一种综合的本体相似度计算方法*)

张忠平[1] 田淑霞[1] 刘洪强[2]

(燕山大学信息科学与工程学院 秦皇岛 066004)[1] (南京邮电大学计算机学院 南京 210003)[2]

摘 要 本体相似度计算是本体映射的关键环节。本体的实例、关系、属性、结构等信息是相似度计算需要考虑的重要因素。针对目前本体映射过程中相似度计算所存在的问题,提出了一种综合的相似度计算方法。首先判断不同本体之间是否存在相关性。若相关,则充分考虑各种相关因素,从语义和概念两个层面来进行比较,然后给出了本体的综合相似度计算方法。最后采用了两组测试数据对该方法进行实验,并与 GLUE 系统的概率统计方法进行了实验对比。实验结果表明,该方法能够有效确保相似度计算的准确性。

关键词 本体,本体映射,相关度,相似度,本体相似度,概念相似度

Compositive Approach for Ontology Similarity Computation

ZHANG Zhong-ping[1] TIAN Shu-xia[1] LIU Hong-qiang[2]

(College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China)[1]

(College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)[2]

Abstract Similarity computation among ontologies is the critical tache in the process of mapping. The information about instances, relations, sturctures and attributes are also important factors. Aiming at the current problems, put forward a compositive approach of similarity computation. The relativity among different domain ontologies was judged first. And then in this subsumption, based on semantic level and concept level, a comprehensive similarity measuring method was proposed, after taking a full consideration about relative factors. This method was tested with two datasets and compared with probability statistical method of GLUE system. Experimental results indicate that this approach can significantly improve the precision.

Keywords Ontology, Ontology mapping, Ontology relativity, Ontology similarity, Concept similarity

1 引言

近年来,本体(Ontology)已经成为知识工程、语义 Web、人工智能、数据集成、信息检索等研究领域的热门课题。随着本体的广泛应用,为实现资源共享,各领域纷纷定义相应的本体标准。然而本体的建立一直没有一个统一的规范来进行约束,由此产生了诸如系统异构、结构异构、语义异构等许多问题。本体映射(Ontology Mapping)的研究正是为了解决这些异构问题。本体映射是本体集成(Integration)、本体串联(Alignment)、本体合并(Merging)等的技术基础,是解决知识共享与重用的有效途径。而本体相似度(Ontology Similarity)计算是本体映射的关键环节。已有的相似度计算方法通常是通过概念的相似度得到。然而在复杂应用环境中,仅仅考虑概念的相似度是远远不够的,本体的实例、关系、属性、结构等信息也是相似度计算需要考虑的重要因素。

本文针对目前本体映射过程中相似度计算所存在的问题,引入了相关度(Relativity),提出了一种综合的相似度计算方法。首先判断不同本体之间的相关性。若两本体相关,则充分考虑本体结构、本体组成元素(概念、属性、实例等)以及元素语义等相关因素,分别给出本体语义相似度以及基于名称、实例、属性、结构等相关信息的概念相似度计算方法,然后将两种相似度合并,得到本体的综合相似度。

2 相关术语

2.1 本体

本体最初起源于哲学领域,称为本体论、存在论,反映的是事物本质的科学内涵。20 世纪 80 年代,科研人员把本体引入人工智能(Artificial Intelligence)领域,并赋予其新的含义。在计算机领域,"本体是一个对共享概念的形式化的、显性的规范说明"[1]。本体的形式化定义有很多,本文采用文献[2]中的本体表示形式,即采用 OWL(Web Ontology Language)的子语言 OWL Lite 表示。本体主要包括概念(concepts)、关系(relations)、实例(instances)以及公理(axioms),可表示为一个七元组:$O:=(C,H_C,R_C,H_R,I,R_I,A)$。

其具体含义如下:概念 C 包含于层次 H_C 中;两个独立的概念之间存在关系 R_C;关系(属性)处于层次 H_R 中;实例 I 是特殊化概念,由属性实例 R_I 关联;另外,公理 A 用逻辑语言表示,可从已有的知识推导出新的知识。公理在本体映射推理过程中应用得较多。

2.2 本体映射

本体映射是指两本体之间存在语义级的概念关联,通过语义关联,实现将源本体的实体映射到目标本体的过程[3]。

*)基金项目:国家自然科学基金(60773100),教育部科学技术研究重点项目(205014),河北省教育厅科研计划项目(2006143)。**张忠平** 博士,博士后,副教授,硕士生导师,CCF 会员(E20-0006458S),主要研究领域为网格技术、语义网、本体论、XML 数据库、数据挖掘;**田淑霞** 硕士研究生,研究方向为语义网、本体论;**刘洪强** 硕士研究生,研究方向为人工智能、语义网。

