DOI:10.16208/j.issn1000-7024.2014.01.037

基于模板和领域本体的 Deep Web 信息抽取研究

顾韵华[1,2]，高 原[1,2]，高 宝[1,2]，杜 杰[1,2]

（1. 南京信息工程大学 江苏省网络监控中心，江苏 南京 210044；
2. 南京信息工程大学 计算机与软件学院，江苏 南京 210044）

摘 要：为简化模板的抽取规则、提高抽取的准确率，提出了一种基于双模板和领域本体的 Deep Web 信息抽取方法。该方法采用 DIV 块模板和表格模板结合的方法，建立双模板。利用基于中文分词的网页预处理结果，在领域本体知识的指导下，通过 C4.5 决策树算法来训练分类模型，筛选出待抽取的 DIV 块序号，构建 DIV 块模板，从而可以精确定位到数据块。利用 XML 技术构建 XSLT 文档，得到表格模板的抽取规则，从而抽取出数据片段。选取天气领域进行 Deep Web 信息抽取实验，实验结果表明，抽取准确率和召回率都可以达到 95%以上，取得了较好的抽取效果。

关键词：Deep Web；信息抽取；模板；领域本体；决策树

中图法分类号：TP311 **文献标识号**：A **文章编号**：1000-7024（2014）01-0327-06

Research on Deep Web information extraction based on template and domain ontology

GU Yun-hua[1,2], GAO Yuan[1,2], GAO Bao[1,2], DU Jie[1,2]

(1. Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science and Technology, Nanjing 210044, China; 2. College of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract: To simplify the extraction rules for the template to improve the extraction accuracy, an algorithm based on template and domain ontology was presented to extract Deep Web information. The combinations of DIV block template and table template are used. Using the result of web page pretreatment based on Chinese word segmentation, under the guidance of domain ontology knowledge, by the algorithm of C4.5 decision tree to train the classifier, the number of extracted DIV blocks is selected, and the template of DIV blocks is built which can locate the data area. Then XSLT document is constructed using the technology of XML, and forming the table template helps extracting the data fragment. The result of the Deep Web information extraction experiment in the field of weather, show that average accuracy rate and recall rate can achieve above 95% and better extraction effect is obtained.

Key words: Deep Web; information extraction; template; domain ontology; decision tree

0 引 言

Deep Web 相对于表面网（surface web）而言，蕴含着更加丰富而专业的数据资源[1]。据统计，中国 Deep Web 大约有 24000 个站点，28000 个后台数据库和 74000 个查询接口[2]，目前仍在快速增长。有效的利用 Deep Web 的丰富信息资源，能够更好地满足人们学习和查找知识的需求。

Deep Web 信息抽取的目的是从 Deep Web 结果页面中抽取出有价值的信息[3]。虽然目前的抽取技术已经发展到自动化程度，但抽取数据的准确率较低且抽取规则的适应性较差。手工编写规则可以达到很高的准确率，但是规则繁琐，代价也很大。本文引入 DIV 块和表格双重模板，同时，考虑信息内部联系，引入领域本体来指导模板的建立，可减少无关信息，简化模板的抽取规则，提高抽取的准确

收稿日期：2013-04-22；修订日期：2013-06-28

基金项目：国家自然科学基金项目（61103142）

作者简介：顾韵华（1965-），女，江苏泰州人，教授，CCF 会员，研究方向为 Web 数据挖掘、信息安全；高原（1988-），女，江苏扬州人，硕士研究生，研究方向为 Web 数据挖掘；高宝（1989-），女，江苏徐州人，硕士研究生，研究方向为 Web 数据挖掘；杜杰（1979-），男，江苏南京人，副教授，研究方向为智能计算、信息安全。E-mail：yhgu@nuist.edu.cn

率。此双重模板是基于 DIV 块和表格构建而成的，在具体的抽取过程中，这两种模板是先后使用的关系，先使用 DIV 块模板进行粗粒度的信息抽取，再使用表格模板进行细粒度的信息抽取。

