基于马尔科夫逻辑网络的实体解析改进算法

楼俊杰 徐从富 郝春亮
(浙江大学人工智能研究所 杭州 310027)

摘 要 实体解析(Entity Resolution,ER)是数据挖掘过程中关键而又费时的一个步骤。华盛顿大学的Domingos和Singla提出了基于马尔科夫逻辑网络(Markov Logic Networks,MLNs)的ER算法。基于此算法,在原有的MLNs体系中,引入了一个可变权重的规则,试图解决原有系统无法处理的实体二义性问题。实验证明,新算法能够有效缓解数据记录的二义性问题,并且在一定程度上提高了原始算法的精度。

关键词 ER,MLNs,可变权重
中图法分类号 TP181 **文献标识码** A

Improvement of Entity Resolution Based on Markov Logic Networks

LOU Jun-jie XU Cong-fu HAO Chun-liang
(Artificial Intelligence Research Institute, Zhejiang University, Hangzhou 310027, China)

Abstract Entity Resolution is a crucial and expensive step in the data mining process. Domingos and Singla of University of Washington proposed of well-founded, integrated solution to the entity resolution problem based on Markov Logic. This paper tried to improve Domingos and Singla's solution by adding a formula with a changeable weight to it, to handle the problem of ambiguity of entities that the original system cannot distinguish. The new algorithm can effectively handle ambiguity of entities, and improve accuracy compared with the original algorithm, which is proved by experiments.

Keywords ER, MLNs, Changeable weight

1 引言

实体解析(Entity Resolution,ER)是数据挖掘、信息融合等领域的难点问题。通常,多个数据源发送的记录都包含了同一组实体集的信息,但它们往往未用统一的标识符来表示这些实体。例如,数据源A用John Smith表示某公司的一个雇员,而数据源B用J. Smith表示该公司的同一个雇员。ER算法试图寻找这些指代同一个实体的标识符,并把相应的记录整合在一起。如何正确地整合这些记录,从而为后续的数据处理(如挖掘、融合等)提供高质量的数据,是一个必需且困难的工作。近年来,ER问题及ER算法已引起国内外人工智能、数据挖掘、信息融合等领域的学者们的高度关注[1,2]。

1959年Newcombe等人首次提出ER问题[3]。1969年Fellegi和Sunter给出其统计学表述形式[4],提出了Fellegi-Sunter模型,该模型将ER问题视为传统的分类问题:给出两个实体符号各自的属性集,据此判断这两个实体符号是否指代同一个实体,这种判断也称为匹配决策。对于一组实体符号来说,每个需要鉴定的候选实体符号对,比如{(John Smith, Address: 上海, Tel: 84954356, Birthday: 09/02/1971), (J. Smith, Address: Shanghai, Tel: 021-84954356, Birthday: 1971年9月2日)},都需要进行匹配决策,并做出相应的调整。然后,应用传递闭包性和Logistic回归(Logistic Regression,LR)模型[5],以保证数据记录的一致性。当前,ER研究领域有两个重要研究方向:一是如何在大型数据库中避免进行平方级的判断次数[6-9];二是运用主动学习(Active Learning,AL)技术减少对已知分类数据的依赖[10-12]。同时,许多研究者为ER问题设计了若干相似性度量方法,并进行了比较研究[13-15]。此外,一些研究者还提出了区别于Fellegi-Sunter模型的其它统计学表述形式[16]。如今,ER算法已在许多领域得到了重视和应用[17,18],且应用于文本[19]和图像[20]等不同类型的数据集中。

近年来有人认为,孤立地对数据库中的一个记录进行匹配决策是不可取的[21-23]。换言之,将数据库中的实体记录都视作独立且均匀分布(independent and identically distributed, i.i.d)的观点是错误的。事实上,关系数据库中的记录之间并非孤立,而是存在内在联系。虽然这种内在联系增加了推理和学习的复杂度,但是通过利用这些以前被忽视的数据属性,可有效地改进传统的ER算法。不难发现,上述Fellegi-Sunter模型即是孤立地进行决策匹配,因此,学者们提出了一些替代该模型的方法。Domingos和Singla[21],Dong[22],Culotta

