文章编号:1000-582X(2010)08-036-06

Markov 逻辑网在重复数据删除中的应用

张玉芳[a],黄 涛[a],艾东梅[a],熊忠阳[a],唐蓉君[b]
(重庆大学 a.计算机学院;b.网络中心,重庆 400044)

摘 要:为了解决和突破现阶段重复数据删除方法大多只能针对特定领域,孤立地解决问题的某个方面所带来的不足和局限,提出了基于 Markov 逻辑网的统计关系学习方法。该方法可以通过计算一个世界的概率分布来为推理服务,从而可将重复数据删除问题形式化。具体采用了判别式训练的学习算法和 MC-SAT 推理算法,并详细阐述了如何用少量的谓词公式来描述重复数据删除问题中不同方面的本质特征,将 Markov 逻辑表示的各方面组合起来形成各种模型。实验结果表明基于 Markov 逻辑网的重复数据删除方法不但可以涵盖经典的 Fellegi-Sunter 模型,还可以取得比传统的基于聚类算法和基于相似度计算的方法更好的效果,从而为 Markov 逻辑网解决实际问题提供了有效途径。

关键词:重复数据删除;Markov 逻辑网;Markov 网;统计关系学习;机器学习
中图分类号:TP391 **文献标志码**:A

Markov Logic Networks with its application in De-duplication

ZHANG Yu-fang[a], *HUANG Tao*[a], *AI Dong-mei*[a], *XIONG Zhong-yang*[a], *TANG Rong-jun*[b]
(a. College of Computer Science; b. Center of Information and Network,
Chongqing University, Chongqing 400044, P. R. China)

Abstract: In order to solve the limitation that the traditional De-duplications are mostly used for a specific field and only address one aspect of a problem, a scheme based on Markov Logic Networks (MLNs) is proposed, which is a new Statistical Relational Learning (SRL) model. With its advantage of computing the probability distribution of worlds to serve for the inference, the De-duplication is formalized. Discriminative learning algorithm is adopted for Markov Logic Networks weights, MC-SAT algorithm is adopted for inference. It shows how to capture the essential features of different aspects in De-duplication with a small number of predicate rules and also combines these rules together to compose all kinds of model. The experiment results prove that the method based on Markov Logic Networks not only covers the original Fellegi-Sunter model, but also achieves a better result than the traditional methods based on Clustering Algorithms and Similarity Measures in De-duplication. It reveals that the Markov Logic Networks can play an important part in practical application.

Key words: de-duplication; markov logic networks; markov networks; statistical relational learning; machine learning

---

收稿日期:2010-01-02
基金项目:重庆市自然科学基金资助项目(CSTC 2008BB2021);中国博士后科学基金资助项目(20070420711)
作者简介:张玉芳(1965-),女,重庆大学副教授,主要从事数据挖掘、网络入侵检测方向研究,(Tel)15826106116;(E-mail)ram.tntht@gmail.com。

统计关系学习（statistical relational learning，SRL）又称概率逻辑学习（probabilistic logical learning，PLL），是人工智能、机器学习和数据挖掘交叉研究的产物，旨在将关系（逻辑）表示、似然推理（不确定性处理）和机器学习（数据挖掘）结合起来，以获取关系数据中的似然模型[1]。统计关系学习方法由似然关系模型和学习算法组成。似然关系模型是指关系的似然表示形式，学习是指基于数据来调整似然关系模型的过程，学习分为参数学习和结构学习 2 个过程。统计关系学习的早期研究多集中于归纳逻辑程序设计（inductive logic programming，ILP）[2]，随着对 SRL 研究的不断深入，陆续提出许多非 ILP 的统计关系学习方法，Markov 逻辑网就是基于 Markov 网的 SRL 方法。

