ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.8, August 2008, pp.2149−2158
DOI: 10.3724/SP.J.1001.2008.02149

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

基于混合跳链条件随机场的异构Web记录集成方法*

黄健斌[1,2+], 姬红兵[2], 孙鹤立[3]

[1](西安电子科技大学 计算机学院,陕西 西安 710071)

[2](西安电子科技大学 电子工程学院,陕西 西安 710071)

[3](西安交通大学 计算机科学与技术系,陕西 西安 710049)

Integration of Heterogeneous Web Records Using Mixed Skip-Chain Conditional Random Fields

HUANG Jian-Bin[1,2+], JI Hong-Bing[2], SUN He-Li[3]

[1](School of Computer Science and Technology, Xidian University, Xi'an 710071, China)

[2](School of Electronic Engineering, Xidian University, Xi'an 710071, China)

[3](Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China)

+ Corresponding author: E-mail: jbhuang@xidian.edu.cn

Huang JB, Ji HB, Sun HL. Integration of heterogeneous Web records using mixed skip-chain conditional random fields. *Journal of Software*, 2008,19(8):2149−2158. http://www.jos.org.cn/1000-9825/19/2149.htm

Abstract: An improved sequence labeling model named Mixed Skip-Chain Conditional Random Field is presented to solve the problem of schema matching between semi-structured Web records and relational database. The proposed model can be trained on mixed samples set which consists of labeled samples and unlabeled relational database records to reduce the dependence on manually labeled training data. Moreover, it provides a novel way to incorporate the long-distance dependencies between different state variants. Experimental results using a large number of real-world data collected from diverse domains show that the proposed method can improve the performance of schema matching significantly.

Key words: mixed skip-chain conditional random fields; Web data integration; schema matching

摘 要: 提出了一种混合跳链条件随机场序列统计学习模型,以实现异构 Web 记录与关系数据库的模式匹配.该模型可以在由手工标注样本和关系数据库记录组成的联合样本集上进行训练,减少了对繁琐手工标注样本的依赖.此外,通过在线性链条件随机场模型上增加对跳边的支持,使得该模型能够有效地处理状态变量间的长距离依赖.在多个领域的真实数据集上的实验结果表明,所提出的方法能够显著提高异构 Web 记录语义模式匹配的性能.

关键词: 混合跳链条件随机场;Web 数据集成;模式匹配

中图法分类号: TP393 **文献标识码**: A

随着 WWW 的不断发展,Web 网页中已经存放了涵盖各个领域的大量有价值的信息.Web 记录正是这样一

* Supported by the National Natural Science Foundation of China under Grant No.60202004 (国家自然科学基金); the Doctoral Innovation Foundation of Xidian University of China under Grant No.05013 (西安电子科技大学博士创新基金)

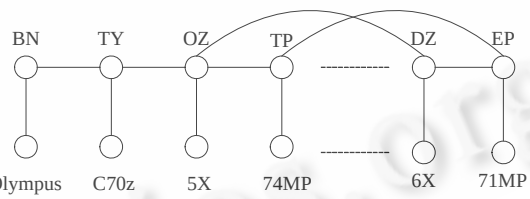
Received 2006-10-14; Accepted 2007-03-08

类由多个数据字段及可选的属性标签按照特定模式组织在一起的半结构化数据对象[1].目前,已经开发出一些能够完全自动地从 HTML 网页中识别并抽取 Web 记录的包装器(wrapper)软件,如 Depta[1],Omini[2],STAVIES[3]等.然而,由于不同 Web 网站通常使用不同的显示模板,使得抽取的 Web 记录具有明显的异构特征,例如,图 1 显示了两种嵌入不同网页的 Web 数码相机记录.将这些从多个网站抽取的描述同一事物的异构 Web 记录集成到同一个关系数据库中是很有意义的,它可以为用户提供访问这些数据的统一接口,进而产生其他增值服务,例如客户信息收集、比价导购、智能问答等.

Fig.1 Heterogeneous digital camera records embedded in different Web pages

图 1 嵌入不同网页中的异构数码相机记录

数据集成的关键问题是异构数据源的模式匹配,这个问题在数据库领域已经研究了多年,提出了许多有效的模式匹配方法[4,5].但 Web 数据集成还是一个新的课题,由于 Web 数据没有固定的模式结构,因此很难找到简单、有效的模式匹配解决方案.Web 记录中的数据元素彼此之间不是孤立的,具有强烈的序列特征,如图 2 所示,网页中数码相机记录的“类型”数据通常紧跟在“品牌”数据之后,而“光学变焦”数据通常在“感光元件”数据的后面显示.McCallum 等人[6−8]提出了一种条件随机场(conditional random field,简称 CRF)序列标注学习模型,利用数据元素的状态和序列特征对序列数据进行语义分割与标注,显示了良好的性能.基于此,本文采用 CRF 模型来解决异构 Web 记录与关系数据库记录的集成问题.