其实质就是概念层上语义相关的两个本体的实体根据语义关系进行转换的过程，以便用户能使用通用的接口，对同一事物形成共同的理解。

Ehrig 与 Staab 归纳出本体映射 6 个过程[4]。这一过程表明，本体相似度的计算是本体映射的关键环节。

2.3 相似度

相似度即两个对象相似的程度，形式化定义为

$$\mathrm{Sim}(e_{i1},e_{i2})=\frac{\alpha}{d+\alpha}$$

其中，$\mathrm{Sim}(e_{i1},e_{i2})$表示元素 e_{i1}（$e_{i1}\in O_1$）和 e_{i2}（$e_{i2}\in O_2$）的相似度，取值范围是[0,1]，如果两个元素是完全相似的，则相似度为 1；如果两个元素无任何共有特征，则相似度为 0；α 为可调节的参数，d 是一个整数，关于取值采用如下策略[5]。

(1)若 $e_{i1}=e_{i2}$，则令 $d=0,\alpha\neq 0$，即 $\mathrm{Sim}(e_{i1},e_{i2})=1$；

(2)若 $e_{i1}\neq e_{i2}$，则令 $\alpha=0,d\neq 0$，即 $\mathrm{Sim}(e_{i1},e_{i2})=0$；

(3)$\mathrm{Sim}(e_{i1},e_{i2})=\mathrm{Sim}(e_{i2},e_{i1})$，具有对称关系；

(4)若 E_{i1}，E_{i2} 分别表示元素 e_{i1}，e_{i2} 的特征集合，$(E_{i1}\cap E_{i2})\neq\phi$，则令 $d=1$，其相似度值为 $\alpha/(1+\alpha)$，且计算方法如下：

$$\mathrm{Sim}_{set}(e_{i1},e_{i2})=\frac{|E_{i1}\cap E_{i2}|}{|E_{i1}\cap E_{i2}|+\lambda\left|\frac{E_{i1}}{E_{i2}}\right|+(1-\lambda)\left|\frac{E_{i2}}{E_{i1}}\right|},$$

$\lambda\in[0,1]$

E_{i1}/E_{i2} 表示集合 E_{i1} 中除去集合 E_{i2} 具有的特征，E_{i2}/E_{i1} 反之类似；λ 是非公共特征的相对重要程度，可以调节。

(5)若 $\mathrm{Sim}_1(e_{i1},e_{i2}),\cdots,\mathrm{Sim}_n(e_{i1},e_{i2})$ 表示度量得到的多个不同方面的相似度值，则相似度迭代的计算方法如下：

$$\mathrm{Sim}_{agg}(e_{i1},e_{i2})=\frac{w_1*\mathrm{Sim}_1(e_{i1},e_{i2})+w_2*\mathrm{Sim}_2(e_{i1},e_{i2})+\cdots+w_n*\mathrm{Sim}_n(e_{i1},e_{i2})}{w_1+w_2+\cdots+w_n}$$

$(w_i>0),w_1+w_2+\cdots+w_n=1$

w_i 为可调参数，其取值由相似度值的重要性决定。

3 综合的本体相似度计算

为了比较两个本体并计算它们之间的相似度，本文首先引入相关度，通过相关度来判断两本体的相关性，然后在相关的基础上依据 Alexander Maedche 和 Steffen Staab[6] 从不同层面上进行相似度计算的基本思想，从语义和概念两个层面来进行比较，即分别计算语义相似度和概念相似度。

3.1 相关度

相关度指概念之间相关的程度。相关性是一个比相似性更普遍的概念，相似意味着词汇所表达的概念在某些特征方面有重合，而相关性表明概念间具有相似性，但概念所表达的一些特征不直接重合。因此，相似只是相关的一个特殊方面。例如，汽车和汽油紧密相关但不相似，而汽车与自行车在功能上相似但不是紧密相关。相似的实体通常都可以因为它们的相似性而被认为是相关的，如固定电话-无线电话。但不相似的实体仍然可能具有很强的语义相关性，如企鹅-南极洲。本文统一用一个在[0,1]区间取值的实数来表示度量的结果，该实数的取值 0(1)表示两个实体完全不相关。具体方法采用 Hirst-St-Onge 语义相关度算法[7,8]，其基本思想是：当两个词在 WordNet 同义词集(synset)中有一条较短的路径相连时，在语义上就具有较大的相关度。当此路径不存在时，$Rel_{HS}(w_1,w_2)=0$。