重复数据删除（de-duplication）问题首次由 NewCombe 等人提出[3]，用于判断数据库中哪些记录指代的是同一个记录，将冗余的重复记录进行删除。之后 Fellegi 和 Sunter 给出了统计学描述[4]，现在的大多数方法都是基于 Fellegi-Sunter 模型的，重复数据删除问题可视为二分类问题。常用的方法有罗杰斯特回归模型[5]，贝叶斯算法[6]，记录链[7]，概率模型[8]等，这些方法主要集中在姓名的比较上。此外，还有采用主动学习技术[9]，利用相似度计算[10]，聚类算法[11-14]等方法。但大多是针对特定领域的应用，只能孤立地解决问题的某一方面。给出了基于 Markov 逻辑网的统计关系学习方法，将重复数据删除问题形式化，并用 Markov 逻辑表示。实验基于 Cora 数据集[11]，采用五折交叉验证（5-fold cross-validation），结果表明 Markov 逻辑网用于解决重复数据删除问题是行之有效的。

1　Markov 逻辑网简介

Markov 逻辑网最早是由华盛顿大学 Domingos 等人提出[12]，之后经过 Domingos、Kok、Singla 等人进一步完善。Markov 逻辑网是公式附加权值的一阶逻辑知识库，且可作为构建 Markov 网的模板[12]。从概率的角度来看，Markov 逻辑网为大型 Markov 网提供一种简洁的描述语言，并能够灵活地将大量领域知识采用模块化的形式引入到 Markov 网中。从一阶逻辑的角度来看，Markov 逻辑网不仅可以处理不确定性，还可以允许不完整和矛盾的知识。

在一阶逻辑中，通常一个世界只要违反了一个公式，该世界发生的概率就为 0（即一阶逻辑知识库作为可能世界的强约束）。Markov 逻辑网的基本思想是软化这个约束：当一个世界违反了知识库中的一个公式，原本在一阶逻辑中不可能发生的世界，在 Markov 逻辑网中可能发生，但只是发生的概率降低了。而违反的公式越少，则发生的概率越大。公式上的权值体现了该公式的限制强度，权值越大，满足该公式的世界的发生概率与不满足该公式的世界的发生概率之间的差就越大。随着公式上权值的增加，Markov 逻辑网逐渐向纯逻辑知识库靠拢。

Markov 逻辑网的算法分为 2 类[15]，一类是学习算法，一类是推理算法。学习算法主要有伪似然估计方法[10]（maximum pseudo-likelihood estimation，MPLE）和判别式训练方法[11]（discriminative training）2 种。推理算法[10]主要采用马尔科夫链蒙特卡洛方法（markov chain monte carlo，MCMC），该方法采用不同的转移核将导致不同的 MCMC 方法。文献[15]，对这 2 类算法及其相关的方法进行了详细介绍。在这里采用的 MC-SAT[10]是一种切片抽样的 MCMC 算法。

2　重复数据删除问题的 Markov 逻辑表示

2.1　Markov 逻辑中的各种等价关系

大多数一阶逻辑的推理系统都有唯一名称假设，即不同的个体常项代表了不同的对象。这一假设可通过引入等价谓词（Equals（x，y）或简写为 $x = y$）来消除，该谓词满足以下逻辑公式

自反性：$\forall x \quad x = x$；

对称性：$\forall x, y \quad x = y \Rightarrow y = x$；

传递性：$\forall x, y, z \quad x = y \wedge y = z \Rightarrow x = z$；

谓词等价：对于每个二元谓词 R 有 $\forall x_1, x_2, y_1, y_2; x_1 = x_2 \wedge y_1 = y_2 \Rightarrow (R(x_1, y_1) \Leftrightarrow R(x_2, y_2))$，对于多元谓词有类似的公理。

反向谓词等价：对于每个二元谓词 R 有 $\forall x_1, x_2, y_1, y_2; R(x_1, y_1) \wedge R(x_2, y_2) \Rightarrow (x_1 = x_2 \Leftrightarrow y_1 = y_2)$，该公式可以转换为下面 2 个子句的形式：$\forall x_1, x_2, y_1, y_2; R(x_1, y_1) \wedge R(x_2, y_2) \wedge x_1 = x_2 \Rightarrow y_1 = y_2$ 和 $\forall x_1, x_2, y_1, y_2; R(x_1, y_1) \wedge R(x_2, y_2) \wedge y_1 = y_2 \Rightarrow x_1 = x_2$。

