基于数据实例分布特征的自动模式匹配方法

李 由[1] 刘东波[2,3] 张维明[1]

(国防科技大学信息系统与管理学院 长沙 410073)[1] (华中科技大学计算机科学与技术学院 武汉 430074)[2]

(中国电子设备系统工程研究所 北京 100039)[3]

摘 要 模式匹配已经成为信息集成、数据仓库、电子商务等很多应用领域中的基本问题。现有的模式匹配工作仍是以人工方式为主,这种方法费时、易出错,代价很高。本文提出了一种基于神经网络的模式匹配方法 SMDD,通过分析模式元素所包含数据实例的分布规律,自动完成模式匹配。SMDD 既可独立使用,也可与其他模式匹配方法结合使用,从数据内容的角度提高匹配质量。

关键词 映射,匹配,神经网络,模式

A Method for Automatic Schema Matching Using Characteristic of Data Distribution

LI You[1] LIU Dong-Bo[2,3] ZHANG Wei-Ming[1]

(Department of Management Science & Engineering, National University of Defense Technology, Changsha 410073)[1]

(College of Computer Science & Technology, Huazhong University of Science & Technology, Wuhan 430074)[2]

(Institute of China Electronic System Engineering, Beijing 100039)[3]

Abstract Schema mapping has been a basic problem in many application domains, such as data integration, data warehouse, E-business. This paper introduces a schema matching method called SMDD which is based on neural network. By analyzing the characteristics of data distribution, SMDD automatically fulfills the task of schema mapping. It can be used independently or combined with other schema mapping methods to improve the accuracy of schema matching from the view of data contents.

Keywords Mapping, Match, Neural network, Schema

1 引言

由于不同的背景环境而造成的数据源(关系数据库、面向对象数据库、XML、HTML 等)异构问题已经成为信息集成的主要障碍。实现异构数据源无缝集成的首要问题是定义异构数据源模式元素之间的语义映射关系,即所谓模式匹配。模式匹配作为模式的基本操作已成为信息集成、数据仓库、电子商务等许多应用领域的基本问题。例如,在数据仓库中,用模式匹配技术发现数据源模式与数据仓库模式之间的映射关系,以完成对数据源数据的抽取和转换;在电子商务应用中,需要通过模式匹配实现不同信息系统之间的数据转换。

现有的模式匹配工作仍以人工(领域专家、系统开发人员或 DBA)定义方式为主,费时、费力且容易出错。随着数据源数量的增多和规模的增大,模式匹配已经成为影响上述应用的主要瓶颈。例如,文[1]中引用了美国通用电话与电子设备公司(GTE)的数据库集成项目,该公司需要集成的 40 个数据库中共包含大约 27,000 个元素(即关系表中的属性),该项目的规划者估计,假如没有数据库开发人员的参与,定义元素之间的语义映射关系需要 12 个(人·年)以上。

近年来,模式匹配作为数据管理应用中的基础性问题受到了普遍关注,出现了多种针对特定应用领域的模式匹配方法。Cupid[2] 结合了名称匹配和结构匹配,是一种典型的混合匹配方法。LSD[3] 及其扩展 GLUE 利用模式信息和数据实例信息,通过机器学习机制进行模式匹配,其计算结果的准确性很大程度上依赖于训练集的选取。SF[4] 提出了一种基于图形的通用模式匹配方法,将模式信息转化为置标有向图,并通过定点计算匹配图中的相关节点。COMA[7] 提供了一种基于组合方法的模式匹配机制,通过匹配器组合方法完成异构数据源的模式匹配,等等。

由于数据源模式为模式匹配提供了大量的结构和语义信息,现有的大部分匹配方法主要利用了数据源的模式信息,对数据内容的利用相对有限。本文通过分析模式元素所包含数据实例的分布规律,利用神经网络的模式识别功能,提出了一种基于数据实例分布特征的模式匹配方法 SMDD(Schema Mapping method based on Data Distribution)。该方法既可独立使用,也可作为其它模式匹配方法的补充,有助于从数据内容的角度优化匹配结果,提高匹配的质量。

2 SMDD 模式匹配方法

现有的模式匹配方法大多采用了硬编码的方式,即使用有限的匹配准则并有固定的计算顺序。在模式匹配应用中,由于数据源的多样性和模式元素之间语义关联程度的模糊性,这种硬编码的方式很难满足实际应用的需求。近年来得到广泛应用的神经网络具有良好的学习能力及由此而来的泛化能力,为模式匹配问题提供了一个较好的解决途径。