$$Rel_{HS}(w_1,w_2)=c-len(w_1,w_2)-k*turns(w_1,w_2)\quad(1)$$

其中，c 和 k 是两个常参数，$turns(w_1,w_2)$代表在 synset 中的路径转向次数，$len(w_1,w_2)$是路径长度。

3.2 语义相似度

语义相似度是指概念间自身语义的相似程度，也就是考虑概念的意思来计算相似度，语义相关度和语义相似度成正比关系。本文采用 WordNet 语义词典来计算概念的语义相似度。

WordNet 是由 Princeton 大学开发的一个庞大的语言知识库系统，是一部树状的英语语义字典。其中每个结点 s 表示一个词义，结点中保存了多个同义词或者短语，每个单词或短语又可以存在于多个语义结点中(即表明该单词有多个词义)。基于语义词典的语义相似度的基本思想是：两个单词通过上位关系(hypernym)连接的距离越近，它们的相似度越大；反之，它们的相似度越小。如果它们在有限上位层次中没有共同的父结点，则 $\mathrm{Sim}_d(w_1,w_2)=0$。Lin 等人在 WordNet 中定义了两个词义的相似度[9]：

$$\mathrm{Sim}_d(s_1,s_2)=\frac{2*\log p(s)}{\log p(s_1)+\log p(s_2)}\quad(2)$$

其中，$p(s)=count(s)/total$ 表示在 WordNet 中词义结点 s 及其子结点所包含的单词个数在整个词典中所占的比例，$total$ 是 WordNet 的单词总数。另外 $w_1\in s_1,w_2\in s_2$，表示单词 w_1 和 w_2 分别位于结点 s_1 和 s_2 中，结点 s 是 s_1 和 s_2 的公共祖先结点。

令 $s(w_1)=\{s_{1i}\mid i=1,2,\cdots,m\}$ 和 $s(w_2)=\{s_{2j}\mid j=1,2,\cdots,n\}$ 分别表示单词 w_1 和 w_2 的所有词义，则两个单词的相似度定义为它们之间词义相似度的最大值：

$$\mathrm{Sim}_d(w_1,w_2)=\max(\mathrm{Sim}_d(s_{1i},s_{2j}))$$
$$s_{1i}\in s(w_1),s_{2i}\in s(w_2)\quad(3)$$

3.3 概念相似度

为了准确计算概念之间的相似度，本文充分考虑概念的名称、实例、属性、结构等因素，并在现有技术的基础上，多方面、多角度地给出概念相似度的综合计算。

3.3.1 基于概念名称

使用概念名称来计算相似度是映射过程中最直接的也是最基本的方法，这种方法不考虑概念的语义，仅仅考虑两个概念在语言文字上的相似度。通常可将概念看作是不同的字符串，进行字符串之间的比较。常用的方法有 Edit Distance，N-gram，Humming Distance 等方法。

本文采用编辑距离(Edit Distance)方法来计算概念名称之间的相似度。Edit Distance[10] 又称 Levenshtein Distance，是由 Levenshtein 提出的，用来比较两个字符串(后扩展到语句)的相似度。它测量从一个字符串转换到另一个字符串所需的插入、删除、替换等的最小操作数目。计算公式如下：

$$\mathrm{Sim}_{gr}(c_i,c_j)=\max\left(0,\frac{\min(|c_i|,|c_j|)-ed(c_i,c_j)}{\min(|c_i|,|c_j|)}\right)\quad(4)$$

其中，C_1，C_2 分别为本体 O_1，O_2 中的概念集合，$c_i\in C_1$，$c_j\in C_2$，则可以定义两本体之间的名称相似度为

$$\mathrm{Sim}(C_1,C_2)=\frac{1}{|C_1|}\sum_{c_i\in C_1}^{n}\max_{c_j\in C_2}\mathrm{Sim}(c_i,c_j)\quad(5)$$