在一阶逻辑中，这些公式可能是错误的，因为同一谓词的不同闭谓词无法表示对象的相同元组。然而，在 Markov 逻辑网中则不矛盾。附加权值的一阶逻辑公式，反映了重要的统计规律：如果 2 个对象在某种关系上是相同的，那它们有可能就是同一个对象。例如，有 2 篇重复的引文，它们的标题相同，在一阶逻辑中无法得到“标题相同则引文相同”这样的结论；然而在 Markov 逻辑网中，则会对“标题相同则引文相同”赋予权值，用来反映统计上“标题相同则引文有可能相同”这一规律。

2.2 De-duplication 问题的 Markov 逻辑表示

为简单起见，假定 Markov 逻辑表示中仅包含二元关系。这一假设并不失一般性，因为一个 n 元关系可以表示成 n 个二元关系。例如，引文数据库中包含谓词 Paper(title, author, venue)，可以通过 Hastitle(paper, title)、HasAuthor(paper, author)和 HasVenue(paper, venue)来替代。假定知识库中的谓词是有类型的，比如 HasAuthor(paper, author)的第一个类型 Paper、第二个类型 Author。De-duplication 的目标就是对同类型的 2 个常量 (x_1, x_2) 判断是否表示同一个实体，即判断是否 $x_1 = x_2$。给出了针对引文 De-duplication 的谓词定义和不同情况下的谓词公式。

2.2.1 谓词定义

根据引文的特征，定义了 4 种类型，即 Bible(文献)、Author(作者)、Title(标题)、Venue(来源)，其中作者、标题、来源也为数据库中的字段(Field)。根据这 4 种类型，则有以下 3 种非查询谓词

Author(bib, author)：表示文献 bib 的作者是 author；

Title(bib, title)：表示文献 bib 的标题是 title；

Venue(bib, venue)：表示文献 bib 的来源是 venue；

De-duplication 的任务就是判断以下查询谓词的真假值：

SameBib(bib1, bib2)：表示文献 bib1 和 bib2 是同一篇文献；

SameAuthor(author1, author2)：表示作者 author1 和 author2 是同一作者；

SameTitle(title1, title2)：表示标题 title1 和 title2 是同一标题；

SameVenue(venue1, venue2)：表示来源 venue1 和 venue2 是同一来源；

此外，假定 3 种字段的值都可分为一个或多个字符串，则对不同字段可定义非查询谓词如下

HasWordAuthor(author, word)：表示作者 author 包含字符串 word；

HasWordTitle(title, word)：表示标题 title 包含字符串 word；

HasWordVenue(venue, word)：表示来源 venue 包含字符串 word；

谓词定义部分简记为 P。

2.2.2 单元子句

单元子句(unit clause)是仅包含一个文字的子句，也叫单一谓词规则(single predicate rules)。单元子句的权值可以大致上反映相应谓词的边缘概率分布，而非单元句子则反映谓词间的依赖关系。由于查询谓词的大多数闭谓词真值为假(例如，大多数文献都不是相同文献)，所以使用单元子句的否定形式。即

! SameBib(b1, b2)

! SameAuthor(a1, a2)

! SameTitle(t1, t2)

! SameVenue(v1, v2)

