不确定模式匹配研究综述

翁年凤[1,2] 刁兴春[2] 曹建军[2] 冯 径[3]
(解放军理工大学指挥自动化学院 南京 210007)[1] (总参第六十三研究所 南京 210007)[2]
(解放军理工大学气象学院 南京 211101)[3]

摘 要 模式匹配是数据集成、语义Web等研究领域的重要研究内容,需要依据一定的启发式信息发现模式元素之间的对应关系。鉴于启发式信息处理方法的不同,对模式匹配方法进行了分类,并从模式匹配结果集结方法的角度,介绍了综合模式匹配方法。不确定性是模式匹配过程固有的特性,介绍了建模模式匹配过程中不确定性的数据模型,在此基础上介绍了处理模式匹配过程中不确定性的模式匹配方法。最后对模式匹配研究进行了展望。

关键词 数据集成,模式匹配,不确定性
中图法分类号 TP311 **文献标识码** A

Survey of Uncertain Schema Matching

WENG Nian-feng[1,2] DIAO Xing-chun[2] CAO Jian-jun[2] FENG Jing[3]
(Institute of Command Automation, PLA University of Science & Technology, Nanjing 210007, China)[1]
(The 63rd Research Institute of PLA General Staff Headquarters, Nanjing 210007, China)[2]
(Institute of Meteorology, PLA University of Science & Technology, Nanjing 211101, China)[3]

Abstract Schema matching is one of the research directions of data integration and semantic Web. Its work is to discover correspondence of schema elements using heuristic information. We proposed a new classification of schema matching based on the way that heuristic information is processed. An overview of composed schema matching was given according to different aggregation methods of schema matching results. Uncertainty is the inherent character of schema matching. We surveyed data models modeling uncertainty of schema matching, and on this basis we introduced uncertain schema matching. At last we prospected the research directions.

Keywords Data integration, Schema matching, Uncertainty

模式匹配(Schema Matching)是数据集成、语义Web等领域的重要研究内容,它以两个待匹配模式作为输入,根据启发式信息发现模式元素之间的对应关系,输出模式映射(Schema mappings)[1]。数据集成系统是对传统数据库系统的扩展,通过将数据源的本地模式映射到全局模式,为用户提供一个统一的查询接口,以实现在多个系统之间共享数据[2]。Web上包含大量可通过动态查询而不是静态链接访问的结构化数据,Deep Web数据集成通过发现相关的数据源并建立查询接口之间的映射来集成隐藏于查询接口背后的数据源[3,4]。在语义Web中描述Web资源,进而使计算机可以发现并理解Web资源,实现资源的共享,以及资源的动态分配和调度,实现这些目标需要建立Web资源描述之间的映射,与模式匹配相近的问题是本体的匹配问题[5,6]。Web Services使用XML描述接口和消息,实现应用系统的解耦和动态组合,以快速响应业务需求的变化,Web Services之间的动态组合需要建立XML消息之间的映射[7,8]。

进行模式匹配需要准确判断模式及其元素的语义,而模式是对数据的抽象描述,由符号和结构组成,它们很难完全体现数据所包含的丰富语义,因此,模式匹配工作一般需要领域专家或者模式设计者来完成[9,10]。由于模式匹配需要消耗大量的人力和时间,而且容易出错,因此人们开始使用半自动的模式匹配工具来辅助完成模式匹配,用户通过干预来指导匹配过程或者验证匹配结果。随着数据集成研究的发展以及Web应用的兴起,模式匹配研究需要考虑新的应用需求,如语义Web环境下Web资源和Web服务的机器可理解、无监督地发现并组合Web服务等,靠人工或者半人工完成模式匹配已经变得不现实,更无法保证领域专家或者模式设计者参与模式匹配,因此部分研究工作开始致力于全自动的模式匹配。自动模式匹配面临诸多挑战[1]:首先,可以利用的启发式信息非常有限,这些信息通常包括模式元素名称、模式结构等;其次,启发式信息不完备,即通过这些启发式信息无法准确判断模式元素的语义,不足以确定模式元素之间的映射关

