基于非对称属性描述的网格资源匹配算法

陈 建, 黄宏斌, 马武彬, 薛永奎
(国防科学技术大学 信息系统与管理学院 信息系统工程重点实验室, 长沙 410073)

摘 要: 针对网格资源匹配过程逐渐复杂化,以语义相似度查找机制为基础,提出了一种基于非对称资源属性描述的网格资源匹配算法 BARM。BARM 利用两次匹配对匹配过程加以边界约束,通过调整权重和阈值控制匹配结果的查准率和查全率,从而满足用户个性化的需求。

关键词: 资源匹配; 非对称属性; 语义相似度
中图分类号: TP311 **文献标志码**: A **文章编号**: 1001-3695(2012)06-2102-03
doi:10.3969/j.issn.1001-3695.2012.06.025

Grid resource matching algorithm based on asymmetric properties description

CHEN Jian, HUANG Hong-bin, MA Wu-bin, XUE Yong-kui
(*Key Laboratory of Information System Engineering, College of Information System & Management, National University of Denfense Technology, Changsha* 410073, *China*)

Abstract: For grid resource matching process becoming more and more complex, this paper proposed a description matching algorithm(BARM) based on asymmetric grid resource properties, which was based on semantic similarity search mechanisms. BARM used two times match to restrict the boundary constraints. It could control the precision rate and recall rate to meet the needs of users by adjusting the weights and threshold.

Key words: resource matching; asymmetric properties; semantic similarity

0 引言

网格资源匹配是把网格资源与资源请求者联系起来的重要环节,是网格资源管理的关键技术。目前,网格中的资源匹配多数采用基于对称属性的匹配机制。在进行匹配的过程中,资源提供者和请求者必须遵循共同的语法,即同样的属性名以及属性描述。这种精确匹配使得这些系统不够灵活,引入新的概念或属性也变得相对困难,并且在网格环境中,资源与用户跨越多个组织,确保资源与请求使用相同的属性名也变得困难;另外,也不能确保所有用户和资源提供者用同一种方式对同一属性进行语义解释[1]。本文以语义相关度查找机制为基础,提出了一种基于非对称资源属性描述的网格资源匹配算法 BARM。

1 相关研究

语义相似度计算是领域本体应用中的基础问题,用于表示本体中两个概念之间的语义接近程度,以便提高知识检索、服务匹配、本体映射等过程的性能[2]。国内外在关于面向本体和模式的相似度方面已经进行了大量的研究,下面对一些常用的相似度计算方法和相关系统进行介绍。

计算两个概念之间的语义相似度,目前有很多种方法。其中基于标志符的方法是现有比较常见的相似度计算方法,其使用语法驱动技术和构词法相似性来寻找表示术语的字符串之间的相似度。基于同义词词典的方法[3],词语语义距离还可以根据同义词词典来计算,同义词词典将所有的词组织在一棵或几棵树状的层次结构中。在一棵树状图中,任何两个节点之间有且只有一条路径,因而用这条路径的长度来度量两个词汇之间的距离。如 Fellbaum[4] 利用 WordNet 计算词语的语义相似度;Formica[5] 提出了一种基于形式概念分析的相似度算法,基于概念的上下文计算两者的相似度;吴奎等人[6] 提出了一种基于贝叶斯估计的概念语义相似度算法;江敏等人[7] 以加权和的方式综合数种方法计算语义相似度。本文以 Lin 等人[8] 的基于本体论知识的方法为基础,提出了一种基于非对称资源属性描述的匹配算法 BARM。

2 基于非对称资源属性描述匹配算法 BARM

2.1 问题描述

Lin 等人提出一个相似性函数 $S(W_1, W_2) = 1/(1+p)$ 来评价两个不同词(W_1 和 W_2)之间的相似性,其中 p 表示 W_1 和 W_2 之间的最短距离,但是这种方法在资源的层次化组织结构中计算相似度不够精确。假设本体武器层次资源树结构如图 1 所示。其中,95 式自动步枪到 87 式自动步枪之间的最短距离是 2,95 式自动步枪到步枪之间的最短距离也是 2。根据上面提出的相似度公式计算这两对资源的相似度都是 $S = 1/3$,这是

