ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.2, February 2009, pp.327−338
doi: 10.3724/SP.J.1001.2009.03203

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

SKM:一种基于模式结构和已有匹配知识的模式匹配模型*

申德荣+, 余恩远, 张 旭, 寇 月, 聂铁铮, 于 戈

(东北大学 信息科学与工程学院,辽宁 沈阳 110004)

SKM: A Schema Matching Model Based on Schema Structure and Known Matching Knowledge

SHEN De-Rong+, YU En-Yun, ZHANG Xu, KOU Yue, NIE Tie-Zheng, YU Ge

(College of Information Science and Engineering, Northeastern University, Shenyang 110004, China)

+ Corresponding author: E-mail: shenderong@ise.neu.edu.cn

Shen DR, Yu EY, Zhang X, Kou Y, Nie TZ, Yu G. SKM: A schema matching model based on schema structure and known matching knowledge. *Journal of Software*, 2009,20(2):327−338. http://www.jos.org.cn/1000-9825/3203.htm

Abstract: To make up the limitations of existing schema matching methods based on schema structure information, a schema matching model called SKM (schema and reused knowledge based matching model) is proposed based on schema structure information and known matching knowledge. In this model, neural network influence procedure is imitated to realize semantic matching reasoning. The known matching knowledge is reused to mine the deep semantic relation between two schemas. It is also reused to curtail uncertain threshold interval automatically to specify the threshold for decreasing manual intervention. A simple approach of specifying matching relation between two matching elements is given. In the meantime, a self-learning adaptive and iterative model is presented to mine and enrich the known matching knowledge. Experimental results show that the SKM is feasible.

Key words: schema matching; knowledge reuse; semantic inference; data integration; data mining

摘 要: 针对已有基于模式结构的模式匹配方法的局限性,提出了一种利用模式结构信息和已有匹配知识的模式匹配模型——SKM(schema and reused knowledge based matching model).在该模型中,借鉴神经网络元之间的影响过程实现语义匹配推理;通过重用已有匹配知识深入挖掘模式元素之间的深层语义关系;基于已有匹配知识自动缩减不确定阈值区之间来确定匹配阈值,有效减少人工干涉;给出了简单的确定模式元素之间匹配关系的方法;同时通过自适应式迭代模型,进一步挖掘求精已有匹配知识.实验结果表明,SKM 模型切实可行.

关键词: 模式匹配;知识重用;语义推理;数据集成;数据挖掘

中图法分类号: TP181 **文献标识码**: A

模式匹配是获取不同模式之间语义关联关系的技术.它在许多应用中都起着关键性的作用,如数据挖掘中

* Supported by the National Natural Science Foundation of China under Grant No.60673139 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2008AA01Z146 (国家高技术研究发展计划(863))

Received 2007-02-10; Accepted 2007-10-09

正确地挖掘模式之间的语义映射关系、数据集成中异构数据源的模式匹配、电子商务中的异构消息映射、数据网格中的数据资源发现等.目前,典型的模式匹配的研究[1]有基于有限的待匹配模式信息的模式匹配[2−4]和基于模式信息集或实例集的模式匹配[5−7]两类.前者主要面对某一特定领域的应用,其主要思想是尽可能地利用所能获取的一切模式信息进行匹配推理.该方法存在的主要不足是:模式自身携带信息的有限性导致了匹配结果具有一定的局限性;需要领域专家进行手工干预模式匹配处理;当模式信息改变时,导致匹配知识无效.基于大数据量模式信息集的模式匹配通常采用统计分析的方法实现,需要正确的训练数据集,且初始规则难以确定.

本文针对目前基于模式结构信息的模式匹配方法存在的主要问题进行研究,提出了一种基于模式结构和已有匹配知识的模式匹配模型——SKM(schema and reused knowledge based matching model).首先基于模式结构信息进行初始匹配和语义推理,然后重用已知匹配知识深入挖掘模式元素语义匹配关系.同时,通过已有匹配知识自推理和自协调,最大限度地减少人工参与度,并为后续模式匹配提供更精确的匹配知识.

本文第 1 节介绍相关工作.第 2 节介绍 SKM 模型的总体结构.第 3 节给出初始匹配矩阵模型.第 4 节介绍基于模式结构的语义推理模型以及对应的影响策略.第 5 节给出重用已知模式匹配知识的重用模型.第 6 节讨论已有匹配知识的高收敛阈值确定策略.第 7 节针对模式及其匹配关系的变化,给出一种迭代的自适应策略.第 8 节介绍相似匹配对之间的匹配类型确定策略.第 9 节为实验结果与分析.第 10 节总结全文.

1 相关工作

有关模式匹配理论的研究按照基于知识的不同主要分为两类:第 1 类是基于待匹配模式元素的标签和模式自身的结构信息进行语义匹配关系推理;第 2 类是针对大量的模式信息集和对应的数据实例集进行统计分析,以推理出语义模式匹配关系.

第 1 类具有代表性的研究工作主要有:(1) Microsoft 的 Cupid 方法[2]:基于模式结构信息进行匹配推理,并应用辅助信息处理同义词、缩略词、首字母缩写等.(2) Leipzig 大学的 COMA(a system for flexible combination of schema matching approaches)方法[3]:采用复合式方法,灵活地组合不同的匹配算法及结果,以显著地提高匹配效率.(3) Stanford 大学的 Similarity Flooding 方法[4]:基于相邻元素之间的相似传递性进行推理,如果两个模式元素的邻近元素相似,它们就趋于相似.(4) Washington 大学的 GLUE 方法[8]:基于机器学习实现模式匹配,并采用多策略学习方法自动合并不同匹配器的匹配结果.(5) LSD 方法[9]:基于机器学习实现新数据源模式到全局模式之间的映射.(6) IBM 的 Clio 方法[10]:半自动化的模式匹配系统.此类方法的主要优点是基于模式自身携带的有限结构信息挖掘模式匹配关系.但是,模式自身携带语义信息的有限性决定了最后匹配结果具有一定的局限性,主要表现为:(1) 当模式元素匹配对信息量不足或模式元素匹配对之间的区分度较小时,匹配效果会急剧下降.(2) 匹配的前提是相邻匹配对彼此相互影响,当候选匹配对之间缺乏绝对层次关系时,比如 html 页面表单中的两个标签元素在语义上有相邻关系但在模式文本中的距离却相距很远时,此类方法无法有效处理.(3) 只能对一对一型匹配关系挖掘,没有考虑一对多匹配关系,而在实际的模式匹配中存在着大量的一对多匹配关系.

