基于集成学习和二维关联边条件随机场的Web数据语义标注方法

丁艳辉 李庆忠 董永权 彭朝晖
(山东大学计算机科学与技术学院 济南 250014)

摘 要 大规模Web信息抽取需要准确、自动地从众多相关网站上抽取Web数据对象.现有的Web信息抽取方法主要针对单个网站进行处理,无法适应大规模Web信息抽取的需要.调查研究表明,有效地实现Web数据语义自动标注,结合现有的包装器生成技术,可以满足大规模Web信息抽取的要求.文中提出一种基于集成学习和二维关联边条件随机场的Web数据语义自动标注方法,首先,利用已抽取的信息和目标网站训练页面中呈现的特征构造多个分类器,使用Dempster合成法则合并分类器结果,区分训练页面中的属性标签和数据元素;然后,利用二维关联边条件随机场模型对Web数据元素间的长距离依赖联系和短距离依赖联系进行建模,实现数据元素的自动语义标注.通过在多个领域真实数据集上的实验结果表明,所提出的方法可以高效地解决Web数据语义自动标注问题,满足大规模Web信息抽取的需要.

关键词 Web信息抽取;语义标注;集成学习;条件随机场;长距离依赖
中图法分类号 TP393 **DOI号**: 10.3724/SP.J.1016.2010.00267

Semantic Annotation of Web Data Based on Ensemble Learning and 2D Correlative-Chain Conditional Random Fields

DING Yan-Hui LI Qing-Zhong DONG Yong-Quan PENG Zhao-Hui
(*School of Computer Science and Technology, Shandong University, Jinan* 250014)

Abstract Large-scale Web information extraction needs to extract information from many Web sites accurately and automatically. However, most current Web information extraction methods place emphasis on single Web site, which causes that they can't meet the need of large-scale Web information extraction. The empirical study shows that automatic semantic annotation of Web data, combined with current wrapper learning techniques, may meet the need of large-scale Web information extraction. In this paper, a method based on ensemble learning and two-dimensional Correlative-Chain Conditional Random Fields (2DCC-CRFs) is proposed to solve the problem of automatic semantic annotation of Web data. Firstly, several classifiers based on different kinds of features can be built by analyzing the previously extracted data and sample Web pages; Then, attribute tags and Web data elements can be identified by combining multiple classifiers using Dempster-Shafer theory of evidence; Finally, 2DCC-CRFs is built to do semantic annotation of Web data element automatically, which extends a classic model, 2DCRFs, by adding correlative edges. Experimental results using a large number of real-world data collected from diverse domains show that the proposed approach can do automatic semantic annotation of Web data efficiently, which can meet the need of large-scale Web information extraction.

收稿日期:2009-07-15;最终修改稿收到日期:2009-09-17.本课题得到国家自然科学基金(90818001)、山东省自然科学基金(Y2007G24)资助.**丁艳辉**,男,1981年生,博士研究生,主要研究方向为Web信息集成、Web信息抽取.E-mail: dingyanhui@gmail.com.**李庆忠**(通信作者),男,1965年生,教授,博士生导师,主要研究领域为信息集成.E-mail: lqz@sdu.edu.cn.**董永权**,男,1979年生,博士研究生,主要研究方向为Web信息集成.**彭朝晖**,男,1978年生,博士,讲师,主要研究方向为信息检索.

Keywords Web information extraction; semantic annotation; ensemble learning; conditional random fields; long distance dependencies

1 引 言

随着 WWW 的快速发展,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 数据对象正是这样一类由多个数据元素及可选的数据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]. 通过 Web 信息集成技术将来自不同网站的 Web 数据对象进行有效的集成,可以进一步提高互联网上信息的利用率,提高其利用价值. 大规模 Web 信息抽取是 Web 信息集成系统中的关键步骤之一,大规模 Web 信息抽取与传统的 Web 信息抽取区别在于:待抽取的网站众多,自动化程度要求比较高. 如何准确、自动地从众多相关网站上抽取 Web 数据对象,为 Web 信息集成系统提供必要的数据支持,是一个亟待解决的问题.

很多研究学者和研究机构对 Web 信息抽取开展了大量的研究工作,按照 Web 信息抽取的自动化程度可以分为 3 类:手工建立、手工标记自动学习和全自动抽取程序[2]. 但是,现有的大部分方法主要针对单个网站进行处理,对于大规模 Web 信息抽取来说,主要存在以下几点不足:

(1) 手工建立方法和手工标记自动学习方法均需要人工参与,对于大规模 Web 信息抽取来说,代价太高,不适合使用.

(2) 全自动抽取程序对目标网站具有一定的要求,要求待抽取页面具有相同的展示模板[3];或要求待抽取页面为列表页面[1]. 但是,对于 Web 信息集成系统来说,一方面,列表页面和详细页面都需要进行信息抽取;另一方面,由于来自相同网站的网页可能由多个模板产生,导致页面间的同模板检测代价过高[4],不适合应用到大规模 Web 信息抽取中.

(3) 从一个特定网站中学到的包装器(Wrapper),不能够被应用到新网站中,即使它们属于同一领域. 这反映了现有抽取方法的适应性不强[5],不能满足大规模 Web 信息抽取的需要.

综上所述,现有的大部分 Web 信息抽取方法不适合完成大规模 Web 信息抽取任务. 调查研究表明[5],通过对目标网站上的训练页面进行语义标注,即为识别出的每个数据元素分配一个有意义的标签来表示该数据元素的语义,可以方便地得到目标网站的训练样例,结合现有的包装器生成技术,可以自动地为目标网站生成包装器,完成信息抽取任务.

为了适应大规模 Web 信息抽取的需要,本文提出一种基于集成学习和二维关联边条件随机场的 Web 数据语义自动标注方法,通过利用已抽取信息和目标网站训练页面中呈现的特征,自动、准确地完成对众多网站上训练页面的语义标注,为下一步自动构造包装器,完成大规模 Web 信息抽取提供保证. 本文的创新点主要体现在以下几个方面:

(1) 提出一种适合大规模 Web 信息抽取的 Web 数据语义自动标注方法;

(2) 提出基于集成学习的判别方法,充分利用已抽取信息的特征和目标网站训练页面中的特征,完成对 Web 数据对象中属性标签和数据元素的区分;该方法具有良好的扩展能力;

(3) 提出 2DCC-CRFs 模型,综合利用 Web 数据元素间的长距离依赖联系和短距离依赖联系,提高语义标注的准确性;

(4) 通过在多个领域真实数据集上的实验表明,所提出的方法能够较好地完成大规模 Web 信息抽取任务.

2 相关工作

在 Web 信息抽取领域,已经提出了许多自动/半自动的生成 Wrapper 的方法. 文献[1]和文献[6]是两种独立于模板的 Wrapper 生成方法,Lerman 等[6]利用详细页面中的信息分隔列表页面中的信息,并构造相应的 Wrapper 进行抽取. Zhai 等[1]利用字符串匹配以及一些视觉特征来挖掘页面中数据记录. 但是,文献[1]和文献[6]抽取出的数据均不包含语义标签. Embley 等[7]利用本体加上一些启发式规则的方法在包含多条 Web 数据记录的文档中自动地抽取数据,并进行语义标注. 但是,对于不同领域的本体必须手工创建. Mukherjee 等[8]利用一些展示规则和相关元素的空间位置关系进行语义标注,这个过程仍然依赖于领域知识. Arlotta 等[9]提出一种完全自动地对搜索结果中的数据项标注有意义标签的方法,利用结果页面中距离数据项最近的标签进行标注. 但是,这个方法具有一定的局限性,因为很多网站没有将相关标签在结果页面中显示出来.