NO	WS	BN	TY	ST	EP	TP	OZ	DZ
23 781	www.itsogo.com	Nikon	COOLPIX S10	CCD	61.8MP	61MP	10X	4X
23 793	www.itsogo.com	Samsung	L70	CCD	74MP	72MP	3X	5X
43 659	www.westdig.com	OLYMPUS	SP510 uz	CCD	71MP	74MP	10X	5X

Fig.2 Integrate Web records in relational database

图 2 集成 Web 记录到关系数据库中

为了解决线性链条件随机场模型需要大量手工标注样本和难以处理数据元素间复杂依赖关系的缺点[9,10],本文提出了一种混合跳链条件随机场模型(mixed skip-chain CRF,简称 MSCRF).该模型通过将最大熵模型(maximum entropy,简称 ME)[11]与线性链 CRF 混合,可以在手工标注样本和无须标注的关系数据记录组成的联合训练样本集上进行训练,从而减少了对手工标注样本的依赖;其次,通过在线性链模型上增加对跳边的支持,使得模型能够处理状态变量间的长距离依赖;最后,使用 Gibbs 抽样近似推理算法,提高了模型的收敛速度.实验结果表明,该模型改进了传统模型的语义标注性能,能有效地解决异构 Web 记录与关系数据库的集成问题.

本文第 1 节介绍混合跳链条件随机场模型及其参数估计和推理方法.第 2 节阐述模型在异构 Web 记录模式匹配中的应用.第 3 节给出综合实验结果及分析.最后总结全文.

1 混合跳链条件随机场模型

1.1 线性链条件随机场

条件随机场是在最大熵模型和隐马尔可夫模型(hidden Markov model,简称 HMM)的基础上提出来的一种判别式概率无向图学习模型,由于使用全局优化技术,它克服了最大熵马尔可夫模型的标注偏置问题,是目前处理序列数据分割与标注问题的最好的统计机器学习模型.条件随机场的一般定义如下:

定义 1. 设 $G=\langle V,E\rangle$是一个无向图,$Y=\{y_v|v\in V\}$是以图 G 中结点 v 为索引的随机变量 y_v 构成的集合.如果每个随机变量 y_v 相对于图 $p(y_v|\{y_w\}_{w\neq v},X)=p(y_v|y_u,X,(u,v)\in E)$服从马尔可夫属性,则称$\langle X,Y\rangle$是一个条件随机场[9].

设 $C=\{\langle x_c,y_c\rangle\}$是图 G 中所有的团构成的集合,根据随机场的基础理论(由 Hammersley 和 Clifford 于 1971 年提出),在给定观测序列 x 的条件下标记序列 y 的概率分布 $p(y|x)$为

$$p_\Lambda(y\mid x)=\frac{1}{Z(x)}\prod_{c\in C}\exp\left(\sum_k\lambda_k f_k(y_c,x_c)\right) \tag{1}$$

其中,$f_k(y_c,x_c)$是特征函数,$Z(x)=\sum_y\prod_{c\in C}\exp(\lambda_k f_k(y_c,x_c))$ 是归一化因子,模型参数是一个由实数构成的特征函数的权值集合$\Lambda=\{\lambda_k\}$.

当用该模型来建模序列数据时,图 $G=\langle V,E\rangle$中状态变量 y 的形状最简单且最常用的是一条一阶链.这条链中的团是其中的结点和边.因此,我们在整个观测序列上可以定义两类特征函数:状态特征函数 $g(i,y_i,x)$和转移特征函数 $f(i,y_{i-1},y_i,x)$.给定训练样本集$\{\langle x^{(k)},y^{(k)}\rangle\}$和预定义的特征函数,可以从样本集中学习一个 CRF 模型.模型参数 Λ可以使用极大似然、极大后验或 Quasi-Newton 等方法[12]估计.

对于一个输入测试序列 x,则可以使用训练得到的 CRF 模型来推断它对应的标注序列,x 最可能的标记序列 $\hat{y}$ 表示为

$$\hat{y}=\arg\max_y p_\Lambda(y\mid x)=\arg\max_y\sum_{c\in C}\sum_k\lambda_k f_k(y_c,x_c) \tag{2}$$

$\hat{y}$ 可以使用与 HMM 中相同的动态编程 Viterbi 算法来查找.

1.2 混合跳链条件随机场

与很多机器学习方法一样,训练一个 CRF 模型需要大量手工标注的样本数据,这无疑是一件非常耗时的工作.然而,存放 Web 记录的关系数据库中往往通过其他途径积累了大量的同类 Web 记录.我们一般认为,关系数据库表中的字段是没有先后次序的,因此无法提取其中的序列特征.但是,每个字段中存放的数据与 Web 记录中的相关数据元素应该有近似的语法模式特征,这些特征有助于区分 Web 记录中不同数据元素的语义类型.为此,我们提出了混合 CRF 模型.它可以同时利用手工标注样本和关系数据记录中的数据特征来辅助新抽取 Web 记录的属性标注.

定义 2. 设 $X=\{\langle x_{i1},x_{i2},\ldots,x_{im}\rangle|i,m\in N\}$是由序列观测数据构成的集合,$R=\{\langle a_{j1},a_{j2},\ldots,a_{jn}\rangle|j,n\in N\}$是模式为 $R(A_1,A_2,\ldots,A_n)$的一个关系,标注集$\alpha=\{A_1,A_2,\ldots,A_n\}$由关系 R 中的所有属性名构成.令 $X'=X\cup R$.在 X 上定义一组转移特征 $F(y',y,x)=\{f_k(i,y_i,y_{i-1},x)\}$并且在 X'上定义一组状态特征 $H(y,x')=\{h_l(i,y_i,x')\}$.这样,得到 CRF 与 ME 的混合模型,该模型的联合概率分布 $p(Y|X,X')$为

$$p_M(Y\mid X,X')=\frac{1}{Z_{\lambda,\mu}(X,X')}\exp(\lambda\cdot F(y',y,x)+\mu\cdot H(y,x')) \tag{3}$$

其中,$Z_{\lambda,\mu}(X,X')=\sum_y\exp(\lambda\cdot F(y',y,x)+\mu\cdot H(y,x'))$ 是归一化因子,$M=\{\lambda,\mu\}$是模型参数.