第 2 类匹配方法主要是基于大数据量模式集合、重复实例元组及组合已有的模式匹配技术进行模式匹配的研究,代表性的工作有:(1) Corpus 方法[5]:应用信息检索中的方法对大数据量文本集进行分析,挖掘模式元素之间的语义关联关系.(2) 应用统计分析方法对大量模式知识进行分析[6],挖掘模式中“隐含”的对象模型,并基于此模型进行模式匹配.(3) Duplicates 方法[7,11]:对大量的数据集进行深度挖掘,并使用数据集中重复的数据列实例挖掘模式之间的语义匹配关系.(4) SemInt 方法[12]:应用从关系数据库系统中得到的元数据和从实例数据获得的统计信息强化匹配效果.此类方法的主要优点是针对大量模式知识进行分析处理,能够挖掘出隐藏较深并具有普遍意义的匹配知识.此类方法存在的主要不足有:(1) 不易获取正确的匹配训练数据集,其工作量不亚于手工进行模式匹配处理;(2) 一对多型匹配关系的挖掘规则难以确定;(3) 只能针对特定领域进行处理,不具有通用性;(4) 阈值难以确定.

综上所述,已有的模式匹配方法各具优点,同时也存在着一些亟待解决的问题.SKM 模型是模拟神经网络中神经元之间的影响过程而提出的.它充分利用有限的模式结构知识进行语义匹配推理,并通过高效、合理地

重用已有匹配知识提高模式匹配精度.通过采用自推理的启发式阈值确定策略,最大限度地减少人工干预,并给出一对多匹配关系的简单确定策略.同时,获得的匹配知识通过自迭代动态求精,可为后续的模式匹配提供更可信的知识基.

2 SKM模型框架

SKM 模型借鉴神经元理论、智能推理机制和启发式思想,有效地实现了模式匹配和已有匹配知识的融合,提高了匹配模型的准确度和匹配效率.SKM 模型的主要思想是:(1) 充分利用有限的模式结构信息,达到最大获取知识的目的;(2) 重用已有匹配知识,提高模式匹配的准确度和匹配效率;(3) 最大程度地减少人工干预.SKM 模型框架如图 1 所示,主要由初始匹配矩阵模型(initial matching matrix model,简称 IMMM)、结构化语义推理模型(structural semantic matching reasoning model,简称 SMRM)、匹配知识重用模型(knowledge reuse model,简称 KRM)、匹配知识自适应迭代模型(self-learning iterative adaptive model,简称 SIAM)、阈值确定策略(threshold specify strategy,简记为 ThresholdSpecifyS)和匹配类型确定策略(match type specify strategy,简记为 MatchType SpecifyS)组成.

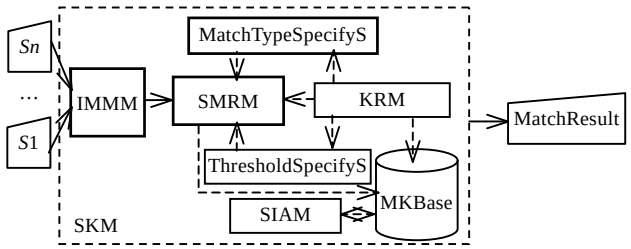


Fig.1 Overall structure of SKM

图 1 SKM 模型整体框架

IMMM 是 SKM 模型的基础.首先参照辅助信息库(可选)如 Wordnet,将模式元素分离成词条向量,然后基于向量匹配计算模型计算各模式元素的匹配度,进而生成初始匹配矩阵(initial matching matrix,简记 IMM).SMRM 是 SKM 模型的核心,是模拟神经元之间影响规律的语义推导模型.IMM 经过结构化语义匹配推理模型作用后,得到的匹配矩阵简记为 SMM.基于 KRM 重用历史知识,进一步挖掘模式之间的匹配关系,精化后的 SMM 简记为 RMM.经 ThresholdSpecifyS 确定的最终阈值(threshold)过滤后的匹配矩阵,简记为 FMM.最后,基于 MatchTypeSpecifyS 确定模式元素之间的最终匹配关系.模式匹配知识库(MKbase)中的已有模式匹配知识随着模式信息的改变,基于 SIAM 进行自协调.具体匹配流程如图 2 所示.

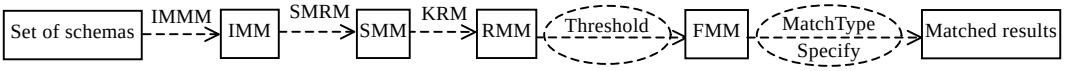


Fig.2 Schema matching flow

图 2 模式匹配流程

3 初始匹配矩阵模型

SKM 基于初始匹配矩阵模型(IMMM)构建初始匹配矩阵(IMM).本节介绍 SEP-XML 树型模式结构及其初始匹配矩阵模型.

3.1 SEP-XML树型模式结构

基于 XML 描述模式信息已成为首选标准,并采用典型的 start-end 标识法表示层次结构.然而,start-end 无法直接表达亲兄弟关系和左右顺序关系.为增强模式元素的结构信息,我们提出 start-end-plus 表达方式(简写为 SEP),*SEP*:=(*s*,*t*,*l*),其中,*s*,*t*,*l* 分别为起始标号、结束标号和父节点标号.

本文结合 SEP 和 XML 描述模式元素信息,即 SEP-XML 树型模式结构,也称为 SEP-XML 模式元素表达式.

定义 1(SEP-XML 模式元素表达式). SEP-XML 模式元素表达式定义为

$$SEP\text{-}XML := SchemaID{:}SchElemLable[SEP]\{[DataType][ValueRange]\} \quad (1)$$

其中,*SchemaID* 为模式 ID;*SchElemLable* 为模式标签;{}中为可选信息,用于验证匹配推理的结果.

3.2 初始匹配矩阵模型

初始匹配矩阵用于描述两模式之间的基本匹配关系.首先基于Wordnet对模式元素进行预处理[2].之后基于词条的相似性评估对预处理后得到的模式元素进行相似性评价.根据模式元素类型(单词条模式元素和多词条模式元素(也称为向量模式元素))分别进行处理,具体定义[7]如下:

定义 2(单词条模式元素相似匹配度(*Sim*(*a*,*b*))). 基于编辑距离定义相似度,具体定义为

$$Sim(a,b)=1-\frac{ed(a,b)}{\max\{|a|,|b|\}} \tag{2}$$

其中,*a*,*b* 分别为两个元素词条;max{|*a*|,|*b*|}表示词条 *a*,*b* 中较长一个字符的长度.