基于 D-S 证据理论的集成学习策略已经被多次提出[10]. Altincay 和 Demirekler[11] 提出利用 D-S 证据理论对基于等级的分类器进行组合. 同时，D-S 证据理论也被应用于对以 boosting 方法构造的分类器进行组合[12]. 在本文中，将该策略应用于解决大规模 Web 信息抽取问题.

条件随机场是利用序列特征处理序列数据分割与标注问题的经典机器学习方法，在自然语言理解、信息提取等多个领域得到了广泛的应用. 针对 Web 数据元素间的二维序列特征，Zhu 等[13] 提出二维条件随机场模型(2DCRFs)，用于解决 Web 数据语义标注问题. 但是，2DCRFs 并没有考虑 Web 数据元素间的长距离依赖联系[14].

发现并利用数据间的长距离依赖联系的想法，已经被一些方法提出. Sutton 等[15] 提出 Skip-CRFs 模型，利用文档中相同的单词之间的长距离依赖联系，提高实体识别的准确率. 但是，对于单条 Web 记录来说，并不具备类似的长距离依赖联系，导致该方法不能直接应用于 Web 数据语义标注. 黄健斌等[16] 针对 Web 记录集成中的模式匹配问题，提出了混合链条件随机场模型(MSCRFs)，通过在内容相似的 Web 数据元素之间建立跳边，加强了对长距离依赖关系的处理. 但是，在很多情况下，单条 Web 数据记录中并不包含内容相似的数据元素对，导致无法有效地建立 Web 数据元素间的长距离依赖联系.

针对大规模 Web 信息抽取任务，目前开展的工作还比较少. Wong 等[5] 提出一种 Web 信息自适应抽取方法，通过利用贝叶斯模型实现网站间包装器的适应性学习以及发现新属性. 但这种方法仅能将页面中存在的属性标签赋予相应的数据元素，对于页面内没有相应属性标签的数据元素不做标注. 而本文提出的方法，将对两种情况都进行标注.

3　Web 数据语义标注

本文将主要研究如何准确、自动地完成对众多网站中待标注页面(训练页面)的自动语义标注，为大规模 Web 信息抽取提供基础支撑. 大规模 Web 信息抽取是 Web 信息集成系统的关键步骤之一，将为其提供必要的数据基础.

如图 1 所示，在大规模 Web 信息抽取系统中，包含众多待抽取网站，需要 Web 信息抽取系统自动地完成相应网站的信息抽取工作. 首先，从待抽取网站中选择多个页面作为训练页面；其次，自动完成训练页面中 Web 数据对象的语义标注，形成训练样例；然后，利用自动标注后的训练样例，结合包装器生成技术[17] 自动地生成相应网站的包装器；最后，利用包装器完成对相应网站的自动信息抽取.

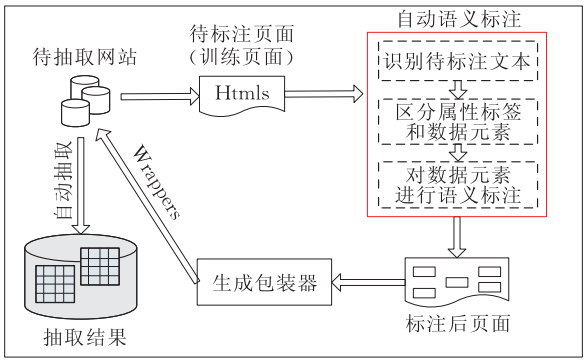


图 1　大规模 Web 信息抽取流程图

Web 数据语义标注(图 1 中右上矩形框所示)主要分为 3 个阶段：第 1 阶段，找出训练页面中的主数据区域[18]，识别出待标注文本；第 2 阶段，对待标注文本进行自动判断，区分出属性标签和数据元素；第 3 阶段，利用训练页面中包含的属性标签和已抽取信息中包含的属性名称，对数据元素进行自动语义标注. 第 1 阶段的任务比较简单，利用现有的技术可以完成；第 2 阶段的任务，利用本文提出的集成学习策略完成；第 3 阶段的任务，利用本文提出的二维关联边条件随机场模型完成.

3.1　识别待标注文本

Web 网页中包含了大量的信息，其中一部分是用户感兴趣的 Web 数据对象信息，例如图书的基本信息、招聘的基本信息等；而另一部分是杂质信息，例如广告、导航信息、交互表单等. 为了提高 Web 信息抽取的准确性和效率，将首先定位网页中的主数据区域(MainBlock)[18]，主数据区域中包含着用户感兴趣的 Web 数据对象信息. 将网页解析成 DOM 树，主数据区域对应 DOM 树中的一棵子树 T，子树 T 中所有文本叶子节点上的内容被称为待标注文本.

3.2　区分属性标签和数据元素

识别出待标注文本后，要对待标注文本进行判断，区分出 Web 数据对象中的属性标签和数据元素. 本文提出一种基于集成学习的判别方法，首先，利用已抽取信息的特征和目标网站训练页面中呈现的特征构造多个分类器；然后，利用构造的多个分类器对待标注文本进行独立判断；最后，利用基于 D-S 证据理论的合成法则，对所有的分类器结果进行合并，得到每一个待标注文本的最终分类结果，完成属

性标签和数据元素的区分.

3.2.1 构建分类器

在集成学习中,强调分类器的多样性.相关研究[19]指出,基于不同特征的分类器组合会产生较好的结果.在本文中,将利用已抽取信息的特征和待标注页面中呈现的特征,构造基于不同特征类型的分类器,对待标注文本进行判断.

3.2.1.1 基于已抽取信息的分类器

Web信息集成通常是对某一领域内的信息进行集成,不同网站上的信息具有一定的联系.通过利用已抽取信息的特征,可以对目标网站的信息抽取提供指导.

假设Web信息抽取系统利用现有的抽取方法(自动的方法或手工标注的方法)已经对个别网站完成了准确的抽取工作,那么将获得领域内的大量结构化信息,包括已抽取的Web数据对象的属性名称列表以及对应的数据元素值.通过对现有的属性名称列表和数据元素值进行分析,可以获得现有数据的特征.例如,对于图书来说,ISBN是属性名称,对应的属性值是一个10位或13位的数字;出版日期是日期型字段等;在表1中,列出了部分在实验中使用到的已抽取信息的特征.

表1 已抽取信息的特征

对象	特征类型
属性名称	与已抽取的属性名称相同
	与已抽取的属性名称相似
属性值	与已抽取的属性值相同
	与已抽取的属性值的模式相同
	平均长度
	数据类型

3.2.1.2 基于目标网站训练页面中特征的分类器

在同一网站中,训练页面中的Web数据对象会呈现出一定的特征,例如,属性标签的展示格式趋于相同;数据元素的展示格式趋于相同;属性标签通常会在多个训练页面中重复出现;而不同Web数据对象中的数据元素趋于不同等.在表2中,列出了部分在实验中使用到的特征.

