第 33 卷 第 9 期 计 算 机 工 程 2007 年 5 月
Vol.33 No.9 Computer Engineering May 2007
·博士论文· 文章编号：1000—3428(2007)09—0028—03 文献标识码：A 中图分类号：TP182

分布式环境下基于语义相似的案例检索

李 锋[1]，魏 莹[2]
（1. 华南理工大学工商管理学院，广州 510640；2. 香港中文大学系统工程与工程管理系，香港）

摘 要：分布式环境下的异构案例表达制约了案例检索过程中案例属性之间的可比性，进而成为分布式环境下案例推理系统成败的一个关键问题。该文提出基于语义相似的案例检索，通过利用 Ontology 技术来理解案例属性的内在含义，在此基础上定义并计算属性之间的相似程度。对原型系统的初步测试证明了基于语义相似的案例检索有效性。

关键词：语义；相似度；案例检索；分布式环境

Approach to Case Retrieval in Distributed Environment Based on Semantic Similarity Calculation

LI Feng[1], WEI Ying[2]
(1. School of Business Administration, South China University of Technology, Guangzhou 510640;
2. Department of Systems Engineering and Engineering Management, The China University of Hong Kong, Hong Kong)

【Abstract】 Heterogeneous case representation in distributed environment decides whether the two features of two cases are the same and comparable in process of case retrieval. It is one of key problems of case-based reasoning (CBR) systems in distributed environment. An approach of semantic based case retrieval exploits a method to understand features of case using ontology technology. And based on the understanding of features, similarity between the two features is defined and computed. The experiment result of the prototype system verifies the efficiency of semantic based case feature mapping and case retrieval.

【Key words】 Semantic; Similarity; Case retrieval; Distributed environment

基于案例的推理(Case-based Reasoning，简称案例推理)，是通过调整历史问题的解决方案，从而得到当前新问题的解的一种推理方法[1]。案例推理流程可以用“4R”来概括：案例检索，案例重用，案例调整以及案例学习[2]。其中，案例检索的主要任务是从案例库中检索出与当前新问题最相似的历史案例(集)，为新问题求解提供有用的经验和知识。

建立在“相似问题具有相似解”假设基础上的案例推理方法的应用中，检索出与新问题最相似的历史案例是其成功的关键要素之一。

1 研究现状

案例检索的核心是如何正确和客观地定义和量化案例之间的“相似程度”。当前案例检索有许多不同的技术选择，如最近相邻算法、归纳法、知识导引法、神经网络法等[3,4]。其中，在工程中，应用最广泛的案例相似度计算方法是最近相邻算法(K-Nearest Neighbors)。

最近相邻算法是用案例属性相似度的加权和表征新问题案例与历史案例间的相似度的一种算法。其计算公式为

$$S_{(X,Y)}=\sum_{i=1}^{n}\omega_i\times sim(f_i^X,f_i^Y)\Big/\sum_{i=1}^{n}\omega_i \tag{1}$$

其中，$S_{(X,Y)}$表示案例X和案例Y的相似度，ω_i表示案例的第i个属性在整个案例属性集合中所占的权重，$sim()$为属性相似度计算公式，f_i^X和f_i^Y分别表示案例X和案例Y的第i个属性的属性值。

最近相邻算法要求参与计算相似度的两个案例的属性/指标集合(属性个数、度量等)完全一致并一一对应，且两个案例的属性权重取值完全相同。一般来说，最近相邻算法适用于具有相同案例表达结构的案例库以及案例推理系统。但是对于分布式环境下具有不同案例表达结构的案例库来说，需要进行一定的预处理将异构的案例转化为具有相同案例结构才能应用最近相邻检索算法。

本文以最近相邻算法为例，通过引入 Ontology 技术，设计基于语义相似的属性相似度计算公式，实现异构案例库集成和检索算法。

2 基于语义相似的案例检索

英语单词“Ontology”最初起源于哲学领域，中文翻译为“本体论”或“存在论”，其定义为“对世界上客观存在物的系统地描述”。近 20 年来，计算机领域，特别是人工智能领域赋予了Ontology新的含义和应用。其中，广泛被接受的定义为：以一种机器可以理解的方式，正式地、清楚地描述共享的概念和概念之间的联系[5]。

应用 Ontology 技术于案例检索，其核心思想是：基于领域 Ontology 知识，发现不同案例表达形式下各自属性所代表的真实含义，并在此基础上建立属性之间的映射关系。由于基于 Ontology 技术的案例属性间映射是建立在对案例属性的语义理解之上，因此本文又称之为基于语义相似的案例检索。