定义 3(词条向量模式元素相似匹配度(*Sim*(*V*$_1$,*V*$_2$)). 综合考虑模式向量中各词条的作用和词条之间相对独立的位置关系,具体定义为

$$Sim(V_1,V_2)=\sum_{\langle t,s\rangle\in Close(\theta,V_1,V_2)} w(V_1,t)\cdot w(V_2,s)\cdot Sim(t,s) \tag{3}$$

其中,V_1 和 V_2 满足$Close(\theta,V_1,V_2)$,$Close(\theta,V_1,V_2)=\{\langle a,b\rangle \mid \exists a\in V_1\wedge\exists b\in V_2\wedge Sim(a,b)>\theta\}$;$w(V,r)$表示词条 r 在模式向量元素 V 中的权重,$w(V,r)=\log(tf_{V,r}+1)\cdot\log\left(\frac{N}{df_r}+1\right)$;$tf_{V,r}$表示词条 r 在 V 中出现的频率,N 表示该模式对应的模式元素个数,df_r 表示词条 r 在该模式所对应的所有模式元素向量中出现的频度.

定义 4(初始匹配矩阵模型(IMMM)). 存在模式 S_x,S_y,$S_x=\{e_{x1},e_{x2},\ldots,e_{xn}\}$,$S_y=\{e_{y1},e_{y2},\ldots,e_{ym}\}$,$e_{xi}$,$e_{yj}$ 分别是模式 S_x 和模式 S_y 中的任意两个元素,则模式 S_x 和模式 S_y 的初始矩阵匹配模型($IMMM(S_x,S_y)$)定义为

$$IMMM(S_x,S_y)=\begin{bmatrix} Sim(e_{x1},e_{y1})\ldots Sim(e_{x1},e_{ym}) \\ Sim(e_{x2},e_{y1})\ldots Sim(e_{x2},e_{ym}) \\ \ldots\quad\ldots\quad\ldots \\ Sim(e_{xm},e_{y1})\ldots Sim(e_{xn},e_{ym}) \end{bmatrix} f \tag{4}$$

其中,$Sim(e_{xi},e_{yj})$表示 e_{xi} 和 e_{yj} 的相似度,基于式(2)和式(3)计算得到.

基于初始匹配矩阵模型得到初始匹配矩阵(IMM_{xy}).如图 3 所示,IMM_{xy} 为基于 IMMM 构造的有关模式 S_x,S_y 之间的初始匹配矩阵,其中 $Sim(e_{x1},e_{y1})=0.8$.

为了便于说明,初始匹配矩阵和文中后续基于初始匹配矩阵产生的各种矩阵统称为匹配矩阵.针对匹配矩阵,给出如下定义:

定义 5(候选匹配对). 匹配矩阵中相似度大于 0 的模式元素对($Sim(e_{xi},e_{yj})>0$)为候选匹配对,描述为$\langle S_x{:}e_{xi},S_y{:}e_{yj}\rangle$,若上下文中模式名已明确,则可简写为$\langle e_{xi},e_{yj}\rangle$.如图 3 中,$\langle S_x{:}e_{x1},S_y{:}e_{y2}\rangle$为模式 S_x,S_y 之间的一个候选匹配对.

S_x	S_y			
	e_{y1}	e_{y2}	…	e_{ym}
e_{x1}	0.8	0.2	…	0
e_{x2}	0.01	0.75	…	0.1
…	…	…	…	…
e_{xn}	0	0	…	0.65

Fig.3 Initial matching matrix (IMM_{xy}) between schema S_x and S_y

图 3 模式 S_x 与 S_y 的初始匹配矩阵(IMM_{xy})

定义 6(不相容匹配对). 匹配矩阵中处于同一行或同一列的候选匹配对互为不相容候选匹配对,它们彼此具有不相容匹配关系.如图 3 中,$\langle e_{x1},e_{y1}\rangle$与$\langle e_{x1},e_{y2}\rangle$是不相容匹配对,彼此具有不相容匹配关系.

定义 7(相容匹配对). 匹配矩阵中不处于同一行和同一列的候选匹配对为相容匹配对,它们彼此具有相容匹配关系.如图 3 中,$\langle e_{x1},e_{y1}\rangle$与$\langle e_{x2},e_{y2}\rangle$是相容匹配对,彼此具有相容匹配关系.

定义 8(相似匹配对). 确定具有相似关系的候选匹配对称为相似匹配对,彼此具有匹配关系.如$\langle$*Book*:*auther*,*eBook*:*writer*$\rangle$被确定为相似匹配对.

定义 9(非相似匹配对). 确定不具有相似关系的候选匹配对称为非相似匹配对,彼此具有非匹配关系.如$\langle$*Electronic Device*:*notebook*,*Official Material*: *notebook*$\rangle$为非相似匹配对,彼此具有非匹配关系.

定义 10(确定匹配对和不确定候选匹配对). 相似匹配对和非相似匹配对统称为确定匹配对,其他候选匹配对称为不确定候选匹配对,分别称彼此具有确定的匹配关系和具有不确定的匹配关系.

4 结构化语义匹配推理模型(SMRM)

初始匹配矩阵(IMM)只考虑了模式元素之间的基本匹配关系,实际上,模式元素之间还存在一定的相互影响.本节给出了基于初始匹配矩阵的相互影响模型,目的是最大程度地利用有限的模式结构信息.

4.1 SMRM模型定义

通过分析大量模式之间的匹配关系,并结合 Cupid 和 SF 中的设计思想,提出如下匹配对之间的影响规则:模式 S_x,S_y 中候选匹配对$\langle e_{xi},e_{yj}\rangle$在受到 S_x,S_y 之间的其他候选匹配对影响的同时,若其相似度大于一定值,则其也反过来影响其余候选匹配对.

基于以上影响规则,我们提出了一种模拟神经元之间相互影响的结构化语义匹配推理模型,并且只有相容匹配对之间才会发生影响作用.如图 4 所示,其描述候选匹配对 L 与其相容匹配对之间彼此互相影响.

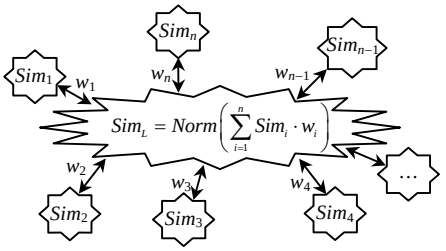


Fig.4 Influence diagram of the similar pair L

图 4 相似匹配对 L 的匹配影响示意图

定义 11(结构化语义匹配推理模型(SMRM)). 借鉴数值分析中的迭代插值算法,结构化语义匹配推理模型具体定义为

$$SMRM:=\{IMM_{xy},Sim_k^r\} \tag{5}$$

$$Sim_k^{r+1}=Normal\left(Sim_k^r+f\left(\sum_i^m(Sim_i\cdot w_i),\lambda\right)\right) \tag{6}$$

其中,IMM_{xy} 是模式 S_x 和 S_y 之间的初始匹配矩阵. Sim_k^r 代表第 k 对相容匹配对第 r 次迭代的相似匹度值,w_i 表示第 i 对相容匹配对对第 k 对相容匹配对的影响加权因子.其取值源于模式匹配之间影响关联关系的直观认知:相隔越近的模式匹配对之间的相互影响作用越大,反之,相互影响越小.为此,w_i 取值为第 i 对相容匹配对与第 k 对相容匹配对之间跨越的候选匹配节点对个数加 1 的倒数. $f(i,\lambda)=\begin{cases}i, & i\geq\lambda\\ 0, & i<\lambda\end{cases}$,模拟神经元的影响模型,相似度值大于输出阈值$\lambda$(取经验值 0.5)的相容匹配对将影响其余相容匹配对.