到稿日期:2011-01-19 返修日期:2011-05-07 本文受国家自然科学基金(61070174),中国博士后科学基金特别资助金(201003797),中国博士后科学基金(20090461425),江苏省博士后科研资助计划项目(0901014B)资助。
翁年凤(1983-),男,博士生,CCF学生会员,主要研究方向为数据工程,E-mail:wengnf@gmail.com;刁兴春(1964-),男,研究员,博士生导师,主要研究方向为网络及信息技术等;曹建军(1975-),男,博士后,主要研究方向为数据质量、进化计算等;冯 径(1962-),女,教授,博士生导师,主要研究方向为计算机网络、分布式计算等。

系；另外，对语义等价或者相似的判断具有一定的主观性，依赖于特定领域上下文，因此判断两个模式元素是否等价或者相似的标准很难准确描述，具有不确定性。不确定性是模式匹配过程所固有的、不可避免的。而现有的模式匹配方法大多是基于确定语义的，在动态变化、需要迅速反应的环境中，这些方法效果不佳、缺乏可扩展性[11,12]。因此，相关的研究工作试图对模式匹配过程中的不确定性建模，设计考虑不确定性的模式匹配方法，并取得了一些重要的研究成果。

1 模式匹配

1.1 启发式信息

文献[13]对模式匹配方法进行了综述，根据所依据的模式信息和层次的不同将模式匹配方法分为模式层和实例层匹配、元素级和结构级匹配、基于语言和基于约束的匹配。文献[11]将现有模式匹配方法分为4类：第一类是基于信息检索技术的匹配方法；第二类是基于机器学习算法的匹配方法；第三类是使用图理论的匹配方法；第四类是整合前三类方法，对不同方法的匹配结果进行加权平均。模式匹配可以利用的启发式信息包括模式元素名称、模式元素数据类型、模式元素的值域、模式结构、约束条件、数据实例、描述文档、领域上下文等。这些启发式信息大体可分为3类：文本信息、结构信息、约束信息。

利用文本信息的模式匹配方法，使用信息检索领域中的文本处理方法，来处理模式元素名称、数据实例、描述文档等启发式信息，以此判断模式元素的相似性。首先对文本进行规范化，然后计算规范化文本的相似度，常用的计算方法是文本中最大匹配字符串的长度与最大字符串长度的比值。为了减少文本比较的次数，Cupid[14]对模式元素进行了分类，只比较相容类别中的元素，文本相似度由模式元素的名称相似度乘以其所属类别的最大名称相似度得到。Harmony[15]在对文本进行预处理的同时对单词进行计数，出现频率越低的单词其价值越高，每个单词的权重是其出现频数的倒数；元素的相似度为规范化后的文本中匹配单词的权重之和。

利用结构信息的模式匹配方法，使用模式的结构信息来判断模式元素之间的相似性，即两个元素的部分相似度传递到了各自的邻居。SF[16]首先将使用词频的字串匹配相似度进行随机增减调整后作为初始相似度，然后使用不动点迭代计算进行精确匹配，每次迭代中相似度等于原相似度加上邻居的相似度与传递系数的乘积。Cupid[14]采用TreeMatch算法，首先对叶子节点计算初始相似度，然后后序遍历两棵模式树，计算模式元素之间的结构相似度，非叶子节点的结构相似度记为其叶子节点匹配数与叶子节点总数的比值。如果两个非叶子节点相似，则增加其叶子节点之间的相似度，否则降低叶子节点间的相似度。

利用约束信息的模式匹配方法，使用各种约束条件来判断两个模式元素之间是否相似。例如，如果两个模式元素的数据类型相同，且其取值范围也相近，则这两个模式元素可能存在相似关系。文献[17]利用模式元素值域范围的分布情况计算模式元素之间的相似度，文献[18,19]则使用完整性约束来发现模式元素之间的对应关系。Cupid[14]将树模型扩展为有根的图模型，用一条路径定义一个上下文，以便根据上下文进行匹配，为每一个参照约束引入新的节点，从而可以将参照约束引入匹配过程。Harmony[15]首先识别明显的顶层实体对应关系，然后考察模式元素的值域，以发现新的对应关系。文献[12]引入了应用语义来增强匹配过程，认为Web页面中词条的排列顺序表明了业务规则中的时间约束，其分别计算位于给定词条前面和后面的词条的相似度，然后进行加权平均，计算相似度。