“!”表示否定，将单一谓词规则部分简记为 S。

2.2.3 传递闭包规则

针对不同的查询谓词，可得到如下 4 种传递闭包规则：

SameBib(b1, b2) ^ SameBib(b2, b3) => SameBib(b1, b3)；

SameAuthor(a1, a2) ^ SameAuthor(a2, a3) => SameAuthor(a1, a3)；

SameTitle(t1, t2) ^ SameTitle(t2, t3) => SameTitle(t1, t3)；

SameVenue(v1, v2) ^ SameVenue(v2, v3) => SameVenue(v1, v3)。

将传递闭包规则部分简记为 T。

2.2.4 反向谓词等价规则

根据上述反向谓词等价可知，该类规则是基于二元谓词的，故对于不同的查询谓词可以得到不同种类的规则。若 2 篇文献相同，则作者、标题、来源等字段也有可能相同。故可定义如下规则：Author(b1, a1) ^ Author(b2, a2) ^ SameBib(b1, b2) => SameAuthor(a1, a2)。其他字段类似作者字段，由于该类规则得到不同字段的查询谓词，故将该部分简记为 CF。

此外，若作者、标题、来源等字段相同，则也有可能是同一篇文献。类似的可定义如下规则：Author(bc1, a1) ^ Author(bc2, a2) ^ SameAuthor(a1, a2) => SameBib(bc1, bc2)。其他字段类似作者字段，该类规则得到是否为同一文献的查询谓词，故将该部分简记为 CB。

2.2.5 字段比较规则

根据谓词定义，对于不同字段有谓词 HasWord(field, word)，表示某个字段包含某个词。对于该二元谓词应用反向谓词等价，则有：HasWordAuthor(a1, w) ^ HasWordAuthor(a2, w) => SameAuthor(a1, a2)。该规则表示若作者 a1 包含词 w，且作者 a2 也包含词 w，则 a1、a2 有可能是同一作者。

换言之，含有相同词的字段值可能是同一字段值。反向谓词等价应用到二元谓词 HasWord()上，

在某种程度上实现了字段之间的相似度比较，与 Bilenko 文中提到的方法[10]类似。此外，可以给 HasWord( ) 添加否定形式，有如下规则：! HasWordAuthor(a1, w) ^ HasWordAuthor(a2, w) => SameAuthor(a1, a2)和 HasWordAuthor(a1, w) ^ ! HasWordAuthor(a2, w) => SameAuthor(a1, a2)。该规则表示作者 a1 不包含词 w，作者 a2 包含词 w，而 a1、a2 仍有可能是同一作者。若否定形式的规则学习到的权值低(表示该规则成立的可能性小)，而非否定形式的规则学习到的权值高(表示该规则成立的可能性大)，则可进一步提高推理的准确性。若否定形式的规则学习到的权值也高，则表明该词 w 对于判断“是否为同一作者”没有太大贡献。

其他字段的规则类似作者字段，将字段比较规则部分简记为 FR。

2.2.6　Fellegi-Sunter 模型

Fellegi-Sunter 模型[4]采用的是朴素贝叶斯算法，通过字段之间的比较来预测 2 个记录是否相同。如果预测的是字段之间的匹配，二元谓词 R 采用 HasWord()的形式，则 Fellegi-Sunter 模型是反向谓词等价规则的特例。若二元谓词 R 表示字段和记录之间的关系，即 HasAuthor(paper, author)这种形式，则 Fellegi-Sunter 模型可以表示为如下谓词公式：Author(b1, a1) ^ Author(b2, a2) ^ HasWordAuthor(a1,w) ^ HasWordAuthor(a2,w) => SameBib(b1, b2)。其他字段类似作者字段，将 Fellegi-Sunter 模型部分简记为 FS。

3　实　验

3.1　数据集

实验采用了预标注的 Cora 数据集，由 McCallum[11]提供，且有 Bilenko[10]等人使用过。该数据集收集了来自 Cora 计算机科学论文引擎的 1295 条计算机科学研究方面的引文。最初的数据集包含的是未分字段的英文字符串，Bilenko 使用信息抽取系统将每条引文分为几个字段(Author、Title、Venue、Year 等)，这里只采用 3 个最重要的字段：作者、标题、来源(来源包含会议、期刊、学位论文等)，下面给出了 Cora 数据集中的一条引文示例。

