SMGM:一种基于模式结构和已有匹配知识的模式匹配模型*)

余恩运 申德荣 张 旭 王广奇 于 戈
(东北大学计算机科学系 沈阳 110004)

摘 要 模式匹配是确定模式间语义匹配关系的技术,它在许多应用中起着重要的作用,如数据集成中异构模式信息整合、本体知识映射、电子商务中消息映射等。针对已有模式匹配方法的局限性,本着最大限度地减少人工干预使模式匹配自动化的原则,本文提出一种利用模式结构信息和已有匹配知识的模式匹配模型 SMGM。它借鉴神经网络元间影响作用过程实现语义匹配推理;通过重用已有匹配知识,补充、精化匹配知识,自动缩减不确定阈值区间;并给出一种自适应式迭代挖掘求精已有匹配知识的自学习型模式匹配模型。实验表明:SMGM 模型切实可行。

关键词 模式匹配,重用,阈值区间,推理,神经元网络,自学习

SMGM:One Schema Matching Model Based on Schema Structures and Known Matching Knowledge

YU En-Yun SHEN De-Rong ZHANG Xu WANG Guang-Qi YU Ge
(Department of Computer Science, Northeast University, Shenyang 110004)

Abstract Schema matching is the task of finding semantic correspondences between elements of two schemas. It is critical in many applications, such as data integration, data warehouse loading and XML message mapping, etc. Against the limitations of existed schema matching methods, with the aim of reducing the amount of user effort as much as possible to automatic schema matching, based on the schema structure information and known matching knowledge, we propose a novel approach to schema matching method called SMGM. It imitates the influence procedure between neurons to realize the semantic matching reasoning. By reusing the known matching knowledge to supplement and dive the matching knowledge and curtail the uncertain threshold interval automatically, and presented a self learning schema matching model which can mine and dive the known matching knowledge adaptively and iterately. The result of our experiment shows that the SMGM is feasible.

Keywords Schema matching, Reuse, Threshold interval, Inference, Neural network, Self learning

1 前言

模式匹配是获取不同模式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。如数据挖掘中正确地挖掘模式间的语义映射关系;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等。目前,有关模式匹配的研究[1]典型地为基于有限的待匹配的模式信息的模式匹配和基于模式信息集的模式匹配两类。

第一类具有代表性的研究工作主要有:Microsoft 的 Cupid[2]、Washington 大学的 GLUE 方法[3]、LSD[4]、Leipzig 大学的 COMA 方法[5]、Stanford 大学的 Similarity Flooding[6]方法和 IBM 的 Clio 方法[7]。此类方法的优点是能对模式自身携带的有限的结构信息进行深入挖掘以获取模式匹配关系。但由于模式自身携带语义信息的有限性决定了匹配结果的局限性。

第二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:Corpus 方法[8]、使用统计分析方法对大量模式知识进行挖掘分析的模式匹配方法[9]、SemInt 方法[10]和 Duplicates 方法[11]。第二类方法的优点是针对大量模式知识进行分析处理,挖掘出隐藏较深并具有普遍意义的匹配知识。主要不足有:(1)正确的匹配训练数据集不好获取;(2)一对多型匹配关系的发掘规则难以确定;(3)只能针对特定领域进行处理,不具有通用性;(4)阈值难于确定。

本文结合近年来已有模式匹配理论的相关性研究,针对可参考知识匮乏所导致的匹配结果的脆弱性、模式信息变动导致的匹配知识的无效性,以及模式匹配的智能性等问题进行研究,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(SMGM)。该模型首先基于模式结构进行初始匹配,之后重复利用已知匹配知识深入挖掘模式元素语义匹配关系,目的是合理地构造一种灵活、高效且可扩展的匹配模型,并最大限度地减少人工参与。

2 模型概览

SMGM 模型(Schema Matching Graph Based Model)借鉴神经元理论,采用智能推理机制,结合启发式思想,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。SMGM 模型如图 1 所示,主要由初始匹配矩阵模型、结构化语义推理模型、已知匹配知识重用模型、匹配知识自适应迭代模型、阈值确定策略和匹配类型选择策略组成。首先参照辅助信息库将模式元素分离成词条向量,之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。结构化语义匹配推理模型是 SMGM 模型的核心,是模拟神经元间影响作用的语义推导模型。其基于匹配知识重用模型重用历史知识,基于高收敛阈值确定模型策略确定最终阈值,并基于匹配类型选择策略确定模式元素之间的最终匹配关系,进而获得最终的匹配结果。模式匹配知识库中的模式匹配信息被组织成图结构,基于自适应迭代模型,对模