$\mathrm{Sim}(C_1,C_2)$是对称计算。显然，$\mathrm{Sim}(C_2,C_1)$可能与 $\mathrm{Sim}(C_1,C_2)$完全不同。比如说，如果 C_2 不仅包含 C_1 中的所有概念，而且包含其它一些概念，则 $\mathrm{Sim}(C_1,C_2)=1$，而 $\mathrm{Sim}(C_2,C_1)$则可能趋近于 0。因此需要定义一个相关系数(the relative number of hits)[2]：

$$SetHit(C_1,C_2)=\frac{|C_1\cap C_2|}{|C_1|}$$

为了使 $\mathrm{Sim}(C_1,C_2)$ 在任何情况下都能运算正确，定义 $SetHit(C_1,C_2)\leqslant 1$。

3.3.2 基于概念实例

在需要映射的两个本体中，可以用概念的具体实例计算概念相似度。一个概念的实例也是它祖先概念的实例。基于实例计算相似度的理论依据是：如果概念所具有的实例全部都相同，那么这两个概念是相同的；如果两个概念具有相同实例的比重是相同的，那么这两个概念是相似的。对于概念 A，B 的具体实例，可利用 Jaccard 系数[11]来计算相似度，公式为

$$\mathrm{Jaccard\_sim}(A,B)=P(A\cap B)/P(A\cup B)=\frac{P(A,B)}{P(A,B)+P(A,\overline{B})+P(\overline{A},B)} \tag{6}$$

基于实例计算概念相似度涉及到 3 个概率：$P(A,B)$，$P(A,\overline{B})$，$P(\overline{A},B)$。其中 $P(A,B)$ 表示一个实例在某本体中既属于概念 A 又属于概念 B 的概率；$P(A,\overline{B})$ 表示一个实例在某本体中属于概念 A 但不属于概念 B 的概率；同理 $P(\overline{A},B)$ 表示实例在某本体中不属于概念 A 但属于概念 B 的概率。在计算 $P(A,B)$，$P(A,\overline{B})$，$P(\overline{A},B)$ 时，要用到概念 A 和 B 在各自本体中的实例个数。用 U_i 表示本体 O_i 中的实例集，$N(U_i)$ 表示实例集中的实例个数。用 $N(U_i^{A,B})$ 表示在 U_i 中既属于 A 又属于 B 的实例个数，其它类似。以计算 $P(A,B)$ 为例，具体过程如下 6 个步骤[12]：

(1) 对于本体 O_1 的实例集 U_1，把它分成属于概念 A 的实例集 U_1^{A} 和不属于概念 A 的实例集 $U_1^{\overline{A}}$。

(2) 把这两个实例集中的实例分别作为正反样本，用机器学习方法来训练对于概念 A 的学习器 L。

(3) 对于本体 O_2 的实例集 U_2，把它分成属于概念 B 的实例集 U_2^{B} 和不属于概念 B 的实例集 $U_2^{\overline{B}}$。

(4) 使用学习器 L 对实例集 U_2^{B} 中的实例进行分类，分成两个实例集 $U_2^{A,B}$ 和 $U_2^{\overline{A},B}$。同样用 L 把实例集 $U_2^{\overline{B}}$ 分成 $U_2^{A,\overline{B}}$，$U_2^{\overline{A},\overline{B}}$。

(5) 将本体 O_1 和本体 O_2 的位置调换，重复以上各步，这样最终可以获得实例集 $U_2^{A,B}$，$U_2^{\overline{A},B}$，$U_2^{A,\overline{B}}$ 和 $U_2^{\overline{A},\overline{B}}$。

(6) 从各步计算中求得 $N(U_1)$，$N(U_2)$，$N(U_1^{A,B})$ 和 $N(U_2^{A,B})$，并利用式(7)来计算：

$$P(A,B)=\frac{N(U_1^{A,B})+N(U_2^{A,B})}{N(U_1)+N(U_2)} \tag{7}$$

采用同样的步骤方法来计算 $P(A,\overline{B})$，$P(\overline{A},B)$，计算公式分别为(8)和(9)：

$$P(A,\overline{B})=\frac{N(U_1^{A,\overline{B}})+N(U_2^{A,\overline{B}})}{N(U_1)+N(U_2)} \tag{8}$$

$$P(\overline{A},B)=\frac{N(U_1^{\overline{A},B})+N(U_2^{\overline{A},B})}{N(U_1)+N(U_2)} \tag{9}$$