SMDD 方法是基于假设"对来自不同数据源的模式元素,如果它们所包含的数据实例分布特征相似,则它们也可能相似"。SMDD 通过分析模式元素所包含数据的分布特征,利用

神经网络模式识别功能找出具有相似数据分布规律的元素集合，计算模式元素之间的相似度，最终向用户推荐候选映射。

图 1 表示了 SMDD 方法的模式匹配过程，主要包括提取数据分布特征向量、分布特征向量聚类、训练神经网络、计算元素相似度和向用户推荐候选映射集合五个阶段。

步骤一：提取数据分布特征向量。SMDD 利用特征提取器提取数据源 S_1 中模式元素所包含的部分数据实例，分析其分布规律，生成数据分布特征向量。

步骤二：分布特征向量聚类。SMDD 利用聚类算法将输入的数据分布特征向量动态分为 M 类，并计算聚类中心。分类数 M 由凝聚层次聚类算法根据相似度域值动态确定，不要求用户预先设定。

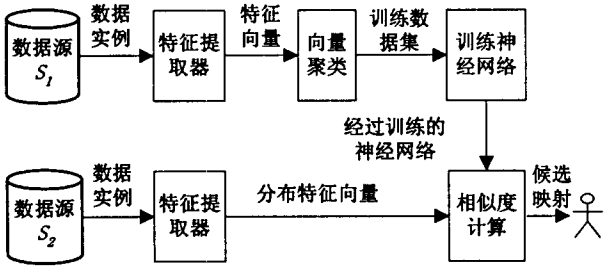


图 1　SMDD 模式匹配过程

步骤三：训练神经网络。SMDD 生成三层前馈神经网络，将聚类中心作为训练样本，利用 BP 学习算法反复迭代计算，调整各突触权重以适应输入模式的激励，最终识别类 C_i。经过训练的神经网络中每个输出节点代表一个类。

步骤四：计算元素相似度。利用特征提取器提取数据源 S_2 中的模式元素的数据分布特征向量，将其输入经过训练的神经网络，计算输入向量与数据源 S_1 中各分类的相似度。

步骤五：选择相似度较高的元素对作为候选映射，向用户推荐。

2.1　提取数据分布特征向量

对数据源 S_1（或 S_2）中的每个模式元素 e_i，SMDD 利用特征抽取器以随机抽样的方式抽取 n 个数据实例 $R_i=\{r_{1i}, r_{2i}, \cdots, r_{ni}\}$，同时采用等距离划分法对 R_i 进行离散化，以生成数据分布特征向量。具体地，对于数值型元素，设 $r_{\max}=\max(R_i)$，$r_{\min}=\min(R_i)$，SMDD 首先将区间 $[r_{\min}, r_{\max}]$ 等分为 N 个子区间 Δ_k，$k=1,\cdots,N$，并计算落在每个小区间 Δ_k 内的数据频率 x_{ki}，生成数据分布特征向量，记为 $X_i=[x_{1i}, x_{2i}, \cdots, x_{Ni}]^T$（数据源 S_1），或 $X'_i=[x'_{1i}, x'_{2i}, \cdots, x'_{Ni}]^T$（数据源 S_2）。对于离散型元素，计算数据实例 r_{ki} 在 R_i 中出现的频率 x_{ki}，选择频率较大的前 N 个 x_{ki}，按照从大到小顺序依次排列，生成数据分布特征向量 X_i 或 X'_i。

2.2　分布特征向量聚类

SMDD 利用凝聚层次聚类法对数据源 S_1 的特征向量聚类，动态生成 M 个分类 C_i，计算所有聚类中心 c_i，$i=1,2,\cdots,M$。符号 δ 表示特征向量 X_i 与 X_j 之间的欧氏距离：

$$\delta(X_i,X_j)=\sum_{k=1}^{N}((x_{ki}-x_{kj})^2)^{1/2}$$

算法 3.1　分布特征向量聚类算法。

Step1　每个输入向量自成一类，即

$C_j=X_j, \forall j\in I, I=\{j\mid j=1,2,\cdots,N_0\}$

其中，C_j 表示各分类，N_0 为输入的样本数量。

Step2　在集合 $\{C_j\mid j\in I\}$ 中找到一对满足 $\Delta(C_i,C_k)=\min_{\forall j,l\in I}\Delta(C_j,C_l)$ 的聚类集合 C_i, C_k

