter 进行匹配。示例中仅给出 Article 和 Book 的匹配过程。

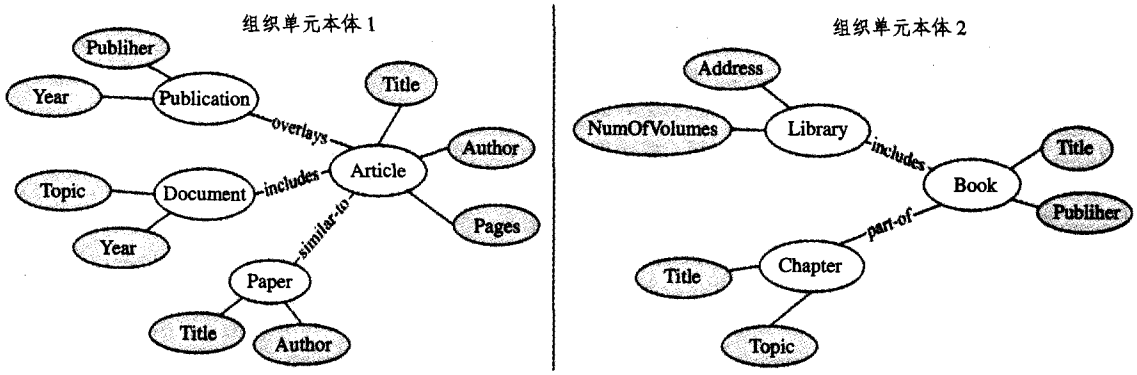

图2 组织单元本体

有 Article 的概念上下文：

$$Ctx(Article)=\left\{\begin{array}{l}(Title,property)\\(Author,property)\\(Pages,property)\\(Publication,overlays)\\(Document,\ includes)\\(Paper,similar\text{-}to)\end{array}\right\}$$

Book 的概念上下文：

$$Ctx(Book)=\left\{\begin{array}{l}(Title,property),(Publisher,property),\\(Library,includes),(Chapter,part\text{-}of)\end{array}\right\}$$

Article 与 Book 概念上下文的名称相似度和语义关系相似度如表 2。

表 2 Article 与 Book 概念上下文的名称相似度和语义关系相似度

LA()	Title	Author	Pages	Publication	Document	Paper
Title	1.0	0.25	0.25	0.4	0.4	0.4
Publisher	0.25	0.0	0.25	0.5	0.5	0.5
Library	0.25	0.25	0.0	0.5	0.5	0.5
Chapter	0.25	0.25	0.25	0.64	0.8	0.5

RA()	property	property	property	overlays	includes	similar-to
property	1.0	1.0	1.0	0.3	0.5	0.8
property	1.0	1.0	1.0	0.3	0.5	0.8
includes	0.5	0.5	0.5	0.8	1.0	0.7
part-of	0.7	0.7	0.7	0.6	0.8	0.9

那么，有计算结果：

$LA(Article,Book)=0.5$；

$CA(Article,Book)=0.553$。

$SA(Article,Book)=0.5\cdot 0.5+0.5530.5=0.5265$。

$\widetilde{A}_1(0.5265)=0$，$\widetilde{A}_2(0.5265)=0.9649$，$\widetilde{A}_3(0.5265)=0.0351$。

可知，Article 与 Book 为语义次关联。

用户在资源请求时可以指定资源匹配的类型{弱关联，次关联，强关联}。假设要求返回与资源请求目标 Article 强关联的概念，Book 就不满足条件。用户可以根据反馈的结果调整资源匹配的类型，扩大或缩小查询结果范围。本例中如果将匹配类型改成次关联，查询将返回 Book 相应的数据。

