DOI:10.3969/j.issn.1673-4785.201405061
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20150302.1106.004.html

模式匹配不确定性的多因素集结度量

胡文彬,潘祝山,纪兆辉
(淮海工学院 计算机工程学院,江苏 连云港 222005)

摘 要:为了能够有效度量模式匹配的不确定性,提出了一个模式匹配不确定性的度量模型,根据不确定性因素间的关系提出了一个集结算子。使用全知熵度量语义匹配和属性匹配的不确定性,引入过程不确定性的度量方法度量匹配决策过程的不确定性。使用多因素集结算子判断各因素的影响程度,并可合成各度量结果。实验证明,所提模型和方法能够有效度量模式匹配的不确定性,且具有高效性和可扩展性。

关键词:模式定义;模式分析;模式匹配;不确定性分析;数据不确定性度量;度量方法;决策分析;熵;集结评估方法
中图分类号: TP18;TP391 **文献标志码**:A **文章编号**:1673-4785(2015)02-0286-07

中文引用格式:胡文彬,潘祝山,纪兆辉. 模式匹配不确定性的多因素集结度量[J]. 智能系统学报, 2015, 10(2): 286-292.
英文引用格式:HU Wenbin, PAN Zhushan, JI Zhaohui. Uncertain measure for schema matching based on the aggregation of uncertain factors[J]. CAAI Transactions on Intelligent Systems, 2015, 10(2): 286-292.

Uncertain measure for schema matching based on the aggregation of uncertain factors

HU Wenbin, PAN Zhushan, JI Zhaohui
(School of Computer Engineering, Huaihai Institute of Technology, Lianyungang 222005, China)

Abstract: To measure efficiently uncertainty of schema matching, a measure model based on all uncertain factors was proposed and an aggregation operator was given according to the relations of uncertain factors. A measure method of semantic matching and attribute matching based on all known entropy uncertain ratio was designed. A measure algorithm of process uncertainty was introduced to measure uncertainty of a decision making process. The aggregation operator based on relationships between uncertain factors was proposed to determine influence degree of uncertain factors and merge all measure values in the measure process. The real world examples illustrate that the proposed model and methods can completely reflect three factors of uncertainty and can measure efficiently uncertainty for schema matching. The proposed methods are efficient and scalable.

Keywords: schema definition; schema analysis; schema matching; uncertainty analysis; measured data uncertainty; measurement method; decision analysis; entropy; aggregation estimation method

模式匹配是许多领域的关键操作,是模式对象间的映射或相应关系的识别[1]。由于模式对象间的语义不能完全来源于数据和元数据信息,因此模式匹配中存在固有的不确定性,而且其不确定性会影响模式集成的整个过程[2],被认为是开展大规模数据集成的一个关键瓶颈,不确定性管理是未来的挑战之一[3]。通常,自动或是半自动模式匹配的方法都是耗时和难于实施的,尤其是进行大规模模式匹配就更困难了,但若能在具体模式匹配实施前,对整个过程进行不确定性度量,将会为模式匹配在语义 Web、模式集成、无线网络和电子商务等诸多领域中的高效应用提供决策参考。

收稿日期:2014-06-06. **网络出版日期**:2015-03-02.
基金项目:国家自然科学基金资助项目(60903027);江苏省自然科学重大研究项目资助项目(BK2011023);江苏省自然科学基金资助项目(BK2011370).
通信作者:胡文彬. E-mail:hwb1008@163.com.