1 相关研究

目前，对 Deep Web 信息抽取的研究成果大多数集中在 DOM 树的挖掘上，包括基于 DOM 树的 Deep Web 实体抽取、基于重复模式的 Deep Web 信息抽取、基于 DOM 与模板的结合和基于视觉特征的 Deep Web 信息抽取等方法。

文献［4］提出了一种基于 DOM 树的 Deep Web 实体抽取机制，采用基于 DOM 树的自动实体抽取策略，利用 DOM 树中的文本内容和层次结构来确定数据区域和实体区域。该方法在多个实体显示在 Web 页面中的同一行时会造成各个实体的 DOM 树结构互相参杂，实验效果中针对电子商务领域的抽取性能相对较差。

文献［5］提出了一种 Deep Web 数据源下重复记录识别模型，在数据预处理模块中将所抽取的数据生成实体记录形式，在异构记录处理模板中利用在同构记录处理模块所得到的权重，计算各实体记录的相似度，得到重复记录。

文献［6-8］提出了基于模板的抽取方法。通过对产生于同一模板的网页的对比分析总结出一个通用的抽取模板，从而免去对众多网页进行重复处理的繁琐。文献［6］将网页模板表示为一个正则表达式。首先利用网页的树状结构特点计算子树的相似度生成一种特殊的树，接着利用此树生成模板，再利用一系列合并规则对模板进行修剪。此类模板的生成过程比较复杂。文献［7，8］依赖于 XPath 表达式进行待抽取信息节点的定位。对于专利信息等有规律且更新不频繁的网站，这类模板比较清晰，易于实现。而对于复杂的网页来说，XPath 表达式就会变长，越长就越不稳定。

文献［9］提出了基于视觉的方法，利用深层网页的视觉功能，以实现 Deep Web 数据提取，包括数据记录提取和数据项提取。视觉特征包括字体的颜色和大小、文本的长度等。但是网页设计的多样性给基于视觉特征的抽取方法增加了难度。

通过上述研究发现，现有的 Deep Web 信息抽取技术并不能完全准确而自动地抽取网页信息。为了尽量减少人工干预和复杂性，本文从模板和语义的角度出发，建立“DIV＋Table”双模板，对 Deep Web 页面中有意义的信息进行准确定位和抽取。对基于领域的中文 Deep Web 网站的信息抽取有着实用的意义，另一方面，加入了语义信息之后，有利于 Deep Web 信息集成以及语义数据的处理。

2 基于模板与领域本题的 Deep Web 信息抽取

2.1 Deep Web 信息抽取框架

本文所设计的抽取框架，分为模板构建和目标页面信息抽取两个部分。模板构建是在领域本体的指导下构建 Deep Web 站点的模板，并将模板存至模板库，为页面信息抽取而服务。页面信息抽取则是从模板库中选择匹配的模板，利用模板对应的抽取规则进行信息抽取。该框架如图 1 所示。其中网页预处理的目标是将 HTML 文档处理成以 DIV 块为基本单元，并含有中文分词结果的数据集合。此数据集合经过适当的筛选，即可作为决策树分类模型的训练数据集。