其中，$\Delta(C_j,C_l)=\min_{\substack{X_i\in C_j\\ X_j\in C_l}}\delta(X_i,X_j)$。

Step3　将 C_i 并入 C_k，删除类 C_i。

Step4　将 i 从指标集 I 中除掉，即 $I=I\backslash\{i\}$。当 $|I|=2$ 时，终止计算，否则转步骤 2。

Step5　用相似度域值 δ_0 截取生成的嵌套层次图，生成类集 $\{C_i\mid i=1,2,\cdots,M\}$，采用平均值法计算聚类中心 $\{c_i=1,2,\cdots,M\}$，$i=1,2,\cdots,M$。

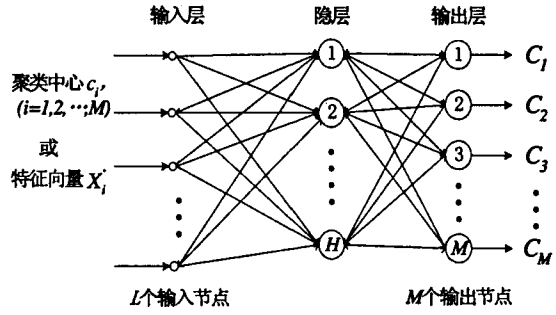


图 2　SMDD 中的三层前馈神经网络结构

2.3　训练神经网络

在训练神经网络阶段，系统生成三层前馈神经网络 NN，将聚类中心 $\{c_i\mid i=1,2,\cdots,M\}$ 作为训练样本，利用 BP 学习算法反复迭代计算，不断调整突触权重和域值，以适应输入模式的激励，最终识别类 C_i。经过训练的神经网络中每个输出节点代表一个类。

例如，对于包含三个输入节点和四个输出节点的三层前馈神经网络，假设误差域值 $\xi=0.005$，对于类 C_1 的聚类中心输入向量 $(0.2\quad 0.4\quad 0.4)^T$，神经网络的理想输出目标结果是 $(1\quad 0\quad 0\quad 0)^T$。如果实际计算的输出为 $(0.95\quad 0.05\quad 00.1)^T$，则神经网络继续迭代计算，不断调整突触权重和域值，直到误差能量的平均值 E_{AV} 小于误差域值 ξ 为止。

算法 3.2　SMDD 中的 BP 学习算法

Step1　构建包含 L 个输入节点、H 个隐藏节点以及 M 个输出节点的三层前馈神经网络 NN，M 为聚类算法生成的分类数。初始化权值矩阵 $W_1=[w_{ij}]_{H\times L}=0$；$W_2=[w_{ij}]_{M\times H}=0$。其中，$W_1$ 为输入层与隐层之间的权值矩阵，W_2 为隐层与输出层之间的权值矩阵。

Step2　前向计算。设输入向量为 Y_0，$Y_0=c_i$，$i=,1,2,\cdots,M$。net_1 和 net_2 分别为隐层和输出层的输入向量，Y_1 和 Y_2 分别为隐层和输出层的输出向量。f_1 和 f_2 分别为隐层和输出层各神经元的激活函数。图 3 为 SMDD 采用的激活函数曲线。

$net_1=W_1Y_0$，$Y_1=f_1(net_1)$

$net_2=W_2Y_1$，$Y_2=f_2(net_2)$

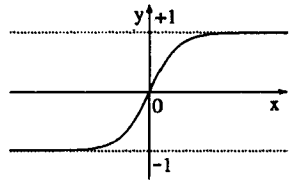


图 3　$y=\text{tansig}(x)$

Step3　反向计算局部梯度 δ。d 为期望响应，δ_2、δ_1 为局域梯度。

$$\delta_2=(d-Y)*f'_2(net_2),\delta_1=W_2^T\delta_2^* f'_1(net_1)$$

Step4 计算误差梯度矩阵。

$$\frac{\partial E}{\partial W_1}=-\delta_1 Y_0^T,\frac{\partial E}{\partial W_2}=-\delta_2 Y_1^T$$