表2 训练页面中呈现的特征

特征类型	说明
待标注文本的路径熵	在不同页面中,路径相同而文本值不同的待标注文本,可能是属性标签;否则,可能是数据元素
待标注文本的信息熵	反映了待标注文本在训练页面集中的重复出现情况

(续 表)

特征类型	说明
在网页中的位置	页面中的坐标值
展示格式	字体、颜色、大小等
两个待标注文本的相对位置	是否相邻;是否在同一行上(属性名称通常在对应属性值的上边或左边)
包含超链接	是/否
与页面内已知的属性标签具有相同的路径	是/否
与页面内已知的数据元素具有相同的路径	是/否

3.2.2 分类结果合并

由于在Web信息集成中需要对众多相关网站进行抽取,信息的特征会得到不断的补充,例如,属性名称列表在不断丰富;新属性值的特征被不断发现,这些将导致为不同分类器固定权重的方法不再适合,需要动态调整不同分类器的权重,为其分配在当前情况下最适合的权重值.由于D-S证据理论[20]可以有效地调整分类器的权重值,所以,本文在分类结果合并策略中,选择基于D-S证据理论的合并方法.

3.2.2.1 Dempster-Shafer证据理论

Dempster-Shafer证据理论[20]是由Dempster首先提出,并由Shafer进一步发展起来的一种处理不确定性的理论.在D-S证据理论中,首先将待识别对象有可能结果的集合所构成的空间识别框架记作Θ,并把Θ中所有子集组成的集合记作2^{Θ}.对于2^{Θ}中任何假设集合A,有$m(A)\in[0,1]$,并且

$$m(\varnothing)=0 \tag{1}$$

$$\sum_{A\in 2^{\Theta}} m(A)=1 \tag{2}$$

其中,$\varnothing$为空集,m称为2^{Θ}上的概率分配函数,$m(A)$称为A的基本概率分配(Basic Probability Assignment,BPA).

D-S证据理论定义了信任函数Bel和似然函数Pls来表示问题的不确定性,即

$$Bel:\ 2^{\Theta}\rightarrow[0,1],\ Bel(A)=\sum_{B\subseteq A} m(B) \tag{3}$$

$$Bel:\ 2^{\Theta}\rightarrow[0,1],\ Pls(A)=\sum_{B\cap A\neq\varnothing} m(B) \tag{4}$$

在有多证据存在的情况下,可以使用Dempster组合法则对多个BPA进行合成,即

$$m(A)=\frac{\sum_{\cap A_i=A}\ \prod_{1\leq i\leq n} m_i(A_i)}{1-K} \tag{5}$$

其中$K=\sum_{\cap A_i\neq\varnothing}\ \prod_{1\leq i\leq n} m_i(A_i)$,$m_1,m_2,\cdots,m_n$为$n$个

BPA.

3.2.2.2　基于 D-S 证据理论的多分类器集成

首先定义识别框架 $\Theta=\{O_1,O_2,\cdots,O_n\}$，其中 O_i 代表不同的分类类别；在系统中，将构造多个基本分类器(Basic Classifier)，分别表示为 $BC_1,BC_2,\cdots,BC_m$. 每一个分类器具有 n 个输出，每个输出 j 指向类别 O_j.

在分类器集成系统的构建阶段，将为每个基本分类器的任一输出赋予一个 BPA，一个 BPA 赋予输出 j，代表了当一个待标注文本 text 进入集成系统时，被分类为 O_j 的可信程度. 每一个 BPA 表示为 $m_i(O_j)$，其中 i 代表分类器 BC_i，j 代表输出类型 O_j. $m_i(O_j)$ 的值有多种计算方法，在本文的实验中，采用查准率作为 BPA 的值.

在分类器集成系统构建阶段，完成了对基本分类器性能的第一次评价，分类器的性能评价基于在训练集中正确分类的个数和错误分类的个数. $m_i^{\text{Training}}(O_j)$ 用于记录基本分类器 BC_i 的任一输出 j 的 BPA. 为了使评价更为准确，在测试集上对每个基本分类器的任一输出进行二次评价，获得每个基本分类器在任一输出上的 BPA，记为 $m_i^{\text{Test}}(O_j)$. 通过利用 D-S 证据理论对于同质特征的合并规则[10]，得到一个唯一的 BPA，使用以下公式计算出来，

$$m_i(O_j)=1-(1-m_i^{\text{Training}}(O_j))(1-m_i^{\text{Test}}(O_j)) \tag{6}$$

基本分类器加上推理机构成完整的集成学习分类系统[10]. 当一个新的待标注文本输入系统时，各个分类器将产生多个输出类别，所有这些输出类别将通过推理机进行综合处理. 如果多个分类器均指向同一输出类型，那么不需要进行任何推理工作；当多个分类器指向不同的输出类型，需要进行推理，计算出最终输出类型. 推理过程分成两步进行，第 1 步，各个基本分类器独立进行，使用的组合规则是 D-S 证据理论基于冲突证据的组合规则[10]. 第 2 步，将不同分类器的分类结果利用 Dempster 组合法则(式(5))进行合并.

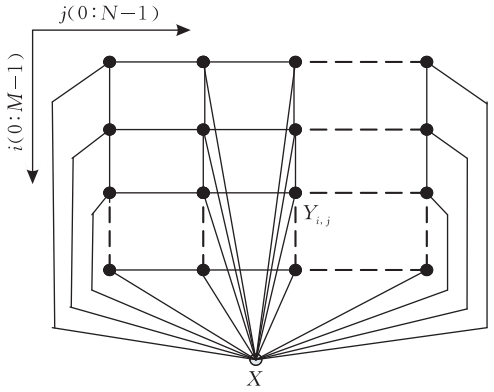
3.3　数据元素的语义标注

当 Web 数据对象中的数据元素和属性标签成对出现时，利用已有的一些启发式规则，可以很好地进行标注. 例如，属性标签和数据元素在页面上位置相邻；属性标签一般位于数据元素的前方或上方等. 但是，当数据元素和属性标签未成对出现在页面上时，现有标注方法的标注准确率将会降低. 例如，2DCRF[13] 是进行语义标注的经典方法，但是，它仅考虑了 Web 数据元素间的短距离依赖联系，当页面中出现同模式的数据元素时，准确率会大大降低.

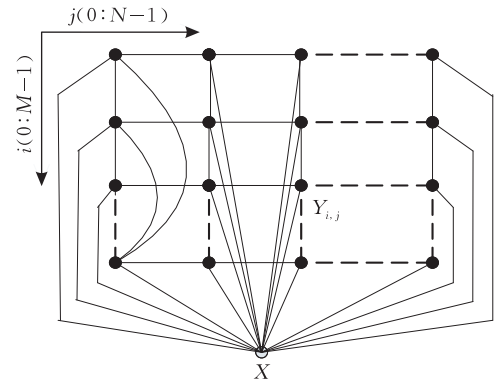
本文提出一种基于二维关联边条件随机场的数据元素语义标注方法，通过在 2DCRF 模型的基础上增加关联边，实现对 Web 数据元素间的长距离依赖联系和短距离依赖联系的充分建模，提高 Web 数据元素语义标注的准确率.