1.2 综合匹配方法

单个匹配方法往往只关注某些预期要解决的问题，因此缺乏通用性。同时，模式匹配非常困难，各种方法差异较大，单个匹配器很难得到准确的匹配结果[20,21]。现有的匹配工具往往根据特定的领域背景，同时使用多种匹配方法以提高匹配准确性。综合匹配方法可分为3类：第一种是混合型匹配方法，即综合使用多种匹配标准或者模式属性，从而克服了使用单个匹配标准的片面性，最终的匹配结果是各种匹配标准所给出的相似度的加权平均，如SemInt[22]、Cupid[14]等。第二种是组合型匹配方法，即先根据不同的匹配算法（可以是混合型匹配算法）分别进行匹配，然后将由不同匹配算法所得到的匹配结果整合到一起。COMA[20]通过不同的匹配器得到一个相似度立方体，经过相似度集结得到一个相似度矩阵，再经过过滤策略选择最佳候选匹配。Harmony[15]中源模式元素和目的模式元素之间的关系用映射矩阵表示，多个表决器使用不同策略来识别匹配对，每个表决器为每个匹配对赋以置信分值，表决合并器将多个表决器给出的置信分值合并为一个置信分值，并根据置信分布偏离的程度为不同的表决器给定的置信分值赋以权重。LSD[23]基于多策略学习方法，将模式匹配看作分类问题，基本学习器从训练样本中学习，为对象属于每个分类提供信任分值，元学习器通过训练学习各个基本学习器的权重，最后的匹配结果是各个基本学习器给出的置信分值的加权平均。第三种是元模式匹配，其思想来源于元搜索[24]，即多个匹配器分别对候选模式映射进行排序，元模式匹配器则对多个排序结果进行整合，得到一个综合排序结果，最后从最终的排序结果中得到模式映射，如文献[21,25]利用Web搜索和数据库中间件领域中的排序集结算法进行模式匹配。上述3种综合匹配方法的区别在于，混合型匹配直接根据不同的启发式信息计算模式元素之间的相似度；组合型匹配则由各个匹配器分别计算相似度，然后将多个相似度集结为一个相似度；元匹配则是集结多个匹配方法输出的候选匹配的排序结果。

2 不确定数据模型

数据集成系统首先构建一个全局模式，然后建立数据源本地模式与全局模式之间的映射。用户基于全局模式提交查询，系统利用语义映射将查询转化为针对数据源本地模式的查询。通常可以使用加权二分图模型来建模模式映射：二分图中的两个子图分别代表两个模式，二分图中的节点表示模式元素，二分图中的边代表模式元素之间的映射，边的权重则表示模式元素之间映射的置信度[26]。尽管数据集成研究取得了很多成果，但建立和维护数据集成系统仍然需要大量的前期和后续工作。如何减少数据集成系统的投入，已经成为数据集成研究领域关注的热点问题之一[27]。相关研究工作致力于从少量的或者不精确的语义映射开始，然后根据需求逐步改进这些映射，即所谓量入为出（pay-as-you-go）[27-29]的

方法。这需要系统具备自引导的能力，即在没有前期准备的情况下，提供诸如搜索、查询、浏览等基础服务，然后按照量入为出的方式改进语义映射的质量，从而提高查询响应的质量。一般的模式匹配方法假设置信度高的属性映射更好，通过设置阈值过滤置信度较低的映射。然而，设置阈值的方法只适用于优劣分明的情形，而且阈值如何选取也是一个难题。现实中的模式很难和数据语义保持高度一致。现有的数据集成方法都是确定性的，无法有效管理语义差异，从而会导致查询响应质量降低，因此模式匹配过程需要考虑一定的不确定性[11]，而面向确定语义的模式匹配方法总是试图去除这种不确定性，或者使用人工干预的方式排除某些候选匹配。事实上，那些被排除的候选匹配在某些情况下具有更高的价值。