混合 CRF 利用关系数据库记录作为部分状态观测样本,以减少对手工标注样本的依赖.但是,线性链 CRF 只能建模满足马尔可夫属性的序列数据,即每个标记状态变量仅依赖于其前一个状态变量.对于以 HTML 标签子树的形式嵌入网页中的 Web 记录,数据元素彼此之间存在复杂的次序关系,例如层次邻接关系、兄弟邻接关系等.因此,抽取后的序列数据中存在长距离依赖,不能看作是简单的链状结构.虽然理论上 CRF 模型可以处理

状态变量间的任意图结构.但是,复杂的 CRF 模型又会带来参数估计困难和推理过程的收敛性问题.为此,我们提出了以下跳链 CRF 模型,通过在线性链模型上叠加跳边来处理数据元素间的长距离依赖.

定义 3. 设 $G=\langle x,x',y\rangle$是一个混合条件随机场,x 是序列观测数据,x'是状态观测数据,y 是状态标注序列.如果存在 $y_u,y_v\in y,|u-v|>1$ 使得 y_v 依赖于 y_u,则称边(y_u,y_v)是一条跳边,并称含跳边的混合条件随机场模型为混合跳链条件随机场(mixed skip-chain CRF,简称 MSCRF).令 $I=\{(u,v)\}$是跳边集合,则 MSCRF 的概率分布 $P_S(Y|X,X')$为

$$p_M(Y|X,X')=\frac{1}{Z_{\lambda,\mu,\gamma}(X,X')}\exp(\lambda\cdot F(y',y,x)+\mu\cdot H(y,x'))\prod_{(u,v)\in I}\exp(\gamma\cdot G(y_u,y_v,x)) \tag{4}$$

MSCRF 模型可以看作是线性链 CRF,ME 和跳链 CRF 层叠形成的.图 3 给出了 MSCRF 模型的示意图.理论上,在增加的跳边上可以引入高阶转移特征函数,一个端点的状态标注会影响另一个端点的状态标注.

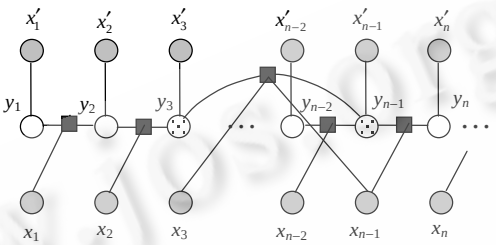


Fig.3 Graphical structure of a MSCRF

图 3 混合跳链条件随机场的结构示意图

1.3 MSCRF模型的参数估计和推理

MSCRF 模型的精确概率分布较难计算,但是,由于我们仅利用所得的模型来进行标注推理,因此可以利用模型的可分解特性,对状态特征函数、一阶转移特征函数和高阶转移特征函数的参数分别进行估计.而将跳链 CRF 模型看作是线性链 CRF 模型和跳链的简单叠加.由于一阶和高阶转移特征函数的估计过程相同,下面仅给出混合 CRF 模型参数估计的过程.

设 $\tilde{p}(x,y)$ 表示 x,y 的经验概率分布,定义对应于 $\tilde{p}(x,y)$ 和条件分布 $p(y|x,x',\Lambda)$的对数似然模型 $L(\Lambda)$为

$$L(\Lambda)=\prod_{x,y}\tilde{p}(x,x',y)\log p(y|x,x',\Lambda)=\tilde{p}(x,y)\sum_k\log p_\lambda(y^{(k)}|x^{(k)})+\tilde{p}(x',y)\sum_n\log p_\mu(y^{(t)}|x'^{(t)}) \tag{5}$$

我们使用该模型对以上提出的混合 CRF 模型的参数$\Lambda=\{\lambda_1,\lambda_2,\ldots,\mu_1,\mu_2,\ldots\}$进行估计.因为函数 $L(\Lambda)$是凹的,它能够收敛到全局最大点,下面对每个参数求偏导数得到其梯度矢量.

$$\frac{\partial L(\Lambda)}{\partial\lambda_k}=\sum_{x,y}\tilde{p}(x,y)\sum_{i=1}^{n}f_k(y_{i-1},y_i,x)-\sum_{x,y}\tilde{p}(x,y)p(y|x,\Lambda)\sum_{i=1}^{n}f_k(y_{i-1},y_i,x)=E_{\tilde{p}(x,y)}[f_k]-E_{p(x,y)}[f_k] \tag{6}$$

$$\frac{\partial L(\Lambda)}{\partial\mu_k}=\sum_{x,y}\tilde{p}(x',y)\sum_{i=1}^{n}h_k(y_i,x')-\sum_{x,y}\tilde{p}(x',y)p(y|x',\Lambda)\sum_{i=1}^{n}h_k(y_i,x')=E_{\tilde{p}(x',y)}[h_t]-E_{p(x',y)}[h_t] \tag{7}$$

当每个偏导数均等于 0 时,函数 $L(\Lambda)$取得最大值.可以看出,每个特征对模型的约束为“特征的样本期望值等于其模型期望值”.