$Normal\left(Sim_j(x,y)\right)=\dfrac{Sim_j(x,y)}{\max\limits_{i\in N}\left(Sim_i(x,y)\right)}$,用于归一化相似度值,以提高收敛速度.

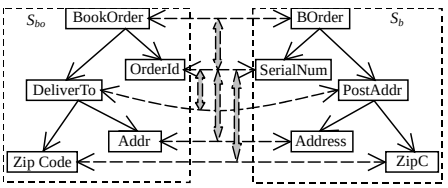


Fig.5 Matching relation diagram between schema S_{bo} and S_b

图 5 模式 S_{bo},S_b 之间匹配关系示意图

当所有模式匹配对节点都满足 $sim_k^{i+1}-sim_k^i<\varepsilon$ (ε取经验值 0.05),或迭代次数超过 100 次时,迭代停止.例如,在图 5 中,模式 S_{bo},S_b 之间存在如下相容匹配对:$\langle BookOrder,BOrder\rangle$,$\langle OrderId,SerialNum\rangle$,$\langle DeliverTo,PostAddr\rangle$,$\langle Addr,Address\rangle$,$\langle Zip\ Code,ZipC\rangle$,用虚线表示.双向宽箭头表示影响关系,如$\langle OrderId,SerialNum\rangle$在受到其余相容匹配对影响的同时也对其余的相容匹配对产生影响.

初始匹配矩阵(IMM)经结构化语义匹配模型的语义推理得到的矩阵即 SMM.

4.2 算法迭代收敛性与复杂度分析

SMRM 中迭代算法收敛过程的计算可以说是一个特征向量的计算[13].令 T 为基于待匹配模式 A 和 B 之间构建的相似匹配度矩阵.若匹配对$\langle i,j\rangle$之间的匹配度值为 c,则令矩阵 T 中的项 t_{ij} 取值为 c.按照平均遍历定理(mean ergodic theorems) 理论,当 T 为非周期且不可约矩阵时,计算过程一定收敛.而当且仅当图 G 为强联通图时,矩阵 T 就是不可约的.为确保这一性质,可以将加数σ^0 引入收敛性计算等式,即 $\sigma^{i+1}=normalize(\sigma^0+\varphi(\sigma^i))$.如果$\sigma^0$ 赋给模式 A 与 B 之间的每一对映射对,则等价于将 G 修改为所有节点之间都彼此相连并且匹配度不小于σ^0 的图 G',设 G'对应的相似匹配矩阵为 T'.

在 SMRM 中,模式元素匹配树结构具有强联通性,并由公式(6)保证了相似矩阵不可约.因此,SMRM 的迭代算法具有收敛性.

由于迭代计算的迭代次数与图 G 中边的个数成比例,因而与模型 A 和 B 中边数量的乘积也成比例.设 N_A 和 N_B 表示 A 和 B 中的节点个数,并且 A 和 B 中的节点是完全互相连通的,则 A 和 B 中的边的数目为 $O(N_A^2)$ 和 $O(N_B^2)$, G 中边数为 $O(N_A^2N_B^2)$.因此,最坏情况下每次迭代的复杂度为 $O(N_A^2N_B^2)$ 或 $O(\|A\|\cdot\|B\|)$, 其中 $\|A\|$ 和 $\|B\|$ 分别指 A 和 B 中的边数.然而,在许多场景中,每次迭代的平均复杂度为 $O(N_AN_B)$.

5 匹配知识重用模型(KRM)

当模式自身结构信息不够充分时,仅基于模式的结构信息挖掘其匹配关系是不够的.为此,本文在结构化语义匹配推理模型基础上提出,通过重用已有模式匹配知识提高匹配精度和匹配效率.已有的典型的 COMA 方法[3]的重用机制是整个模式层次的简单重用和模式片断的直观重用,没有进行深入挖掘,也没有给出系统的重用规则.本文给出了匹配知识重用模型 KRM 的详细描述,并定义了 3 种重用规则:匹配关系重用、非匹配关系重用、混合型重用.

定义 12(模式匹配图(SMG)). 已有的匹配知识以模式匹配图(schema matching graph,简称 SMG)形式存储.SMG 具体定义如下:

$$SMG:=(V,E) \tag{7}$$

其中,V 是图中的点组成的集合,$V=\{S_1,S_2,\ldots,S_n\}$是模式的有穷非空集合;E 是图的边,$E=\{SMx_{ij}|\forall(S_i,S_j)\in V\wedge SMx_{ij}\neq\varnothing\}$是模式匹配矩阵的有穷集合,$SMx_{ij}$ 是模式 S_i 和 S_j 的相似匹配矩阵.

如图 6 所示,S_1 和 S_5 具有匹配关系,SMx_{15} 是模式 S_1 和 S_5 的相似匹配矩阵.

定义 13(匹配知识重用模型(KRM)). SMG 是已有知识的模式匹配图,RU 是重用规则集,RU={匹配关系重用,非匹配关系重用,混合型重用},则匹配知识重用模型定义为

$$KRM:=\{RU,SMG\} \tag{8}$$

匹配知识重用模型中具体规则定义如下:

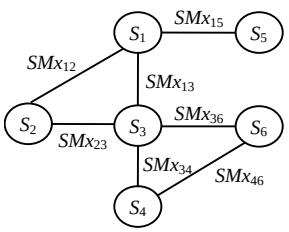


Fig.6 Schema matching diagram

图 6 模式匹配示意图

规则 1(匹配关系重用). 利用相似匹配对的匹配关系的传递性进行匹配关系推理.若已知相似匹配对:$\langle S_a{:}r,S_1{:}a\rangle$, $\langle S_1{:}a,S_2{:}b\rangle,\langle S_2{:}b,S_3{:}c\rangle,\ldots,\langle S_{n-1}{:}x{:}S_n{:}y\rangle,\langle S_n{:}y,S_b{:}t\rangle$, 则 $Sim\langle S_a{:}r,S_b{:}t\rangle=[Sim\langle S_a{:}r,S_1{:}a\rangle+Sim\langle S_1{:}a,S_2{:}b\rangle+\ldots+Sim\langle S_n{:}y,S_b{:}t\rangle]/(n+1)$.