文献[30]提出了不确定数据集成的概念，其中不确定性的来源包括数据源本地模式与全局模式之间的语义映射、关键字查询转换为结构化查询、从非结构化数据源中抽取结构化数据。考虑到客观存在的不确定性，引入了概率模式映射，其具有基于表和基于元组两种语义：基于表的语义表明两个模式之间只有一个映射是正确的，且适用于所有的数据；而基于元组的语义表明两个模式之间存在多个正确映射，且每一个映射只适用于部分元组。给定两个模式 $\overline{S}$ 和 $\overline{T}$，概率映射 $pM=(S,T,m)$，其中 $S\in\overline{S}$，$T\in\overline{T}$ 为两个关系，$m=\{(m_1,\Pr(m_1)),\cdots,(m_l,\Pr(m_l))\}$，并且对于 $i\in[1,l]$，m_i 是一个 S 与 T 之间的映射，对于任意 $i,j\in[1,l]$，$i\neq j\Rightarrow m_i\neq m_j$；$\Pr(m_i)\in[0,1]$，且 $\sum_{i=1}^{l}\Pr(m_i)=1$。概率模式映射 $\overline{pM}$ 是模式 $\overline{S}$ 和 $\overline{T}$ 之间概率映射的集合，且模式中的每个关系只在一个概率映射中出现。概率模式映射描述了各种可能模式映射的概率分布，扩展了模式匹配。

文献[21]使用序关系来建模候选匹配之间的关系，而不是先将候选匹配映射到一个实数值，然后从候选匹配中选取最佳匹配。给定源模式 S 和目的模式 S'，令 $\Sigma=2^{S\times S'}$ 为所有可能的候选匹配，匹配器 A 以两个模式 S、T 以及领域约束 Γ 作为输入，输出模式映射 $\Sigma_\Gamma=\{\sigma\in\Sigma\mid\Gamma(\sigma)=1\}$ 和序关系 $\leq_A$。对于映射 $\sigma_1,\sigma_2\in\Sigma_\Gamma$，$\sigma_1\leq_A\sigma_2$ 表示匹配器 A 判断 σ_2 至少和 σ_1 一样好。

文献[26]使用带标记边的二分图建模模式匹配，边上的标记表示某种程度的确定性，并利用模式元素映射的确定性计算整个模式映射的确定性。要保证整个模式匹配的相似度最大即为最佳匹配，相似度集结算子必须具备单调性，即确定性等级高的映射优于确定性等级低的映射。使用准确性来评价模式匹配结果，观察结果显示准确性在[0,1]之间离散取值。在每个准确性水平上，确定性呈钟形分布，且确定性均值随着准确性的降低而降低。但由于每个准确性水平中方差的不同，不同准确性水平的确定性分布彼此重叠，这一点违背了单调性原则。为此提出了统计单调性，即确定性随着准确性的增加而增加。统计单调性可以解释模式匹配中的特定现象，也可以作为模式匹配算法的指导原则。

3 不确定模式匹配

3.1 基于模糊集的方法

文献[27]试图构建一个不需要人工干预的自动数据集成系统，提供“尽最大努力”的查询应答，并允许以量入为出的方式改进系统。为了支持“尽最大努力”查询应答和持续改进映射，引入了一种概率数据模型。一个概率模式映射包含一个映射集合，集合中的每个映射都附有一个概率。首先从所有数据源属性中构造一个加权图，节点表示属性，标记边表示属性之间的相似度。只保留相似度超过某个设定阈值的边，并且允许阈值存在一定的误差，这样会出现两种边：一种是相似度大于阈值加上误差的确定边，另一种是相似度介于阈值减去误差和阈值加上误差之间的不确定边。在源属性和全局属性之间计算加权匹配对，在得到的加权匹配对基础上可能得到多个概率映射，然后为每个概率映射以熵最大化的原则分配概率。

文献[11]使用模糊框架建模匹配过程中的不确定性，每个属性匹配对被赋予一个信任度。根据信任度的计算方法，信任度高的映射不一定最准确，通过单调性来保证信任度高的匹配器准确性也高。使用不准确性建立置信度与单调性之间的关系，将映射收益和成本定义为从一个映射结果转到另一个映射结果置信度差值的函数。严格单调性指从不准确性等级高的映射转到不准确性等级低的映射的收益大于代价，证明了使用加权平均算子集结模式映射置信度，不准确性等级越低置信度越高。而除非两个模式的属性之间高度相关，否则不能使用三角准则计算置信度。实际应用中，可以使用两种弱单调性：配对单调性指从一个映射转到准确映射的收益大于代价；统计单调性指低不准确性等级映射的置信度的期望较大。实验结果表明，满足统计单调性的情况下，不准确性是置信度的主要因素，而在满足配对单调性的情况下，准确匹配总是排序靠前的。