引文各字段：

authors: Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby;

title: Information, prediction, and query by committee;

venue: Advances in Neural Information Processing System;

3.2　模型

根据描述的各种 Markov 逻辑表示，实验采用了以下几种模型。

3.2.1　基本模型

P+S+FS：因为 Markov 逻辑网的学习过程就是为每个公式学习权值，而含有 FS 的各种模型，因为包含了 HasWord 的谓词，所以会对该谓词中每个词生成一个公式，然后学习权值。该模型也即 Fellegi-Sunter 模型(朴素贝叶斯模型)。

P+S+FR：公式针对 SameAuthor、SameTitle、SameVenue 3 种谓词，同样因为包含了 HasWord 的谓词，所以会对该谓词中每个词生成一个公式，然后学习权值。对每个词学习公式的权值。

P+S+FR+FS：结合了以上 2 种基本模型，理论上效果会优于以上 2 种情况。

3.2.2　扩展模型

在 3 种基本模型的基础上，分别添加 2 种不同的反向谓词等价规则，形成以下几种扩展模型：

P+S+FS+CB

P+S+FR+CB

P+S+FS+CF

P+S+FR+CF

P+S+FR+FS+CB

P+S+FR+FS+CF

P+S+FR+FS+CB+CF。

此外，还可通过添加传递闭包规则形成如下 2 种模型：

P+S+FR+FS+T

P+S+FR+FS+CB+CF+T。

3.3　实验方法

Markov 逻辑网相关的权值学习和概率推理在 Alchemy[13]下进行。Alchemy 是 Domingos 等人开发的基于 Markov 逻辑表示的软件包，提供了相关的统计关系学习和概率逻辑推理方面的算法。实验具体过程分为以下几个步骤：

1)数据集预处理，将 1295 条引文随机分为 5 个部分，各部分的情况如表 1 所示。

属性值的重复表示某个属性下，看似不同的文献可能是同一文献。比如针对 Cora1 部分，有 259 篇文献，假设编号 Bib1、Bib2、…、Bib259。若 Bib1、Bib2、Bib3、Bib4 是相同文献，则重复文献数就是 12(Bib1、Bib2 和 Bib2、Bib1 认为是不同的)。表 1 是根据随机分成的 5 部分数据库得到的。

表1 数据集中引文各部分情况

数目 \ 数据	Cora1	Cora2	Cora3	Cora4	Cora5
文献	259	267	262	263	244
重复文献	5231	6593	7054	5651	6442
作者	43	28	35	34	34
重复作者	209	122	151	158	172
标题	62	42	41	45	50
重复标题	162	140	113	129	154
来源	94	97	97	105	78
重复来源	368	771	581	747	550

2)De-duplication 的 Markov 逻辑表示。即上节提到的各种模型,根据给出的表示,构建不同的.mln 文件。

3)知识库构建。根据谓词定义,生成包含各种不同谓词的知识库。

4)权值学习。采用判别式训练方法。

5)谓词推理。根据上一步学习到的权值和测试知识库,采用 MC-SAT 算法进行推理。

3.4 结果分析

表2给出了不同模型下,引文及各个字段识别结果的 F_1 值;图1是不同模型下,引文及各个字段识别结果的 F_1 值比较情况。

表2 不同模型下的重复数据删除结果

模型 \ 字段	Bib	作者	标题	Venue
P+S+FS	0.889	0.916	0.532	0.434
P+S+FR	0.631	0.980	0.896	0.571
P+S+FR+FS	0.901	0.984	0.917	0.653
P+S+FS+CB	0.895	0.911	0.820	0.632
P+S+FR+CB	0.720	0.977	0.749	0.598
P+S+FS+CF	0.865	0.971	0.741	0.606
P+S+FR+CF	0.739	0.990	0.824	0.552
P+S+FR+FS+CB	0.857	0.973	0.895	0.593
P+S+FR+FS+CF	0.788	0.972	0.799	0.550
P+S+FR+FS+CB+CF	0.642	0.911	0.647	0.511
P+S+FR+FS+T	0.656	0.929	0.754	0.444
P+S+FR+FS+CB+CF+T	0.645	0.917	0.561	0.479