Step5 修正权值矩阵。η_i 为用户设定的后向传播算法的学习率参数，t 为时间变量，α 为动量常数，通过增添动量项可以使突触权值的变化轨迹相对光滑。

$$\Delta W_i(t+1)=-\eta_i\frac{\partial E}{\partial W_i}+\alpha_i\Delta W_i(t)(i=1,2)$$

2.4 计算元素相似度

对于数据源 S_2，SMDD 利用特征向量抽取器提取模式元素 a_i 所包含的部分数据实例，生成数据分布特征向量 X'_i、神经网络计算输入的特征向量 X'_i 与各聚类中心的相似度。

SMDD 根据计算结果向用户推荐候选映射。SMDD 支持以下两种选择策略：

最大值策略：对每次输入特征向量的相似度计算结果，选取拥有最大值的元素对作为候选映射。

域值选择策略：对每次输入特征向量的相似度计算结果，选择大于用户设定的域值参数 θ 的元素对作为候选映射。

2.5 匹配质量度量

SMDD 采用 F 方法[8]度量匹配质量，该方法根据用户对匹配结果的反馈信息对匹配质量做出评价。

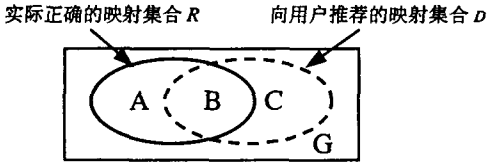


图 4 推荐映射与实际正确的映射比较关系

图 4 表明了推导出的映射集合 D 与实际正确的映射集合 R 的比较关系。F 方法综合考虑了查全率和查准率两种因素，在度量匹配质量的过程中，两者的相对重要性通过参数 α 来刻画。例如，当

$$\text{F-Measure}(\alpha)=\frac{|B|}{(1-\alpha)*|A|+|B|+\alpha*|C|}$$

$(0\leqslant\alpha\leqslant1)$ $\alpha=0.5$ 时，表明在度量匹配质量时查全率与查准率是同等重要的。

3 实验

我们将 SMDD 算法应用于数据集成中的异构数据源模式匹配任务，以检验算法的匹配质量。

某校对现有的两个异构人员关系数据库进行数据集成，分别记为数据源 S_1 和数据源 S_2。SMDD 利用特征提取器分析数据源 S_1 中部分元素 age、major、score、Student-score、sex（关系表中的属性）所包含的数据实例，生成数据分布特征向量。实验中对每个属性随机抽取 100 个数据项，将其分为 8 ($N=8$)个子区间，统计数据分布情况直方图，如图 5 所示。

层次聚类算法将其分为四类（设 $\delta_0=0.3$）：$C_1=\{\text{age}\}$，$C_2=\{\text{major}\}$，$C_3=\{\text{score},\text{Student-score}\}$，$C_4=\{\text{sex}\}$。由于 $\delta(\text{student-score},\text{score})=0.113<\delta_0$，归为一类。

利用 Matlab 提供的神经网络工具箱，构建如图 2 所示的三层 BP 神经网络。本例中设输入节点 $L=8$，隐藏节点 $H=12$，输出节点 $M=4$，学习率参数 $\eta=0.9$，动量常数 $\alpha=0.05$，误差域值为 0.005，网络训练误差情况如图 6 所示。图 7 为 SMDD 抽取的数据源 S_2 中的部分属性 S_score、Faculty_name、S_sex 的数据分布特征直方图，经过神经网络计算得出结果量 Y_1、Y_2、Y_3。根据域值选择策略（$\theta=0.8$），可得出 SMDD 匹配结果，如表 1 所示。黑体部分表示实际正确的匹配。