绝大部分模式匹配方法都是基于相似度的，即根据不同的相似度指标发现对应关系。每个相似度指标由基本匹配器计算，而且结合不同基本匹配器的结果可以得到比使用单个基本匹配器更佳的结果。通常使用加权平均作为集结算子，权重由人为指定或者通过机器学习得到。人为指定权重非常困难，而通过机器学习则需要丰富的训练数据集。FOAM[31]使用 OWA 算子来集结由不同相似度指标得到的相似度值，OWA 算子中权重不是赋给某个相似度指标，而是赋给特定的排序位置。

3.2 基于概率统计的方法

在模式匹配模型中，模式映射是属性映射的集合。模式匹配过程输出一个相似度矩阵，需要找到最满足该相似度矩阵的模式映射。不完全和不确定信息会引入噪声，因此从相似度矩阵中将产生两个概率分布：一个是不正确的属性映射，另一个是正确的属性映射。文献[32]中的实验结果表明这两个概率分布都可以用 β 分布拟合，通过估计 β 分布的两个参数，可以得到属性映射的概率分布。给定相似度矩阵，采用朴素贝叶斯方法将属性映射分类到两个分布中。给定由两个属性构成的属性对，各个匹配器给出的相似度构成特征向量。在由特征向量构成的特征空间中，目标函数就是将观察到的数据采样映射到一个二值函数。根据贝叶斯公式，目标函数可分为两个部分：一部分可从训练数据中取值的频率得到；另一部分为给定映射与映射分类的相似度，可以从两个估计分布中得到。

文献[33,34]假设同一领域内存在潜在的模式模型，以某种概率从有限的属性词汇表中生成模式实例。给定输入模式作为观测到的模式实例，构造一个与观测到的模式实例统计

一致的模型，利用该模型发现同义词，进而可以发现模式元素的对应关系。一个模式模型由一个4元组构成，包括词汇表、概念划分、概念概率函数和属性概率函数。词汇表是模式属性的集合，概念划分是对词汇表的一个划分，概念概率函数确定在生成实例时包含某个概念的概率，属性概率函数确定当概念被选择时该概念下的属性被选择的概率。MGS框架首先指定一个参数化的估计模型结构，生成所有可能生成观测到的模式实例的模型，最后选择与模式实例统计一致的模型。算法包括初始假设生成和迭代两个阶段：初始假设生成阶段建模目标问题，选择待考虑的属性，并构造假设空间，然后生成可能的假设模型；迭代阶段由参数估计、假设选择交替进行，使用 χ^2 假设检验选择与模式实例一致的候选模型。

3.3 基于 Top-K 的方法

文献[25]使用 Top-K 方法作为管理不确定性的工具。模式匹配可分为两个步骤：第一步使用名称匹配、值域匹配、结构匹配等方法计算元素之间的相似度，第二步选择使目标函数最优的映射作为最佳映射。实际准确映射比匹配器识别的最佳映射包含更少的映射，由于不确定性的存在，无法通过设置阈值过滤来得到准确映射。实验结果表明，随着 K 值的增加，保持不变的准确映射占所有准确映射的比例保持稳定，而保持不变的不准确映射占所有不准确映射的比例变化明显，即准确的属性映射在 Top-K 中总是出现的，因此利用 Top-K 映射可以帮助发现准确映射。