规则 2(非匹配关系重用). 利用单个模式内不相同的元素标签意义必然不同的特性,将需要探测的候选匹配对集合缩小.若$\langle S_1{:}x,S_2{:}y\rangle$和$\langle S_3{:}w,S_2{:}v\rangle$为候选匹配对,而 $S_2{:}y$ 和 $S_2{:}v$ 属于同一模式的不同词条,则 $S_1{:}x$ 与 $S_3{:}w$ 确定为非相似匹配对,即具有非匹配关系.

规则 3(混合型重用). 结合“匹配关系重用”和“不匹配关系重用”两种重用规则的混合型重用规则.

在 SKM 模型中,通过重用已有匹配知识解决如下问题:(1) 通过重用匹配知识,减少待探测模糊匹配对的数

量,提高匹配效率;(2) 通过两模式的关联匹配元素自动挖掘其模式匹配关系,提高匹配精度;(3) 提供一个自动阈值紧缩算法,减少人工介入次数,降低匹配代价;(4) 基于重用机制实现已有模式匹配知识的自动化求精,使系统能够自动挖掘隐含的模式匹配信息,进一步提高模式匹配的正确性和匹配效率.

SMM 矩阵通过重用已有的知识,进一步精化了矩阵中的模式元素之间的匹配关系,得到 RMM 矩阵.

6 阈值确定策略(ThresholdSpecifyS)

模式之间可能存在高度的语义异构,这决定了完全自动化式的模式匹配是不可能实现的.该阈值确定策略的思想是,通过重用已有匹配知识,达到有效减少操作员介入次数的目的.

在已有的模式匹配模型中,大多数是完全由操作人员手工干预确定最终阈值.这要求操作员对模式标签的语义非常熟悉,但也无法保证结果的正确性[12].为此,我们给出一种改进的阈值确定策略,即利用已知模式匹配知识进行一定的自动缩减,最大限度地减少人工参与度.首先确定初始阈值区之间和不确定的候选匹配对队列,然后基于历史匹配知识,由阈值推理机自动缩减阈值区之间的待探测范围,最大限度地减少交互次数,并通过不断地缩减阈值区之间获得最终的阈值.

6.1 不确定候选匹配对队列定义

令 $CZone$=[$floor$,$ceil$]为初始的可能性探测区,根据历史经验确定.在 RMM 中,候选匹配对的相似匹配度值小于 $floor$ 的匹配对确定为非相似匹配对;相似匹配度值大于 $ceil$ 的候选匹配对为相似匹配对;而相似度值处于可能性探测区内的候选匹配对称为不确定候选匹配对(或待探测的候选匹配对),其组成的队列为不确定候选匹配队列($CndList$).具体定义为

$$CndList:=\{Pair_i|Sim(Pair_i)\in CZone\wedge Pair_i\in RMM\} \tag{9}$$

$$Sim(Pair_i)=\begin{cases}0, & Sim(Pair_i)<floor\\ Sim(Pair_i), & Sim(Pair_i)\in CZone\\ 1, & Sim(Pair_i)>ceil\end{cases} \tag{10}$$

6.2 阈值推理机处理过程

设可能性探测区之间 $CZone$ 为[$floor$,$ceil$],将不确定候选匹配序列按照相似度值由高到低进行排序,设相似度最高值为 U,最低值为 L,则待探测的候选匹配对的阈值区之间为 $realZ$=[L,U],而实际需要探测的不确定阈值区之间 dZ=[min{L,$floor$},min{U,$ceil$}]=[LL,UU].阈值处理具体步骤如下:

步骤 1. 确定不确定候选匹配队列 $CndList=\{Pair_i|Sim(Pair_i)\in dZ\wedge Pair_i\in RMM\}$.取最接近 $CndList$ 中之间的且具有确定的匹配关系的候选匹配对 $Pair_k$,其相似匹配度值为 V_k.

步骤 2. 若 $CndList$ 中不存在这样的匹配对,即 $Pair_k$ 为空,则推理机停止,跳出,交给人工交互确定匹配关系;转回步骤 3.

步骤 3. 若 $Pair_k$ 为非相似匹配对,则转步骤 4;反之,则转步骤 5.

步骤 4. 将 $CndList$ 中所有相似度值小于 V_k 的候选匹配对都置为非相似匹配对,执行 $CndList=CndList-\{Pair_r|Sim(Pair_r)<V_k\}$;将待探测阈值区之间缩减为 $dZ=[V_k,\max\{V_k,UU\}]$;转步骤 6.

步骤 5. 将相似匹配度值大于 V_k 的候选匹配对置为相似匹配对,执行 $CndList=CndList-\{Pair_r|Sim(Pair_r)>V_k\}$操作;将待探测阈值区之间缩减为 $dZ=[\min\{LL,V_k\},V_k]$.

步骤 6. 若 $CndList$ 为空,则推理机停止处理;否则,转步骤 1.

可见,该阈值处理过程是利用确定的匹配关系来辅助确定待匹配关系中不确定的匹配部分.

7 匹配类型确定策略(MatchTypeSpecifyS)

确定最终的阈值以后,RMM 中相似度值高于此阈值的候选匹配对为相似匹配对,并且矩阵中只存在相似匹配对和非相似匹配对,得到的矩阵为 FMM.其相应的相似匹配对集合表示为 $RsltSet=\{Pair_i|Sim(Pair_i)\geq$

$Threshold \wedge Pair_i \in FMM$}.但此集合中有 1:1 关系匹配对,也有 1:$m$ 匹配关系对.对于 1:1 型匹配关系的匹配对,直接归入匹配结果集合;对于 1:m 的匹配对,需要进一步处理,以确定最终匹配关系.

有关复杂性匹配关系的确定一直是一个难题,具有权威性的典型解决方案是 iMap 匹配方法[14].其最好的匹配效果是 58%,并且解决方案较为复杂.本文给出了一种基于模式自身结构和重用已有匹配知识确定模式匹配关系的简单策略,用于确定多对多的匹配关系.例如,模式 S_A 和 S_B 之间存在匹配关系,S_A 中包含模式元素 A_1,A_2 和 A_3,其中 A_2 和 A_3 为 A_1 的子节点,A_2 和 A_3 分别对应匹配 S_B 中的模式元素 B_2 和 B_3,由 S_A 模式自身的性质可知,$A_1 \leftrightarrow \{A_2, A_3\}$,结合 $A_2 \leftrightarrow B_2$ 和 $A_3 \leftrightarrow B_3$,可得出 $A_1 \leftrightarrow \{B_2, B_3\}$,与此类似,可用于已知复杂型模式匹配关系的传递性上.反过来,基于此思想也可以否定一些复杂型模式匹配候选匹配对.