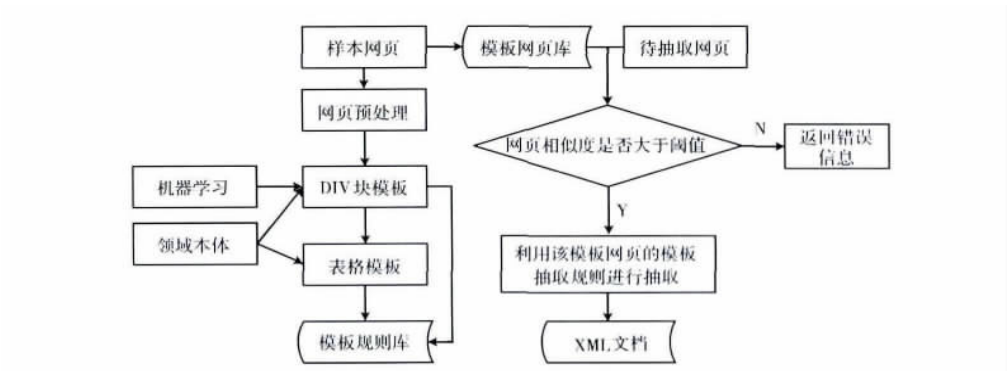

图 1 Deep Web 信息抽取框架

2.2 领域本体及其构建

领域本体作为某个领域内不同主体之间进行交流的语义基础，在模板构建过程中能够起到优化的作用，减少模板中出现与领域不相关的内容[10]。

本体的构建需要完整的工程化、系统化的方法来支持。很少有通用的大规模本体，大多数的本体只是针对某个具体应用领域构建的。本文借鉴斯坦福大学医学院开发的七步法[11]的思想，构建天气和图书领域的本体知识库。

针对天气和图书领域，对国内多个 Deep Web 网站进行调查分析，从中提炼出一些核心概念、概念之间的关系及相关实例。定义领域本体为一个六元组。

定义 1 领域本体 $O=\{C, H, R, P^D, P^O, I\}$。其中 O 代表本体的名称，C(concepts 或者 class) 为概念的集合，H(hierarchy) 为概念之间层次的集合，R(relationship) 为

概念之间关系的集合，P^D（datatype property）为数据属性的集合，P^O（object property）为概念属性的集合，I（instances）为实例的集合。图 2 展示了天气领域本体层次结构。

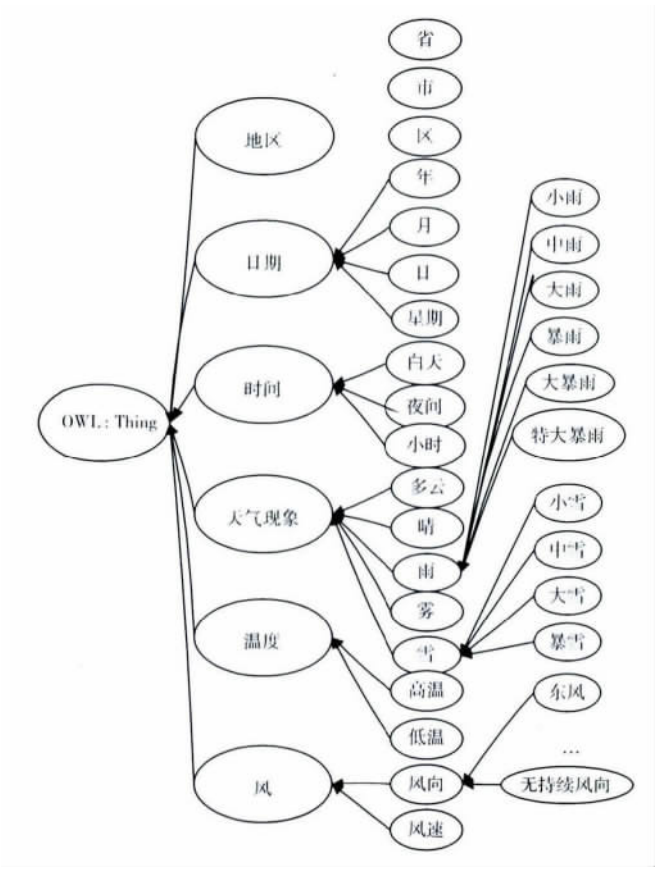

图 2　天气领域本体的层次

2.3　基于领域本体指导的模板构建

2.3.1　引入双重模板

Deep Web 的信息抽取任务不仅是要识别出数据块，更重要的是抽取出数据片段。这样抽取出的数据才有意义。本文以这两个任务为出发点，先通过 DIV 块模板定位到数据块，再通过表格模板定位到数据片段。图 3 中分别标识了一个详细信息页面中的数据块和数据片段。