假设存在 m 个匹配器，源模式 S 和目的模式 S' 中分别包含 n 个和 n' 个属性，模式匹配过程可分为3个步骤完成：第一步，匹配器 A 根据领域约束为每一对属性映射赋一个相似度，得到一个 $n\times n'$ 的相似度矩阵 $M^{(A)}$，其中 $M^{(A)}_{i,j}$ 表示源模式中的第 i 个属性与目的模式中的第 j 个属性的相似度；第二步，使用局部集结算子 $f^{(A)}(\sigma,M^{(A)})=f^{(A)}(M^{(A)}_{1,\sigma(1)},\cdots,M^{(A)}_{n,\sigma(n)})$ 集结相似度矩阵中的相似度，从而对不同的模式映射 σ 进行量化评估；第三步，使用全局集结算子 $\langle\vec{f},F\rangle(\sigma)=F(f^{(1)}(\sigma,M^{(1)}),\cdots,f^{(m)}(\sigma,M^{(m)}))$ 集结各个局部集结算子的评估结果，得到最终的评估结果。给定源模式 S、目的模式 S'、约束 Γ 和 $K\geqslant 1$，Top-K 模式匹配方法就是要找到 K 个模式映射 $\{\sigma^1,\cdots,\sigma^K\}\subseteq\Sigma_\Gamma$，使得 $\sigma^i=\arg\max_{\sigma}\{\langle\vec{f},F\rangle(\sigma)\mid\sigma\in\Sigma_\Gamma\setminus\{\sigma^1,\cdots,\sigma^{i-1}\}\}$。在具体模式匹配算法中，第一步计算相似度，生成加权二分图；第二步，对于目的模式中的每个节点，对连接到该节点的所有边按权重降序排列，计算每个节点对应的边列表中最大权重边与下一条边的权重差，并将权重差值插入到最小堆中，然后进行迭代。在第 i 次迭代中将第 $i-1$ 个映射中权重差最小的边替换为该边所在列表中的下一条边，并将下一条边的权重差插入到最小堆中，从而得到第 i 个映射。假设匹配器是单调的，则给定 Top-K 映射，计算 K 个映射中每条边出现的次数。如果出现次数小于设定阈值，则将该边的权重置0，最终得到 Top-1 最佳映射。

文献[21]在 Web 搜索和数据库中间件领域的 FA[35] 和 TA[36] 排序集结算法的基础上派生出 MD、MDB 和 Cross-Threshold 算法来解决不同匹配器模式匹配结果的集结问题。其中 TA 首先出现在数据库中间件领域，其时间复杂度与模式规模呈指数关系，MD 则具有多项式时间复杂度。但给定一个具体问题，无法确定 TA 与 MDB 是否优于对方，MDB 派生自 MD 和 TA，适用于 MD 不适用的场合。CrossThreshold 算法是 TA 和 MDB 的混合，且优于 TA 和 MDB。

(1)FA 算法

(i)按序、并发地访问 m 个经过排序的列表 L_i，直到构造出一个包含 K 个对象的集合 H，使得 H 中的每个对象在每个列表中都已经被访问到；

(ii)分别从各个列表 L_i 中访问 H 中每个对象 R 的属性；

(iii)根据对象的每个属性值计算该对象的分值，然后按照分值高低输出对象。

(2)TA 算法

(i)按序、并发访问 m 个经过排序的列表 L_i，当在某个列表中访问到某个对象 R 时，从其他列表中访问该对象的属性值，然后计算该对象的得分。如果该对象的得分是目前所看到的位于前 K 的，则记录下该对象及其得分；

(ii)对于每个列表，假设 x_i 为该列表中最后被访问的对象，定义阈值 t 等于该对象的得分。如果已经有 K 个对象的得分大于或者等于 t，则算法停止；

(iii)输出得到的 K 个对象。

TA 算法要求全局集结算子 F 是单调的，即给定一个相似度矩阵，局部集结得分高的映射，其全局集结得分也高。TA 是实例最优的，即给定任意数据集和排序集结算法，TA 的时间复杂度与该算法处于同一量级。但 TA 是一个通用的算法，可以利用模式匹配问题特有的属性来降低复杂度。

(3)MD 算法

所有匹配器使用相同的局部集结算子，且局部集结算子与全局集结算子在给定相似度矩阵上可互换，即 $\langle f,F\rangle(\sigma)=\langle F,f\rangle(\sigma)$。首先集结相似度矩阵，然后对相似度进行集结、排序。

(4)MDB 算法

存在集结函数 $\langle h,H\rangle$ 在相似度矩阵上可互换，且在相似度矩阵上优于 $\langle f,F\rangle$，使用 $\langle h,H\rangle$ 和 MD 算法取 Top-K，然后按序取 Top-K 中的映射，分别取每个匹配器的给分再集结，即 $\langle f,F\rangle(\sigma)$，并将结果按序排列。此时检查结果列表中是否有 K 个大于当前映射的 MD 算法结果。如果已有，则算法停止。