MatchTySpecifyS 具体处理步骤如下:

步骤 1. 找出 1:m 中右方的 m 个模式元素之间的结构关系.设左方模式元素为 t,右方 m 个模式元素组成集合 $Mset=\{r_1,r_2,\ldots,r_m\}$.

步骤 2. 基于原始模式树将 $Mset$ 分为内部节点集合($InnerSet$)和叶子节点集合($LeafSet$),令 $InnerSet=\{i_1,i_2,\ldots,i_p\}$,$LeafSet=\{L_1,L_2,\ldots,L_q\}$,其中 $p+q=m$.

步骤 3. 进一步对 $Mset$ 进行分类,按照节点是否有邻居关系划分为孤立节点集($IndepSet$)和依赖节点集($DepSet$),令 $DepSet=\{(u_i,u_j,\ldots,u_k) | u_i,u_j,\ldots,u_k \in Mset \wedge u_i,u_j,\ldots,u_k$ 存在邻居关系$\}$,$IndepSet=\{v_i | v_i \in Mset$,但 v_i 的邻居节点都不属于 $Mset\}$.

步骤 4. 1:1 匹配关系集(1-1Set)定义为 $1\text{-}1Set=\{\langle t,a_i\rangle | a_i \in (IndepSet \cup (DepSet \cap InnerSet))\}$;1:$m$ 匹配关系集(1-mSet)定义为 $1\text{-}mSet=\{\langle t,a_i\rangle | a_i \in (DepSet \cap LeafSet)\}$.

例如,存在相似匹配对$\langle S_1{:}b,S_2{:}bd\rangle$,$\langle S_1{:}b,S_2{:}be\rangle$,$\langle S_1{:}b,S_2{:}bf\rangle$,$\langle S_1{:}b,S_2{:}bg\rangle$,则模式 S_1 中元素 $b(S_1{:}b)$与模式 S_2 中匹配对元素 $S_2{:}\{bd,be,bf,bg\}$的匹配关系确定示例如图 7 所示.

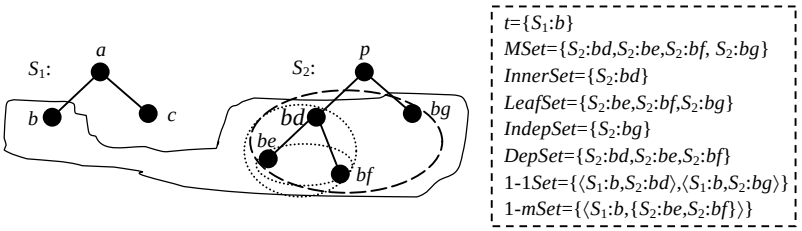


Fig.7 Specify matching relation diagram

图 7 确定匹配关系示例

8 匹配知识自适应迭代算法

通过重用已有匹配知识,提高模式匹配精度.新产生的模式匹配关系反过来也将精化系统中已有匹配结果.当模式信息或模式之间的匹配关系发生了增加、删除或更改变化时,基于匹配知识自适应地迭代模型协调 SMG.

定义 14(自适应迭代模型(SIAM)). SIAM 具体定义如下:

$$SIAM:=\{SMG,IterFresh,S\} \tag{11}$$

其中,$IterFresh$ 是自协调 SMG 的刷新算法,S 是新加入或修改或删除的模式集合.

如图 6 所示,假设已有模式 S_1,S_3,S_4 和 S_5,以及相应确定的匹配关系 SMx_{13},SMx_{34} 和 SMx_{15},当加入 S_2,S_6 时,已知确定的匹配关系是 SMx_{23} 和 SMx_{36},则 SMx_{12} 和 SMx_{46} 是通过 IterFrash 算法刷新获得的.由于确定的匹配矩阵中只存在相似匹配对和非相似匹配对(相似度值为 0),因此保证了由 IterFrash 算法刷新得到的新模式匹配关系(SMx_{12} 和 SMx_{46})的正确性.IterFrash 算法具体见算法 1.

算法 1. IterFresh.

输入:SMG 的当前状态.

输出:SMG 迭代后的精化状态.

步骤:

```
1.    初始化 count 为 0;   //用于记录迭代次数
2.      Vector smx_v:=SMG.getSmx();   //取 SMG 中所有相似矩阵集合存入 smx_v 中
3.      Vector smx_v_copy:=smx_v;   //复制一份 smx_v 存入 smx_v_copy 中
4.      for (i->smx_v.size()){   //按顺序访问矩阵集 smx_v 中所有相似矩阵
5.            Smx sx:=smx_v.get(i);   //从向量 smx_v 中取一个相似匹配矩阵存入 sx
6.            Schem left:=sx.getLeftSchema();   //取其左方模式放入 left
7.            Schem right:=sm.getRightSchema();   //取其右方模式放入 right
          //从 SMG 中将所有关联 left,right 的无环路径集找出,放入向量 lrPath 中
8.            Vector lrPath:=SMG.getPaths(left,right);
9.          if (sx 为非确定性匹配关系矩阵)
10.             for (int j=0;j<lrPath.size();j++){   //遍历路径集中所有路径
11.                 Path ph:=lrPath.get(j);   //取一条路径放入 ph
12.                 MatrixChain mc:=ph.getMtxes();   //获取 ph 上的相似矩阵链 mc
13.                 sx=sx.affected(mc);   //由 mc 影响修正 sx
14.           }
15.   }
16.     SMG.writeBack(smx_v_copy);   //将经过修正的状态写回 SMG
17.     if (smx_v 与 smx_v_copy 相似度大于阈值)   // ++count<100) goto 2;
```

设整个 *SMG* 包含模式个数为 *m*,每个模式平均包含模式元素个数为 *n*,则算法时之间复杂度为 $O(m^2n^2)$.在实现中,我们对算法作了如下优化:(1) 对所有匹配元素做聚集并以 Trie 索引为倒排链.设模式元素标签平均长度为 *l*,则可以在 $O(l)$时之间内查询到满足匹配关系的模式元素集.(2) 采用经典的动态规划算法对一条路径上的多个 *Smx* 矩阵链做运算,将时之间复杂度上界优化为 $O(n^3)$,空之间复杂度为 $O(n^2)$.由于模式匹配知识库的迭代处理为离线型服务操作[8],对用户操作的时之间效率没有直接影响.然而,精确的模式匹配知识可以有效地提高匹配结果的准确性和全面性.

9 实验分析

(1) 实验环境

硬件环境:P4 2.6G,512M RAM,80G HARD DISC.

实验数据集:http://metaquerier.cs.uiuc.edu/repository/和 http://www-db.stanford.edu/~melnik/mm/sfa/上提供的数据模式,通过分类和xml格式重写等预处理后得到3个领域的模式数据:AutoMobile,BookStore和Computer,并增加了部分领域知识.