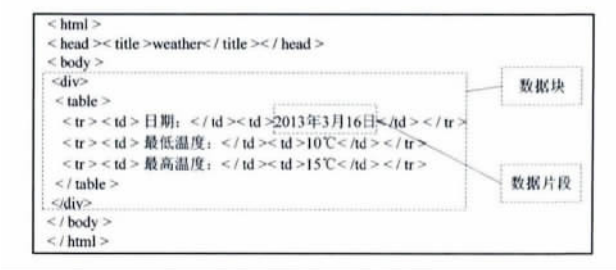

图 3　数据块和数据片段的定义

2.3.2　DIV 块模板的定义和构建

网页模板是指一种网页框架，决定了网页的基本结构和文档设置。目前大多数的网页布局通常采用“DIV＋CSS”方式。“＜div＞”标签用于把文档分割成独立的、不同的 DIV 块。对于一个网页设计者来说，首先要考虑的是页面内容的语义和结构。因此，需要分析 DIV 块以及每个 DIV 块服务的目的。

Deep Web 查询结果页面具有基于 DIV 块的模板化的特征。这些页面可以分为不变和可变部分，不变的部分是网页中内容块的组织顺序、语义说明和静态信息，可变的部分是经过查询所得到的动态结果，这也正是所要抽取的内容，它们存在于一个或者多个 DIV 内容块中。因此，可以将 DIV 模板定义为所要抽取的 DIV 块的集合。用 DIV 块在 HTML 文档中的序号进行形式化定义。

定义 2　DIV 块模板 $M=\{Name, Type, \{D_i, D_j, \ldots\}, Number, Time\}$，其中 Name 是指 Deep Web 站点的名称；Type 是指这个站点的模板种类；$\{D_i, D_j, \ldots\}$ 是指所要抽取的 DIV 块的集合，下标 i 代表这个 DIV 块在 HTML 代码中的序号；Number 代表共有多少个 DIV 块构成了一个完整的抽取内容；Time 代表了模板的建立时间，便于能定期更新模板。保证模板的有效性。

例如　$M=\{weatherchina, one, \{D_{20}, D_{23}\}, 2, 2012.11.1\}$，代表 2012 年 11 月 1 日建立了用于抽取“中国天气网”中的一种查询结果页面模板，共有 2 个 DIV 内容块构成了抽取内容，分别为第 20 和 23 个 DIV 内容块。

本文将构建 DIV 块模板的过程看作是识别所要抽取的 DIV 数据块的过程。将网页预处理的结果作为训练数据集，结合预先构建好的领域本体知识，采用决策树学习算法来学习分类模型，分类模型将 DIV 块分为需要抽取的和不需要抽取的这两类。通过此分类模型就可以对新的 DIV 块集进行分类。

决策树算法采用自顶向下的方式将从一组训练数据中学习到的函数表示为一颗分类决策树。这种算法适用于分类数据和归纳决策规则，具有简化处理流程，算法复杂度低的优势。常用的决策树算法有 ID3、C4.5 等。ID3 算法最初的定义是假设属性值是离散值，但在实际环境中，有很多属性是连续的，不能用一个确定的标准来对其进行划分。C4.5 使用一系列处理过程将连续的属性划分成离散的属性，进而达到构建决策树的目的。C4.5 算法的优点在于产生的分类规则易于理解，准确率较高。

C4.5 采用信息增益率作为度量选择属性的指标。信息增益（gain ratio）的概念能表述选择某一个属性后再选择其它属性时信息量的变化。信息增益是基于熵来度量信息的增量。熵作为数据混杂度的衡量指标，其值越小代表数据越纯。式（1）描述的是数据集 D 信息熵的计算方法，式（2）描述的是信息论中的熵，式（3）描述的是属性 A_i 信息增益率，Entropy（D）表示区分前的熵，$Entropy_{A_i}$（D）表示根据属性 A_i 划分后的熵

$$Entropy_{A_i}(D) = \sum_{j=1}^{|v|} \frac{|D_j|}{|D|} \times Entropy(D_j) \quad (1)$$

$$Entropy(D) = \sum_{j=1}^{|C|} \Pr(C_j)\log_2 \Pr(C_j) \quad (2)$$

$$Gain(D,A) = Entropy(D) - Entropy_{A_i}(D) \quad (3)$$

为了训练分类模型，需要将经过网页预处理得到的数据进行适当筛选，作为训练数据集。训练集中属性是根据领域本体知识进行选取的，需要选择领域本体中定义的若干词汇。

以天气领域的 DIV 块分析为例，见表 1。其中 num __day 表示“白天”出现的次数，has __else 表示是否有类似于“版权”、“旅游”和“防晒”等词，morecity 表示是否有多个地名，hasde 表示是否含有词“的”，lessthan5 表示分词个数是否小于 5，IsNeedDIV 代表类别，指明是否是需要的 DIV 块。