3.3.1　二维关联边条件随机场

由于 2DCRFs(图 2(a))没有考虑 Web 数据元素间的长距离依赖联系，本文提出一个新的二维关联边条件随机场(two-Dimensional Correlative-Chain CRFs，2DCC-CRFs)模型，如图 2(b)所示，通过在 2DCRFs 模型上叠加关联边来处理数据元素间的长距离依赖联系.



(a) 2DCRFs的图模型



(b) 2DCC-CRFs的图模型

图 2　2DCRFs 和 2DCC-CRFs 的图模型

定义 1.　设 $G=\langle X,Y\rangle$ 是一个二维条件随机场，X 是序列观测数据随机变量，Y 是状态标注序列随机变量. $Y_{i,j}$ 是 Y 在位置 (i,j) 上的组成元素. 如果存在 $Y_{i,j}$，$Y_{m,n}$，且 $Y_{i,j}\in Y$，$Y_{m,n}\in Y$，$|i-m|>1$，$|j-n|>1$，使得 $Y_{m,n}$ 依赖于 $Y_{i,j}$，则称边 $(Y_{i,j},Y_{m,n})$ 是一条关联边，并称包含关联边的二维条件随机场模型为二维关联边条件随机场.

在本文提出的模型中，关联边分为两种类型：

CU 型关联边和 UU 型关联边，具体定义如下.

定义 2. 设$(Y_{i,j},Y_{i',j'})$是一条关联边，在推理过程之前，如果 $Y_{i,j}$ 具有确定的语义标签 $y_{i,j}$，而 $Y_{i',j'}$ 不具有确定的语义标签，那么称关联边$(Y_{i,j},Y_{i',j'})$为 Certain-Uncertain 型关联边，简称为 CU 型关联边.

定义 3. 设$(Y_{m,n},Y_{m',n'})$是一条关联边，在推理过程之前，如果 $Y_{m,n}$ 和 $Y_{m',n'}$ 都不具有确定的语义标签，那么称关联边$(Y_{m,n},Y_{m',n'})$为 Uncertain-Uncertain型关联边，简称为 UU 型关联边.

令 $E'=\{(Y_{u,v},Y_{u',v'})\}$是关联边的集合，则在给定观测序列 x 的条件下，标注序列 y 的概率分布 $p(y|x)$ 为

$$p(y|x)=\frac{1}{Z(x)}\exp\Big(\sum_{e\in E,k}\lambda_k f_k(e,y|_e,x)+\sum_{v\in V,k}\mu_k g_k(v,y|_v,x)+\sum_{e'\in E',k}\gamma_k h_k(e',y|_{e'},x)\Big) \tag{7}$$

其中，f_k，g_k 和 h_k 是定义在不同团(例如，普通边、结点和关联边)上的特征函数，λ_k，μ_k 和 γ_k 是特征函数的权重值，将利用训练集训练得到. $Z(x)$ 是归一化因子，表示为

$$Z(x)=\sum_{y}\exp\Big(\sum_{e\in E,k}\lambda_k f_k(e,y|_e,x)+\sum_{v\in V,k}\mu_k g_k(v,y|_v,x)+\sum_{e'\in E',k}\gamma_k h_k(e',y|_{e'},x)\Big) \tag{8}$$

3.3.2 基于二维关联边条件随机场的 Web 数据元素语义标注

利用二维关联边条件随机场模型完成对 Web 数据元素的语义标注，需要完成以下 3 方面工作：(1)建立关联边；(2)参数估计；(3)推理.

3.3.2.1 关联边

关联边将在推理过程之前建立，构造关联边的关键在于选择关联边的始末结点. 通过找出能够确定语义标签的数据元素，就可以构造出所有的关联边，包括 CU 型关联边和 UU 型关联边. 目前，存在很多种方法[21]可以判断出 Web 数据元素 $R_{i,j}$ 标注语义标签 l_t 具有较高的置信度.

在 2DCC-CRFs 模型中，将主要利用已有数据库的结构化信息和记录特征以及训练集中样本数据的文本特征来判断某一 Web 数据元素标注不同语义标签的置信度[22]. 如果 Web 数据元素 $R_{i,j}$ 与数据库中的数据记录具有很好的匹配，包括内容匹配或模式匹配；或者 Web 数据元素 $R_{i,j}$ 对于某个语义标签具有一个足够高的发射概率，那么可以认为 Web 数据元素 $R_{i,j}$ 标注语义标签 l_t 具有较高的置信度. 例如，如果存在以下统计结果，

$$p(l_t=\text{"conference"}|R_{i,j}\text{ contains "in proceedings of"})=0.99,$$

那么当 Web 数据元素 $R_{i,j}$ 为"in proceedings of SIGMOD08"时，$R_{i,j}$ 标注为"conference"这个语义标签.

构建关联边的算法如下.

算法 1.

输入：一条 Web 数据记录 R 和语义标签集合 L

输出：CU 型关联边集合 E'_{cu} 和 UU 型关联边集合 E'_{uu}

1. 对于 $\forall l_t\in L$，$\forall R_{i,j}\in R$，计算数据元素 $R_{i,j}$ 标注语义标签 l_t 的置信度 $c_{i,j}(l_t)$；

2. 如果 $R_{i,j}$ 标注语义标签 l_t 的置信度大于阈值，即 $c_{i,j}(l_t)>\omega$，那么将 $R_{i,j}$ 添加至语义标签已确定元素的集合 *CertainElements*；否则，将 $R_{i,j}$ 添加至语义标签未确定元素的集合 *UncertainElements*；

3. 对于 $\forall R_{i,j}$，$R_{i',j'}$ 且 $R_{i,j}\in R$，$R_{i',j'}\in R$，如果 $R_{i,j}\in$ *CertainElements*，$R_{i',j'}\in$ *UncertainElements*，那么将$(R_{i,j},R_{i',j'})$添加至 CU 关联边集合 E'_{cu}；

4. 对于 $\forall R_{m,n}$，$R_{m',n'}$ 且 $R_{m,n}\in R$，$R_{m',n'}\in R$，如果 $R_{m,n}\in$ *UncertainElements* 且 $R_{m',n'}\in$ *UncertainElements*，那么将$(R_{m,n},R_{m',n'})$添加至 UU 型关联边集合 E'_{uu}；

5. 返回 E'_{cu} 和 E'_{uu}.

3.3.2.2 参数估计

在 2DCC-CRFs 模型中，参数估计分为两个部分，一部分是针对普通边(不包括关联边在内的邻接边)和结点上特征函数的参数估计，另一部分是针对关联边上特征函数的参数估计.