在 CRF 序列标注模型的参数估计中,一般使用二维矩阵组和前向-后向算法来计算特征的期望值与条件概率.设α是标注集,在观测序列 x 中的每个位置 i 上定义一个$|y|\times|y|$的矩阵 $M_i(x)$.当我们对 CRF 模型的参数λ进行估计时,由于只使用了一阶转移特征,于是有

$$M_i[y',y]=\exp\left(\sum_k\lambda_k f_k(i,y',y,x)\right) \tag{8}$$

当对输入观测序列进行测试时,我们将在 x 和 x'上得到的状态观测特征也加入到该矩阵中,得到 $M'_i(y,y')$

$$M'_i[y',y]=M_i[y',y]+\exp\left(\sum_t\mu_t h_t(y,x)\right) \tag{9}$$

此时,归一化函数为

$$Z_{\lambda,\mu}(x,x')=(M_1'(x)M_2'(x)...M_{n+1}'(x)) \tag{10}$$

在以上定义的二维矩阵组上,使用动态编程 Viterbi 算法可以很容易地查找出最优标记状态序列.对任意的标记序列 y 的条件概率为

$$p(y\mid x)=\frac{\prod_{i=1}^{n}M_i'(y_{i-1},y_i\mid x)}{\prod_{i=1}^{n}M_i'(x)} \tag{11}$$

为了计算边界概率,我们在每个位置 i 定义前向矢量$\alpha_i(y|x)$和后向矢量$\beta_i(y|x)$如下

$$\alpha_i(y\mid x)=\begin{cases}\alpha_{i-1}(y\mid x)M_i'(x), & 0<i\le n\\ 1, & i=0\end{cases},\ \beta_i(y\mid x)^T=\begin{cases}M_{i+1}'(x)\beta_{i+1}(y\mid x)^T, & 1\le i<n\\ 1, & i=n\end{cases} \tag{12}$$

2 基于MSCRF的Web记录与关系数据库模式匹配

下面给出特征定义、模型训练以及模式映射的实现细节.

2.1 特征模板的定义

CRF 模型的最大优点是可以学习样本数据中的各种特征来完成对观测序列数据类型的辨别,并且不同特征之间不要求相互独立.因此,模型的标注性能在很大程度上依赖于特征的选取.

特征的实质是一个二值或实值函数,例如$f:(y,x)\to\{0,1\}$.用于训练 MSCRF 模型的特征函数分为 3 类:状态特征、一阶转移特征和高阶转移特征.状态特征用于描述观测序列与某个状态变量之间的关系.例如,对于上述的模式匹配问题,我们可以定义以下状态特征函数:

$$f(y_i,x_i)=\begin{cases}1, & x_i=\text{“2}x\text{”}\wedge y_i=\text{“光学变焦”}\\ 0, & \text{否则}\end{cases}.$$

它表示当观测序列中的元素 x_i 的值为“2x”且状态标记 y_i 的值为“光学变焦”时,特征函数 f 的值为 1;否则为 0.转移特征则用于描述观测序列与某个状态转移之间的关系,这类特征关联两个相邻的状态变量.例如,以下是一个一阶转移特征函数

$$f(y_{i-1},y_i,x_i)=\begin{cases}1, & x_i=\text{“2}x\text{”}\wedge y_{i-1}=\text{“数码变焦”}\wedge y_i=\text{“光学变焦”}\\ 0, & \text{否则}\end{cases}.$$

高阶转移特征与一阶转移特征的唯一不同点在于它关联的是两个非邻接的状态变量.

2.2 模型训练

若定义了 k 个特征,这些特征对概率分布 p 就产生了 k 个约束.接下来的模型训练就是完成一个满足这 k 个约束的最优解问题.我们使用 L-BFGS 算法[12]对模型的参数进行估计,模型的训练过程如下:

(1) 从系统抽取的多网站领域 Web 记录中随机选取若干个记录作为样本数据,手工标注样本数据中每个元素的属性标记,将样本数据保存在磁盘文件中.

(2) 定义模型训练使用的特征模板,将其保存在磁盘文件中.;

(3) 创建一个特征产生器,这个特征产生器可以加载特征模板文件,并顺序读取相应的训练样本数据,识别其中包含的预定义的数据特征.

(4) 创建一个 CRF 模型对象,使用特征产生器中的状态特征和一阶转移特征从训练样本数据中得到一个混合 CRF 模型.

(5) 从手工标注序列样本数据中估计特征产生器中的每个高阶转移特征的参数,保存到模型文件中.

2.3 Web记录模式匹配算法

当我们使用训练得到的 MSCRF 模型来测试输入的 Web 记录时,对于每个测试记录,我们选择条件概率最高的作为输出标注序列,从而得到其中每个数据元素的属性字段标注.具体模式匹配算法如算法 1 所述.

算法 1. 异构 Web 记录与关系数据库的模式匹配.

输入:混合 CRF 模型文件 f,标注集 $\alpha=\{A_1,A_2,\ldots,A_m\}$,Web 记录集合 $X=\{x_1,x_2,\ldots,x_n\}$.

输出:关系记录集 R.

(1) 加载文件 f,建立模型对象 m.