基于语义相似的案例相似度计算主要包括以下两个的过

基金项目：国家自然科学基金资助项目(70472041)
作者简介：李 锋(1975－)，男，讲师、博士，主研方向：知识工程，供应链管理；魏 莹，博士生
收稿日期：2006-06-12 **E-mail：**fenglee@scut.edu.cn

程：Ontology 库的建立和案例属性之间的映射。

2.1 Ontology 库的建立

Ontology 库中知识的质量以及丰富性是基于语义相似的案例检索成功的关键之一。同时，Ontology 库还需要从案例推理机制角度上满足以下几个技术指标：

(1)快速性：所建立的 Ontology 库应该能够快速地检索出属性的相关信息，而不能成为整个案例推理系统的瓶颈。

(2)可扩展性：所建立的 Ontology 库应该能够易于实现浏览、修改等知识库维护操作。

综合考虑以上两个因素，采用传统关系型数据库存储和维护 Ontology 知识，并利用关系数据库完善、快速的数据操作函数实现 Ontology 的维护操作。

Ontology 是由概念、概念的属性，以及概念和属性之间的关系组成的网状结构，其描述了领域内概念之间的相互关系。因此，建立如图 1 所示的关系数据库表的结构，用于存储和维护 Ontology。

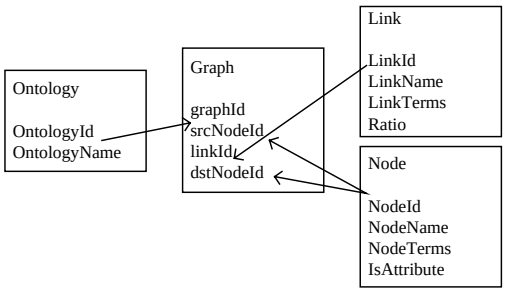

图 1 数据库表结构

表“ Node ”用于定义 Ontology 图描述下的节点信息。由于描述一个节点的术语可能有多个同义词，因此设置字段“ NodeTerms ”存储该节点的所有同义词，可以减少数据库存储资源和提高检索效率。同理，在用于描述节点之间的关系表“ Link ”中也定义了类似功能的字段“ LinkTerms ”。

关系表“ Link ”中不仅定义了概念之间的“ 定性 ”的关系，还包括以字段“ Ratio ”描述的“ 定量 ”的关系。而这字段在后续的属性映射处理中非常有用。

由于 Ontology 库中维护了多个相关的 Ontology，因此，创立表“ Ontology ”用于索引每个 Ontology。

相比之下，依托关系数据库而建立的 Ontology 库搜索引擎，其效率要高于以文本文件形式，如 RDF(http://www.w3.org/RDF)构建的 Ontology 库。

2.2 案例属性映射

异构案例表达下的案例属性之间的关系有以下 3 种可能：(1)属性之间存在某种联系，如属性“ 质量 ”与属性“ 重量 ”之间的关系；(2)属性之间不存在任何联系，如属性“ 质量 ”与属性“ 颜色 ”;(3)属性之间存在某种不能识别的关系(受到 Ontology 库完备性约束)。

案例属性映射主要任务包括是发现并定义不同案例表达下案例属性之间的关系。除因受到系统 Ontology 库限制而不能识别的属性之间的相互关系外，案例属性映射要求尽可能地识别并定义属性之间的相互关系。其主要包括两个子任务：确定属性之间是否存在某种联系-定性分析；量化属性之间的联系-定量分析。

2.2.1 定性分析

属性之间的相互关系可能是直接的，即这类关系可以从数据库表“ Link ”中检索出。然而更多的属性之间的关系为间接关系的，需要从由多个属性/概念，及其关系形成的关系链中推理得到。

(1)如果两属性之间关系为直接关系，且其关系为“ equivalentProperty ”关系，则这两个属性存在可映射的联系。

(2)如果两属性之间关系为间接关系，且其关系链中不存在非“ equivalentProperty ”关系或“ equivalentClass ”关系，则这两个属性存在可映射的联系。