然后利用式(6)计算概念 A 和概念 B 基于实例的相似度。

3.3.3 基于概念属性

本体中，概念的属性对概念的描述具有十分重要的作用。基于属性计算概念相似度的理论依据是：如果两个概念的属性是相同(相似)的，那么这两个概念是相同(相似)的；如果两个属性的域和范围是相同的，那么这两个属性也是相同的。在本体中，属性有两种类型：对象属性和数据类型属性。通常对象属性是由个体关联到个体，而数据类型属性是个体关联到数据类型的值。属性有属性名称、属性数据类型、属性实例数据等要素，因此判断两个属性是否相似主要从这 3 个要素进行考虑。

属性名称、属性类型本身是文本类型，是字符串，因此可以采用字符串相似度计算方法进行判定。例如用 humming distance[13]来比较两字符串。设两字符串 s 和 t，则 s,t 之间的相似度可由式(10)给出：

$$\mathrm{Sim}(s,t)=1-\frac{\left(\sum_{i=1}^{\min(|s|,|t|)}f(i)\right)+||s|-|t||}{\max(|s|,|t|)} \tag{10}$$

其中，若 $s[i]=t[i]$，则 $f(i)=0$，否则 $f(i)=1$。

字符串相似度的计算也可以采用其它方法，前面已有说明。由于每个概念的实例对该概念的每一个属性都分配了一个相应的值，因此对于其它类型的数据，可以采用基于实例的方法进行计算。

设概念 A 的属性为 a_i，概念 B 的属性为 b_j，两个属性之间的相似度记为 $\mathrm{ASim}(a_i,b_j)$，则属性 a_i,b_j 的相似度计算公式为：

$$\mathrm{ASim}(a_i,b_j)=w_1*\mathrm{Sim}(a_{i_{name}},b_{j_{name}})+w_2*\mathrm{Sim}(a_{i_{datatype}},b_{i_{datatype}})+w_3*\mathrm{Sim}(a_{i_{value}},b_{j_{value}}) \tag{11}$$

其中 w_1,w_2,w_3 是权重，代表属性名称、数据类型、属性实例数据对属性相似度计算的重要程度，且 $w_1+w_2+w_3=1$。

设概念 A,B 之间共计算出 m 个 $\mathrm{ASim}(a_i,b_j)$ 并设置相应的权值 $w_{attribute}^{k}$，则概念 A 和概念 B 之间基于属性的相似度为

$$\mathrm{Sim}_{attribute}(A,B)=\frac{\sum_{k=1}^{m}w_{attribute}^{k}\mathrm{ASim}(a_i,b_j)}{\sum_{k=1}^{m}w_{attribute}^{k}} \tag{12}$$

由于一个概念可能有多个属性，每个属性对概念的描述程度和作用也各不相同。若每个都考虑，则计算量相当大。所以在计算属性相似度时，可以先依据机器学习方法给出属性的信息增益[14,15]，并以此为依据来确定各个属性的优先级。最后只选取几个信息增益较大的属性进行计算。

3.3.4 基于概念结构

OWL 中本体概念主要有 4 种基本关系：kind-of，part-of(或 part-whole)，instance-of 和 attribute-of。由于本体中的实例都是单独给出，因此关系 instance-of 可以不作考虑。关系 attribute-of 表示一个概念是另一个概念的属性，其相似度计算也就是概念属性的相似度计算，在此也不作考虑。本体可以依据概念的 kind-of 和 part-of 关系刻画出概念间的层次结构，因此本文中通过考虑概念的父子关系、兄弟关系并结合一些启发规则来给出结构相似度的计算方法。

本文涉及到的启发规则主要有以下几条：

(1) 如果两个概念相似，那么它们的子概念在一定程度上也相似。

(2) 如果两个概念的子概念都相似，那么这两个概念也相似。

(3) 如果两个概念具有相似的兄弟概念，则这两个概念也相似。

(4) 如果所有子概念都与某概念 X 相似，那么它们的父概念也与 X 相似。

(5) 如果某概念的兄弟概念都与某一概念 Y 相似，那么该概念与 Y 也可能相似。

(6) 同一本体中，如果两个概念属于同一个父概念，那么这两个概念是相似的，即兄弟概念是相似的。

(7) 如果概念对中的概念 A 和概念 B 都有多个子概念，其中概念 A 有子概念$\{A_1,A_2,\cdots,A_n\}$，概念 B 有子概念$\{B_1,B_2,\cdots,B_n\}$，并设定一个阈值 δ。若概念 A 中有大于 δ 个子概念与概念 B 中的子概念相似，则可认为概念 A 与概念 B 是相似的。