表 1 用于分类的数据

NO	num __day	has __else	morecity	hasde	lessthan5	IsNeedDIV
1	3	yes	no	no	no	DIV
2	4	no	no	no	no	DIV
…	…	…	…	…	…	…
k	0	yes	no	no	yes	NOTDIV
18	9	yes	yes	yes	no	NOTDIV

经过训练，得到决策树分类模型如图 4 所示。

```
hasde = no
|   lessthan5 = no
|   |   num_day <= 4: DIV(7.0/1.0)
|   |   num_day > 4: NOTDIV(2.0)
|   lessthan5 = yes: NOTDIV(3.0)
hasde = yes: DOTDIV(6.0)

Number of Leaves :  4

Size of the tree :  7
```

图 4 决策树分类模型

可以看出分类模型的准确率是 0.952。当准确率达到一定要求的时候，就能确保 DIV 块的判断不会出错。这比完全凭借启发式规则更加可靠。

2.3.3 表格模板的定义和构建

为了抽取出数据片段，还需要构建另一种模板，也就是表格模板。本文中对数据片段的抽取是利用 XML 技术的。使用的解析页面模板为 XML 文件，而模板中的抽取规则是基于 DOM 和 Xpath 的表格节点定位。可以将 Xpath 理解为 XML 的 SQL 语句。它基于 XML 文档的逻辑结构，用 Path 来确定 XML 文档中某部分位置。

XSLT 是一种对 XML 文档进行转化的语言。XSLT 指令通常与 XPath 表达式结合使用。XSLT 包含一组称为模板的规则，模板规则用 xsl:template 元素表示，每个＜xsl:template＞元素包含当一个特定节点匹配时所应用的规则。从网页抽取的角度，将 XSLT 文档看作抽取规则。

结合上述 XML 的相关知识，可以将表格模板定义为如下形式：

定义 3 表格模板 T＝{Name，Type，Path}，其中“Name”、“Type”与 DIV 块模板中的定义是一致的，这也便于最终将两种模板结合在一起来抽取 Deep Web 网页。Path 是指数据片段在 DOM 树中的路径表达式。

本文在生成表格模板的过程中，采用的流程如图 5 所示。

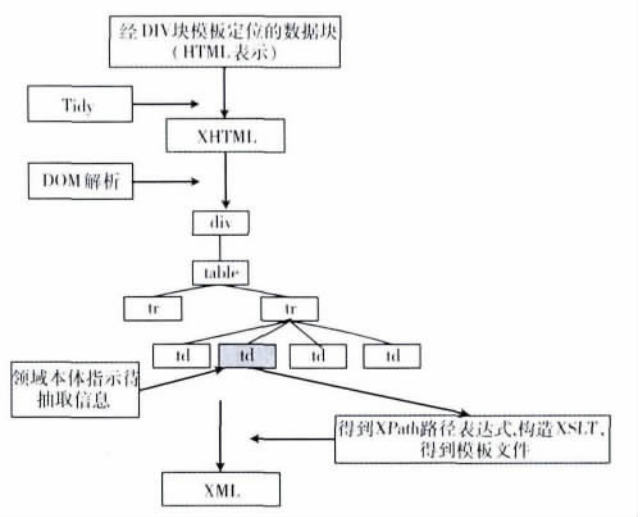

图 5 表格模板的生成流程

以“中国天气网”为例，以下是根据数据片段的 Xpath 得到的部分 XSLT 文件，其中“/div[1] /div[1] /table[2] /tr[1] /td[4] /text()”代表待抽取数据片段在 XML 文档中的路径信息。

```
<day>
  <condition><xsl: value－of select=" /div[1] /div[1] /table[2] /tr[1] /td[4] /text( )" /></condition>
  <maximum temperature><xsl:value-of select="..." /></maximum temperature>
  <minimum temperature><xsl:value-of select="..." /></minimum temperature>
  <wind><xsl:value-of select="..." /></wind>
  <windpower><xsl:value-of select="..." /></windpower>
</day>
```