到稿日期:2009-09-23 返修日期:2009-12-09 本文受国家自然科学基金(No. 60970081),国家863计划专题课题(No. 2007AA01Z197),"十一五"国防预研项目资助。

楼俊杰(1982—),男,硕士生,主要研究方向为人工智能、机器学习等,E-mail:junjie.lou@163.com;徐从富(1969—),男,博士,副教授,硕士生导师,主要研究方向为人工智能、机器学习、数据挖掘、信息融合等;郝春亮(1986—),男,硕士生,主要研究方向为人工智能、机器学习等。

和 McCallum[23]都考虑到了实体之间的内在联系，且能从一种类型的实体(如 Name)解析中获取对另一种相关实体(如 Birthday)解析的有用信息。例如，若经过匹配决策得到 John Smith 和 J. Smith 是指同一人，则其生日记录必然相同；类似地，生日记录相同也可作为住址可能相同的证据。2006 年，Singla 和 Domingos[24]基于马尔科夫逻辑网络(Markov Logic Networks，MLNs)理论，提出了一套简洁而完备的 ER 问题解决模型，它吸收了统计关系学习(Statistical Relational Learning，SRL)领域中的最新研究成果[25]，从而为非独立且非均匀分布(non i. i. d)类型的数据提供了高效的推理和学习算法。然而，该模型未能解决有关实体二义性的问题。例如，当两条记录中出现的"John Smith"其实并非指代同一人时，该模型就会产生错误结果，而这种错误是不能容忍的。本文基于 Singla 和 Domingos 提出的 MLNs 理论及 ER 算法，引入新规则及对应的权重，在一定程度上解决了实体解析中存在的二义性问题。

本文第 2 节简要介绍了马尔科夫网络及马尔科夫逻辑网络的基本概念，以及推理和学习算法；第 3 节分析了 Singla 和 Domingos 提出的基于马尔科夫逻辑网络的实体解析算法；第 4 节论述了对 Singla 和 Domingos 实体解析算法的具体改进；第 5 节给出并分析了改进算法的实验结果；最后，指出了未来的研究工作。

2 马尔科夫逻辑网络

2.1 MNs

马尔科夫网络(Markov Networks，MNs)也称马尔科夫随机场(Markov Random Field，MRF)[26]，是一组变量集合 $X=(X_1,X_2,\cdots,X_n)\in\chi$ 联合分布率的一种表示模型。它是由一个无向图 G 和定义于 G 上的一组势函数 ϕ_k 组成。无向图的每一个节点都代表一个随机变量。对于无向图中每个"团"(Clique)都相应地定义了一个势函数，用以表示该团的一个状态，它是一个非负实函数。MNs 所代表的变量集的联合分布率表示如下：

$$P(X=x)=\frac{1}{Z}\prod_k\phi_k(x_{\{k\}}) \tag{1}$$

式中，$x_{\{k\}}$ 表示 MNs 中第 k 个团的状态，即对应于第 k 个团中，所有变量的取值状态。Z 是归一化因子，且 $Z=\sum_{x\in\chi}\prod_k\phi_k(x_{\{k\}})$。通常，在概率图模型(Probabilistic Graphical Models，PGMs)中式(1)表示为对数线性模型。若把 MNs 中每个团的势函数表示为 e 的 n 次方，其中 n 为对应团的加权特征量，可得：

$$P(X=x)=\frac{1}{Z}\exp\{\sum_j\omega_j f_j(x)\} \tag{2}$$

在 MNs 中，最常用的近似推理算法是马尔科夫链蒙特卡洛(Markov Chain Monte Carlo，MCMC)方法[27]，其代表是吉布斯采样(Gibbs Sampling)。当需要计算边缘概率时，比如 $P(X=x)$，仅需要在所有抽到的样本中计算满足 $X=x$ 的样本数 m，然后除以样本总数 n。当需要计算条件概率时，比如 $P(X=x|Y=y)$，就需要先设置变量 Y 的值为 y，然后进行采样。