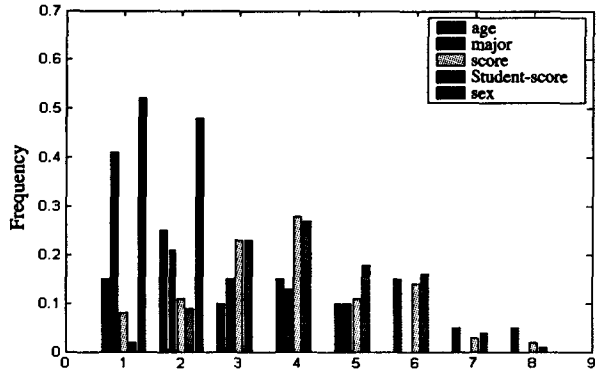


图 5 数据源 S_1 的部分属性数据实例分布直方图

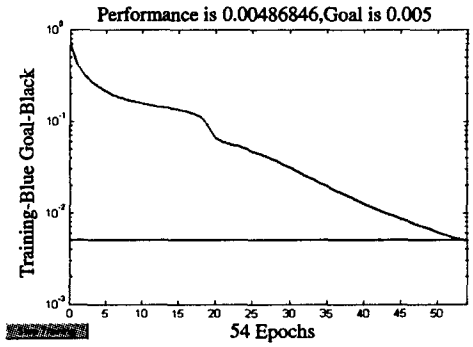


图 6 网络训练误差

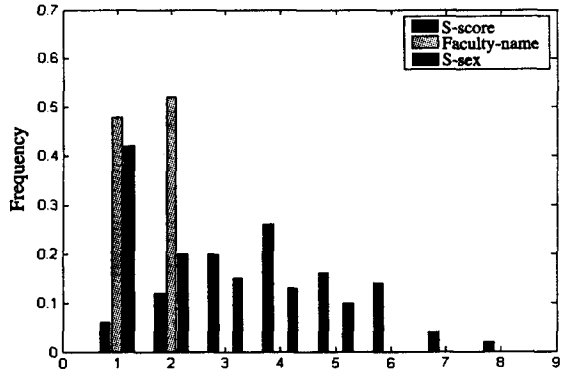


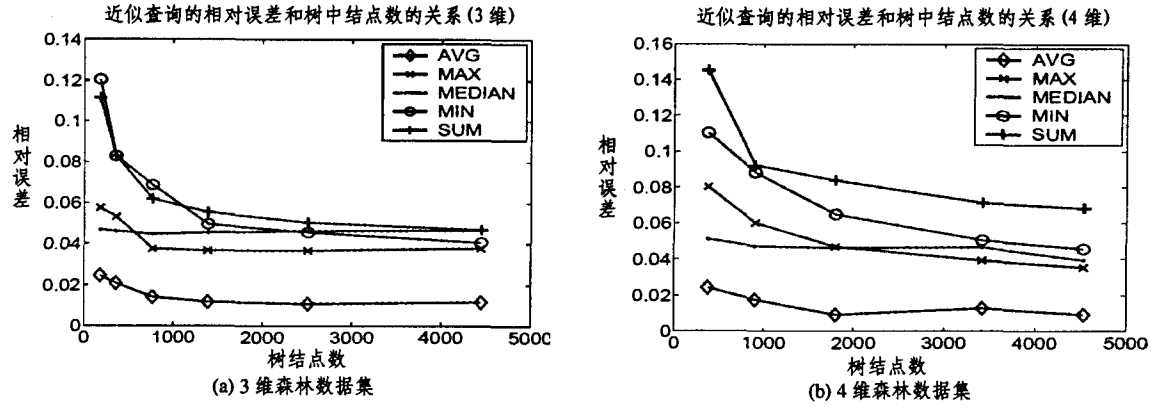
图 7 数据源 S_2 的部分属性数据实例分布直方图

$$Y_1=\begin{bmatrix}0.4260\\-0.0960\\\underline{0.9199}\\-0.2786\end{bmatrix}\quad Y_2=\begin{bmatrix}0.1087\\\underline{0.8506}\\0.0099\\-0.0788\end{bmatrix}$$

$$Y_3=\begin{bmatrix}-0.0789\\-0.0835\\-0.0413\\\underline{0.9272}\end{bmatrix}$$

（下转第 103 页）

上述两种方法。



(a) 3 维森林数据集　　(b) 4 维森林数据集

图 8　AdenTS 的性能

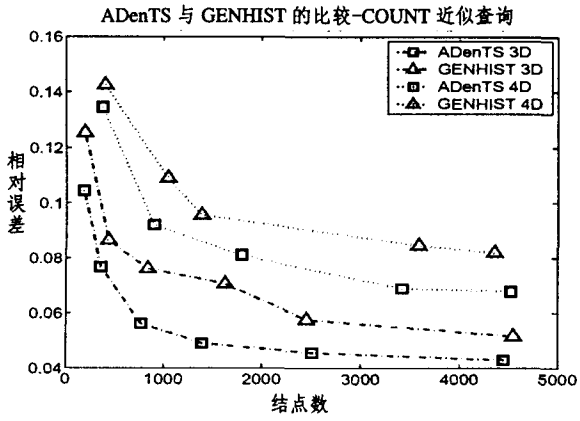


图 9　ADenTS 与 GENHIST 比较

参 考 文 献

1 Chakrabarti M N, Garofalakis R R, Shim K. Approximate Query Processing Using Wavelets. In: Proc 26th Int Conf. on Very Large Data Base (VLDB '00), Cairo, Egypt, 2000