模式匹配实质上是一多属性决策过程[4],其过程中需要考虑一定的不确定性。对模式匹配不确定性的度量研究目前比较少,相关领域主要针对科学数据库和确定数据库进行基于确定语义的匹配操作,主要目的是尽量提高映射结果的正确率。不确定模式匹配的相关研究中,基于不确定语义映射的模式集成[2]、基于相似度计算的方法[5]、基于概率映射的模式匹配方法[6]、基于 by-table 和 by-tuple 的数据集成方法[7]和 Top-K 方法[8]等均是在匹配结果上尽量提高输出正确率,而抛弃掉一些不确定性信息和结果,因此会丢失一些对用户有用的信息,并且这些研究工作中均未对整个匹配的不确定性进行度量,未考虑匹配过程中不确定性因素对结果的综合影响。由 B.Liu 在 2007 年提出的不确定性度量适用于不精确数量数据的度量[9-10]。与本文相近的工作有 AMUR 算法和粗糙集(rough set, RS)的不确定性度量,AMUR 算法处理的对象是 RFID 数据[11]。RS 理论是由波兰科学家 Pawlak[12]在 1982 年提出的一种有效处理不确定性的工具,对 RS 不确定性度量的研究是近年来的研究热点,在经典 RS 理论中,产生不确定性的原因有集合的粗糙性和知识(概念)的不确定性[13]。基于信息熵的度量方法[14]能够反映出产生不确定性的 2 个因素,但不能够全面地反映出知识不确定性,该方法被应用在粗糙集的异常值发觉中;基于不确定熵的度量方法[13]综合了粗糙熵、精确度和包含度 3 种基本方法,能够反映出粗糙性和知识的不确定性;基于知识粒度的不确定性度量方法[15]适于解决集合的粗糙性;定位服务的不确定性度量方法[16]运用粗糙集和证据理论进行不确定性度量。这些方法的实际应用范围有限,容易受系统规模的影响,且未详细讨论不确定性因素对度量结果的影响程度。

本文针对不确定性模式匹配的处理过程[17],提出了一个多因素集结的模式匹配不确定性度量模型,根据语义匹配和属性匹配不确定性因素的特点,运用全知熵度量其中的不确定性,并引入过程不确定性度量方法对匹配决策的不确定性进行了度量。根据不确定性因素间的相互关系,给出了一个集结算子,用于判断各不确定性因素的影响程度和合成度量结果以生成总不确定率。所提出的模型和方法能够有效度量模式匹配的不确定性,能够综合各不确定性因素产生的影响,能够处理大规模模式匹配的不确定性度量,为复杂系统的不确定性度量奠定了基础。

1　模式匹配中的不确定性

根据不确定性模式匹配的处理过程,模式匹配(schema matching,SM)的不确定性主要出现在语义匹配、属性匹配和匹配决策过程中,存在于其中的不确定性表现[17],可归结为语义因素、属性因素和过程决策因素,这 3 个不确定性因素具有源发性和主导性。

定义 1　不确定语义匹配。2 个模式 S_1 和 S_2 的不确定语义匹配是一个三元组 $\langle S, O, \mathrm{UM}\rangle$,其中 S 是模式有限集,$S_1, S_2 \in S, O \in S_i$ 是模式对象有限集,$\mathrm{UM} = \{\langle r_{11}, m_{11}\rangle, \langle r_{12}, m_{12}\rangle, \cdots, \langle r_{1k}, m_{1k}\rangle, \langle r_{21}, m_{21}\rangle, \cdots, \langle r_{nk}, m_{nk}\rangle\}$ 是模式对象间的不确定匹配关系集,$r_{ij} \in R, i = 1,2,\cdots,n, j = 1,2,\cdots,k, n = |S_1|$ 和 $k = |S_2|$ 为所包含模式对象的个数,$R = \{$相等,包含,相交,超集,不相交,不相容$\}$ 是 6 种相互排斥的语义关系集,m_{ij} 为 r_{ij} 的不确定率。

定义 2　不确定属性匹配。2 个模式对象的不确定属性匹配是一个三元组 $\langle A, T, \mathrm{UD}\rangle$,其中 A 是属性集,$T = \{\mathrm{ANM}, \mathrm{ATM}, \mathrm{KRM}, \mathrm{DIM}\}$ 是匹配类型集,ANM 是属性名匹配,ATM 是属性数据类型匹配,KRM 是关键字约束匹配,DIM 是数据实例匹配,$\mathrm{UD} = \{\langle A_1, \mathrm{UD}_1\rangle, \langle A_2, \mathrm{UD}_2\rangle, \langle A_3, \mathrm{UD}_3\rangle, \langle A_4, \mathrm{UD}_4\rangle\}$ 是各类属性匹配的不确定率集,$A_i \in A$。

定义 3　不确定决策过程。不确定决策过程 UDP 是一个四元组 $\langle T, \mathrm{ST}, P, f\rangle$,其中 T 是任务集,ST 是状态集,P 是不确定度集,$f: T \times \mathrm{ST} \to P$ 是一个决策函数。