基于以上定义，只需要考虑那些具有相同含义的属性之间的映射关系，而不需涉及其它复杂的属性之间的关系，如父子关系“ subPropertyOf ”。

定义 Ontology 的邻接矩阵 A：设集合 $N=\{n_1,n_2,\ldots,n_m\}$ 表示 Ontology 中的节点(包括概念和属性)，则 $m\times m$ 矩阵 A 的元素 a_{ij} 为

$$a_{ij}=\begin{cases}1, & R(n_i,n_j)\in\{equivalent\ \mathrm{Pr}operty,\\ & \quad equivalentClass\}\\ 0, & R(n_i,n_j)\notin\{equivalent\ \mathrm{Pr}operty,\\ & \quad equivalentClass\}\end{cases}$$

其中，$R(n_i,n_j)$ 表示从数据库表“ Link ”中检索出的节点 n_i 与节点 n_j 之间的相互关系。同样定义 Ontology 的可达矩阵 R 的元素为

$$r_{ij}=\begin{cases}1 & n_i\text{可达}n_j\\ 0 & n_i\text{不可达}n_j\end{cases},\text{且}r_{ij}=1$$

由邻接矩阵可以计算出可达矩阵，计算公式为：$R=(A\cup I)^n$，矩阵 I 为对角线元素为 1，其它元素为 0 的单位矩阵。

在可达矩阵中，如果属性 n_i 与属性 n_j 之间存在可映射的关系，则其对应的矩阵元素 r_{ij} 为 1，否则 r_{ij} 为 0。

2.2.2 定量分析

对于已经确认存在可映射的属性对需要进一步的量化其映射关系。

在量化属性之间关系时需要考虑：属性之间可能存在多条映射关系链。并且，映射关系具有方向性，即属性“ 质量 ”到属性“ 重量 ”的映射关系不等同于属性“ 重量 ”到属性“ 质量 ”的映射关系。

为此，定量分析的计算步骤如下：

(1)从邻接矩阵出发，利用 Dijkstra 算法搜索出一条最短路径。由于邻接矩阵中各个节点之间的距离为 1 或 0，因此 Dijkstra 算法搜索出的最短路径实质上是计算次数最少的映射关系链(路径)；

(2)从源属性出发，依次处理关系链上的每个关系：从数据库表“ Link ”中检索出节点之间的映射关系字段“ Ratio ”并累计；

(3)最终给出源属性与目的属性之间的映射关系。

2.3 案例相似性计算

识别并量化属性之间的映射关系后，就可以应用最近相邻算法计算案例之间的相似度。变化后的计算公式为

$$S_{(X,Y)}=\frac{\sum_{i=1}^{n}\omega_i\times sim(f_i^X,F(f_i^X,f_i^Y))}{\sum_{i=1}^{n}\omega_i}\qquad(2)$$

式(2)中函数 F()表示基于语义映射后得到的案例 Y 的属性值

f_i^Y。而属性之间相似度计算函数 sim()则采用文献[4]中定义的计算函数。

3 系统与测试

结合所承担的研究课题，笔者开发了一套案例推理原型系统用以测试基于语义相似的案例检索能力。该系统目标是从众多笔记本电脑供应商提供的笔记本电脑中选择一款最能够满足客户需要的产品。为此，建立一整套描述笔记本电脑价格、性能的指标系统。同时，建立一个笔记本电脑案例库用于存储和查询不同型号笔记本电脑的各种参数。

原型系统采用 Java 语言开发，编译环境为 J2SE 5.0，后台 Ontology 库存储在 MySQL 4.1 关系数据库中。系统框架如图 2 所示。

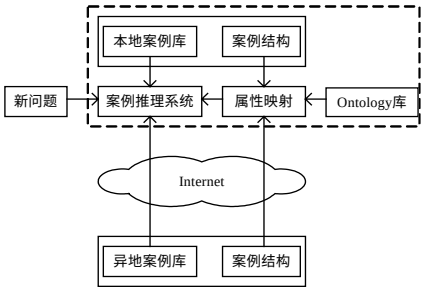

图 2 系统结构

图 2 在原有的案例推理系统外增加了属性映射模块和 Ontology 库，利用异构案例库中案例结构信息计算出属性之间的映射关系，并将结果提交给案例推理系统中的案例检索模块。

在研究与开发的开始阶段，参考笔记本电脑 Thinkpad 指标参数，建立了本地的笔记本电脑案例和案例库。

此案例库以及案例推理系统能够较好地满足用户的部分需要，即从 Thinkpad 笔记本电脑中选择一款最能够满足用户需要的笔记本电脑。但是，当用户需要从多种笔记本电脑品牌中选择一款最适合的笔记本电脑时，案例推理系统因为无法理解其它品牌笔记本电脑的指标系统而满足用户需要。为此，以多个品牌笔记本指标系统为基础建立描述笔记本电脑的 Ontology，并以戴尔公司(DELL)笔记本电脑为测试案例，测试戴尔笔记本电脑指标属性集合的映射比率。