(5)CrossTheshold 算法

采用交叉阈值的方法将 MDB 和 TA 算法结合，即 M 个匹配器和 $\langle h,H\rangle$ 分别并行地排序，MDB 和 TA 两者之一停止则算法停止。

结束语 对模式匹配问题的研究由来已久，也取得了一系列研究成果。但由于模式匹配问题的复杂性，现有的研究工作并没有达到令人满意的效果，还有较大的提升空间。

(1)没有哪一种匹配器能适应所有应用的需求，各个匹配器所采用的领域约束可能是彼此相关的，或者遗漏掉了某些重要的领域约束信息，导致匹配结果存在偏见。因此，如何根据实际应用的需要选择合适的匹配器，值得研究。另外，可以考虑发现更多的启发式信息，对匹配器进行分组，从而构建不同的匹配策略，以便进一步提高模式匹配的准确性。

(2)评价一个模式匹配方法或者工具的优劣，需要客观公正的评价标准。而基于学习的模式匹配需要知道训练样本中哪些是正确的匹配、哪些是错误的匹配，对匹配结果正确与否的评判标准更加重要。目前常用的模式匹配结果评价指标有 precision、recall、F-Measure、F-Measure(σ) 以及 overall 等[37]。

在考虑模式匹配过程中不确定性的情况下，对匹配结果的评价指标不是单一的，根据上下文的不同，需要采用不同的指标或者指标的综合，因此考虑不确定性的模式匹配方法的结果评价问题还有待进一步研究。

(3)不确定性是模式匹配过程所固有的特性，对模式匹配过程中不确定性的建模将直接影响到模式匹配的效果。模式匹配问题实际上是多属性决策问题，即在事先无法确定何种启发式信息占主导地位的情况下，根据不同的启发式信息对候选匹配选择问题进行决策。多属性决策领域对决策过程中不确定性的研究已经取得了丰富的研究成果[38,39]，而在模式匹配研究领域相关研究成果还不多见。

参 考 文 献

[1] Doan A H, Halevy A Y. Semantic Integration Research in the Database Community: A Brief Survey [J]. AI Magazine, 2005, 26(1): 83-94

[2] Almarimi A, Pokorny J. Schema Management for Data Integration: A Short Survey [J]. Acta Polytechnica, 2005, 45(1)

[3] He Bin, Zhang Zhen, Chang K C-C. Knocking the Door to the Deep Web: Integration Web Query Interfaces [C]// ACM SIGMOD. 2004

[4] 刘伟，孟小峰，孟卫一. Deep Web 数据集成研究综述[J]. 计算机学报，2007，30(9)

[5] Shvaiko P, Euzenat J. Ten Challenges for Ontology Matching [C]// Proceedings of the 7th International Conference on Ontologies, Data Bases, and Applications of Semantics (ODBASE). 2008

[6] Choi N, Song I-Y, Han H. A Survey on Ontology Mapping [J]. SIGMOD Record, 2006, 35(3)

[7] Srivastava B, Koehler J. Web Service Composition-Current Solutions and Open Problems [C]// Proceedings of ICAPS. 2003

[8] Aulbach S, Grust T, Jacobs D, et al. Multi-tenant Databases for Software as a Service: Schema-Mapping Techniques [C]// ACM SIGMOD. 2008

[9] Halevy A Y. Structures, Semantics and Statistics [C]// Proceedings of the 30th VLDB Conference. 2004

[10] Halevy A Y. Why Your Data Don't Mix: Semantic Heterogeneity [J]. Queue, 2005, 3(8): 50-58

[11] Gal A, Anaby-Tavor A, Trombetta A, et al. A framework for modeling and evaluating automatic semantic reconciliation [J]. The VLDB Journal, 2003

[12] Gal A, Modica G, Jamil H, et al. Automatic Ontology Matching Using Application Semantics [J]. AI Magazine, 2005, 26(1): 21-31

[13] Rahm E, Bernstein P A. A survey of approaches to automatic schema matching [J]. The VLDB, 2001, 10: 334-350

[14] Madhavan J, Bernstein P A, Rahm E. Generic Schema Matching with Cupid [C]// Proceedings of the 27th VLDB Conference. 2001