定义 4　模式匹配的不确定性度量。模式匹配的不确定性度量是满足系统不确定性度量[18]中 4 个条件的不确定性度量。

2　模式匹配的不确定性度量

2.1　不确定性度量模型

模式集成中的模式匹配不确定性度量模型由模式对象清洗(schema object cleanout,SOC)、语义匹配不确定性度量(uncertainty measure of semantic matching,UMSM)、属性匹配不确定性度量(uncertainty measure of attribute matching,UMAM)、决策过程不确定性度量(uncertainty measure of process,UMP)和不确定性度量合成器(uncertainty measure synthesizer,UMS)5 个模块组成。模式匹配不确定性度量模型的框架如图 1 所示。

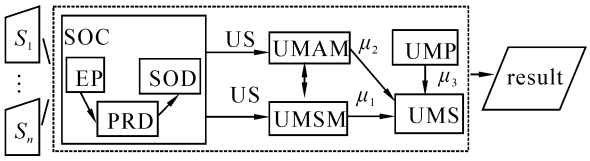


图 1 模式匹配不确定性的度量模型

Fig.1 Uncertainty measure model for schema matching

待匹配模式集作为输入,由 SOC 清洗掉确定模式对象,不确定部分(uncertainty set, US)由 UMSM 和 UMAM 进行语义匹配和属性匹配不确定性的度量,再由 PUM 对整个决策过程的不确定性进行度量,最后由 UMS 合成各度量结果而得到总不确定率。

2.2 模式对象清洗

模式匹配的复杂度会随数据集成规模的增大而增大,对输入模式进行预处理至关重要。在 SOC 中使用属性约减方法[19]对输入模式所包含的模式对象进行等价类划分(equipollence partition, EP)后,再进行正域约减(positive region deduction, PRD)和模式对象约简(schema object deduction, SOD),得到可能存在不确定性的部分——模式对象约简集。模式匹配不确定性的度量规模经过 SOC 的处理后明显缩小。

2.3 基于全知熵的不确定性度量

粗糙性是指由于知识的不完备性或不精确性,导致对象与对象之间不可分辨,从而使得对象与概念之间的关系具有不确定性[12]。信息熵是信息理论中用于分析不确定程度的一种重要度量,以所需信息量的多少来衡量不确定性的程度[20]。基于信息熵的度量方式中全知熵不确定率对系统的不确定性比较敏感,能够较为准确地反映不确定性的变化规律[21]。模式匹配的执行过程能够表达其系统内的条件属性知识和决策属性知识,模式匹配的不确定性结构和程度可由属性知识完全确定,因此基于全知熵的度量方式适于度量模式匹配的不确定性。

2.3.1 模式匹配的全知熵不确定率

定义 5 模式匹配的全知熵不确定率。四元组 $DS=(O, MA, V, f)$ 为模式匹配决策系统,其中,O 为模式对象有限集;$MA=C\cup D$ 是匹配属性的集合,C 为不确定匹配关系集,D 为决策属性集,$C\cap D=\phi$,$a\in C\cup D$;$V=\cup V_a$ 是属性的值域,$f: O\times MA\rightarrow V$ 是一个信息函数。模式匹配的全知熵不确定率定义为

$$\mu_{all}=1-(H_{all}(C\rightarrow D)-H(D))/(\log(|O|)-H(D))$$

式中:$H_{all}(C\rightarrow D)=H(C)+H(D|C)$ 为全知熵[21],$H(C)$ 为 C 在 U 上的信息熵,且 $H(C)=-\sum_{1\leqslant i\leqslant n}p(X_i)\log(p(X_i))$ ($O/\mathrm{IND}(C)=\{X_1, X_2, \cdots, X_n\}$, $n=|O/\mathrm{IND}(C)|$),$H(D|C)$ 为条件熵,且 $H(D|C)=-\sum_{1\leqslant i\leqslant n}p(X_i)\sum_{1\leqslant j\leqslant m}p(Y_j|X_i)\log(p(Y_j|X_i))$ ($O/\mathrm{IND}(D)=\{Y_1, Y_2, \cdots, Y_m\}$, $m=|O/\mathrm{IND}(D)|$),$H(D)$ 为 D 在 U 上的信息熵。

可将模式匹配看做一个决策系统,C 的元素为定义 1 中 R 所包含的元素,$D=\{0$(不是),1(是),2(不确定)$\}$。所定义的不确定率满足粗糙集不确定性度量的基本准则。