*)基金项目:该课题得到国家 863 项目(编号:2003AA414210)资助,国家自然科学基金(编号:60573090)。余恩运 硕士研究生,主要研究方向为数据网络,Web 信息处理。申德荣 教授,博士,主要研究方向为分布式系统,数据网络以及 Web 服务技术。于 戈 教授,博士生导师,主要研究方向为数据库技术,数据挖掘,Web 技术等。

式匹配知识进行精化和深入挖掘，为结构化语义匹配推理和阈值区间自动缩减等操作提供快速准确的指导。

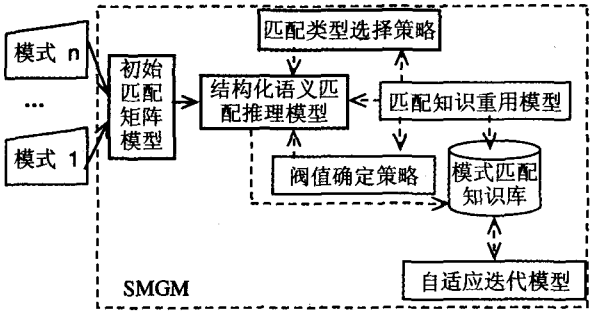

图 1　SMGM 模型整体框架

3　匹配影响作用模型

通过分析大量模式间的匹配关系，并结合 Cupid 和 SF 中的设计思想，可总结出匹配对间的影响规则如下：模式 Sx、Sy 中候选匹配对〈a，b〉在受到 Sx、Sy 间的其它候选相似匹配对的影响的同时，也反过来影响其余候选匹配对。

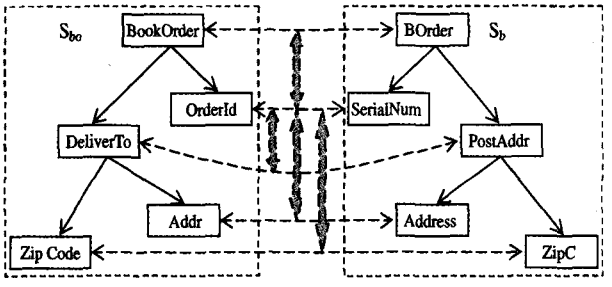

图 2　模式 S_{bo}、S_b 间匹配关系示意图

基于以上影响规则，我们提出了一种模拟神经元间相互影响的结构化语义匹配推理模型。其主要思想为候选匹配对 L 与其相容的且相似度值大于特定阈值的相似候选匹配对之间彼此互相影响。采用迭代插值方法的迭代过程描述为：

$$Sim_k^{r+1}=Normal(Sim_k^r+f(\sum_{i\in nbrNode\&\&i!=k}^{i}(Sim_i^r*w_i),\lambda))$$

其中，Sim_k^r 代表第 k 对相似匹配对第 r 次迭代的相似匹配度值，w_i 表示第 i 对匹配对对第 k 对相似匹配对的影响加权因子，取值为第 i 对相似匹配对与第 k 对相似匹配对间跨越的相似匹配结点对个数加 1 的倒数，如图 2 中〈OrderId，SerialNum〉与〈DeliverTo，PostAddr〉间隔了匹配对〈BookOrder，Border〉，其影响作用因子为 $1/(1+1)=0.5$。

$f(i,\lambda)=\begin{cases} i & i\geq\lambda \\ 0 & i<\lambda \end{cases}$ 表示只有候选匹配对相似度值大于输出阈值时，才对其余的候选相似匹配对产生迭代影响。

$$Normal(Sim_j(x,y))=\frac{Sim_j(x,y)}{\max\limits_{i\in N}(Sim_i(x,y))}$$

用于归一化相似度。当所有节点对满足 $sim_{i+1}-sim_i<\varepsilon$（$\varepsilon$ 取经验值 0.05），或迭代次数超过 100 次时，迭代停止。