(2) 评价标准

借用信息检索中经典的评估方法进行评价.*m*:*n* 型匹配看作 *m* 个 1:*n* 型匹配关系,1:*m* 型看作 *m* 个 1:1 型匹配关系.设模式 S_1 和 S_2 之间实际正确的匹配对数量为 *R*,本模型处理得出的匹配关系数量为 *P*,其中正确的匹配对数量为 *T*,则错误的匹配对数量为 *F=P–T*,没有挖掘出来的正确匹配对数量为 *M=R–P*.评价指标定义如下:查准率:*Precision=T/P*;查全率:*Recall=T/R*;综合考虑二者:*F-Measure*=2·*Recall*·*Precision*/(*Precision*+*Recall*).

Precision表示结果的可信度;Recall表示模型获取正确结果的覆盖程度;*F*-Measure综合考虑前两者,关注加上丢失的正确匹配关系和舍弃错误的匹配关系需要付出的代价.

(3) 实验及结果分析

本文设置了多轮实验全面评估 SKM 模型的效能,具体测试如下:

① 只基于模式结构信息时的匹配性能;② 重用已有模式匹配知识的匹配性能;③ 已有模式知识在阈值缩减过程中的作用;④ 迭代算法的收敛性验证.

实验 1. 只基于模式结构信息时的匹配性能

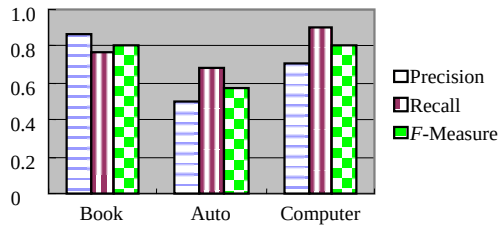


Fig.8 Precision, Recall and *F*-Measure of SMRM

图 8 SMRM 模型的 Precision,Recall 和 *F*-Measure

本实验观察没有重用已有匹配知识时的模型匹配性能,即只测试结构化语义匹配推理模型(SMRM)的匹配性能.不借助任何辅助知识,也没有重用历史匹配知识.同时对比 SMRM 和 SF[4]匹配方法的性能.测试结果如图 8 和图 9 所示.横坐标的 Book,Auto 和 Computer 表示 3 个不同的领域,纵坐标表示匹配性能值.

实验结果表明,在 Book 和 Computer 领域,SMRM 的匹配效果较好,优于 SF 算法;但在 Auto 领域相对较差,主要是因为 Auto 领域中的模式大多是直接从关系模式转化而来的 2 层树型结构,结构化信息较少,导致实验效果不佳.可见,本模型更适合于复杂的模式(层次较多的模式结构).

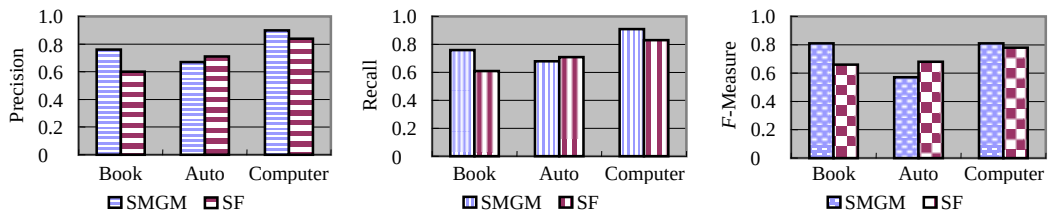


Fig.9 Performance comparison between SMRM and SF

图 9 SMRM 和 SF 性能比较

实验 2. 重用已有匹配知识的模式匹配性能

本实验在结构化语义匹配推理模型(SMRM)的基础上,重用已有匹配知识,测试 SKM 模型的模式匹配性能.测试结果如图 10 和图 11 所示.

实验结果表明,通过重用已有模式匹配知识,显著地提高了模式匹配效果.并且已知匹配知识越多,模式匹配结果越精确.

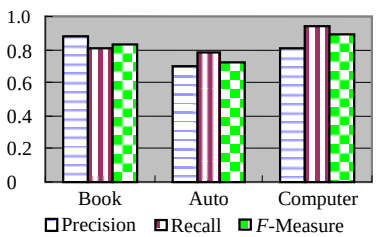


Fig.10 Matching results with reuse known matching knowledge

图 10 重用已有匹配知识的匹配结果

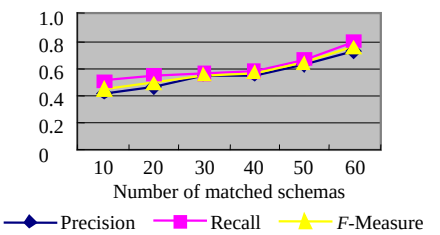


Fig.11 Performance variation with number of schemas increasing

图 11 模式数量增多时性能变化情况

实验 3. 已有匹配知识在阈值缩减过程中的作用

本实验测试由手工确定阈值和采用自动缩减策略确定阈值时的操作人员介入次数情况和重用知识数量对操作人员所需介入次数的影响情况.测试结果见表 1 和图 12.

在表 1 中,Task 为匹配任务,Manual 为手工匹配时的操作人员介入次数,Automatic 为采用本文的自动缩减策略时的操作人员需要手工介入的次数,CurtailRate 为 Automatic 与 Manual 的平均比值.图 12 中,横轴表示模式

知识库中所含有的已知匹配模式个数,纵轴代表需要人手工交互的次数.

实验结果表明,通过重用已有匹配知识,可以使操作员的交互次数减少为原来交互次数的 64%~82%不等,并且其交互次数随着知识基的不断积累而逐渐降低.

Table 1 Comparison of times of manual intervention

表 1 手工介入次数对比

Domain	Book					Auto					Computer				
Task	B_1	B_2	B_3	B_4	B_5	A_1	A_2	A_3	A_4	A_5	C_1	C_2	C_3	C_4	C_5
Manual	7	12	21	6	13	6	13	12	14	22	21	16	18	18	16
Automatic	5	8	15	6	6	6	8	12	6	10	17	16	10	18	12
CurtailRate	$\frac{6+8+...+10}{7+12+...+22}=0.68$					$\frac{17+10+...+7}{21+16+...+16}=0.64$					$\frac{17+16+...+12}{21+16+...+16}=0.82$				

实验 4. 迭代算法收敛性的实验验证

本实验通过实践来验证我们提出的收敛算法的有效性.SMRM 中迭代算法的运行情况见表 2,已有匹配知识的迭代精化算法的运行情况见表 3.

在表 2 中,Iterate Times 为迭代次数,Average 为各个领域 5 个匹配任务的迭代次数之和的平均值.结果表明,SMRM 中迭代算法的自动迭代的次数约在 20~50 次之间停止.可见,SMRM 具有收敛性.