收稿日期: 2011-10-11; **修回日期**: 2011-11-28

作者简介: 陈建(1984-),男,湖南湘乡人,硕士研究生,主要研究方向为信息资源管理、智能辅助决策(hncj123@126.com);黄宏斌(1975-),男,副教授,博士,主要研究方向为信息集成、CPS、智能决策;马武彬(1986-),男,博士研究生,主要研究方向为信息管理、资源调度;薛永奎(1988-),男,硕士研究生,主要研究方向为数据挖掘、数据管理、社会网络.

不符合实际的[9],因为 95 式自动步枪与 87 式自动步枪的相似性明显要大于 95 式自动步枪与步枪。产生这种错误的原因是这个公式中只考虑了资源之间的路径长度,而没有考虑资源之间的层次化深度、本体论信息量等。为更好地解决这个问题,本文提出了基于非对称资源属性描述的匹配算法 BARM。

2.2　BARM 算法描述

2.2.1　关键词相似度

定义 1　Q 表示资源请求,Qw_i 表示资源请求描述信息;P 表示资源提供,Pw_i 表示资源提供描述信息。Qw_i 与 Pw_i 之间的相似度就是资源之间的匹配度。

在 BARM 算法中将资源关键词之间的路径长度 p、深度 d 和关联度 r 作为计算关键词相似度的参数。

定义 2　路径长度 p 表示两个资源之间的最短路径,p 越大相似性越小,p 越小相似性越大。长度转换函数用单调递减函数 $f_p(p)=e^{-\frac{p}{a}}$(a 为调节因子)[9]表示。

如图 1 所示,步枪到 87 式自动步枪的路径数目为 2,所以 p 就等于 2。

定义 3　深度 d 表示资源在层次化结构中的层次数,层次越深表示资源分得越细,所以相似性越大。用 $\Delta d=|d_1-d_2|$ 表示两个资源深度差,则深度转换函数用单调递减函数 $f_d(d)=e^{-\frac{\Delta d}{b}}$($b$ 为调节因子)表示[10]。

如图 1 所示,95 自动步枪和 87 自动步枪、自动步枪和半自动步枪的路径长度 p 都是 2,但是前者的相似度明显大于后者。

定义 4　关联度 $r=Q\cap P$,表示两个资源 Q 和 P 共同拥有的父概念节点数,$N=Q\cup P$,表示从资源树顶点到两个资源 Q、P 的资源节点并集。共同父概念越多,r 越大,相似度越大,则关联度转换函数用单调递增函数 $f_r(r)=\frac{r+c}{N+c}$(c 为调节因子)[10,11]。

图 1 所示半自动步枪与 95 自动步枪的关联度 $r=2$。

$$\begin{aligned}\mathrm{sim}_{\mathrm{key}}(Qw_i,Pw_i)&=f(p,d,r)=\\&\alpha f_p(p)+\beta f_d(d)+\chi f_r(r)=\\&\alpha e^{-\frac{p}{a}}+\beta e^{-\frac{\Delta d}{b}}+\chi\left(\frac{r+c}{N+c}\right)\end{aligned}\tag{1}$$

其中:
$$\begin{cases}p\geqslant 0,\Delta d\geqslant 0,r\geqslant 0,a,b,c\geqslant 1;\\p=0,f_p(p)=1;\\d=0,f_d(d)=1;\\0\leqslant\alpha,\beta,\chi\leqslant 1,\alpha+\beta+\chi=1\end{cases}$$

2.2.2　非对称资源属性相似度

定义 5　$|C_q|$ 和 $|C_p|$ 表示资源描述信息 Qw_i 和 Pw_j 属性概念集合。$\mathrm{LCS}(C_q,C_p)$ 表示属性概念 C_q 和 C_p 的最小公共集合[12]。

$$\mathrm{LCS}(C_q,C_p)=|C_q\cap C_p|\tag{2}$$

$$\mathrm{sim}_{\mathrm{attribute}}(Qw_i,Pw_i)=2\frac{\mathrm{LCS}(C_q,C_p)}{|C_q|\cup|C_p|}\tag{3}$$