2.2 MLNs

2004 年 Richardson 和 Domingos 提出了 MLNs，并论证了 MLNs 可作为统计关系学习的统一框架[34]。当前，国际人工智能界普遍公认 MLNs 是一种简单且较完美地结合一阶谓词逻辑和概率图模型的逻辑结构表达方式，具有重要的研究价值和广阔的应用前景，已成为人工智能、机器学习、数据挖掘等领域的研究热点。在国内，对 MLNs 理论及其应用的研究才刚刚起步[35,36]。

MLNs 的定义如下。

定义 1[28] 马尔科夫逻辑网络 L 是一组二元项 (F_i,ω_i)，其中 F_i 表示一阶逻辑规则，ω_i 是一个实数。这组二元项 (F_i,ω_i) 和有限常量集 $C=\{c_1,c_2,\cdots,c_n\}$ 定义了一个如下的马尔科夫逻辑网络 $M_{L,C}$：

(1) L 中出现的任意一个原子的任何可能的常量取值(grounding of atom)，都对应了 $M_{L,C}$ 中的一个二值节点。若此常值原子为真，则对应的二值节点取值为 1；若为假，则取值为 0。

(2) L 中任意一个规则 F_i 的任何可能的常量取值(grounding of formula)，都对应着一个特征值，若此常值规则为真，则对应的特征量为 1；若为假，则为 0。并且这个特征量的权重为二元项中该规则 F_i 对应的权重 ω_i。

从上述定义和式(1)、式(2)可知，一个常量 MNs 中所蕴含的可能域 x 的概率分布为：

$$P(X=x)=\frac{1}{Z}\exp\{\sum_i\omega_i n_i(x)\}=\frac{1}{Z}\prod_i\phi_i(x_{\{i\}})^{n_i(x)} \tag{3}$$

式中，$n_i(x)$ 表示关于规则 F_i 的取值为真的对应常量规则的个数，$X_{\{i\}}$ 表示出现在规则 F_i 中的原子集合的状态，并且 $\phi(x_{\{i\}})=e^{\omega_i}$。式(3)的第一个等式给出的是 MLNs 的对数线性模型，而式(3)的第二个等式采用了势函数乘积的形式表示 MLNs，这两种方式是等价的。

2.3 MLNs 的推理以及学习

最大可能性问题是 MLNs 推理过程中涉及的重要内容，可通过应用一个加权可满足性解决器(如 MaxWalkSAT[29])来高效地解决。此外，计算边缘概率和条件概率是 MLNs 推理过程中涉及的另一重要内容，通常采用吉布斯采样(Gibbs Sampling)[30,31]方法来解决。

给定一组规则，其权重可由两种方式学习得到[28]，即伪最大似然估计学习和区别式训练(Discriminative training)[32]。

假定将训练 MLNs 的数据库分割成两个集合，即证据谓词 X 和查询谓词 Y，那么，通过对条件概率：

$$P_\omega(y|x)=\frac{1}{Z_x}\exp(\sum_{i\in F_Y}\omega_i n_i(x,y)) \tag{4}$$

求最大似然来学习权值 ω_i。对式(4)的条件对数似然函数(CLL)求偏导，可得：

$$\frac{\partial}{\partial\omega_i}\log P_\omega(X=x)=n_i(x,y)-E_\omega[n_i(x,y)] \tag{5}$$

式中，$n_i(x,y)$ 是数据库中第 i 个从句的常量从句的真值个数，$E_\omega[n_i(x,y)]$ 为在所有可能的数据域中，基于当前权重向量 $\omega=\{\omega_1,\omega_2,\cdots,\omega_i,\cdots\}$，对第 i 个从句的常量从句的真值个数的期望。MLNs 的结构可用任意的归纳逻辑程序设计(Inductive Logic Programming，ILP)技术来学习或者改进[33]。

3 基于 MLNs 的实体解析过程

一般地，MLNs 都有"名字唯一性"的假设[31]，即在 MLNs 中不同的常量代表不同的对象实体，然而，该假设可通

过引入一个“等价”谓词来移除[28]。该等价谓词及其性质如下。

等价谓词 Equals(x,y)

等价谓词的性质如下：

自反性：$\forall x, x=x$ (6)

对称性：$\forall x,y, x=y \Rightarrow y=x$ (7)

传递性：$\forall x,y,z, x=y \cap y=z \Rightarrow x=z$ (8)