以惠普公司(HP)、东芝公司(TOSHIBA)、Thinkpad 和索尼(SONY)公司笔记本电脑信息建立“笔记本电脑”的 Ontology。部分 Ontology 信息如图 3 所示。

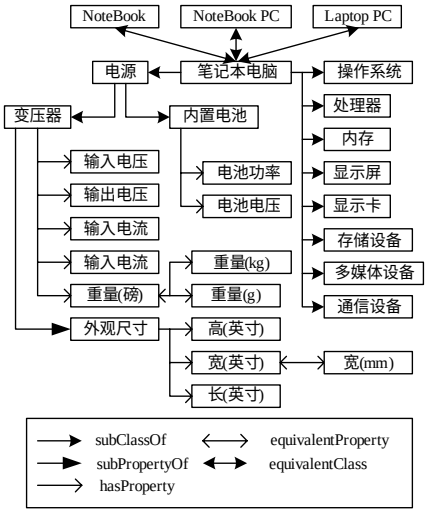

图 3 “笔记本电脑”Ontology

图 3 中给出了 Ontology“笔记本电脑”的部分信息。结合文章 2.2.1 小节给出的定义可知，“变压器”的属性“重量(磅)”与属性“重量(kg)”存在映射关系，而“内置电池”属性“电池电压”与“变压器”的属性“输出电压”没有映射关系。

对比集成与未集成 Ontology 库，案例属性映射对比见表 1。从表 1 数据显示，集成 Ontology 后，案例推理系统对 DELL 笔记本电脑属性识别率增长了 1 倍多(30%~63%)。那些未能被识别并映射的 DELL 笔记本电脑属性主要是由 Ontology 库信息不完全所致，例如：戴尔笔记本电脑“通信设备”子系统中“网卡”定义为“Network Interface”，而在 Thinkpad 中定义为“Ethernet”，在惠普中定义为“Network Interface Controller”，在东芝中定义为“Ethernet Controller”，在索尼中定义为“Ethernet”。

表 1 属性映射对比表

DELL 笔记本电脑	集成前	集成后
系统信息(11 个指标)	3	6
屏幕与显示(2 个指标)	0	2
存储与多媒体(5 个指标)	1	3
重量与尺寸(4 个指标)	4	4
接口与连接(2 个指标)	0	1
通信设备(3 个指标)	0	1

进一步完善关于“笔记本电脑”的 Ontology 后，以戴尔公司型号为 XPS M140 的笔记本电脑为例计算其与不同品牌不同型号笔记本电脑的相似度(由于在未使用 Ontology 之前，无法识别其它品牌笔记本电脑的属性，因此其相似度为 0)。表 2 显示了几款价格与 XPS M140 价格(＄999)相似的笔记本电脑的相似度计算结果(其中假设各个指标权重相同)。

表 2 案例相似度表(单位：%)

相似度	戴尔 XPS M140
戴尔 Inspiron 700m(＄999)	86.66
东芝 Satellite R10(＄999.2)	89.29
索尼 VAIO-FS790(＄989.99)	94.31
IBM Thinkpad R50E(＄949)	80.89

4 结论

本文尝试在案例推理系统中通过集成一个 Ontology 库和案例属性映射模块，发现不同案例库中案例属性的内在含义，据此建立不同表达下案例属性的内在联系。这样通过建立不同案例表达下属性之间的映射关系，能够实现异构案例库的无缝集成，使得网络环境下的案例推理成为可能。同时，对结合研究课题背景而开发的原型系统的测试结果证明了该方法的有效性。

参考文献

1 Reisbeck C K, Schank R C. Inside Case Based Reasoning[M]. Hillsdale: Lawrence Erlbaum Associates, 1989.

2 Aamodt A, Plaza E. Case Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches[J]. AI Communications, 1994, 7(1): 39-59.

3 李 锋, 冯 珊. 基于人工神经网络的案例检索与案例维护[J]. 系统工程与电子技术, 2004, 26(2): 234-236.

4 李 锋, 周凯波, 冯 珊. 基于统计特征的属性相似度计算模型[J]. 华中科技大学学报(自然科学版), 2005, 33(6): 80-82.

5 Gruber T R. A Translation Approach to Portable Ontologies[J]. Knowledge Acquisition, 1993, 5(2): 199-220.