定理 模式匹配的全知熵不确定率满足粗糙集不确定性度量的基本准则。

证明 全知熵不确定率 $\mu_{all}=1-(H_{all}(C\rightarrow D)-H(D))/(\log(|O|)-H(D))=1-(H(C)+H(D|C)-H(D))/(\log(|O|)-H(D))$。$R_1$ 和 R_2 是 U 上的 2 个等价关系。

1) $0\leqslant(H(C)+H(D|C)-H(D))/(\log(|O|)-H(D))\leqslant 1$,因此 $0\leqslant\mu_{all}\leqslant 1$ 非负;

2)若 $R_1\approx R_2$,则 $H(C_1)=H(C_2)$,$H(D_1)=H(D_2)$,所以 μ_{all} 满足不变性;

3)若 $R_1<R_2$,则根据文献[22]的定理 7 有 $H(C_1)<H(C_2)$,$H(D_1)<H(D_2)$,所以 μ_{all} 满足单调性。

综上所述,模式匹配的全知熵不确定率满足粗糙集不确定性度量的基本准则[23]。

2.3.2 语义匹配的不确定性度量

语义匹配是用于确定各模式及其模式对象间匹配程度的过程之一,由于模式对象的语义不能完全来源于数据和元数据信息,并且识别确定的语义映射是非常困难的,因此语义匹配中产生的不确定性是模式匹配中存在不确定性的主因之一[24]。语义匹配是一决策过程,语义匹配的不确定率如下:

$$\mu_1=1-(H_{all}(C_1\rightarrow D_1)-H(D_1))/(\log(|O_1|)-H(D_1))$$

式中:C_1 是语义匹配条件属性集,D_1 是语义匹配决策属性集,$O_1\in O$ 是模式对象集。

2.3.3 属性匹配的不确定性度量

属性匹配不确定性的度量由 UMAM 来完成,实现属性名匹配(ANM)、数据类型匹配(ATM)、关键字约束匹配(KRM)和数据实例匹配(DIM)的不确定性度量,计算出属性匹配的总不确定率。属性匹配过程同样是一决策过程,其中的条件属性集合均为{相等,包含,相交,超集,不相交}。

根据定义 5 属性匹配的不确定率如下:

$$\mu_2 = \sum_{i=1}^{4} (\mu_{2i}^2 / \sum_{j=1}^{4} \mu_{2j})$$

式中:$\mu_{2i} = 1 - (H_{\text{all}}(C_{2i} \rightarrow D_{2i}) - H(D_{2i})) / (\log(|U_{2i}|) - H(D_{2i}))$,$i=1,2,3,4$,分别是 4 种属性匹配的不确定率,$\mu_{2i} / \sum_{j=1}^{4} \mu_{2j}$,$j=1,2,3,4$,是各属性匹配不确定率的权值。

2.4　匹配决策过程的不确定性度量

模式匹配是根据运行时管理者的决策或过程数据有条件执行的,因而其过程中存在不确定性是毫无疑问的[25]。用 Petrinets 表示模式匹配的决策过程(decision process,DP),如图 2。

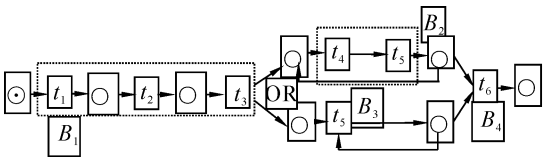


图 2　模式匹配的过程模型

Fig.2　Process model of schema matching

t_i为任务块,B_1为 SOC 的执行过程,B_2为语义匹配和属性匹配的顺序执行过程,B_3为属性匹配的执行过程,B_4为匹配结果合并过程。图 2 可转换为图 3 形式。

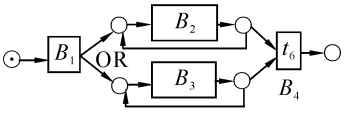


图 3　转换后的过程模型

Fig.3　Converted process model

决策过程不确定性的计算通式如下:

$$K(\text{DP}) = -\sum_{k=1}^{M} p(\text{BS}_k) \log_2 p(\text{BS}_k) + \sum_{g=1}^{N} p(B_g) U(B_g)$$

式中:N 是过程中任务的总数量,M 是可能被执行任务的总数,BS_k是第 k 个可能被执行的任务,$P(\text{BS}_k)$是执行概率,B_g是过程中第 g 个任务,$P(B_g)$是为了完成整个过程而由所有 M 个可执行任务执行的 B_g 的概率,$U(B_g) = \sum_{i=1}^{N} P(B_i) \log_2(B_i)$。