4　重用已有匹配知识模型

本文在结构化语义匹配推理模型基础上，进一步重用已有模式匹配知识，从而挖掘模式元素的深层次的语义关系，以提高匹配效果。主要使用以下三种重用方式：(1)匹配关系重用：利用已有确定性匹配关系的传递性进行匹配关系推理。若已知匹配对：〈S_a：r，S_1：a〉，〈S_1：a，S_2：b〉，〈S_2：b，S_3：c〉…〈S_{n-1}：x：S_n：y〉，〈S_n：y，S_b：t〉，则 Sim〈S_a：r，S_b：t〉=[Sim〈S_a：r，S_1：a〉+Sim〈S_1：a，S_2：b〉+…+Sim〈S_n：y，S_b：t〉]/(n+1)。如图 3，已知〈S_1：r，S_2：i〉和〈S_2：i，S_3：t〉的相似度值分别为 0.92 和 0.94，则可推得〈S_1：r，S_3：t〉的相似度值为(0.92+0.94)/2=0.93。(2)不匹配关系重用。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〈S_1：x，S_2：y〉和〈S_3：w，S_2：v〉为已知确定性匹配关系对，但由于 S_2：y 和 S_2：v 属于同一模式的不同词条，可知 S_1：x 与 S_3：w 肯定为不匹配关系。〈S_1：ab，S_3：bc〉和〈S_3：de，S_2：ae〉都属于匹配关系对，但由于 S_3：bc 和 S_3：de 为一个模式中的不同标签，故它们所代表的含义必然不同，由此可知，即使 S_1：ab 与 S_2：ae 在初始相似矩阵中值不为 0，也不必再进行语义匹配推理。(3)混合型重用。结合“匹配关系重用”和“不匹配关系重用”的混合型重用方式。如图 4，设〈S_1：ab，S_3：bc〉为相似关系，〈S_2：ae，S_4：ef〉、〈S_4：ef，S_3：fg〉为相似关系，但 S_3：bc 与 S_3：fg 为同一模式中的不同标签，则可知{S_1：ab，S_3：bc}与{S_2：ae，S_4：ef，S_3：fg}间没有语义匹配推理的必要。需要指出的是，模式元素间相似关系的传递性要求模式元素的含义具有单一性。

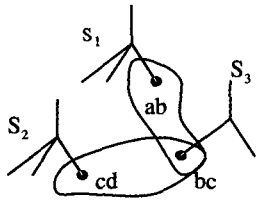

图 3　重用举例一

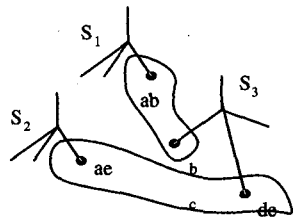

图 4　重用举例二

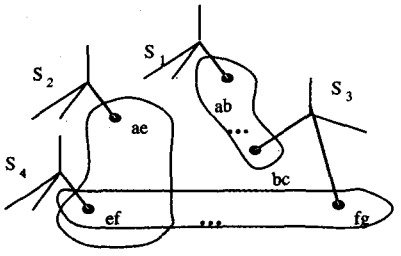

图 5　重用举例三

重用已有匹配知识模型为结构化语义匹配推理模型、阈值确定模型提供基础支持，提高了阈值确定的速度和模式元素匹配关系的准确性、全面性和自动性。

5　实验分析

(1)实验环境。P4 2.6G，512M RAM，80G DISC。数据集来源：http://metaquerier.cs.uiuc.edu/repository/和 http://www-db.stanford.edu/~melnik/mm/sfa/。

（下转封四）

(上接第 169 页)

(2)评价标准。借用信息检索中经典的评估方法。设模式 S1 和模式 S2 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 $F=P-T$,没有发掘出来的正确匹配对数量为 $M=R-P$。定义查准率:Precison$=T/P$,查全率:Recall$=T/R$,综合考虑二者 Overall$=1-(F+M)/R=$Recall$*(2-1/$Precision)。

(3)实验结果。

实验 1 不重用历史匹配知识。仅采用模式结构进行语义匹配关系推理,结果如图 6 所示。实验表明 Book 和 computer 领域的实验结果数据较好,Auto 领域相对较差。实验表明:1)本模型在结构匹配推理模块能力不逊于 SF 算法;2)模式结构越复杂,匹配效果越好。

实验 2 重用已有匹配结果。即对已有模式匹配处理的模式信息进行聚集整合,作为后续模式匹配任务的辅助信息,以协助进行匹配推理。实验效果如图 7 所示。结果表明:本模型定义的重用已有模式匹配知识的方法可大幅提高模式匹配效果。