综上所述,基于非对称资源属性描述的网格资源匹配度为

$$\begin{aligned}&\mathrm{match}_{\mathrm{BARM}}(Qw_i,Pw_i)=\\&\omega_1\mathrm{sim}_{\mathrm{Key}}+\omega_2\mathrm{sim}_{\mathrm{attribute}}=\\&\omega_1\left\{\alpha e^{-\frac{p}{a}}+\beta e^{-\frac{\Delta d}{b}}+\chi\left(\frac{r+c}{N+c}\right)\right\}+\omega_2\times 2\frac{\mathrm{LCS}(C_q,C_p)}{|C_q|\cup|C_p|}\\&0\leqslant\omega_1,\omega_2\leqslant 1,\omega_1+\omega_2=1\end{aligned}\tag{4}$$

2.2.3　BARM 算法设计

非对称资源属性描述匹配算法能够提高异构环境下的资源匹配效率,算法的主要步骤包括:

a)设定权重值,利用关键词匹配度计算资源的分类;

b)设定权重值,利用属性匹配度计算资源的匹配度;

c)设定阈值 ξ_1、ξ_2,返回满足条件资源。

算法的步骤描述如下:

1. inputs resource request
2. initialize $\alpha,\beta,\gamma,\omega_1,\omega_2,a,b,c,\xi_1,\xi_2$
3. get the depth Δd and the min distance p
4. then calculating $\mathrm{sim}_{\mathrm{key}}(Qw_i,Pw_i)$ According to expression 1
5. if ($\mathrm{sim}_{\mathrm{key}}(Qw_i,Pw_i)<\xi_1$)
   return failed
6. else
   get the r and N
   calculating $\mathrm{match}_{\mathrm{BARM}}(Qw_i,Pw_i)$ according to expression 4
7. if($\mathrm{match}_{\mathrm{BARM}}(Qw_i,Pw_i)>\xi_2$)
   return the address of resource
8. else
   return failed}

3　BARM 的实验分析

3.1　领域语义字典

利用 BARM 算法,必须构建一个特定领域语义字典(domain semantic dictionary,DSD)。DSD 定义为一个受限的词汇知识库,其面向的是该领域的常用词汇和专用词汇,其中定义了该领域不同词汇间的语义关系。DSD 采用知网[13]的语义描述结构来进行组织,主要对实词进行描述,以揭示概念与概念之间以及概念所具有的语义属性之间的关系为基本内容的常识知识库。DSD 的信息组织关系模型如图 2 所示。

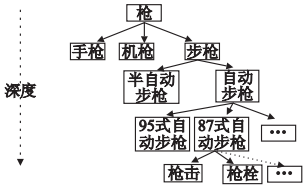

图1　武器(枪)基于本体论的分层属性关系

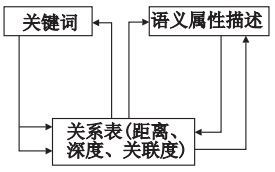

图2　DSD的信息组织关系模型

在实词的描述中,第一个描述式总是一个基本义原,即相当于关键词,这也是对该实词最重要的一个描述式,这个基本义原描述了该实词的最基本的语义特征。在 DSD 中规定了相关基本义原之间的距离、深度、关联度等关系。

3.2　实验过程

本节基于武器领域语义字典,通过在信息资源管理系统上查询匹配来演示 BARM 的工作机制。如表 1 所示(表 1 右列只展示了与资源请求距离较近的资源),实验数据为模拟数据库。算法参数设定为:$\alpha=0.2,\beta=0.2,\gamma=0.6,a=b=c=3,\omega_1=0.1,\omega_2=0.9,\xi_1=0.5,\xi_2=0.60$。