结构相似度的具体计算公式如式(13)所示：

$$\mathrm{Sim}_{A,B}(C_{i_1},C_{i_2})=\frac{\alpha\mathrm{Sim}^{H_f}_{set}(C_{i_1},C_{i_2})+\beta\mathrm{Sim}^{H_b}_{set}(C_{i_1},C_{i_2})+\gamma\mathrm{Sim}^{H_s}_{set}(C_{i_1},C_{i_2})}{\alpha+\beta+\gamma}$$

$$(\alpha\geqslant\beta\geqslant\gamma>0) \tag{13}$$

其中，$\mathrm{Sim}^{H_f}_{set}(C_{i_1},C_{i_2})$ 表示度量概念 C_{i_1}，C_{i_2} 的父概念集 $H^{C_{i1}}_{fi1}$，$H^{C_{i2}}_{fi2}$ 之间的相似度，其父概念集为其特征因子，$\mathrm{Sim}^{H_b}_{set}(C_{i_1},C_{i_2})$(兄弟概念集)和 $\mathrm{Sim}^{H_s}_{set}(C_{i_1},C_{i_2})$(子概念集)与之类似，$\alpha,\beta,\gamma$ 为权重因子。由于在概念的层次结构中，父子、兄弟概念对其相似度的影响是不同的，而父概念占有绝对的权重，因此本文预定为 $\alpha\geqslant\beta\geqslant\gamma$。

3.3.5 综合概念相似度

根据相似度定义(5)，可以将基于名称的相似度、基于实例的相似度、基于属性的相似度以及基于结构的相似度进行合并，得到本体的综合概念相似度，记为 $\mathrm{Sim}_{concept}(A,B)$。公式表示如式(14)：

$$\mathrm{Sim}_{concept}(A,B)=w_1*\mathrm{Sim}_{name}(A,B)+w_2*\mathrm{Sim}_{instance}(A,B)+w_3*\mathrm{Sim}_{attribute}(A,B)+w_4*\mathrm{Sim}_{structure}(A,B)$$

$$(w_i>0,w_1+w_2+w_3+w_4=1) \tag{14}$$

3.4 综合本体相似度

综上，本体的相似度可由本体的语义相似度与概念相似度两部分得到，记为 $\mathrm{Sim}(O_1,O_2)$，即

$$\mathrm{Sim}(O_1,O_2)=w_1*\mathrm{Sim}_{semantic}(O_1,O_2)+w_2*\mathrm{Sim}_{concept}(O_1,O_2) \tag{15}$$

4 实验与分析

4.1 实验数据

本文在两组测试集上对该方法进行了实验：第一组测试数据是 TestData1，该数据集中的两个本体 family. swrl 和 generation 分别针对家庭成员信息作了不同的描述，两个本体的概念和属性名称定义比较相似；第二组测试数据是 TestData2，该数据集中的两个本体 travelOntology 和 travelMessage分别对旅游信息作了不同的描述。表 1 给出了这两组测试数据的具体统计信息。

表 1 测试集的统计数据

数据集	本体	概念	属性	实例
TestData1	Family. swrl	28	15	22
	generation	18	4	7
TestData2	travelOntology	12	6	17
	travelMessage	35	3	20

4.2 实验结果

本文首先利用人工方法计算各相似度并记录结果。然后借助 WordNet 2. 1 和 Protege 3. 3 等工具，使用该方法分别对各相似度进行计算，其中各权值的设定存在一定的主观因素。本文将实验最终所得的部分数据与 GLUE 系统[11]的概率统计方法以及人工方法的计算结果进行比较，见图 1，图中左半部分为 testData1 中两个本体之间各相似度比较，右半部分为 testData2 中两个本体之间各相似度比较。

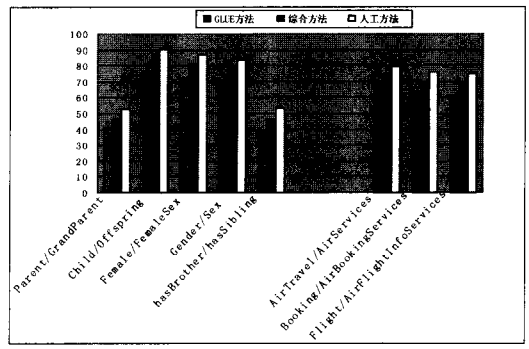

图 1 实验结果对比

4.3 实验分析

与传统 GLUE 系统的概率统计方法相比，综合的相似度计算方法所得到的相似度值更加准确，主要原因在于，本文提出的综合的本体相似度计算方法，首先通过判断相关性，过滤掉了不相关的本体，这样就大大减少了不必要的计算，降低了复杂度。对于相关本体，本文从语义与概念两个层次入手，分别计算本体的语义相似度与概念相似度；而对于每一对概念，则充分考虑概念的名称、实例、属性、结构等相关因素。与已有的计算方法相比，虽然计算量加大了，耗费的时间增多了，但更全面、更准确地反映了概念之间的相似关系。当然这种方法也存在不足，比如计算过程中大量权值设定的存在，对系统存在很大影响。这一点可以在研究过程中通过机器学习(maching learning)或神经网络(neural networks)技术[16]进行改进。