在表 3 中,Schema number 为模式数量,Iterate times 为迭代次数,Average 为迭代次数均值.结果表明,当模式数量增大时,自迭代算法的迭代次数也会相应增加,但增量次数不大.

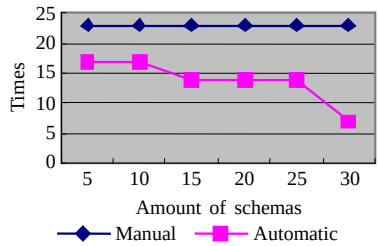


Fig.12 Times of manual interaction with different amount of schemas

图 12 不同模式数量的手工交互次数

Table 2 Times of iterations of matching tasks executing

表 2 各个匹配任务执行的迭代次数

Domain	Book					Auto					Computer				
Task	B_1	B_2	B_3	B_4	B_5	A_1	A_2	A_3	A_4	A_5	C_1	C_2	C_3	C_4	C_5
Iterate times	12	19	30	16	26	17	19	33	16	48	41	38	46	43	53
Average	20.6					26.6					44.2				

Table 3 Iterative times of self-refined mining

表 3 自精化挖掘的迭代次数

Domain	Book				
Schema number	10	15	20	25	30
Iterate times	42	42	47	54	55
Average	48				

10 总结及未来的工作

本文提出的基于模式结构和已有模式匹配知识的模式匹配模型(SKM),通过借鉴神经元影响作用过程,实现了模式匹配的迭代求精;通过高效重用已有匹配知识,提高了模型的匹配精度;通过应用高收敛的阈值确定策略,实现了最少人工参与;给出了确定多对多型模式匹配关系的简单策略;应用自适应迭代模型,使已有模式匹配知识更加准确和完善,为后续模式匹配提供了更精确的匹配知识.本模型达到了预期目的,并具有如下特点:

(1) 基于模式结构进行语义匹配推理时,不必构建繁琐的图结构[4],而是基于简单的相似关系矩阵进行处理;(2) 将匹配对之间的相互影响从局部扩大到了全局范围;(3) 给出了基于模式结构信息的简单的 1:*m* 型匹配关系确定策略;(4) 通过重用已有匹配知识,缓解了仅基于有限的模式信息进行模式匹配的局限性;(5) 采用高收敛的阈值确定策略,有效地提高了模式匹配过程的智能性和匹配效率.

下一步,我们将进一步完善该匹配模型,提高模型的适应性,并深入研究复杂型模式匹配关系的确定策略.

References:

[1] Rahm E, Bernstein PA. A survey of approaches to automatic schema mathcing. VLDB Journal, 2001,10(4):334−350.

[2] Madhavan J, Bernstein PA. Rahm E. Generic schema matching with cupid. In: Apers PMG, Atzeni P, eds. Proc. of the 27th Int'l Conf. on Very Large Data Bases. San Fransisco: Morgan Kaufmann Publishers, 2001. 48−58.

[3] Do HH, Rahm E. COMA—A system for flexible combination of schema matching approaches. In: Bernstein PA, Loannnidis YE, eds. Proc. of the 28th Int'l Conf. on Very Large Data Bases. San Fransisco: Morgan Kaufmann Publishers, 2002. 610−621.

[4] Melnik S, Molina HG, Rahm E. Similarity flooding: A versatile graph matching algorithm. In: Liu L, Reuter A, Whang KY, Zhang JJ, eds. Proc. of the 18th Int'l Conf. on Data Engineering. Los Alamitos: IEEE Computer Society, 2002. 117−128.

[5] Madhavan J, Bernstein PA, Doan A, Halevy A. Corpus-Based schema matching. In: Kitagawa H, Ishikawa Y, eds. Proc. of the 18th Int'l Conf. on Data Engineering. Los Alamitos: IEEE Computer Society, 2005. 57−68.

[6] He B, Chang KCC, Han J. Discovering complex matchings across Web query interfaces: A correlation mining approach. In: Won K, Ron K, Johannes G, William D, eds. Proc. of the 10th Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2004. 148−157.

[7] Bilke A, Naumann F. Schema matching using duplicates. In: Kitagawa H, Ishikawa Y, eds. Proc. of the 18th Int'l Conf. on Data Engineering. Los Alamitos: IEEE Computer Society, 2005. 69−80.

[8] Doan A, Madhavan J, Dhamankar R, Halevy A. Learning to map ontologies on the semantic Web. In: Lawrence S, ed. Proc. of the World-Wide Web Conf. New York: ACM Press, 2002. 662−673.

[9] Doan A, Domingos P, Halvey A. Reconciling schemas of disparate data sources: A machine-learning approach. In: Aref WG, ed. Proc. of the 2001 SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2001. 509−520.

[10] Miller RJ, Hernandez MA, Haas LM, Yan L, Ho CTH, Fagin R, Popa L. The Clio project: Managing heterogeneity. ACM SIGMOD Record, 2001,30(1):78−83.

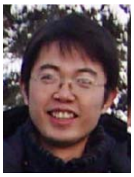
[11] Wang JY, Wen JR, Lochovsky F, Ma WY. Instance-Based schema matching for Web databases by domain-specific query probing. In: Mario AN, *et al*., eds. Proc. of the 30th Int'l Conf. on Very Large Data Bases. San Francisco: Morgan Kaufmann Publishers, 2004. 408−419.

[12] Li WS, Clifton C. Semantic integration in heterogeneous databases using neural networks. In: Bocca JB, Jarke M, eds. Proc. of the 20th Int'l Conf. on Very Large Data Bases. San Francisco: Morgan Kaufmann Publishers, 1994. 1−12.

[13] Wu W, Yu C, Doan A, Meng WY. An interactive clustering-based approach to integrating source query interfaces on the deep Web. In: Weikum G, Konig AC, DeBloch S, eds. Proc. of the 2004 SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2004. 95−106.

[14] Dhamankar R, Lee Y, DoanAH, Halevy A, Domingos P. iMAP: Discovering complex semantic matches between database schemas. In: Weikum G, Konig AC, DeBloch S, eds. Proc. of the 2004 SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2004. 383−394.

申德荣(1964－),女,辽宁沈阳人,博士,教授,CCF 高级会员,主要研究领域为 Web 数据处理,分布式数据库.



余恩运(1982－),男,硕士,主要研究领域为数据集成.

张旭(1982－),男,硕士,主要研究领域为 Web 数据管理.

寇月(1980－),女,博士生,CCF 会员,主要研究领域为 Web 数据管理.

聂铁铮(1980－),男,博士生,CCF 会员,主要研究领域为分布式数据处理.

于戈(1962－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据流,数据挖掘,分布式数据库.