决策过程不确定率 μ_3 的计算公式如下:

$$\mu_3 = \frac{\sum_{i=1}^{N} (T_i \times K(B_i))}{\sum_{i=1}^{N} T_i}$$

式中:T_i是任务 B_i被执行的次数。由于 B_1和 B_4为必须执行的过程,因此可令 $K(B_1)=0$,$K(B_4)=0$。

3　不确定性因素的集结度量

3.1　不确定性因素影响程度的判断

影响模式匹配的不确定因素之间往往不是孤立的,它们之间可能存在着某些关系,这将影响不确定性的准确度量,各因素及其权重的简单线性组合也从一定程度上忽略了这些因素之间存在的相互关系[26]。

定义 6　不确定性的集结度量。$U(\text{SM}) = f(S(\text{SM}), A(\text{SM}), P(\text{SM}))$ 称为模式匹配 SM 不确定性的集结度量,$U(\text{SM}) \in [0,1]$,$U(\text{SM})$越大 SM 的不确定性就越大,其中 f 称为以 $S(\text{SM})$、$A(\text{SM})$和 $P(\text{SM})$为自变量的不确定性度量函数,$f:[0,1]\times[0,1]\times[0,1] \rightarrow [0,1]$。

基于 f 对 SM 不确定性的度量将所有不确定性因素映射到一个[0,1]上,反映了模式匹配的不确定性。

模式匹配中任何不确定因素的细微变化都将影响到整个过程。综合语义匹配、属性匹配和决策过程这 3 个方面的不确定因素以进行模式匹配不确定性的客观度量,而不是各因素的简单线性组合,需要考虑这 3 个因素的内在关系。为了讨论单个因素对不确定性的影响以及多个因素集结起来对不确定性的影响,下面给出相关定义。

定义 7　分别对因素 $S(\text{SM})$,$A(\text{SM})$和 $P(\text{SM})$作如下变换:

$$S'(\text{SM}) = 2S(\text{SM}) - 1$$
$$A'(\text{SM}) = 2A(\text{SM}) - 1$$
$$P'(\text{SM}) = 2P(\text{SM}) - 1$$

令因素 $e \in \{S(\text{SM}), A(\text{SM}), P(\text{SM})\}$,相应地,$z \in \{S'(\text{SM}), A'(\text{SM}), P'(\text{SM})\}$称为因素 e 对 SM 不确定性的影响,$z \in [-1,1]$。若 $z \geqslant 0$,则称 e 为积极因素;若 $z=0$,则称 e 为不变因素;若 $z<0$,则称 e 为消极因素。

定义 8　若给定 $z_1, z_2 \in \{S'(\text{SM}), A'(\text{SM}), P'(\text{SM})\}$,$z_1 \neq z_2$,集结算子定义如下:

$$z_1 \oplus z_2 = \begin{cases} z_1 + z_2 - z_1 \cdot z_2, & z_1 \geqslant 0, z_2 \geqslant 0, \text{或} \\ & z_1 \geqslant 0, z_2 \leqslant 0, |z_1| > |z_2| \\ z_1 + z_2 + z_1 \cdot z_2, & z_1 \leqslant 0, z_2 \leqslant 0, \text{或} \\ & z_1 \geqslant 0, z_2 \leqslant 0, |z_1| < |z_2| \\ 0, \text{其他} \end{cases}$$

若已存在一个不确定因素 e_1,再给定一个不确定因素 e_2,它们的影响分别为 z_1 和 z_2,则 2 个因素 c 之间的内在相互关系有如下特性:

1)增长性,若 e_2 是一个积极因素,即 $z_2>0$。e_1 和 e_2 对 SM 不确定性的集结影响 z 应该满足 $z>z_1$;若 e_1 也是一个积极因素,则 $z>z_2$ 也同样成立。

2)不变性,若 e_2 是一个不变因素,即 $z_2=0$。e_1 和 e_2 对 SM 不确定性的集结影响 z 应该满足 $z=z_1$,即 SM 的不确定性受 e_1 的影响大;若 e_1 也是一个不变因素,则 SM 的不确定性不受加入因素的影响。