2.4 基于 URL 和网页相似度的模板匹配

模板匹配的目的，一方面是为了扩充模板库，另一方

面是为了选择合适的模板对新的待抽取网页执行抽取任务。传统的模板匹配方法仅仅是基于 URL 的，然而这种方法在具体的应用中存在误差。为解决此问题，本文提出将 URL 与网页相似度相结合的算法，可获得更精确的模板匹配结果。网页相似度是衡量不同网页相似程度的指标，本文采用内容与结构相结合的网页相似度计算方法。算法如下：

步骤 1　将待匹配网页 P_A 解析成 DOM 树；

步骤 2　利用 URL 相似度获取模板网页 P_T，同时将模板网页解析成 DOM 树；

步骤 3　计算两个网页的 DIV 块总数。如果 N_T 与 N_A 相等且都为 1，则返回相似度为 1，并结束算法；如果 N_T 与 N_A 不相等，则选取某一个 K 值，继续执行下一步；

步骤 4　采用字符串编辑距离算法分别比较 P_A 和 P_T 中 DIV 块序号为 k(k∈K) 的文本相似度；

步骤 5　将 K 个文本的相似度进行叠加，除以 K，返回网页相似度。如下所示

$$Sim(P_A,P_T)=\frac{\sum_{i=1}^{K} Levenshtein(DIVP_A(i),DIVP_T(i))}{K} \tag{4}$$

在网页相似度的计算方法中，用到了两个阈值 K 和 ε。K 是要比较的最合适的 DIV 块数目，ε 代表选取的最合适的相似度。结合 DIV 块模板，将 K 设为 $(a_{last}-a_{first})$，表示从 DIV 模板数组的第一个一直匹配到数组的最后一个。

为了选取合适的 ε 值，本文进行了以下实验。从 5 个不同的 Deep Web 站点，分别各选取 10 个网页，作为模板网页，再各选取 10 个网页作为待匹配网页。分别计算相似度。每个网站计算的次数为 100 次，统计结果如图 6 所示。可以看出相似和不相似的网页区分度很大，因此设定 ε=0.9。

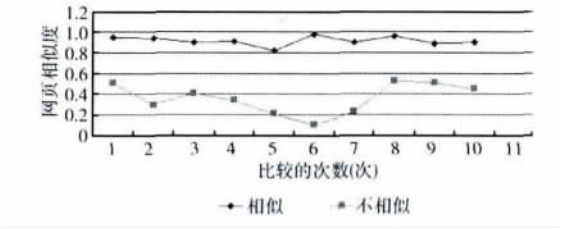

图 6　相似度计算结果

实验表明，若待匹配网页与模板网页结构相似度大于 0.9，则模板匹配，存在抽取规则；否则以不匹配作相应处理。

将本文提出的基于 URL 和网页相似度计算的模板匹配方法，与传统的仅基于 URL 的匹配方法进行对比，统计匹配的正确率。同样选取 5 个 Deep Web 网站，对每个站点只归纳一种模板。再另外选取与这 5 个模板网页的 URL 相似的若干个网页，分别利用两种模板匹配方法，进行实验。对匹配成功的网页利用相应的模板进行抽取，若能抽取出模板设定的结果，则表明匹配正确；否则，表明匹配的不正确。其统计结果见表 2。可以看出，结合了网页相似度的模板匹配能明显提高匹配的正确率。

表 2　两种模板匹配方法的正确率对比

网站	传统的基于 URL 的匹配法	本文方法
Site1	76.6%	95.8%
Site2	60.6%	100%
Site3	75.3%	90.4%
Site4	65%	90%
Site5	72.5%	89.2%