谓词等价：对于任意一个二元谓词 R，

$\forall x_1,x_2,y_1,y_2,R,(x_1=x_2 \cap y_1=y_2) \Rightarrow (R(x_1,y_1) \Leftrightarrow R(x_2,y_2))$ (9)

将上述 4 个规则加入到任意一个 MLNs 体系中，并且赋予这 4 个规则无穷大的权重，那么，该 MLNs 体系将可以处理“名字不唯一”的情况。在文献[24]中，额外引入了 2 个逆反谓词等价规则，即

$\forall x_1,x_2,y_1,y_2,R,(R(x_1,y_1) \Leftrightarrow R(x_2,y_2)) \Rightarrow (x_1=x_2 \cap y_1=y_2)$ (10)

式(10)等价于下面 2 个规则：

$\forall x_1,x_2,y_1,y_2,R,R(x_1,y_1) \cap R(x_2,y_2) \cap x_1=x_2 \Rightarrow y_1=y_2$ (11)

$\forall x_1,x_2,y_1,y_2,R,R(x_1,y_1) \cap R(x_2,y_2) \cap y_1=y_2 \Rightarrow x_1=x_2$ (12)

在通常的一阶谓词逻辑体系中，式(10)－式(12)在逻辑上是不成立的。然而在 MLNs 体系中，如果给式(10)－式(12)加上适当的权重，那么这 3 个式子是合理的，因为它们阐述了一个很重要的统计学规律：如果 2 个实体符号 a,b，跟同一个实体符号 c 有着相同的关系，即 $R(c,a),R(c,b)$，那么 a 和 b 则有可能代表同一个实体。式(6)－式(12)是基于 MLNs 的实体解析过程中所用到的全部必需规则。尽管看起来很简单，但是这种体系却包含了许多目前业已成熟的实体解析算法的必要特征[24]，包括 McCallum 和 Wellner 以及 Singla 和 Domingos 的“集体推理(collective inference)”。

上述问题可通过计算条件概率或边缘概率来解决，从而确定查询谓词即等价谓词 Equal()，而证据谓词则是关系数据库中所有显性和隐性的关系，记为集合 R。集合 R 可以通过 MLNs 的结构学习或者手动添加得到。本文基于 Cora 数据库上的实验，其中的集合 R 为手动添加，MLNs 中的规则权重通过区别式训练学习得到。假设关系数据库中共有 n 种类型的实体(比如 Person 和 Paper 是两种类型的实体)，且记 $x_i[j]$为第 i 种类型的其中一个实体，Equal$_i$()为第 i 种类型的等价谓词。进而为每一个等价谓词生成单位从句[28](此处即为等价谓词加权重)，该单位从句的权重等于第 i 种类型的等价谓词 Equal$_i$()在关系数据库中成立的边缘概率，记为 P_i。那么，通过计算 $P(\text{Equal}_i(x_i[j],x_i[k]) \mid R,M_{L,c})$或者 $P(\text{Equal}_i(x_i[j],x_i[k]) \mid M_{L,c})$，可以得到 $x_i[j]=x_i[k]$的条件概率或边缘概率，记为 p_i。当 $p_i \geqslant P_i$ 时，可以认为 $x_i[j]$和 $x_i[k]$代表的是第 i 种类型的同一个实体。计算条件或边缘概率需得到常量原子的真值分布情况，如 2.3 节所述，采用 MaxWalkSAT 来取得最可能的原子真值分布，并且通过吉布斯采样得到上述的边缘概率或者条件概率。值得注意的是，此处所有涉及的关系或谓词都是二元的，因为多元关系或谓词总是可分解成若干个二元关系或谓词。

4 基于 MLNs 的实体解析算法改进

上述基于 MLNs 的实体解析过程有一个隐含的假设，亦即关系数据库中任意两个相同的实体符号表示的是同一个对象实体。比如，在一个公司的客户关系数据库中，两个记录中出现的 John Smith 表示的是同一个人。然而，实际上两个记录中出现的 John Smith 很有可能不是同一个人。例如，下面两条记录了客户名和住址的记录(John Smith, Shanghai)，(John Smith, Beijing)。当然，上面两条记录中的 John Smith 还是有可能为同一人，因为 John Smith 在上海和北京可能都有房产。上述基于 MLNs 的实体解析过程将完全不能识别实体符号的二义性，即如两个 John Smith 其实指代不同的人。