[15] Mork P, Seligman L, Rosenthal A, et al. The Harmony Integration Workbench [J]. Journal on Data Semantics XI, Lecture Notes in Computer Science, 2008(5383): 65-93

[16] Melnik S, Garcia-Molina H, Rahm E. Similarity Flooding: A Versatile Graph Matching Algorithm and Its Application to Schema Matching [C]// Proceedings of Data Engineering. 2002

[17] 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)

[18] 李国徽，杜小坤，胡方晓，等. 基于函数依赖的结构匹配方法[J]. 软件学报，2009，20(10)

[19] Wang D Z, Dong L, Sarma A D, et al. Functional Dependency Generation and Applications in Pay-as-you-go Data Integration Systems [C]// Proceedings of SIGMOD WebDB. 2009

[20] Do Hong-hai, Rahm E. COMA-A system for flexible combination of schema matching approaches [C]// Proceedings of the 28th VLDB Conference. 2002

[21] Domshlak C, Gal A, Roitman H. Rank Aggregation for Automatic Schema Matching [J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(4): 538-553

[22] Li W-S, Clifton C. SEMINT: A tool for identifying attribute correspondences in heterogeneous database using neural networks [J]. Data & Knowledge Engineering, 2000, 33: 49-84

[23] Doan A H, Domingos P, Halevy A. Reconciling Schemas of Disparate Data Sources: A Machine-Learning Approach [C]// ACM SIGMOD. 2001

[24] 阳小华，刘振宇，谭敏生，等. 元搜索引擎查询结果的合成方法[J]. 计算机科学，2002，29(8)

[25] Gal A. Managing Uncertainty in Schema Matching with Top-K Schema Mappings [J]. Journal on Data Semantics VI, 2006: 90-114

[26] Gal A. Why is Schema Matching Tough and What Can We Do About It? [C]// ACM SIGMOD. 2007

[27] Sarma A D, Dong Xin, Halevy A. Bootstrapping Pay-As-You-Go Data Integration Systems [C]// ACM SIGMOD. 2008

[28] Franklin M, Halevy A, Maier D. From Databases to Dataspaces: A New Abstraction for Information Management [C]// ACM SIGMOD. 2005

[29] Sarma A D, Dong Xin, Halevy A Y. Data Modeling in Dataspace Support Platforms [M]. Borgida A T, eds. Mylopoulos Festschrift. LNCS 5600. 2009: 122-138

[30] Dong Xin, Halevy A Y, Yu Cong. Data Integration with Uncertainty[C]// Proceedings of VLDB Conference. 2007

[31] Ji Qiu, Haase P, Qi Gui-lin. Combination of Similarity Measures in Ontology Matching Using OWA Operator [C]// Proceedings of the 12th International Conference on Information Processing and Management of Uncertatinty in Knowledge-Base Systems. 2008: 243-250

[32] Marie A, Gal A. Managing Uncertainty in Schema Matcher Ensembles [J]. Scalable Uncertainty Management, Lecture Notes in Computer Science, 2007, 4772: 60-73

[33] He Bin, Chang K C-C. Statistical Schema Integration Across the Deep Web [R]. UIUCDCS-R-2002-2304. uilu-eng-2002-1747, 2002

[34] He Bin, Chang K C-C. Statistical Schema Matching Across Web Query Interfaces [C]// ACM SIGMOD. 2003

[35] Fagin R. Combining fuzzy information from multiple systems [J]. J. Comput. System Sci, 1999, 58: 83-99

[36] Fagin R, Lotem A, Naor M. Optimal Aggregation Algorithms for Middleware [C]// PODS. 2001

[37] Do Hong-hai, Melnik S, Rahm E. Comparison of Schema Matching Evaluations [M]. Web, Web-Services, and Database Systems, Erfurt Germany, 2002

[38] 徐泽水. 不确定性多属性决策理论[M]. 北京：清华大学出版社，2004

[39] 徐玖平，吴巍. 多属性决策的理论与方法[M]. 北京：清华大学出版社，2006

[40] Magnani M, Montesi D. Uncertainty in data integration: current approaches and open problem [C]// Proceedings of the MUD Workshop of VLDB Conference. 2007