3)减弱性,若 e_2 是一个消极因素,即 $z_2<0$。e_1 和 e_2 对 SM 不确定性的集结影响 z 应该满足 $z<z_1$;若 e_1 也是一个消极因素,则 $z<z_2$ 也同样成立。一个积极因素和一个消极因素对 SM 不确定性的集结影响取决于绝对值大的因素,且集结影响值小于较大的绝对值。

4)有界性,$z_1 \oplus z_2 \in [-1,1]$,以确保多个因素的集结影响可以通过两两集结来实现。

5)交换率,$z_1 \oplus z_2 = z_2 \oplus z_1$,这可以保证 2 个给定不确定性因素对 SM 不确定性的影响保持不变。

6)结合率,$(z_1 \oplus z_2) \oplus z_3 = z_1 \oplus (z_2 \oplus z_3)$,这表明 2 个以上因素的集结影响与各因素参与计算的次序无关。

3.2 总不确定率

模式匹配的总不确定率主要由 3 部分来确定,分别是语义匹配的不确定率 $S(M)$、属性匹配的不确定率 $A(M)$ 和决策过程的不确定率 $P(M)$。由于 $z_1 \oplus z_2 \oplus z_3 \in [-1,1]$,而模式匹配不确定性度量函数的值域为 $[0,1]$,因此模式匹配不确定性的度量函数(即,总不确定率)定义为

$$\mu_{\text{whole}} = f(S(M), A(M), P(M)) = 1/2\,[z_1 \oplus z_2 \oplus z_3] + 1/2$$

式中:$z_1, z_2, z_3 \in \{S'(M), A'(M), P'(M)\}$,$S(M)=\mu_1$,$A(M)=\mu_2$,$P(M)=\mu_3$。

4 实验与分析

4.1 实验

设计 2 种实验方案:1)较小不确定性的度量。对 2 个模式 S_1(在校生)和 S_2(毕业生)间的匹配进行不确定性度量,所包含模式对象的详细信息分别如图 4、5 所示;2)多模式对象匹配不确定性的度量。对 5 个模式间的匹配进行不确定性度量,共包含 57 个模式对象,实验数据情况见表 1。实验参数见表 2。属性匹配包括属性名匹配(ANM)、属性类型匹配(ATM)、关键字匹配(KRM)和数据实例匹配(DIM)4 个过程,每个过程中条件属性个数 $|C|=6$,条件属性值域 $VC=\{0,1,2\}$,每个过程中的模式个数和模式对象个数见表 3。模式匹配的不确定率见表 4。

Student

PID	pname	sex	class	native	birthday	speciality_id	dep_id
030405	曹杰	男	030404051	江苏南京	02/09/81	102001	02
030506	陈海霞	男	030404051	江苏南京	08/06/82	102002	03
030505	陈珂山	男	030404051	江苏扬州	10/09/82	102001	02

图 4 存储在校生数据的模式截图

Fig.4 Schema data of undergraduate

Student

UID	name	sex	classtype	native_place	birthday	speciality	department
030503	陈振	男	030404051	山东青岛	11/08/81	102003	02
030307	程进	男	030403022	上海	05/04/82	102002	03
030304	杜鹏	男	030403022	江苏苏州	08/05/80	102002	03

图 5 存储毕业生数据的模式截图

Fig.5 Schema data of graduate

表 1 实验数据情况

Table 1 Experimental data

序号	年份	对象个数
1	2005sp	17
2	2006au	17
3	2007au	7
4	2008sp	9
5	2009sp	7

注:表中数据为 2005 年春至 2009 年春江苏省 VFP 二级考试的数据情况

表 2 语义匹配不确定性度量的实验参数

Table 2 Experiment parameters of uncertainty measurement for semantic matching

No.	模式个数 $\|S\|$	模式对象个数 $\|U\|$	条件属性 $\|C_1\|$	条件属性 V_{C1}	决策属性 $\|D_1\|$	决策属性 V_{D1}
1	2	2	6	{0,1,2}	1	0,1,2}
2	5	52	6	{0,1,2}	1	{0,1,2}

注:{0,1,2} 中,0—是,1—不是,2—不确定

表 3 属性匹配不确定性度量的实验参数

Table 3 Experiment parameters of uncertainty measurement for attribute matching

匹配类型	方案 1		方案 2	
	$\|S\|$	$\|U\|$	$\|S\|$	$\|U\|$
ANM	2	12	5	301
ATM	2	16	5	301
KRM	2	2	5	5
DIM	2	300	5	3 501