对于上述现象，需要注意的是：(1)涉及二义性的实体符号很少，即大多数的实体符号都没有二义性；(2)通常，要区分两个同样的实体符号是否有二义性，可以从这二个实体符号涉及的实体所拥有的实体属性来看。然而，如果关系数据库中并没有这些属性，并且因为两个实体符号完全相同，那么就必须由实体与实体之间的关系来近似推断。由于整个 MLNs 的体系是建立在一阶逻辑规则加权的形式之上，为了区分实体符号的二义性，加入如下新规则并学习对应的权重：

$\forall x_1,y_1,y_2,R,R(x_1,y_1) \cap R(x_2,y_2) \cap y_1 \neq y_2 \Rightarrow xs$ *represent different entities* (13)

式(13)在一阶逻辑中是不成立的，因为同一个实体当然可以跟同类型不同的实体对象发生关系。这是因为现实世界中的关系是复杂的，通常两种实体间是多对多的映射关系。而在一个关系数据库中，在封闭的前提下，即不知道两个实体在关系数据库之外是什么关系，有可能实体间一对一或多对一的映射关系并不一定是少数。基于这种假设，引入式(13)。这也反映了 MLNs 是一个容错性较强，甚至可以容忍矛盾的体系。

式(13)的权重可变化，如 $f_i(w_i)=k_i w_i$，其中 w_i 是经过区别式训练得到的权重，i 为关系索引，且当 $i \neq j$ 时，$R_i \neq R_j$。对于该规则组权重的学习，区别式训练的收敛速度要比其它学习算法快得多，因为二义性实体通常占少数。而 k_i 则是相对于关系 R_i 的性质而动态变化，其选取方法为：若关系 R_i 是一对一或多对一的关系，则 k_i 取值为 1；否则，取小于 1 的正实数。对于关系 R_i 来说，参数 k_i 的取值反映了关系数据库中一对一或多对一关系的数量、对于整个体系约束力的大小以及多对多关系的数量对整个体系的影响。由于关系数据库中大多数关系都对应于 $0<k_i<1$ 的情况，因此，有必要对 $0<k_i<1$ 时的 k_i 的取值进行进一步调整。而 $k_i=1$ 则表现为每个一对一或多对一关系对于体系的影响力的基准。对于 $f(w)$的结构和 k 的取值问题，还有待未来进一步研究。

由上所述，根据关系数据库以及式(6)－式(13)生成 MLNs 的完整结构并学习规则相应的权重(MLNs 的结构也可手动写出)。表 1 给出了判断两条记录中的实体符号是否有二义性的完整算法。

表 1 区分二义性算法

步骤 1	视两条记录中的 x 为代表不同实体的实体符号，并分别记作 $x_i[1]$，$x_i[2]$；(假设 x 是第 i 种类型的实体)
步骤 2	在未加入式(13)的 MLNs 结构中，计算 $P(\text{Equal}_i(x_i[1],x_i[2]) \mid R,M_{L,c})$，记为 p_x，且计算 $P(\text{Equal}_i() = \text{true} \mid R,M_{L,c})$，记为 P_i；

步骤 3　若 $p_x<P_i$，则这两条记录中的 x 成为候选对，进入步骤 4。若 $p_x>P_i$，则判定 x 没有二义性；

步骤 4　加入式(13)，计算 $p'(Equal_i(x_i[1],x_i[2]) \mid R,M_{L,c})$，记为 p'_x，且计算 $P'(Equal_i() = true \mid R,M_{L,c})$，记为 P'_i；

步骤 5　若 $p'_x+p_x<1$ && $p'_x<P'_i$，则认为 x 是有二义性的实体；否则，判定 x 无二义性。

记原始算法的 ER 处理过程为 ER(relational database)，原始关系数据库为 RDB，并且记表 1 的过程为 Dis(x)。由表 1 得到基于 MLNs 的实体解析改进算法，如表 2 所列。