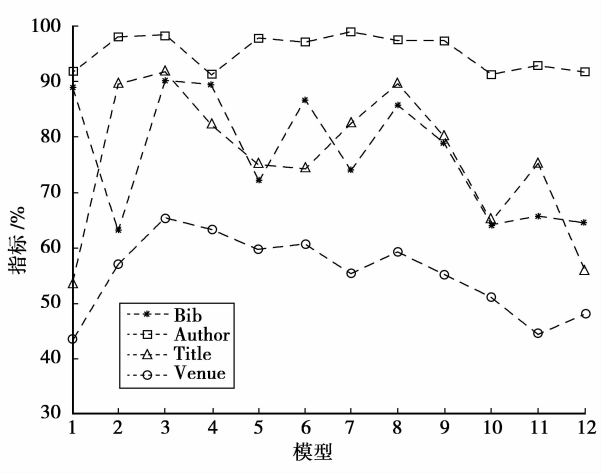

图1 不同模型下的重复数据删除结果对比图

上图横坐标表示不同模型,标号对应模型如下:

1. P+S+FS;2. P+S+FR;3. P+S+FR+FS;4. P+S+FS+CB;5. P+S+FR+CB;6. P+S+FS+CF;7. P+S+FR+CF;8. P+S+FR+FS+CB;9. P+S+FR+FS+CF;10. P+S+FR+FS+CB+CF;11. P+S+FR+FS+T;12. P+S+FR+FS+CB+CF+T。

为了进一步验证,将其分别与文献[10]基于相似度计算的方法以及文献[14]基于聚类算法的方法进行对比。在表3中,给出了 Markov 逻辑网与文献[10]中各种基于相似度计算的方法针对文献识别效果的对比情况。

表3 Markov 逻辑网与基于相似度计算方法的重复数据删除效果对比

方法	F1 值
编辑距离(Edit distance)	0.793
可学习编辑距离(Learned edit distance)	0.824
向量空间(Vector space)	0.867
可学习向量空间(Learned Vector space)	0.803
Markov 逻辑网模型(P+S+FR+FS 模型)	0.901

在文献[14]中,采取了模块化的架构实现重复数据删除,利用聚类算法对可能重复的记录聚类,在记录集上产生若干个聚类,对同一个聚类中的记录进行比较,以推断和检测出重复的记录,主要使用了 Partitioning, CENTER, MERGE-CENTER 这3种聚类算法。通过设定不同的阈值反复实验后,将最佳实验结果与基于 Markov 逻辑网的重复数据删除方法比较,结果如图2所示。

根据以上结果分析可知

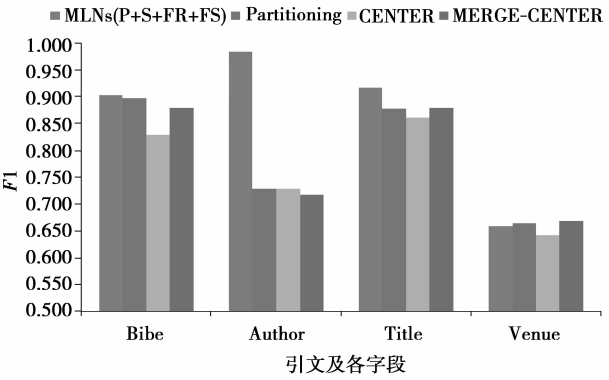

图 2　Markov 逻辑网与基于聚类算法的重复数据删除效果对比

1)由于 FS 是针对文献的识别，FR 是针对不同字段的识别。故基本模型中，P＋S＋FS 的文献识别效果要高于 P＋S＋FR，但字段识别要差。而 P＋S＋FR＋FS 模型综合两者，效果最佳；