根据 BARM 算法得出资源请求匹配情况如表 2 所示。从表 2 的实验结果可得出以下结论:第一个阈值的约束可以节省匹配时间,当关键字不匹配时直接退出,不需要进行资源属性的匹配,有效提高了匹配效率;第二个阈值的约束可以达到精确匹配。本文可以通过修改阈值来约束匹配范围,从而使匹配结果符合请求者要求。

表1 资源描述

请求者资源描述	资源中心描述
1、步枪（自动、口径30 mm、射程300 m）	1、步枪（自动、口径25 mm、射程300 m） 2、步枪（自动、口径30 mm、射程200 m） 3、步枪（自动、口径24 mm、射程300 m） 4、步枪（自动、口径30 mm、射程300 m） 5、步枪（自动、口径25 mm、射程275 m） 6、步枪（半自动、口径25 mm、射程300 m） 7、步枪（半自动、口径30 mm、射程200 m） 8、步枪（半自动、口径24 mm、射程300 m） 9、步枪（半自动、口径30 mm、射程300 m） 10、步枪（半自动、口径25 mm、射程275 m） 11、手枪（口径14.5 mm、射程300 m） 12、手枪（口径12 mm、射程200 m） 13、火炮（自动、口径200 mm、射程1000 m） ……

表2 资源请求匹配表

比较项	匹配值	阈 值	权 重	匹配值与阈值的比较	匹配与否
与资源中心1匹配					
关键字匹配	1	0.5	0.1	>	是
属性匹配	0.5	–	0.9	–	–
资源匹配	0.55	0.6	–	<	否
与资源中心2匹配					
关键字匹配	1	0.5	0.1	>	是
属性匹配	0.5	–	0.9	–	–
资源匹配	0.55	0.6	–	<	否
与资源中心3匹配					
关键字匹配	1	0.5	0.1	>	是
属性匹配	0.5	–	0.9	–	–
资源匹配	0.55	0.6	–	<	否
与资源中心4匹配					
关键字匹配	1	0.5	0.1	>	是
属性匹配	1	–	0.9	–	–
资源匹配	1	0.6	–	>	是
与资源中心5匹配					
关键字匹配	1	0.5	0.1	>	是
属性匹配	0.2	–	0.9	–	–
资源匹配	0.28	0.6	–	<	否
与资源中心6匹配					
关键字匹配	1	0.5	0.1	>	是
属性匹配	0.2	–	0.9	–	–
资源匹配	0.28	0.6	–	<	否
……					
与资源中心12匹配					
关键字匹配	0.25	0.5	0.1	<	否（直接退出）
属性匹配	–	–	–	–	–
资源匹配	–	–	–	–	–
与资源中心13匹配					
关键字匹配	0	0.5	0.1	<	否（直接退出）
属性匹配	–	–	–	–	–
资源匹配	–	–	–	–	–
……					

3.3 实验分析

定义6 P_{relevant}表示相关资源集，P_{return}表示返回资源集，$P_{\text{return}}^{\text{rel}}$表示返回资源中的相关资源集合，则查准率$p$和查全率$r$定义如下[14]：

$$\text{查准率 } p=\frac{P_{\text{return rel}}}{P_{\text{return}}},\ \text{查全率 } r=\frac{P_{\text{return}}^{\text{rel}}}{p_{\text{relevant}}}$$

按照3.2节的实验程序，下面通过调节阈值ξ_2和权重ω_1、ω_2（$\omega_1+\omega_2=1$）的值来分析阈值对资源匹配查准率与查全率的影响，其他参数与3.2节实验相同（$\alpha=0.2,\beta=0.2,\gamma=0.6$，$a=b=c=3$，$\xi_1=0.5$）。

a）图3对10个信息请求进行了匹配，展示了在阈值$\xi_2=0.6$，权重分别取值$\omega_1=0.1,0.2,0.3,\cdots,0.9$时，对平均查准率的影响，其中横轴表示权重$\omega_1$的值，纵轴表示查准率$p$。从图3可看出，随着权重$\omega_1$的增加，$p$递减，这说明第二次匹配资源的属性描述对查准率影响较大。

b）图4展示了$\xi_2=0.6$，权重分别取值$\omega_1=0.1,0.2,0.3,\cdots,0.9$时，对平均查全率的影响，其中横轴表示权重$\omega_1$的值，纵轴表示查全率$r$。从图4可看出，无论权重$\omega_1$的值为多少，查全率都在80%左右，且当权重$\omega_1=0.5$时，查全率最高。