与传统的链式 CRFs 和 2DCRFs 的参数估计相似，2DCC-CRFs 模型中使用最大似然估计来估计针对普通边和结点上的特征函数的权重参数. 即在给定一个具有概率分布为 $\tilde{p}(x,y)$ 的训练集 $D=\{(y^i,x^i)\}_{i=1}^N$ 上，估计参数 $\Phi=\{\mu_1,\mu_2,\cdots;\lambda_1,\lambda_2,\cdots\}$ 的值，使得该训练集数据的对数似然函数 $L(\Phi)$ 达到最大. 似然函数表示为

$$L(\Phi)=\sum_i\tilde{p}(x^i,y^i)\log p(y^i|x^i,\Phi) \tag{9}$$

由于 $L(\Phi)$ 为凹函数，导数为零时为最值点. 故对 Φ 求导，则偏导数公式为，

$$\frac{\partial L(\Phi)}{\partial\lambda_k}=E_{\tilde{p}(x,y)}[f_k]-E_{p(y|x,\Phi)}[f_k] \tag{10}$$

$$\frac{\partial L(\Phi)}{\partial\mu_k}=E_{\tilde{p}(x,y)}[g_k]-E_{p(y|x,\Phi)}[g_k] \tag{11}$$

令式(10)、(11)等于 0，函数 $L(\Phi)$ 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”. $E_{\tilde{p}(y|x,\Phi)}[f_k]$和$E_{\tilde{p}(y|x,\Phi)}[g_k]$如果直接计算需要很大的计算量,可以使用动态规划的方法求解,如向前-向后(Forward-Backward)算法.

为了避免对大量参数估计时出现的过拟合问题,对数似然函数经常需要将参数作先验分布调整,采用高斯先验调整后,式(9)转化为

$$L(\Phi)=\sum_i \tilde{p}(x^i,y^i)\log p(y^i|x^i,\Phi)-\sum\frac{\Phi^2}{2\delta^2} \tag{12}$$

其导数变为

$$\frac{\partial L(\Phi)}{\partial \lambda_k}=E_{\tilde{p}(x,y)}[f_k]-E_{\tilde{p}(y|x,\Phi)}[f_k]-\frac{\lambda_k}{\sigma^2} \tag{13}$$

$$\frac{\partial L(\Phi)}{\partial \mu_k}=E_{\tilde{p}(x,y)}[g_k]-E_{\tilde{p}(y|x,\Phi)}[g_k]-\frac{\mu_k}{\sigma^2} \tag{14}$$

其中,σ^2表示先验方差.于是Φ的参数估计问题可以用最优化方法解决,可以使用 GIS、IIS 等迭代方法.本文的实验使用 L-BFGS 算法[23]实现对目标函数的优化求解. L-BFGS 是一种充分利用以前的梯度和修改值来近似曲率值的二阶方法,可以避免准确的 Hessian 矩阵的逆矩阵的计算.

在 2DCC-CRFs 模型中,针对关联边上的特征函数的参数估计比较简单,CU 型关联边$(Y_{i,j},Y_{i',j'})$上的特征函数的权重值等于语义标签对$\langle y_{i,j},y_{i',j'}\rangle$的互信息[21]$MI(y_{i,j},y_{i',j'})$.

$$MI(y_{i,j},y_{i',j'})=p(y_{i,j},y_{i',j'})\log_2\left(\frac{p(y_{i,j},y_{i',j'})}{p(y_{i,j})p(y_{i',j'})}\right) \tag{15}$$

$$p(y_{i,j},y_{i',j'})=\frac{N(y_{i,j},y_{i',j'})}{N} \tag{16}$$

$$p(y_{i,j})=\frac{N(y_{i,j})}{N} \tag{17}$$

其中,N代表训练集的大小,$N(y_{i,j})$代表语义标签$y_{i,j}$出现的次数,$N(y_{i,j},y_{i',j'})$代表语义标签对$\langle y_{i,j},y_{i',j'}\rangle$同时出现的次数.对于 CU 型关联边,$y_{i,j}$是唯一的,而$y_{i',j'}$可以有多个值,在本文的实验中,通过阈值来控制$y_{i',j'}$的取值个数.

对于 UU 型关联边$(Y_{m,n},Y_{m',n'})$上的特征函数的权重值初始为Δ,在本文的实验中,$\Delta=\frac{1}{|L|}$,其中,$|L|$代表语义标签集的大小.

3.3.2.3　推　理

推理算法的时间复杂度将对模型的性能产生重要的影响.在 2DCC-CRFs 模型中,由于二维表格中包含环,并且环的距离可能比较长或发生重叠,导致精确推理算法的时间复杂度呈指数级增长,所以,精确的推理算法不再适合.本文使用 Loopy Belief Propagation 算法[24]进行近似推理,Loopy Belief Propagation 算法是对向前-向后算法的归纳.向前-向后算法的时间复杂度为$O(n^2T)$,其中n代表状态集合的大小,T代表观察序列的长度.

由于在 2DCC-CRFs 模型中增加了对关联边的处理,并且对单条关联边进行处理的代价等同于向前-向后算法中对单条邻接边的处理代价[24].因此,在 2DCC-CRFs 模型中,Loopy Belief Propagation 算法的时间复杂度变为$O(|L|^2\cdot(T+M))$,其中$|L|$代表语义标签集合的大小,T代表一条 Web 记录中数据元素的个数,M代表关联边的条数,M与T的平方成正比.

Loopy belief propagation 算法是一个迭代算法,尽管它不保证收敛,但是相关研究[23]和本文的实验均表明,其在实际应用中具有较好的近似推理效果,可以有效地推断出最有可能的语义标注序列.

4　实验评价

在本文中,预定义的关系数据库作为已抽取的信息库使用,通过收集多个领域的真实数据集,对提出的方法进行测试,主要通过 4 个方面进行分析评价:(1)基于 D-S 证据理论的集成学习方法与其它方法的比较;(2)2DCC-CRFs 模型与传统条件随机场模型的比较;(3)数据库参与与否对 2DCC-CRFs 模型性能的影响;(4)数据库规模对 2DCC-CRFs 模型性能的影响.

4.1　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(Book dataset,简称 B)

该数据集由从 30 个在线图书网站上收集的 2000 条不同格式的 Web 图书记录组成,数据经手工标注后,随机选择 1000 条作为训练数据,剩余部分作为测试数据.另外,从 http://www.textbookx.com/网站上随机抓取 1 万条图书记录,存入预先定义的关系数据库表中,作为图书数据库.

(2)台式机数据集(Desktop dataset,简称 D)

该数据集由从 FROOGLE 在线购物网站收集

的 3500 条异构台式机记录构成，数据经手工标注后，将其中的 2000 条记录录入预先定义的关系数据库表中，作为台式机数据库. 在其余的 1500 条记录中，随机选择 700 条记录作为训练数据，剩余部分作为测试数据.

(3) 论文参考文献数据集(Paper reference dataset，简称 P)

该数据集由 800 条论文记录构成，是用来评价数据抽取系统的基准数据集之一，数据集来源于 http://www.cs.umass.edu/~mccallum/data. 从该数据集中随机选择 400 条作为训练数据，剩余部分作为测试数据. 另外，从 ACM Digital Library 中随机抓取 2 万条论文记录，作为论文数据库.

4.2 评价标准

本文采用检验 Web 数据语义标注结果的常用标准：查全率、查准率、测度 $F1$ 和实例标注准确率，对实验结果进行评价，具体的定义如下：

设 A 表示待标注的数据元素数；B 表示正确标注的数据元素数；C 表示错误标注的数据元素数.

(1) 查全率($Recall$)、查准率($Precision$)和测度 F_1，其计算公式分别为

$$Recall=\frac{B}{A},\quad Precision=\frac{B}{B+C},$$
$$F1=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{18}$$

(2) 实例标注准确率. 每个数据元素均被正确标记的 Web 数据对象占总的测试对象的比例.