(2) for i=1 to $|X|$

(3) 用对象 m 测试 Web 记录 x_i;

(4) 将 x_i 测试得到的标注序列保存到字符串数组 y 中;

(5) 以 y 为初始标注序列,调用 Gibbs 抽样算法;

(6) 利用得到的最优标注序列 y,将 x_i 中的每个数据元素保存到新增元组 t 的对应分量中;

(7) 将元组 t 插入到关系 R 中;

(8) return R

3 实验评价

3.1 测试数据集

以下是对所提出的方法的性能进行综合评价的真实数据集:

(1) 在线图书数据集(online book dataset,简称 OBD).该数据集包含 20 个在线图书 Web 数据库(http://www.cs.ust.hk/~cswangjy/vldb04_exp.htm),我们从中选择了 2 500 条不同格式的 Web 图书记录,并将其中的 800 条记录录入预先定义的关系数据表中,剩余 1 700 条经手工标注后作为训练和测试用的自由数据.

(2) 数码相机数据集(digital cameras dataset,简称 DCD).该数据集由从 FROOGLE 在线购物网站收集的 4 823 条异构数码相机记录构成(http://froogle.google.com/),预先录入关系数据库表中的记录有 2 000 条,其余 2 823 条记录是自由数据,数据表的模式由手工定义的 19 个基本字段构成.

(3) 论文头部数据集(paper headers dataset,简称 PHD)[8,9,11].该数据集由 935 条论文头部记录构成,是用来评价数据抽取系统的基准数据集之一.学术论文头部是论文首页从标题到论文引言的这部分内容,它包括标题、作者、联系方式、关键词、出版号等 15 个字段.

3.2 评价标准

为了对实验结果进行综合评价,使用了以下多个评价指标:

设 A 表示正确标注的正测试样本数,B 表示错误标注的正测试样本数,C 表示错误标注的负测试样本数,D 表示正确标注的负测试样本数(注:本实验中的测试样本指的是 Web 记录中的数据元素).

(1) 标注精度(precision)、召回率(recall)和 $F1$ 测度,其计算公式分别为

$$Precision=\frac{A}{A+C},\ Recall=\frac{A}{A+B},\ F1=\frac{2\cdot Precision\cdot Recall}{Precision+Recall} \tag{13}$$

(2) 精度、召回率和 $F1$ 的平均值(average).

(3) 标注正确率(accuracy),其计算公式为

$$Accuracy=\frac{A+D}{A+B+C+D} \tag{14}$$

(4) 实例标注正确率(instance accuracy).每个元素均被正确标记的 Web 记录占总的测试记录的比例.

3.3 实验结果与分析

3.3.1 增加数据库记录后的模式匹配结果

我们在 DCD 和 OBD 这两个包含大量异构 Web 记录的数据集上,通过实验分析了在数据库记录参与和不参与两种情况下,对异构 Web 记录模式匹配结果的影响.参与模型训练的样本数据由预先存入数据库中的所有 Web 记录(简称 DB)和随机抽取的 50%自由数据(简称 L)组成,用于测试的是另外 50%的自由数据.实验分为 3 组:第 1 组使用在 DB 样本数据上训练得到的 ME 分类器对测试记录进行属性标注;第 2 组是数据库记录不参与,

用在 L 上训练得到的 SCRF 模型对测试记录进行属性标注;第 3 组使用 MSCRF 模型,它是在 DB 和 L 组成的联合样本集上训练得到的.表 1 给出了在这两个数据集上选择的 13 个典型字段的属性标注结果.每个字段的性能使用精度、召回率和 *F*1 这 3 个指标来评价,同时也计算了它们在每个字段上的平均值.

Table 1 Schema matching results on the datasets OBD and DCD

表 1 数据集 OBD 和 DCD 上的模式匹配结果

Data sets	Fields	Only-DB ME			Only-L SCRF			L+DB MSCRF		
		Precision	Recall	*F*1	Precision	Recall	*F*1	Precision	Recall	*F*1
OBD	Title	68.4	63.6	65.9	76.4	76.6	76.5	81.8	78.9	80.3
	Isbn	73.6	76.4	75.0	85.7	84.4	85.0	95.1	96.7	95.9
	Authors	71.5	63.7	67.9	74.1	69.9	71.9	87.4	89.8	88.6
	Press	66.2	68.5	67.3	80.3	78.5	79.4	84.7	82.6	83.6
	Price	77.5	81.2	79.3	90.6	92.7	91.6	94.2	98.6	96.3
	Publish date	80.3	77.1	78.7	92.4	94.2	93.3	93.2	95.8	94.5
Average	…	72.9	71.7	72.3	83.2	82.7	83.0	89.4	90.4	89.9
DCD	Brand	81.5	85.2	83.3	82.6	81.1	81.8	94.5	89.4	91.9
	Type	79.4	76.3	77.8	85.3	82.3	83.8	85.7	80.9	83.2
	Price	89.6	90.2	89.9	90.4	87.2	88.8	96.3	94.0	95.1
	Optical zoom	69.2	56.7	62.3	83.2	86.7	84.9	89.6	93.6	91.6
	Digital zoom	68.7	49.3	57.4	78.3	72.6	75.3	85.4	89.7	87.5
	Sensor type	86.3	89.8	88.0	92.6	95.5	94.0	99.7	96.7	98.2
	Resolution	82.2	87.4	84.7	82.8	76.4	79.5	94.6	95.2	94.9
Average	…	79.6	76.4	77.6	85.0	83.1	84.0	92.3	91.4	91.8