结束语 资源匹配是资源发现中的重要步骤。为了能够根据组织单元提供的信息更加准确地描述并执行资源发现，考虑更加丰富的语义和上下文信息，本文结合概念名称和概念上下文对语义相似度的影响，提出了一种语义模糊匹配算法。通过在资源的描述和资源的匹配中使用语义信息，消除了语义异构，从而保障用户获取语义上相关联的、更多的数据。同传统的资源匹配比较，本方法可以更加准确地定位资源。同时，引入模糊集的理论，实现了一定程度的资源模糊匹配，可以有效地辅助用户进行选择和决策，为资源发现提供了新的解决方法。未来的工作将研究更加通用的匹配方法。

参 考 文 献

1 Mowshowitz A. Virtual Organization. Communication of the ACM, 1997, 40(9): 30～37
2 赵纯均，陈剑，冯蔚东. 虚拟企业及其构建研究. 系统工程理论与实践，2002，22(10)：49～55
3 董方鹏，龚奕利，李伟，等. 网格环境中资源发现机制的研究. 计算机研究与发展，2003，40(12)：1749～1755
4 Raman R, Livny M, Solomon M. Matchmaking Distributed Resource Management for High throughput Computing. In: Proc. of the Seventh IEEE Intl. Symposium on High Performance Distributed Computing, Chicago, IL, July 1998
5 The portable batch system. http://pbs.mrj.com
6 Liu C, Foster I T. A Constraint Language Approach to Matchmaking. RIDE, 2004. 7～14
7 Tangmunarunkit H, Decker S, Kesselman C. Ontology-Based Resource Matching in the Grid-The Grid Meets the Semantic Web. In: International Semantic Web Conference, 2003. 706～721
8 Chakraborty D, Perich F, Avancha S, et al. Dreggie. Semantic Service Discovery for M－Commerce Applications. In: Proc. of the 20th Symposiom on Reliable Distributed Systems, Workshop on Reliable and Secure Applications in Mobile Environment, 2001
9 Chakraborty D, Perich F, Avancha S, et al. An Algorithm for Matching Contextualized Schemas via SAT: [Technical report]. DIT－03－003. DIT University of Trento, Italy, January 2003. Available at: heep://prints.biblio.unitn.it/archive/00000348/
10 Serafini L, Bouquet P, Magnini B, et al. Matching Techniques for Resource Discovery in Distributed Systems Using Heterogeneous Ontology Descriptions. ITCC (1), 2004. 360～366
11 Zhang X, Freschl J, Schopf J. A Performance Study of Monitoring and Information Services for Distributed Systems. In: Proceedings of IEEE HPDC-12, 2003
12 Keung H, Dyson J, Jarvis S. Predicting the Performance of Globus Monitoring and Discovery Service. In: Proceedings of 4th IEEE/ACM International Workshop on Grid Computing, 2003
13 Miller A G. WordNet: A Lexical Database for English. Communications of the ACM, 1995, 38(11): 39～41

(上接第 127 页)

时，得到的抽取结果为：

```
Wed Sunny Hi 17 Lo 8
```

当所给参数为：

```
文件路径:http://202.117.80.2
抽取方式:Table
关键词 :Wed
```

时(注：同一网页)，得到的抽取结果为：

```
Wed Sunny Hi 17 Lo 8
Thu Partly Cloudy Hi 16 Lo 7
Fri Cloudy Hi 14 Lo 7
Sat Rain Hi 16 Lo 10
```

结束语 本文利用二叉树模型实现了针对表格的信息抽取引擎的开发。该工具具有以下的特征：检索对象是 Web 表格；适用于结构化 Web 文档；高效、通用的信息抽取工具；提供多种信息抽取方式；支持一定的网页容错功能，具有可扩充性。

参 考 文 献

1 田志良，王世普，张皓东，等. 国际网企业网和智能建筑. 昆明：云南大学出版社，1997. 177～215
2 李大友，邱建霞. 计算机网络. 北京：清华大学出版社，1998. 151～166
3 杨明福. 计算机网络. 北京：电子工业出版社，1999. 123～127
4 蔡皖东，陈亚滨. 计算机网络培训教程. 西安：西安电子科技大学出版社，1999. 62～72