4.3 实验结果与分析

4.3.1 基于 D-S 证据理论的集成学习策略与其它方法的比较

RoadRunner[3] 是经典的全自动 Web 信息抽取方法，但它不能对抽取的数据项进行语义标注，所以，本文在实验中将使用 RoadRunner++[5] 进行比较. 首先，选择图书领域的 6 个网站，分别为 DigitalGur.com、Half Price Computer books、Bookpool.com、Amazon.com、Abebooks.com、Discount-PCBooks.com. 然后，基于 D-S 证据理论的集成学习方法(简称为 DSE)和 RoadRunner++分别对这 6 个网站进行抽取和语义标注. 实验结果如表 3 所示.

表 3 DSE 与 RoadRunner++对比结果

站点(URL)	基于 DSE 的结果			基于 RoadRunner++的结果		
	Precision/%	*Recall*/%	*F1*/%	*Precision*/%	*Recall*/%	*F1*/%
www.digitalguru.com	89.7	87.1	88.4	56.7	80.4	66.5
www.halfpricecomputerbook.com	83.1	80.5	81.8	49.4	72.1	58.6
www.bookpool.com	91.7	82.2	86.7	56.1	74.7	64.1
www.amazon.com	85.4	82.6	83.9	51.2	76.3	61.3
www.adebooks.com	83.9	76.4	79.9	40.5	70.3	51.4
www.discount-pcbooks.com	81.5	74.4	77.8	55.1	67.9	60.9

实验结果表明 DSE 具有较 RoadRunner++优越的性能. 从结果中发现，两种方法在查全率方面结果相似，但是 RoadRunner++在查准率方面准确度大大降低. 其主要原因在于 RoadRunner++从目标网站中抽取出大量信息，仅依据编辑距离对其进行语义标注，它不考虑其它特征，例如数据模式、展示特征等，导致较低的查准率.

4.3.2 2DCC-CRFs 与传统条件随机场模型的比较

本文在 3 个数据集上通过实验比较了 2DCC-CRFs 与 Linear-Chain CRFs、2DCRFs 模型在 Web 数据语义标注上的性能. 基于 Linear-Chain CRFs 和 2DCRFs 的语义标注方法详见文献[13]. 表 4、表 5 和表 6 显示了在每个字段上的查全率、查准率、$F1$ 值以及平均 $F1$ 值.

表 4 在数据集 B 上的语义标注结果

字段	Linear-Chain CRFs 标注结果			2DCRFs 标注结果			2DCC-CRFs 标注结果		
	Recall/%	*Precision*/%	*F1*/%	*Recall*/%	*Precision*/%	*F1*/%	*Recall*/%	*Precision*/%	*F1*/%
Title	47.53	46.81	47.16	50.00	47.06	48.49	43.13	79.18	55.84
Author	51.24	50.94	51.08	51.61	52.58	52.09	62.52	49.14	55.02
ISBN	70.25	82.51	75.89	73.33	91.67	81.48	92.94	74.12	82.47
Pages	76.42	61.81	68.34	80.61	66.67	72.98	91.37	64.13	75.36
Publish Dates	76.48	82.58	79.41	81.25	86.67	83.87	81.31	87.29	84.19
Original Price	36.57	53.23	43.35	37.69	55.22	44.81	73.21	73.28	73.24
Current Price	62.27	45.39	52.51	69.78	49.94	58.22	68.17	76.42	72.06
You save	79.57	68.75	73.76	88.05	73.33	79.99	92.02	81.07	86.20
Average *F1*			61.44			65.24			73.05

表 5　在数据集 D 上的语义标注结果

字段	Linear-Chain CRFs 标注结果			2DCRFs 标注结果			2DCC-CRFs 标注结果		
	Recall/%	*Precision*/%	*F*1/%	*Recall*/%	*Precision*/%	*F*1/%	*Recall*/%	*Precision*/%	*F*1/%
Brand	88.34	86.79	87.56	88.89	87.22	88.05	92.44	90.26	91.34
Operating System	65.41	83.17	73.23	66.67	83.33	74.07	74.13	86.89	80.10
DVD/CD Drive	91.84	83.47	87.46	92.71	83.33	87.78	89.50	86.71	88.07
Processor	66.12	65.81	65.33	66.67	67.34	67.01	73.67	71.31	72.47
Main Frequency	91.81	61.72	73.82	93.75	62.50	75.02	84.19	87.39	85.76
RAM	68.28	90.27	77.75	68.75	90.91	78.29	73.64	84.12	78.53
Hard Drive	86.24	91.18	88.64	86.67	92.31	89.40	89.33	93.75	91.49
Average *F*1			79.11		79.94		83.97		

表 6　在数据集 P 上的语义标注结果

字段	Linear-Chain CRFs 标注结果			2DCRFs 标注结果			2DCC-CRFs 标注结果		
	Recall/%	*Precision*/%	*F*1/%	*Recall*/%	*Precision*/%	*F*1/%	*Recall*/%	*Precision*/%	*F*1/%
Author	92.89	76.08	83.65	93.06	76.92	84.22	94.43	80.91	87.15
Title	89.21	57.42	69.87	89.47	56.25	69.07	91.36	79.82	85.20
Editor	44.58	56.71	49.92	44.21	58.16	50.23	47.19	54.89	50.74
Book Title	48.89	74.87	59.15	50.00	75.48	60.15	66.67	63.10	64.84
Date	69.18	76.82	72.80	71.72	77.78	74.63	82.32	91.29	86.57
Journal	55.12	84.92	66.85	55.98	86.21	67.88	62.51	82.65	71.18
Volume	49.75	88.24	63.62	50.51	90.81	64.91	71.67	82.92	79.12
Tech	66.18	62.12	64.09	67.74	62.88	65.22	62.58	68.69	65.49
Institution	72.98	87.92	79.76	73.74	88.67	80.52	79.71	82.31	80.99
Pages	90.59	65.98	76.35	90.76	66.67	76.87	92.18	67.32	77.81
Location	43.19	62.15	50.96	43.29	62.48	51.14	68.65	65.49	67.03
Publisher	56.71	83.49	67.54	57.14	83.15	67.73	60.64	80.61	69.21
Average F1			67.05		67.71		73.78		

从表 4～6 中可以看出，基于 2DCC-CRFs 模型的方法在 Web 数据语义标注上的总体性能要优于基于 Linear-Chain CRFs 和 2DCRFs 模型的方法. 与 2DCRFs 相比，*F*1 的平均值在 3 个数据集上分别提高了 7.81%、4.03%和 6.07%，并且每个字段的 *F*1 值均有所提高. 实验表明，通过增加关联边，可以充分利用 Web 数据元素间潜在的长距离依赖联系，进一步降低 Web 数据语义标注的错误率. 另外，对于一些数据模式完全相同的字段，语义标注准确性的增长尤为明显. 例如，对于数据集 B 上的 Original Pirce、Our Price、You Save 3 个字段，*F*1 值的平均增幅达到 16.2%. 分析其主要原因在于，通过对数据库中的数据记录进行规则挖掘发现，如果在一条记录中同时出现 3 个价格，价格之间存在一定的大小关系，并且两个价格之和等于第 3 个价格，那么可以较准确地确定 3 个数据元素的语义标签.