从测试结果可以看出,MSCRF 模型的标注性能与 ME 和 SCRF 相比均有显著提高.ME 由于仅考虑了样本的语法格式特征,其标注性能在 70%上下.与 SCRF 相比,MSCRF 增加了数据库记录样本,其绝大多数属性的标记精度、召回率和 *F*1 这 3 个指标都有明显提高.*F*1 的平均值提高了约 7%.在这 3 组数据中,DCD 的整体标注性能要略好于 OBD,这是因为 DCD 的训练样本数量要远大于 OBD.由于 CRF 是一种概率统计学习模型,训练样本的数量和分布对测试结果会产生较大影响.另外,从 MSCRF 模型的整体测试结果可以看出,数据库记录的参与提升了模型的整体标注性能,使得减少了对手工标注样本数量的需求.

3.3.2 样本数量对模型性能的影响

我们进一步通过实验分别测试了手工标注样本和数据库记录的增加对模型标注性能的影响.图 4~图 7 是在数据库记录参与和不参与两种情况下,分别在数据集 OBD 和 DCD 上通过增加手工标注样本占总样本的比例得到的 *F*1 值的变化曲线.实验依然采用随机选择的 50%自由数据作为训练样本,另 50%固定用作测试.

实验结果说明,在两种情况下,随着手工标注样本的增加,*F*1 均明显呈上升趋势;但是,当有数据库记录参与时,工标注样本数量的增加对提高 *F*1 值的作用明显降低.

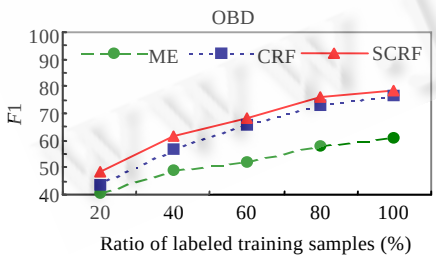


Fig.4 Variation of *F*1with ratio of labeled training samples on OBD without database

图 4 无数据库记录参与,OBD 数据集上标注的 *F*1 值随标注训练样本数量的变化

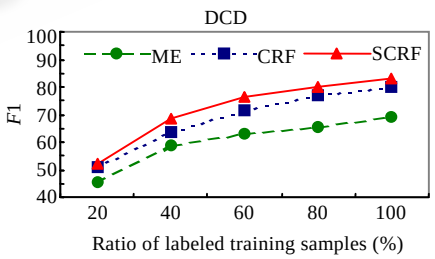


Fig.5 Variation of *F*1with ratio of labeled training samples on DCD without database

图 5 无数据库记录参与,DCD 数据集上标注的 *F*1 值随标注训练样本数量的变化

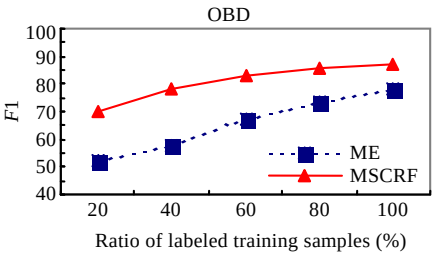


Fig.6 Variation of *F*1with ratio of labeled training samples on OBD with database

图 6 有数据库记录参与,OBD 数据集上标注的 *F*1 值随标注训练样本数量的变化

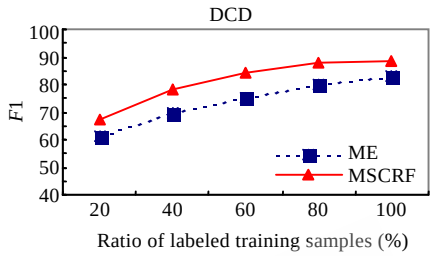


Fig.7 Variation of *F*1with ratio of labeled training samples on DCD with database

图 7 有数据库记录参与,DCD 数据集上标注的 *F*1 值随标注训练样本数量的变化

3.3.3 与其他方法的性能比较

我们在数据集 PHD 上通过实验比较了 MSCRF 与 HMM,SVM,CRF 模型在数据语义标注上的性能.基于 HMM 和 SVM 的标注方法详见文献[13,14].基于 CRF 模型的实验结果是在一阶转移特征、状态数据特征模板上得到的.此时,MSCRF 模型仅使用手工标注的样本,在 CRF 的基础上增加了高阶转移特征模板.所有系统均采用相同的实验方法,即随机选择 500 条记录用于训练,另外 435 条用于测试.表 2 显示了 5 次实验结果的平均值.它反映了在每个数据字段上的标记正确率、*F*1 值和平均 *F*1 值,以及每种方法的实例标注正确率.