2 Gunopulos D, Kollios G, Tsotras V J. Approximating Multi-dimensional Aggregate Range Queries over Real Attributes. In: Proc. ACM SIGMOD 19th Int Conf on Management of Data (SIGMOD '00), Dallas, USA, May 2000

3 Lee J, Kim D, Chung C. Multi-dimensional Selectivity Estimation Using Compressed Histogram Information. In: Proc. ACM SIGMOD 18th Int Conf. on Management of Data (SIGMOD '99), Philadelphia, USA, 1999

4 Lazaridis I, Mehrotra S. Progressive Approximate Aggregate Queries with a Multi-Resolution Tree Structure. In: Proc. ACM SIGMOD 20th Int Conf. on Management of Data (SIGMOD '01), Santa Barbara, USA, 2001

5 Shanmugasundaram J, Fayyad U, Bradley P S. Compressed Data Cubes for OLAP Aggregate Query Approximation on Continuous Dimensions. In: Proc. ACM SIGKDD 6th Int Conf. on Knowledge Discovery and Data Mining (KDD'99), San Diego, USA, 1999

(上接第 87 页)

表 1　SMDD 部分相似度计算结果

S_1	S_2	Similarity
major	**Falculty_name**	**0.8506**
Student-score	**S_score**	**0.9199**
score	**S_score**	**0.9199**
sex	**S_sex**	**0.9272**
name	**S_name**	**0.9942**
name	S_address	0.9942
score	S_height	0.8913
age	S_score	0.4260
age	Falculty_name	0.1087
…	…	…

SMDD 的匹配质量很大程度上依赖于数据源中数据分布的规律性。在数据分布较为规律的情况下，SMDD 匹配质量较好；在数据分布不规律的情况下，SMDD 的计算效果不理想。实验中，我们通过对现有的一些科学试验、商业销售等方面的专业数据库进行分析，显示在数据分布较为规律的状态下，SMDD 对的匹配质量 F-Measure(0.5)最高可达到 0.65；一般情况下，F-Measure(0.5)介于 0.25～0.65 之间。

结束语　近年来，随着数据共享需求的不断提高，模式映射作为数据管理的基本问题已得到了广泛的重视，出现了很多模式匹配方法。现有的大部分模式匹配方法仍是以利用数据源的元素名称、结构等模式信息为主，对于数据内容信息的利用较为有限。本文提出的 SMDD 方法通过挖掘数据内容信息，根据元素的数据分布特征计算模式元素之间的相似度。SMDD 方法适用于数据分布具有一定规律的异构数据源之间的模式匹配，既可以独立使用，也可以与其它模式匹配方法结合使用，互为补充，从数据内容的角度提高匹配质量。

参 考 文 献

1 Doan A H. Learning to Map between Structured Representations of Data: [PhD thesis]. University of Washington

2 Madhavan J, Bernstein P A, Rahm E. Generic Schema Matching with Cupid. VLDB 2001

3 Doan A H, Domingos P, Halevy A. Reconciling Schemas of Disparate Data Sources: A Machine-Learning Approach. In: SIGMOD Record, 2001

4 Melink S, Garcia-Molina H, Rahm E. Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching. In: Proc 18th Intl Conf. on Data Engineering, 2002

5 Dell'Erba M, Fodor O, Ricci F, et al. Harmonise: A Solution for Data Interoperability. IFIP I3E 2002

6 Rahm E, Philip A. Bernstein. A survey of approaches to automatic schema matching. VLDB Journal, 2001, 10: 334～350

7 Do H H, Rahm E. COMA-A system for flexible combination of schema matching approaches. In: Proc. 28th VLDB Conference

8 史忠植. 知识发现. 北京：清华大学出版社，2004