表 2　基于 MLNs 的 ER 改进算法

步骤 1　RDB ← ER(RDB)；
步骤 2　对所有受到怀疑的两条记录中的 x，执行 Dis(x)，更新 RDB；
步骤 3　RDB ← ER(RDB)。

表 2 的步骤 3，是在区分二义性实体以后进行的，期待进一步提高 ER 的精度。实验证明，步骤 3 是必要的。

5　实验结果及分析

5.1　实验结果

实验目标是在基于 MLNs 的 ER 过程[24]基础上，区分原本未能识别的实体二义性，并进一步提高原始 ER 算法[24]结果的精度。

在 Cora[1] 数据库中，抽取了 1000 条形如(author, venue, title, publisher, year)的记录，并且将其中的 L 条记录手工改写为跟其它$(1000-L)$条记录至少有一个域重名的记录，这 L 条记录跟原来的记录集中的一条或若干条有歧义，即二义性实体密度 $\theta\approx 0.1L\%$。实验平台为 Alchemy 系统[2]，该系统是华盛顿大学 Domingos 研究小组 MLNs 的实验平台。下面给出两个实验过程及其结果。

实验 1　对基于原始 ER 算法处理后的关系数据库进行计算，在经过二义性区分算法(见表 1)时，能识别的二义性实体的百分率以及它和二义性实体在关系数据库中的密度 θ 和 k 之间的关系。

在实验 1 中，调整了多对多关系所对应的 $k\in(0,1)$的取值。根据 k 的不同取值以及二义性实体密度 θ 的大小，得到如表 3 所列的实验结果。

表 3　实验 2 中的总体识别百分率

	k=1/10	k=1/25	k=1/5
θ=1%	87.7%	87.9%	87.6%
θ=5%	88.0%	87.9%	88.1%
θ=15%	88.2%	89.7%	90.0%
θ=25%	88.7%	88.6%	88.5%

由图 1 可知，关系数据库中多对多关系的比例对识别二义性实体的影响在 $\theta\approx 15\%$的附近区间影响较大。而随着 θ 的增大，可识别的二义性实体数量呈现抛物线形态分布。

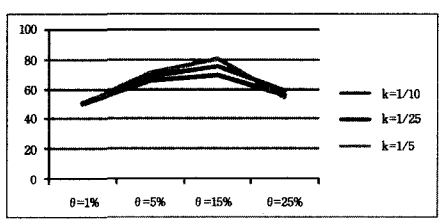

图 1　二义性实体识别百分率(纵轴)与 θ (横轴)及 k 的关系

[1] www.cs.umass.edu/~mccallnum/data/cora-refs.tar.gz

[2] http://alchemy.cs.washington.edu

实验 2　根据表 2 中所给出的基于 MLNs 的 ER 改进算法，得到改进算法总体 ER 识别率(二义性实体识别率不计算在内)。其中，原始算法的总体识别率为 87.50%[24]。

根据 k 的不同取值以及二义性实体密度 θ 的大小得到如表 3 所列的实验结果。由表 3 可知，改进算法的总体识别率比原始算法有所提高。

5.2　实验结果分析

从实验 1 的结果看，首先表 1 中所述的算法是切实有效的，它可以识别原先所不能识别的二义性实体。其次二义性实体的识别精度受到关系数据库中二义性实体密度和 k 的取值的影响。k 的取值实际上代表了关系数据库中，非二义性实体对识别二义性实体的影响。而这种影响实际上是有正负的。因此可以看到，随着 k 的增长，识别精度并非呈线性变化。而二义性密度 θ 的取值，直接影响到式(13)的权重的大小。并且由图 2 的曲线形状可以看出，θ 对总体识别精度的影响也是非线性的。

从实验 2 的结果来看，表 2 的改进算法相对于原始算法，在 ER 识别结果上，即在判定"$x_1=x_2$?，其中 x_1，x_2 在字面上不同"的精度上，也有所提高。这说明表 2 中步骤 3 的确是必要的。而 ER 识别结果的精度的提高程度，大体上随着总体二义性实体的识别率的提高而增大。