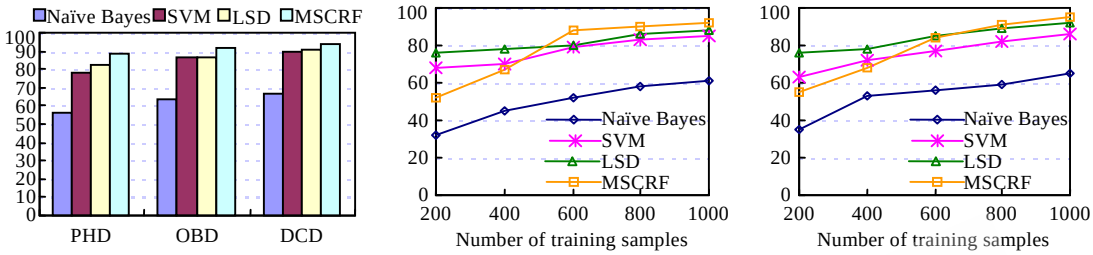
Table 2 Labeling results of different machine learning models on PHD

表 2 PHD 数据集上不同机器学习方法的标注结果

Instance accuracy	HMM (4.13%)		SVM (68.7%)		CRF (73.3%)		MSCRF (78.5%)	
Fields	Accuracy	*F*1	Accuracy	*F*1	Accuracy	*F*1	Accuracy	*F*1
Title	98.2	82.2	98.9	96.5	99.7	97.1	99.7	98.3
Author	98.7	81.0	99.3	97.2	99.8	97.5	99.8	97.8
Affiliation	98.3	85.1	98.1	93.8	99.7	97.0	99.7	97.4
Address	99.1	84.8	99.1	94.7	99.7	95.8	100	98.6
Note	97.8	81.4	95.5	81.6	98.8	91.2	99.2	92.5
Email	99.9	92.5	99.6	91.7	99.9	95.3	100	97.4
Date	99.8	80.6	99.7	90.2	99.9	95.0	100	95.2
Abstract	97.1	98.0	97.5	93.8	99.6	99.7	99.7	99.8
Phone	99.8	53.8	99.9	92.4	99.9	97.9	99.9	97.9
Keyword	98.7	40.6	99.2	88.5	99.7	88.8	99.5	90.1
Web	99.9	68.6	99.9	92.4	99.9	94.1	99.9	96.6
Degree	99.5	68.8	99.5	70.1	99.8	84.9	99.9	85.8
Pubnum	99.8	64.2	99.9	89.2	99.9	86.6	99.9	88.7
Average *F*1		75.6		90.2		93.9		95.1

从表 2 中我们可以看出,基于 CRF 模型的方法在序列数据标注上的性能要明显优于基于 HMM 和 SVM 的方法,每个字段的标注准确率和 *F*1 值几乎均有所提高,并且取得了 70%以上的实例标注正确率.SCRF 模型由于引入了跳边,因此性能在线性链 CRF 的基础上又有一定程度的提高,其中,平均 *F*1 值提高了 1.2%,而实例标注正确率的提高幅度甚至超过了 5%.另一个值得关注的现象是,基于 SCRF 的方法产生了 3 个标注正确率达到 100%的字段,这是其他方法所没有的.

此外,我们还比较了 MSCRF 与其他 Naïve Bayes,SVM 分类器以及 LSD 元分类器方法的平均性能.图 8(a)显示了对于以上 3 个不同领域的平均匹配准确率.结果显示,MSCRF 在 3 个领域都获得了很高的准确率,为 88%~92%.相反,Naïve Bayes 的匹配准确率仅仅是 57~68%,即使是基于 SVM 分类器平均匹配准确率也比 MSCRF 方法低 2%~10%.LSD 方法通过元分类器来组合不同分类器的分类结果,准确率较之基本分类器有所提高.图 8(b)、图 8(c)显示了平均领域准确率随着每个数据源中训练数据量的变化而变化的曲线.结果显示,基于分类器的方法对于训练样本的数量相对敏感,而 LSD 则显示了更好的鲁棒性,可以在相对性小的数据集中工作得更好.



(a) Average matching accuracy (b) Matching accuracy for OBD (c) Matching accuracy for DCD

(a) 平均匹配准确率 (b) ODB 数据集上的匹配准确率 (c) DCD 数据集上的匹配准确率

Fig.8 Average matching accuracy and the average domain accuracy as a function of the amount of training data

图 8 平均匹配准确率和平均领域准确率相对于训练样本数量的函数曲线

4 结 论

本文提出了一种基于混合跳链条件随机场模型的异构 Web 记录集成方法.该方法能够融合数据库中的数据记录和手工标注样本中的诸多特征来实现多个网站抽取的 Web 记录模式到关系模式的语义映射.在不同领域的多个真实数据集上的实验结果表明,本文提出的方法在提高 Web 数据元素的属性标注性能的同时,减少了对繁琐手工标注样本的依赖.

References:

[1] Zhai YH, Liu B. Web data extraction based on partial tree alignment. In: Ellis A, Hagino T, eds. Proc. of the 14th Int'l Conf. on World Wide Web (WWW 2005). New York: ACM Press, 2005. 76−85.

[2] Butter D, Liu L, Pu C. A fully automated object extraction system for the world wide Web. In: Proc. of the 21st Int'l Conf. on Distributed Computing System (ICDCS 2001). Washington: IEEE Computer Society, 2001. 361−370.

[3] Papadakis NK, Skoutas D, Raftopoulos K. STAVIES: A system for information extraction from unknown Web data source through automatic Web wrapper generation using clustering techniques. IEEE Trans. on Knowledge and Data Engineering, 2005,12(17): 1638−1652.

[4] Doan AH, Domingos P, Halevy A. Reconciling schemas of disparate data sources: A machine-learning approach. In: Proc. of the 2003 ACM SIGMOD Int'l Conf. on Management of data (SIGMOD 2003). New York: ACM Press, 2003. 509−520.

[5] Meng XF. An overview of Web data management. Journal of Computer Research and Development, 2001,38(4):385−395 (in Chinese with English abstract).