另外，我们也发现 2DCC-CRFs 模型在部分数据字段的查全率或查准率并没有同时被提高，单个指标出现了下降. 例如，在数据集 B 上，相比 2DCRFs 模型，Author 字段的查全率提高了 10.91%，但是查准率却降低了 3.44%，这反映了在查全率和查准率之间存在着相反的相互依赖关系.

通过观察进一步发现，在数据集 D 和 P 上，Linear-Chain CRFs 和 2DCRFs 模型的语义标注性能相差不大，主要原因在于，数据集 D 和 P 上的数据元素间不存在二维依赖联系. 而数据集 B 上的数据元素间存在二维依赖联系，所以，相对于 Linear-Chain CRFs、2DCRFs 模型在不同属性上的标注准确性都有所提高.

在图 3 中，显示了 2DCC-CRFs 和其它两个模型在实例标注准确率这个指标上的性能比较. 从图中可以看出在 3 个数据集上，相对于 Linear-Chain CRFs 和 2DCRFs、2DCC-CRFs 的实例标注准确率均有不同程度地提高. 相对于 2DCRFs 模型，2DCC-CRFs 模型的实例标注准确率分别提高了 9.43%、5.18%和 6.15%. 其中，数据集 B 上的增幅最为显著，主要原因在于，通过增加对长距离依赖联系的处

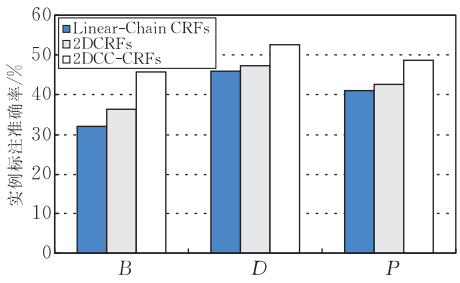


图 3　Linear-Chain CRFs、2DCRFs 和 2DCC-CRFs 模型在不同数据集上的实例标注准确率

理,有效地降低了对 Original Pirce、Our Price、You Save 3 个字段的标记错误率,提高了实例标注准确率.

4.3.3 数据库参与与否对 2DCC-CRFs 模型性能的影响

在数据集 B 上,通过实验分析了在数据库参与和数据库不参与两种情况下,对 Web 数据语义标注性能的影响.参与模型训练的样本数据由随机抽取的 1000 条图书记录(简称 L)组成,用于测试的是另外 1000 条图书记录.图书数据库(简称 DB)主要用于确定部分 Web 数据元素的语义标签,不作为训练集使用.

实验分成 2 组:第 1 组数据库 DB 不参与,用在 L 上训练得到的 2DCC-CRFs 模型对测试记录进行语义标注;第 2 组,数据库 DB 参与,2DCC-CRFs 模型是在 DB 和 L 组成的联合样本数据上训练得到.表 7 给出了在这个数据集上选择的 8 个典型字段的语义标注结果.每个字段的性能使用查全率、查准率和 $F1$ 这 3 个指标来评价,同时也计算了平均 $F1$ 值.

表 7 包含数据库和不包含数据库两种情况下的语义标注结果

字段	L 上的标注结果			L+DB 上的标注结果		
	Recall/%	*Precision*/%	*F*1/%	*Recall*/%	*Precision*/%	*F*1/%
Title	38.27	77.56	51.25	43.13	79.18	55.84
Author	45.16	46.94	46.03	62.52	49.14	55.02
ISBN	92.86	65.22	76.62	92.94	74.12	82.47
Pages	89.44	62.63	73.67	91.37	64.13	75.36
Publish Dates	79.14	76.47	77.78	81.31	87.29	84.19
Original Price	71.13	69.23	70.17	73.21	73.28	73.24
Our Price	66.67	75.86	70.97	68.17	76.42	72.06
You save	91.00	79.31	84.75	92.02	81.07	86.20
Average *F*1			68.89		73.05	

从表 7 中可以看出,第 2 组实验由于包含了数据库,绝大多数属性的查全率、查准率和 $F1$ 这 3 个指标都有明显提高,$F1$ 的平均值提高了约 4.16%.性能提高的主要原因在于,数据库的结构化信息及记录特征较手工标注训练集中的样本数据的文本特征更为准确地确定了相关数据元素的语义标签,进而准确地建立关联边,有效地利用了 Web 数据元素间潜在的长距离依赖联系.

结合表 4 可以发现,与 2DCRFs 模型相比,即使没有数据库的参与,2DCC-CRFs 模型的标注性能也有所提高,$F1$ 的平均值提高了约 3.65%.但是,也发现有一些字段的 $F1$ 值有所降低,例如 Author、publication dates 字段.分析实验数据发现,仅根据手工标注训练集中的样本数据的文本特征,导致一些数据元素的语义标签被错误识别,进而导致生成了一些错误的关联边,导致一些字段的 $F1$ 值下降.

该实验进一步表明,准确地确定关联边,建立 Web 数据元素间正确的长距离依赖联系,对于模型的性能有着重要的影响.

4.3.4 数据库规模对 2DCC-CRFs 模型性能的影响

为了进一步验证使用数据库信息的有效性,通过实验测试了数据库记录的增加对模型标注性能的影响.从论文数据库中随机选择了 0 条、5000 条、10000 条和 20000 条记录来得到不同规模的数据库,在每个数据库上执行同样的实验.图 4(a)和(b)分别展示了$F1$的平均值和实例标注准确率两个指标

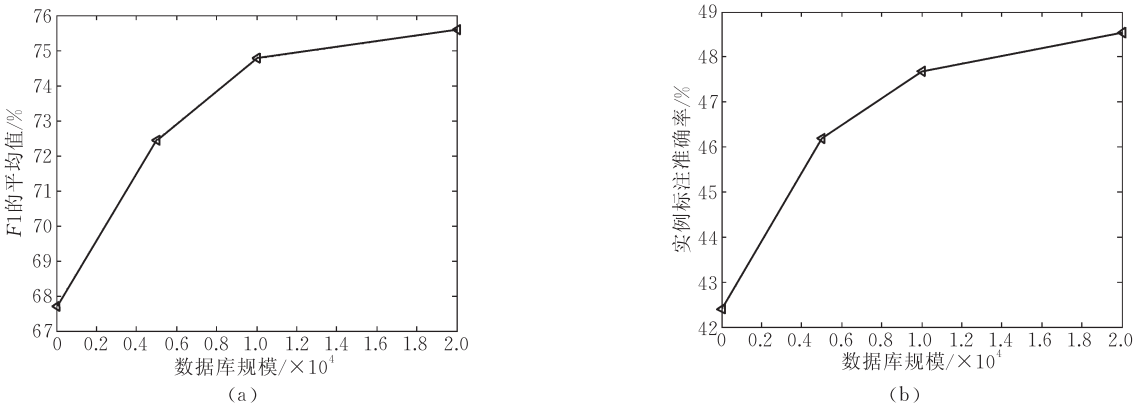


图 4 数据库规模对 $F1$ 的平均值和实例标注准确率的影响(基于数据集 P)

随着数据库规模变化而发生变化的曲线.通过观察发现,F_1 的平均值和实例标注准确率随着数据库规模的增大逐步提高.但是,两条曲线均随着数据库规模的增大而逐渐变得扁平.这个实验表明,是否使用数据库对于 2DCC-CRFs 模型的语义标注性能有着明显的影响;但是,随着数据库规模的继续增大,使用数据库所展示出的有效性在逐步降低.