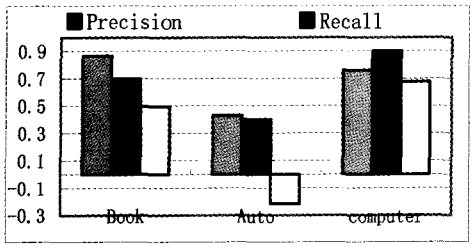

图 6 没有重用已知匹配结果的试验效果

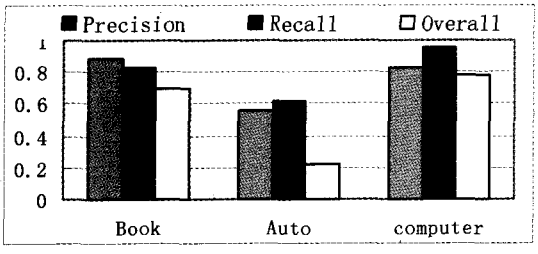

图 7 重用已知匹配结果的试验结果

实验 3 针对 Auto 领域的模式匹配,当模式数量增多时模式匹配的各项评价指标,如图 8 所示。可以看出,已知匹配知识越多,对当前模式匹配的辅助指导作用就越大,匹配结果也越好。

总之,本模型达到了高效重用已有模式匹配关系的初衷,且随着已有匹配知识的丰富,重用效果越好,模式匹配结果也越精确。

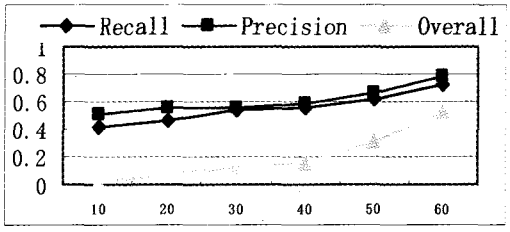

图 8 模式数量增多时各项指标的变化

总结及未来工作 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(SMGM),通过借鉴神经元影响作用过程,实现了模式匹配自适应性迭代求精;通过高效利用已有匹配知识,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;给出了多对多型模式匹配关系解决策略;应用自适应迭代模型,使已有模式匹配知识更准确和完善。

参 考 文 献

1 Rahm B, Bernstein P A. A survey of approaches to automatic schema matching. VLDB Journal, 2001
2 Madhavanl J, Bernstein P A. Rahml E. Generic Schema Matching with Cupid. VLDB, 2001
3 Doan A, Madhavan J, Dhamankar R, et al. Learning to map ontologies on the semantic web. In: Very Large Databases Journal, Special Issue on the Semantic Web, 2003
4 Melnik S, Garcia-Molina H, Rahm E. Similarity Flooding : A Versatile Graph Matching Algorithm. ICDE, 2002
5 Do Hong-Hai, Rahm E. COMA - A system for flexible combination of schema matching approaches. ACM VLDB, 2002
6 Melnik S, Garcia-Molina H, Rahm E. Similarity Flooding : A Versatile Graph Matching Algorithm. ICDE, 2002
7 Miller R J, et al. The Clio project: Managing heterogeneity. SIGMOD Record, 2001, 30(1):78~83
8 Madhavan J. Corpus-based Schema Matching. ICDE, 2005
9 He Bin. Statistical Schema Matching across Web Query Interfaces. SIGMOD, 2003
10 Li W S, Clifton C. Semantic integration in heterogeneous databases using neural networks. In: Proceedings of VLDB, 1994
11 Bilke A. Schema Matching using Duplicates. ICDE, 2005

计算机科学

(1974 年 1 月创刊)
第 34 卷第 03 期 (月刊)
2007 年 3 月 25 日出版

国际标准连续出版物号 ISSN 1002-137X
国内统一连续物出版号 CN50—1075/TP

定价:30.00 元 国外定价:5 美元
邮发代号:78—68
发行范围:国内外公开

主管单位:国家科学技术部
主办单位:国家科技部西南信息中心
编辑出版:《计算机科学》杂志社
重庆市渝中区胜利路 132 号 邮政编码:400013
电话:(023) 63500828 E-mail:jsjkx@sw
网址:www.jsjkx.com
社 长:牟炳林
总 编:彭 丹
主 编:朱宗元
主编助理:徐书令
印 刷 者:重庆科情印务有限公司
总发行处:重庆市邮政局
订 购 处:全国各地邮政局
国外总发行:中国国际图书贸易总公司(北京 399 信箱)
国外代号:6210—MO