结束语 本体是对领域知识概念的抽象和描述，其目的是为了实现知识的共享与重用。但由于构建标准的不统一，语义异构、结构异构等问题日渐突出，因而本体映射越来越成为本体研究的重要课题，作为本体映射基础的相似度计算也越来越成为该课题研究的难点和焦点。本文针对这一问题提出了一种综合的本体相似度计算方法，有效确保了计算的全面性、准确性。但计算过程中，凭经验设定的各个权值对结果造成了一定的误差，因此还需要作进一步改进。后续研究工作中将对这一技术细节作进一步的研究，并将结合本文所提出的相似度计算方法对本体映射技术作更加深入的研究。

参 考 文 献

[1] Gruber T R . A translation approach to portable ontologies . Knowledge Acquisition, 1993, 5(2): 199-220

[2] Wang Zong-jiang, Wang Ying-lin, Zhang Shen-sheng, et al. Effective Large Scale Ontology Mapping. Berlin Heidelberg: Springer-Verlag, LNAI 4092, 2006: 454-465

[3] Ehrig M, Sure Y. Ontology Mapping - An Intergrated Approach [C]// Proceedings of the 1st European Semantic Web Symposium. Heraklion, Greece: Springer, May 2004: 76-91

[4] Ehrig M, Staab S. QOM-Quick Ontology Mapping[C]// ISWC 2004, LNCS 3298. 2004: 683-697

[5] 肖文芳. 基于相似度计算的本体映射研究与实现[D]. 中南大学, 2007

(下转第 182 页)

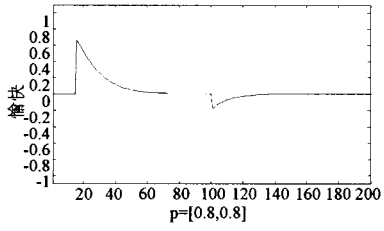

图 11　改变个性后"愉快"的变化曲线

图 12 与图 13 为心境变化曲线。$t+1$ 时刻的心境，除了与 t 时刻的心境有关外，还与最近一段时间内的情绪有关，仿真时定义：$t+1$ 时刻的心境与前面 25 个时间间隔内的情绪相关。初始值设为 $m=[0,0]$，$p=[0.2,0.2]$，施加刺激后，情绪产生变化，随后心境也产生变化，心境变化是多种基本情绪变化的综合体现，而且变化速度与情绪相比相对较慢，最后当情绪恢复平静后，心境也慢慢恢复平静。

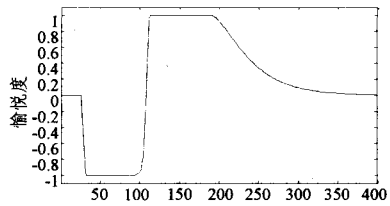

图 12　心境-愉悦度变化曲线图

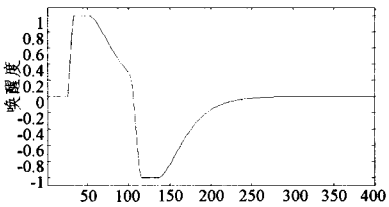

图 13　心境-唤醒度变化曲线

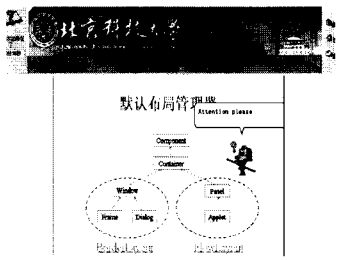

图 14　虚拟教师情绪反应示例

在课题组设计的虚拟教学系统中，基于本模型构建了虚拟教师的模糊情感模型。系统将学生的学习情况作为虚拟教师的情感刺激，如：学生学习时的表情、姿态与学习效果等。针对不同的情感刺激虚拟教师会产生不同的情绪并做出行为反应，例如当学生学习时精神不集中，表现为频繁左右张望、身体后倾等，虚拟教师就会生气，并给出提示，如图 14 所示。