结束语　未来基于 MLNs 的 ER 算法二义性实体识别的研究方向包括如下几个方面：(1)改进式(13)或者引入另外新的规则；(2)更加合理地定义 $f(x)$的结构以及对 k 的取值进行更科学的调整；(3)改进表 2 中步骤 2，高效地确定受到怀疑的候选记录对。

参 考 文 献

[1] McCallum A, Tejada S, Quass D. Mining knowledge from text using information extraction[C]//Proc. of the KDD-2003 Workshop on Data Cleaning, RecordLinkage, and Object Consolidation. ACM Press, 2003

[2] Gehrke J, Ginsparg P, Kleinberg J. KDD cup 2003[EB/OL]. http://www.cs.cornell.edu/projects/kddcup

[3] Newcombe H, Kennedy J, Axford S, et al. Automatic linkage of vital records[J]. Science, 1959, 130: 954-959

[4] Fellegi I, Sunter A. A theory for record linkage[J]. American Statistical Association, 1969, 64: 1183-1210

[5] Agresti A. Categorical Data Analysis[M]. Wiley, New York, NY, 1990

[6] Hernandez M, Stolfo S. The merge/purge problem for large databases[C]//Proc. SIGMOD-95. 1995: 127-138

[7] Monge A, Elkan C. An efficient domain-independent algorithm for detecting approximately duplicate database records[C]//Proc. SIGMOD-97 DMKD Workshop. 1997

[8] McCallum A, Nigam K, Ungar L. Efficient clustering of high-dimensional data sets with application to reference matching[C]//Proc. KDD-00. 2000: 169-178

[9] Cohen W, Richman J. Learning to match and cluster large high-dimensional data sets for data integration[C]//Proc. KDD-02. 2002: 475-480

[10] Tejada S, Knoblock C, Minton S. Learning object identification rules for information integration[J]. Information Systems, 2001, 26(8): 607-633

[11] Sarawagi S, Bhamidipaty A. Interactive deduplication using active learning[C]// Proc. KDD-02. 2002:269-278

[12] Bilenko M, Mooney R. On evaluation and training-set construction for duplicate detection[C]// Proc. KDD-03 Workshop on Data Cleaning, Record Linkage, and Object Consolidation. 2003: 7-12

[13] Cohen W, Ravikumar P, Fienberg S. A comparison of string metrics for matching names and records[C]// Proc. KDD-03 Workshop on Data Cleaning, Record Linkage, and Object Consolidation. 2003:13-18

[14] Tejada S, Knoblock C, Minton S. Learning domain-independent string transformation weights for high accuracy object identification[C]// Proc. KDD-02. 2002:350-359

[15] Bilenko M, Mooney R. Adaptive duplicate detection using learnable string similarity measures[C]// Proc. KDD-03. 2003:39-48

[16] Cohen W, Kautz H, McAllester D. Hardening soft information sources[C]// Proc. KDD-00. 2000:255-259

[17] Noren G, Orre R, Bate A. A hit-miss model for duplicate detection in the WHO Drug safety Database[C]// Proc. KDD-05. Chicago, IL, 2005:459-468

[18] Davis J, Dutra I, Page D, et al. Establishing identity equivalence in multi-relational domains[C]// Proc. ICIA-05. 2005

[19] Li X, Morie P, Roth D. Semantic integration in text: from ambiguous names to identifiable entities[J]. AI Magazine, 2005, 26 (1):45-58

[20] Huang T, Russell S. Object identification : a Bayesian analysis with application to traffic surveillance[J]. Artificial Intelligence, 1998, 103(1/2):77-93

[21] Singla P, Domingos P. Object identification with attribute-mediated dependences[C]// Proc. PKDD-05. Porto, Portugal, 2005: 297-308

[22] Dong X, Halevy A, Madhavan J. Reference reconciliation in complex information spaces[C]// Proc. SIGMOD-05. 2005:85-96

[23] Culotta A, McCallum A. Joint deduplication of multiple record types in relational data[C]// Proc. CIKM-05. 2005:257-258

[24] Singla P, Domingos P. Entity resolution with Markov logic[C]// Proc. of the 6th IEEE International Conference on Data Mining (ICDM2006). Hong Kong, China, December 2006:572-582