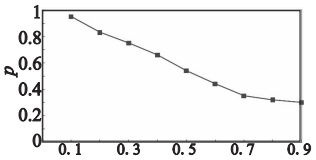

图3 阈值ξ_2=0.6时，p随权重变化的曲线

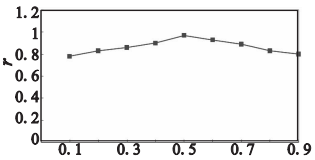

图4 阈值ξ_2=0.6时，r随权重变化的曲线

从上述实验可以看出，合适地选择权重和阈值可以有效提高资源匹配的性能。

4 结束语

本文在分析已有语义相似度算法的基础上，对资源的查找和定位进行了研究，提出了基于非对称资源属性描述的网格资源匹配算法，利用语义概念的关联度、距离、深度来约束关键字的匹配，利用非对称资源属性描述的关联度来约束资源的属性匹配，通过两次匹配有利于提高匹配效率，并且实验验证了通过调整权重和阈值能够控制匹配结果的查准率和查全率，从而满足了用户实际需求。

参考文献：

[1] 陈佳. 网格资源管理关键技术研究[D]. 成都：电子科技大学，2008.

[2] BUDANITSKY A, HIRST G. Evaluating WordNet-based measures of lexical semantic relatedness[J]. Computational Linguistics, 2006, 32(1): 13-47.

[3] AGIRRE E, RIGAU G. A proposal for word sense disambiguation using conceptual distance[C]//Proc of International Conference on Recent Advances in Natural Language Processing. 1995.

[4] FELLBAUM C F. WordNet[M]. Cambridge: MIT Press, 1998: 265-283.

[5] FORMICA A. Ontology-based concept similarity in formal concept analysis[J]. Information Sciences, 2006, 176(18): 2624-2641.

[6] 吴奎，周献中，王建宇，等. 基于贝叶斯估计的概念语义相似度算法[J]. 中文信息学报，2010，24(2)：52-57.

[7] 江敏，肖诗斌，王弘蔚，等. 一种改进的基于《知网》的词语语义相似度计算[J]. 中文信息学报，2008，22(5)：84-89.

[8] LIN Lan-fen, GAO Peng, CAI Ming, *et al*. A knowledge service based model of collaborative manufacturing process planning for networked manufacturing[J]. Journal of Computer-Aided Design & Computer Graphics, 2005, 17(9): 2085-2091.

[9] 彭志平，李晓明，柯文德，等. 基于本体概念群组划分的语义距离计算方法[J]. 模式识别与人工智能，2011，24(2)：194-200.

[10] TAO Fei, HU Ye-fa, ZHAO Dong-ming, *et al*. Study on resource service match and search in manufacturing grid system[J]. Advanced Manufacturing Technology, 2009, 43(3-4): 379-399.

[11] 李文杰，赵岩. 基于本体结构的概念间语义相似度算法[J]. 计算机工程，2010，36(23)：4-6.

[12] BATET M, SANCHEZ D, VALLS A. Ontology-based measure to compute semantic similarity in biomedicine[J]. Journal of Biomedical Informatics, 2011, 44(1): 118-125.

[13] 易丽萍，竹勇，雷小春，等. 知网在词语相似度计算方面的应用[J]. 信息技术与信息化，2005(1)：56-58.

[14] 黄宏斌，刘志忠，张维明，等. 基于层次本体模型（HOM）的语义相似度计算方法[J]. 系统工程与电子技术，2009，31(7)：1750-1754.