表 4　2 种方案下的模式匹配的不确定率

Table 4　Uncertainty ratio of schema matching from two projects

	第 1 种方案	第 2 种方案
USM	$\mu_1 = 0$	$\mu_1 = 0.34$
UAM	$\mu_2 = 0.21$	$\mu_2 = 0.95$
DP	$\mu_3 = 0.17$	$\mu_3 = 0.17$
总计	$\mu_{whole} = 0.19$	$\mu_{whole} = 0.72$

4.2　分析

第 1 种方案中,首先通过 SOC 处理后,匹配规模比[2]中降低 50%。2 个模式的模式对象名语义相同,因此$\mu_1 = 0$。属性匹配中只有属性名匹配具有不确定性。决策过程中的不确定率随模式匹配规模的减小而降低,利用公式计算得到的μ_{whole}值符合实际情况。第 2 种方案中,模式对象规模和属性匹配的规模突然增大,通过 SOC 的处理后,匹配规模降低了近 1/10,同时计算效率也明显提高,不确定率随匹配规模增大而增大符合不确定性度量的基本准则。实验表明度量模式匹配不确定性的模型和不确定率计算方法具有可行性、有效性、可扩展性和高效性。

5　结束语

模式匹配的不确定性研究是国际上相关领域近年来才兴起的热点研究方向,度量原始匹配的不确定性是关键问题。本文根据模式匹配中产生不确定性的主要因素,首次将全知熵不确定率和过程不确定率结合起来,并证明模式匹配的全知熵不确定率满足粗糙集不确定性度量的基本准则,提出了一个多因素集结的模式匹配不确定性度量模型,利用集结算子判断各不确定性因素对模式匹配不确定性的影响程度和合成各阶段的度量结果,实验证明本文提出的方法与已有方法相比可获得更加合理的度量结果。所提模型解决了不确定性度量中规模限制问题,使得大规模模式匹配不确定性的处理复杂度降低。下一步的工作将探讨动态环境下模式匹配不确定性的度量方法及其处理过程中不确定性传播的测算方法。

参考文献:

[1]SHVAIKO P, EUZENAT J. A survey of schema-based matching approaches[J]. Journal on Data Semantics IV, 2005(3730): 146-171.

[2]MAGNANI M, RIZOPOULOS N, BRIEN P, et al. Schema integration based on uncertain semantic mappings[J]. Lecture Notes in Computer Science, 2005(3716): 31-46.

[3]HALEVY A, RAJARAMAN A, ORDILLE J. Data integration:the teenage years[Z]. Seoul, 2006: 9-16.

[4]翁年凤,刁兴春,曹建军,等. 不确定模式匹配研究综述[J]. 计算机科学, 2011, 38(12): 1-5.
WENG Nianfeng, DIAO Xingchun, CHAO Jianjun, et al. Survey of uncertain schema matching[J]. Computer Science, 2011, 38(12): 1-5.

[5]姜芳艽,孟小峰,贾琳琳. Deep Web 集成服务的不确定模式匹配[J]. 计算机学报, 2008, 31(8): 1412-1421.
JIANG Fangjiao, MENG Xiaofeng, JIA Linlin. Uncertain schema matching in deep web integration service[J]. Chinese Journal of Computers, 2008, 31(8): 1412-1421.

[6]MAGNANI M, MONTESI D. Probabilistic data integration[R]. Bologna(italy): UBLCS, 2009.

[7]DONG X L, HALEVY A, YU CONG. Data integration with uncertainty[J]. The VLDB Journal, 2009, 18: 469-500.

[8]AVIGDOR G. Managing uncertainty in schema matching with top-k schema mappings[J]. Journal on Data Semantics, 2006, 6: 90-114.

[9]LIU Baoding. Uncertainty theory[M]. Berlin: Springer-Verlag, 2007: 3-12.

[10]LIU Baoding. Some research problems in uncertainty theory[J]. Journal of Uncertain Systems, 2009, 3(1): 3-10.

[11]王永利,钱江波,孙淑荣. AMUR:一种 RFID 数据不确定性的自适应度量算法[J]. 电子学报, 2011, 39(3): 579-584.
WANG Yongli, QIAN Jiangbo, SUN Shurong. AMUR: an adaptive measuring algorithm of underlying uncertainty for rfid data[J]. Chinese of Journal Electronics, 2011, 39(3): 579-584.