[6] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Brodley C, Danyluk A, eds. Proc. of the 18th Int'l Conf. on Machine Learning (ICML 2001). San Francisco: Morgan Kaufmann Publishers, 2001. 282−289.

[7] Pinto D, McCallum A, Wei X, Croft WB. Table extraction using conditional random fields. In: Proc. of the 26th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2003. 235−242.

[8] Peng FC, McCallum A. Accurate information extraction from research papers using conditional random fields. In: Dumais S, Marcu D, Roukos S, eds. Proc. of the Human Language Technology Conf. of the North American Chapter of the Association for Computational Linguistics (HLT- NAACL 2004). New York: ACM Press, 2004. 329−336.

[9] Zhu J, Nie ZQ, Wen JR, Zhang B, Ma WY. 2D conditional random fields for Web information extraction. In: Proc. of the 22nd Int'l Conf. on Machine Learning (ICML 2005). San Francisco: Morgan Kaufmann Publishers, 2005. 1044−1051.

[10] Zhou JS, Dai XY, Yin CY, Chen JJ. Automatic recognition of Chinese organization name based on cascaded conditional random fields. Acta Electronica Sinica, 2006,34(5):804−809 (in Chinese with English abstract).

[11] Berger AL, Pietra SAD, Pietra VJD. A maximum entropy approach to natural language processing. Computational Linguistics, 1996,22(1):39−71.

[12] Sha F, Pereira F. Shallow parsing with conditional random fields. In: Proc. of the Human Language Technology Conf. and North

American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003). Edmonton: Morgan Kaufmann Publishers, 2003. 213–220.

[13] Seymore K, McCallum A, Rosenfeld R. Learning hidden Markov model structure for information extraction. In: Califf ME, Freitag D, Kushmerick N, Muslea I, eds. Proc. of AAAI'99 Workshop on Machine Learning for Information Extraction. Cambridge: MIT Press, 1999. 31–36.

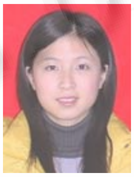
[14] Han H, Giles CL, Manavoglu E, Zha HY, Zhang ZY, Fox EA. Automatic document metadata extraction using support vector machines. In: Proc. of the 3rd ACM/IEEE Joint Conf. on Digital Libraries. New York: ACM Press, 2003. 37–48.

附中文参考文献:

[5] 孟晓峰.Web 数据管理研究综述.计算机研究与发展,2001,38(4):385–394.

[10] 周俊生,戴新宇,尹存燕,陈家俊.基于层叠条件随机场模型的中文机构名自动识别.电子学报,2006,345(5):804–809.

黄健斌(1975－),男,湖北随州人,博士,副教授,主要研究领域为机器学习,Web 挖掘.



孙鹤立(1983－),女,博士生,主要研究领域为信息检索,数据挖掘.

姬红兵(1963－),男,博士,教授,博士生导师,主要研究领域为模式识别,智能系统.

* *

关于推荐 2008 年 CCF 优秀博士学位论文的通知

为推动中国计算机领域的科技进步，鼓励创新性研究，促进青年人才成长，中国计算机学会（CCF）自 2006 年起设优秀博士学位论文奖。2008 年度优秀博士学位论文推荐工作现已启动，现将有关内容通知如下：

一、参评条件

1. 本次优秀博士学位论文的评选范围为 2006 年 7 月 1 日~2008 年 6 月 30 日期间在中国获得计算机科学与技术学科相关专业博士学位的学位论文。

2. 参加评选的博士学位论文须经具有计算机科学与技术学科博士点的高校计算机学院（系）或研究机构推荐，每个具有一级学科博士点单位推荐参评学位论文不超过 2 篇，其他不具有一级学科博士点的单位限推荐 1 篇，已经参评过的论文不得再被推荐。

3. 具体参评条件和约束条件见“中国计算机学会优秀博士学位论文奖条例”。

二、参评申报材料

印刷论文 2 本；电子版论文 1 份；CCF 优秀博士学位论文推荐表（必须有作者答辩时所在单位（如系、院、所等）负责人签字、单位盖章）；其他有关证明材料；评审费：1000 元/篇（CCF 会员 800 元/篇）。

三、申报材料和评审费须于 2008 年 9 月 4 日 17:00 前报送到 CCF，过期无效。

四、评选时间安排

受理：2008 年 7 月 22 日~2008 年 9 月 4 日；格式和资质审查：2008 年 9 月 5 日~9 月 12 日；初评：2008 年 9 月 13 日~10 月 12 日，CCF 组织小同行专家对申报材料进行初评，从中评选出不超过 30 篇入围候选优秀博士学位论文；初评公示：2008 年 10 月 13 日~11 月 12 日；终评：2008 年 11 月 13 日~12 月 12 日，CCF 终评委员会进行终评，评出获奖者。获奖总数不超过 10 篇，另有不超过 5 篇论文获提名奖；终评公示：2008 年 12 月 13 日~2009 年 1 月 12 日。

五、联系方式

联系人：孙文韬，电话：010-62562503-20，E-mail：ccf-ed5@ict.ac.cn

李乐强，电话：010-62562503-14，E-mail：ccf-aw@ict.ac.cn

通信地址：北京 2704 信箱，中国计算机学会，邮政编码：100190

六、详情请浏览中国计算机学会网站（http://www.ccf.org.cn）上的相关网页。