[25] Dzeroski S, Blockeel H, et al. Multi-Relational Data Mining 2004: Workshop Report[C]// Proc. of the KDD-04 Workshop on Multi-Relational Data Mining. Chicago, IL, 2004:140-141

[26] Jordan M I. Graphical models [J]. Statistical Science (Special Issue on Bayesian Statistics), 2004, 19(1):140-155

[27] Baader F, Calvanese D, McGuinness D L, et al. The Description Logic Handbook: Theory, Implementation, Applications [M]. Cambridge, UK: Cambridge University Press, 2003

[28] Richardson M, Domingos P. Markov logic networks[J]. Machine Learning, 2006, 62(1/2):107-136

[29] Gu D, Du J, Pardalos P. The Satisfiability Problem: Theory and Applications[M]. American Mathematical Society, New York, NY, 1997:573-586

[30] Gilks W R, Richardson S, Spiegelhalter D J. Markov Chain Monte Carlo in Practice [M]. London, UK: Chapman and Hall, 1996

[31] Richardson M, Domingos P. Markov logic networks[J]. Machine Learning, 62(1/2):107-136, 2006

[32] Singla P, Domingos P. Discriminative training of Markov logic networks[C]// Proc. AAAI-05. Pittsburgh, PA, 2005:868-873

[33] Kok S, Domingos P. Learning the structure of Markov logic networks[C]// Proc. of the 22nd International Conference on Machine Learning (ICML2005). Bonn, Germany, August 2005: 441-448

[34] Richardson M, Domingos P. Markov logic: a unifying framework for statistical relational learning[C]// Proc. of the ICML-2004 Workshop on Statistical Relational Learning and its Connections to Other Fields. Banff, Alberta, Canada, July 2004:49-54

[35] 刘大有，于鹏，高滢，等. 统计关系学习研究进展[J]. 计算机研究与发展，2008，45(12)

[36] 孙舒杨，刘大有，孙成敏，等. 统计关系学习模型 Markov 逻辑网综述[J]. 计算机应用研究，2007，24(2)

(上接第 213 页)

名规范无法实现对 XML 文档进行联合签名，从而提出了 XML 联合签名的实现方案；并以现有的 XML 加密规范以及 XML 签名规范为基础，为 XML 联合签名进行了语法定义，包括语法结构、XML Schema 以及处理规则。

在下一步工作中，将基于 Eclipse 环境来实现 XML 联合签名平台，用于支持 XML 文档中多个信息联合签名的实现。

参 考 文 献

[1] W3C. Extensible Markup Language(XML) 1.0(Fifth Edition) [EB/OL]. http://www.w3.org/TR/2008/REC-xml-20081126/, 2008

[2] W3C. XML Signature Syntax and Processing(Second Edition) [EB/OL]. http://www.w3.org/TR/2008/REC-xmldsig-core-20080610/, 2008

[3] Walmsley P. Definitive XML Schema[M]. Prentice Hall PTR, 2001

[4] Trappe W, Washington L C. Introduction to Cryptography with Coding Theory[M]. 2nd edition. Prentice Hall, 2005

[5] Mason S. Electronic Signatures in Law[M]. second edition. Tottel, 2007

[6] Visa, MasterCard Inc. Secure Electronic Transaction Specification[EB/OL]. Version 1.0. http://www.visa.com/set, May 1997

[7] Kim H K, Kim T H. Design on Mobile Secure Electronic Transaction Protocol with Component Based Development[C]// Lecture Notes in Computer Science, ICCSA. 2004:461-470

[8] 陈乐君，石锐，李初民. 基于 XML 多重签名的电子病历安全机制[J]. 计算机科学，2007，34(12):136-138，170

[9] Knap T, Mlynková I. Towards More Secure Web Services: Pitfalls of Various Approaches to XML Signature Verification Process[C]// Icws. 2009:534-550

[10] Wang Wei, Li Jun. An XML Firewall on Embedded Network Processor[C]// icns. 2008:1-6

[11] Tan K W, Deng R H. Applying Sanitizable Signature to Web-Service-Enabled Business Processes: Going Beyond Integrity Protection[C]// Icws. 2009:67-74