结束语　本文提出的模糊情感模型用模糊数学描述情绪、心境、个性等心理学概念，构建状态空间方程来描述人类情绪系统。该情感模型考虑了人类情绪变化涉及到的各种因素：刺激、心境、衰减等，能较全面地模拟人类情绪，通过 Matlab 的仿真结果，可以直观地看到本模型模拟的情感变化符合人类情感变化规律。此外该模型结构简单，并且具有较好的可重构性，可以根据实际应用来具体定义情绪、心境和个性，定义可简单也可复杂，因此具有较好的应用性和可实现性，可以用在虚拟人和机器人中，使之产生人类情感。

参 考 文 献

[1] Picard R W. Affective Computing. M. I. T media laboratory perceptual computing section technical report no. 321. november 1995:1-26

[2] 黄崇彬，原田昭. 日本感性工学发展近况与其在远隔控制接口设计上应用的可能性//中日设计教育研讨会论文集. 台湾：国立云林科技大学，1998:17-26

[3] 王志良. 人工心理学——关于更接近人脑工作模式的科学. 北京科技大学学报，2000，10:478-481

[4] Sloman A. What Are the Emotion Theories About //Architecture for modeling emotion: cross-disciplinary foundation American Association for Artificial Intelligence 2004 Spring Symposium. Stanford University，2004:128-134

[5] Canamero L D. A hormonal model of emotions for behavior control. VUB AI-Lab Memo 97-06，Vrije Universiteit Brussels，Belgium. 1997:28-31

[6] Ushida H. Interactive Agents with Artificial Mind. International Journal of Computational Intelligence，2004，1(4):327-323

[7] 王志良，解仑，董平. 情感计算数学模型的研究初探. 计算机工程，2004(21):33-34，167

[8] Kshirsagar S，Magnenat-Thalmann N. A multilayer personality model//Proceedings of 2nd International Symposium on Smart Graphics. ACM Press，2002:107-115

[9] Breazeal C. A Motivational System For Regulating Human-Robot Interaction//Proceedings of the National Conference on Artificial Intelligence. Madison，WI，1998:54-61

[10] Miwa H，Itoh K，Ito D，et al. Introduction of the need model for humanoid robots to generate active behavior // Proceedings of IEEE/RSJ Intl. Conference on Intelligent Robots and Systems. 2003:1400-1406

[11] Mehrabian A. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current Psychology，1996，14:261-292

[12] 郭永玉. 人格心理学. 北京：中国社会科学出版社，2005

(上接第 145 页)

[6] Maedche A，Staab S. Measuring Similarity between Ontologies [C]// Proceedings of the European Conference on Knowledge Acquisition and Management. Madrid，Spain，Oct. 2002:251-263

[7] Budanitsky A，Hirst G. Evaluating WordNet-based Measures of Lexical Semantic Relatedness[C]. Association for Computational Linguistics，2005

[8] 张承立，陈剑波，齐开悦. 基于语义网的语义相似度算法改进[J]. 计算机工程与应用，2006(17):165-166，179

[9] Pantel P，Lin D. Discovering word senses from text. [C]// Proceedings of the 2002 ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Edmonton Alberta，Canada，2002:613-619

[10] Bouquet P，Euzenat J，Franoni E，et al. Speifiation of a common framework for charaterizing alignment. Knowledge Web Deliverable 2. 2. 1v2，University of Karlsruhe，2004

[11] Doan A H，Madhavan J，Domingos P，et al. Learning to Map between Ontologies on the Semantic Web[C]//Proceedings of the 11th International Conference on World Wide Web. New York，USA，2002:662-673

[12] 郑丽萍. 本体映射的研究[D]. 山东科技大学，2005，5

[13] 曹泽文，钱杰，张维明，等. 一种综合的概念相似度计算方法[J]. 计算机科学，2007，34(3):174-175，191

[14] Han Jiawei，Kamber M. 数据挖掘概念与技术. 范明，孟晓锋，译. 北京：机械工业出版社，2007

[15] 郑丽萍，李光耀，梁永全，等. 本体中概念相似度的计算[J]. 计算机工程与应用，2006(30):25-27，61

[16] Li W S，Clifton C. Semantic integration in heterogeneous databases using neural networks // Proceedings of the 20th VLDB. Santiago (CH)，1994:1-12