5　结　论

本文提出了一种基于集成学习和二维关联边条件随机场的 Web 数据语义标注方法,通过结合现有的包装器生成技术,可以有效地解决大规模 Web 信息抽取问题.该方法首先利用已抽取信息和目标网站训练页面中呈现的特征构造多个分类器;接着,利用构造的基于不同特征的分类器对待标注文本进行判断;然后,利用基于 D-S 证据理论的集成学习方法对多个分类器的分类结果进行合并,完成对待标注文本的判断,识别出属性标签和数据元素;最后,利用 2DCC-CRF 模型完成对数据元素的自动语义标注.

参　考　文　献

[1] Zhai Y H, Liu B. Web data extraction based on partial tree alignment//Proceedings of the 14th International Conference on World Wide Web. Chiba, Japan, 2005: 76-85

[2] Chang C H, Kayed M, Girgis M R, Shaalan K. A survey of web information extraction systems. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(10): 1411-1428

[3] Crescenzi V, Mecca G, Merialdo P. Roadrunner: Towards automatic data extraction from large web sites//Proceedings of the Very Large DataBase. Roma, Italy, 2001: 109-118

[4] Nie Zai-Qing, Wen Ji-Rong, Ma Wei-Ying. Webpage understanding: Beyond page-level search. SIGMOD Record, 2008, 37(4): 48-54

[5] Wong Tak-Lam, Lam Wai. Learning to adapt web information extraction knowledge and discovering new attributes via a Bayesian approach. IEEE Transactions on Knowledge and Data Engineering, to appear

[6] Lerman K, Getoor L, Minton S, Knoblock C. Using the structure of web sites for automatic segmentation of tables//Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data. Paris, France, 2004: 119-130

[7] Embley D, Campbell D, Jiang Y et al. Conceptual-model-based data extraction from multiple-record web pages. Data and Knowledge Engineering, 1999, 31(3): 227-251

[8] Mukherjee S, Ramakrishnan I V, Singh A. Bootstrapping semantic annotation for content-rich html documents//Proceedings of the 21st International Conference on Data Engineering. Tokyo, Japan, 2005: 583-593

[9] Arlotta L, Crescenzi V, Mecca G, Merialdo P. Automatic annotation of data extracted from large web sites//Proceedings of the WebDB. San Diego, USA, 2003: 7-12

[10] Reformat M, Ronald R. Building ensemble classifiers using belief functions and owa operators. Soft Computing, 2008, 12(6): 543-558

[11] Altincay H, Demirekler M. Speaker identification by combining multiple classifiers using dempster-shafer theory of evidence. Speech Communication, 2003, 41(4): 531-547

[12] Altincay H. A dempster-shafer theoretic framework for boosting based ensemble design. Pattern Analysis & Applications, 2005, 3(12): 287-302

[13] Zhu J, Nie Z Q, Wen J R et al. 2D conditional random fields for web information extraction//Proceedings of the 22nd International Conference on Machine Learning. Bonn, Germany, 2005: 1044-1051

[14] Zhu Jun, Nie Zai-Qing, Wen Ji-Rong, Zhang Bo, Ma Wei-Ying. Simultaneous record detection and attribute labeling in web data extraction//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data mining. Philadelphia, USA, 2006: 494-503

[15] Sutton C, McCallum A. Collective segmentation and labeling of distant entities in information extraction. England: University of Massachusetts, Technical Report: 04-49, 2004

[16] Huang Jian-Bin, Ji Hong-Bing, Sun He-Li. Integration of heterogeneous of Web records using mixed skip-shain conditional fields. Journal of Software, 2008, 19(8): 2149-2158 (in Chinese)
(黄健斌, 姬红兵, 孙鹤立. 基于混合链条件随机场的异构 Web 记录集成方法. 软件学报, 2008, 19(8): 2149-2158)

[17] Lin W Y, Lam W. Learning to extract hierarchical information from semi-structured documents//Proceedings of the 9th International Conference on Information and Knowledge Management. McLean, USA, 2000: 250-257

[18] Nie Zai-Qing, Wen Ji-Rong, Ma Wei-Ying. Webpage understanding: Beyond page-level search. SIGMOD Record, 2008, 37(4): 48-54

[19] Luo Hui-Lan, Kong Fan-Sheng, Li Yi-Xiao. An analysis of diversity measures in clustering ensembles. Chinese Journal of Computers, 2007, 30(8): 1315-1324(in Chinese)
(罗会兰, 孔繁胜, 李一啸. 聚类集成中的差异性度量研究. 计算机学报, 2007, 30(8): 1315-1324)

[20] Ruthven I, Lalmas M. Using dempster-shafer's theory of evidence to combine aspects of information use. Journal of Intelligent Information Systems, 2003, 19(3): 267-301

[21] Raybaud Sylvain, Lavecchia Caroline, Langlois David, Smaïli Kamel. New confidence measures for statistical machine translation//Proceedings of the International Conference on Agents and Artificial Intelligence. Porto, Portugal, 2009: 61-68

[22] Nie Zai-Qing, Wu Fei, Wen Ji-Rong, Ma Wei-Ying. Extracting objects from the web//Proceedings of the 22nd International Conference on Data Engineering. Atlanta, USA, 2006: 123

[23] Liu D C, Nocedal J. On the limited memory bfgs method for large scale optimization. Mathematical Programming, 1989, 45: 503-528

[24] Kevin P M, Yair W, Michael I J. Loopy belief propagation for approximate inference: An empirical study//Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence. Stockholm, Sweden, 1999: 467-475

DING Yan-Hui, born in 1981, Ph. D. candidate. His current research interests include Web information integration.

LI Qing-Zhong, born in 1965, professor, Ph. D. supervisor. His main research interests include Web information integration and enterprise information integration.

DONG Yong-Quan, born in 1979, Ph. D. candidate. His current research interests focus on Web information integration.

PENG Zhao-Hui, born in 1978, Ph. D., lecturer. His main research interests focus on information retrieval.

Background

Large-scale Web information extraction is one of the key steps of Web information integration. Large-scale Web information extraction needs to extract information from many Web sites accurately and automatically. A lot of research aiming at extracting Web information have been conducted. However, most existing Web information extraction methods only focus on single Web site, which causes that they can't meet he need of large-scale Web information extraction. One major limitation of most existing Web information extraction methods is that they employ supervised learning approaches requiring manually prepared training examples for each Web site. Moreover, the wrapper learned from a particular Web site cannot be applied to other sites for extracting information.

Tak-Lam Wong proposed a Bayesian learning framework for adapting information extraction wrappers with new attribute discovery, reducing human effort in extracting precise information from unseen Web sites. However, the method can only use simple rule to discover the semantic labels of new attribute, which results in low accuracy.

The method proposed in this paper makes good use of the long distance dependencies and the short distance dependencies between Web data elements, which can improve the semantic annotation accuracy of new attribute. In the other hand, our method can solve the problem of large-scale Web information extraction efficiently.

This work is also supported by the National Natural Science Foundation of China under grant No. 90818001 and the Natural Science Foundation of Shandong Province of China under grant No. Y2007G24.