2)由于 CB 是针对文献的识别，CF 是针对不同字段的识别。故在 P＋S＋FS 和 P＋S＋FR2 种模型的基础上添加 CB 或 CF 对文献或字段的识别略有提高；

3)在 P＋S＋FR＋FS 模型上添加 CB 或 CF，识别效果反而会降低。因为 FR 和 FS 是针对每个词学习公式的权值，而 CB 和 CF 则是针对不同实体进行学习，后者学习到的权值对推理的影响大于前者，从而导致识别效果的降低；

4)添加传递闭包规则后，效果同样降低，原因同上；

5)同一数据集下，基于 Markov 逻辑网的方法在重复数据删除应用中效果明显优于文献[10]中的几种基于相似度计算的方法。

6)同一数据集下，基于 Markov 逻辑网的方法对 Bibe、Author、Title 字段的识别效果明显优于文献[14]中基于聚类算法的方法，但对 Venue 字段的识别效果稍弱。这是由于存在一些很相似但关联不同实体的 Venue，聚类算法通过设定阈值可以将其分离；而 Markov 逻辑网在进行推理时可能会误判为重复记录，从而导致识别效果降低。

4　结　论

在统计关系学习中，逻辑(关系)可以用来很好地表示知识，故 De-duplication 问题的 Markov 逻辑表示十分的简洁。将 Markov 逻辑网应用到引文 De-duplication 中，给出了如何用少量的谓词公式来描述问题不同方面的本质特征，并将各方面的 Markov 逻辑表示组合起来形成各种模型。实验结果表明，将该方法用于重复数据删除，其效果明显优于相似度计算的方法，且可以通过 Markov 逻辑网构建此类问题的统一框架。下一步工作考虑将该方法应用到 De-duplication 的其他领域，如中文引文 De-duplication、模式匹配、本体匹配等。

参考文献：

[1] DE R L, KERSTING K. Probabilistic logic learning [J]. ACM-SIGKDD Explorations: Special issue on Multi-Relational Data Mining, 2003, 5(1): 31-48.

[2] DZEROSKI S. Relational data mining [M]. US: Springer, 2005:869-898.

[3] NEWCOMBE H B, KENNEDU J M, AXFORD S J, et al. Automatic linkage of vital records[J]. Science, 1959,130:954-959.

[4] FELLEGI I P, SUNTER A B. A theory for record linkage [J]. Journal of the American Statistical Association,1969, 64(328): 1183-1210.

[5] AGRESTI A. Categorical data analysis(2nd Edition) [M]. NewYork: Wiley, 2002: 372.

[6] HERN M A, STOLFO S J. The merge/purge problem for large databases[J]. IEEE,1995:28-31.

[7] MONGE A, ELKAN C. An efficient domain-independent algorithm for detecting approximately duplicate database records[J]. Tucson, AZ,1997:86-92.

[8] TORVIK V I, WEEBER M, SWANSON D R, et al. A probabilistic similarity metric for Medline records: A model for author name disambiguation[J]. Journal of the American Society for Information Science and Technology, 2005, 56(2): 140-158.

[9] SARAWAGI S, BHAMIDIPATY A. Interactive deduplication using active learning[C]// Edmonton, Alberta. Canada:[s. n.]2002:88-90.

[10] BILENKO M, MOONEY R J. Adaptive duplicate detection using learnable string similarity measures[J]. IEEE, 2003:88-92.

[11] http://www. cs. umass. edu/～ mccallum/data/cora-refs. tar. gz[Z].

[12] RICHARDSON M, DOMINGOS P. markov logic Networks[D]. Seattle, Washington, USA: University of Washington, 2004.

[13] KOK S, SINGLA P, RICHARDSON M, et al. The alchemy system for statistical relational AI [C]// Seattle, WA: Department of Computer Science and Engineering. VSi University of Washington, 2005.

[14] HASSANZADEH O, MILLER R J. Creating probabilistic databases from duplicated data[J]. VLDB J, 2009, 18(5): 1141-1166.

(编辑　侯　湘)