[12]PAWLAL Z. Rough sets[J]. International Journal of Computer and Information Science, 1982, 11(5): 341-356.

[13]QIU Taorong, YOU Min, GE Hanjuan, et al. A method of uncertainty measure based on rough set[Z]. 2008: 544-547.

[14]JIANG Feng, SUI Yuefei, CAO Cungen. An information entropy-based approach to outlier detection in rough sets[J]. Expert Systems with Applications, 2010, 37(9): 6338-6344.

[15]LIANG Jiye, WANG Junhong, QIAN Yuhua. A new measure of uncertainty based on knowledge granulation for rough sets[J]. Information Sciences, 2009, 179(4): 458-470.

[16]IFTIKHAR-U S, ARYYA G. Managing uncertainty in loca-

tion services using rough set and evidence theory[J]. Expert System with Application, 2007, 32(2): 386-396.

[17]胡文彬,李千目,张宏. 基于领域知识的不确定性关系模式集成[J]. 南京理工大学学报:自然科学版, 2010, 34(4): 409-414.
HU Wenbin, LI Qianmu, ZHANG Hong. Uncertain relation schema integration based on domain knowledge[J]. Journal of Nanjing University of Science and Technology: Natural Science, 2010, 34(4): 409-414.

[18]胡文彬,张宏,李千目. 基于全知熵的模式集成不确定性度量模型[J]. 南京航空航天大学学报, 2012, 44(4): 575-579.
HU Wenbin, ZHANG Hong, LI Qianmu. Uncertainty measure model of schema integration based on all known entropy[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2012, 44(4): 575-579.

[19]WANG J G, MENG G Y, ZHENG X L. The attribute reduce based on rough sets and sat algorithm[Z]. 2008: 98-102.

[20]LIANG Jiye, QIAN Yuhua. Information granules and entropy theory in information systems[J]. Science in China Series F: Information Sciences, 2008, 51(10): 1427-1444.

[21]赵军,周应华. 基于粗集理论的系统不确定性度量方式研究[J]. 小型微型计算机系统, 2010, 31(2): 354-359.
ZHAO Jun, ZHOU Yinghua. Study on system uncertainty measures based on rough set theory[J]. Journalof Chinese Computer Systems, 2010, 31(2): 354-359.

[22]YU Daren, HU Qinghua, WU Congxin. Uncertainty measures for fuzzy relations and their applications[J]. Applied Soft Computing, 2007, 7(3): 1135-1143.

[23]胡军,王国胤. 粗糙集的不确定性度量准则[J]. 模式识别与人工智能, 2010, 23(5): 606-615.
HU Jun, WANG Guoyin. Uncertainty measure rule sets of rough sets[J]. Pattern Recognition and Artificial Intelligence. 2010, 23(5): 606-615.

[24]MAGNANI M, MONTESI D. Uncertainty in data integration: current approaches and open problems[M]. Enschede, The Netherlands: the Centre for Telematics and Information Technology, 2007: 26-32.

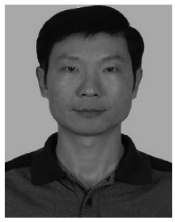
[25]JUNG J Y, CHIN C H, CARDOSO J. An entropy-based uncertainty measure of process models[J]. Information Processing Letters, 2011, 111(3): 135-141.

[26]岳昆,刘惟一,王晓玲. 一种基于不确定性因素叠加的Web服务质量度量方法[J]. 计算机研究与发展, 2009, 46(5): 841-849.
YUE Kun, LIU Weiyi, WANG Xiaoling. An approach for measuring quality of web services based on the superposition of uncertain factors[J].Journal of Computer Research and Development, 2009, 46(5): 841-849.

作者简介:

胡文彬,女,1976 年生,博士,中国计算机学会会员,主要研究方向为数据集成、社会网络、隐私保护,作为主要成员完成省级课题 1 项,参与完成市级课题 2 项。发表学术论文 10 余篇,其中被 EI 检索 3 篇。

潘祝山,男,1968 年生,副教授,主要研究方向为人工智能、确定性理论。参与省市级课题多项。



纪兆辉,男,1971 年生,副教授,中国计算机学会高级会员,主要研究方向为数据挖掘、语义 Web、多 Agent 等。发表学术论文 20 余篇,主持、参与省市级科研课题 10 余项。