3　Deep Web 信息抽取实验

针对天气领域选取了 5 个 Deep Web 站点作为数据的来源。站点的选择依据是 Google PageRank 得出的网站排名。这个排名综合考虑了网站的用户体验和用户数量，属于人们经常关注的网站，信息量比较全，可以为实验提供大量的测试网页。

评价信息抽取的指标是查准率（precision），召回率（recall）以及 F 值（F-measure）。查准率是抽取的信息中正确的点数所占的比率，召回率是测试被正确抽取的信息点的比例，F 指标反映了信息抽取的综合性能。计算公式分别表示如下

$$\text{查准率（P）}=\frac{\text{抽出的正确信息点数（C）}}{\text{所有抽出的信息点数（T）}} \tag{5}$$

$$\text{召回率（R）}=\frac{\text{抽出的正确信息点数（C）}}{\text{所有正确的信息点数（A）}} \tag{6}$$

$$F=\frac{(\beta^2+1)PR}{\beta^2 P+R}\text{（通常将 }\beta\text{ 设为 1）} \tag{7}$$

实验所选取的网页数目，所包含的记录项以及所统计的准确率，召回率和 F 值见表 3。

表 3　天气领域的实验结果

网站	网页数	C	T	A	准确率	召回率	F 值
中国天气网	15	1248	1291	1290	96.67%	96.74%	96.67%
2345 天气预报	15	785	806	840	97.39%	93.45%	95.38%
新浪天气	15	825	865	825	95.37%	100%	97.63%
360 天气	15	510	510	540	100%	94.44%	97.14%
天气 321	15	900	900	900	100%	100%	100%

从表 3 中可以看出准确率和召回率较高，F 指数高于 95%。说明本文所提出抽取方法综合性能较高。对于 F 值较低的网站来说，其原因主要是页面内容块变动较频繁，影响了 DIV 块模板的使用。因为待抽取的极少部分信息所在的 DIV 块被过滤掉。对于这种情况，可以进一步优化分类模型，避免 DIV 块模板的欠缺。

4　结束语

本文主要对 Deep Web 查询结果页面抽取进行了研究。

以模板为主线，提出了双重模板的定义与构建。同时，引入了领域本体来指导模板的建立，减少了无关信息，简化了模板的抽取规则。并且在 URL 模板匹配的基础上，结合网页相似度计算，进行更精确的模板匹配，提高了抽取的准确率。实验表明，该抽取方案取得了较好的效果。该方案适用于 DIV＋CSS 结构的 Deep Web 页面的信息抽取，接下来的工作是考虑与页面内容分析相结合的抽取方法，并解决领域本体属性的进一步约简问题。

参考文献：

[1] He B，Patel M，Zhang Z，et al. Accessing the deep web：A survey [J]. Communications of the ACM，2007，50 (5)：95-101.

[2] ZHAO Pengpeng，CUI Zhiming，GAO Ling，et al. Survey of Chinese Deep Wweb [J]. Journal of Chinese Computer Systems，2007，28 (10)：1799-1802 (in Chinese). [赵朋朋，崔志明，高岭，等. 关于中国 Deep Web 的规模、分布和结构 [J]. 小型微型计算机系统，2007，28 (10)：1799-1802.]

[3] LIU Wei，MENG Xiaofeng，MENG Weiyi. A survey of Deep Web data integration [J]. Chinese Journal of Computer，2007，30 (9)：1475-1489 (in Chinese). [刘伟，孟小峰，孟卫一. Deep Web 数据集成研究综述 [J]. 计算机学报，2007，30 (9)：1475-1489.]

[4] KOU Yue，LI Dong，SHEN Derong. D-EEM：A DOM-tree based entity extraction mechanism for Deep Wweb [J]. Journal of Computer Research and Development，2010，47 (5)：858-865 (in Chinese). [寇月，李冬，申德荣. D-EEM：一种基于 DOM 树的 Deep Web 实体抽取机制 [J]. 计算机研究与发展，2010，47 (5)：858-865.]

[5] Liu Linan，Kou Yue，Sun Gaoshang，et al. Duplicate identification model for Deep Web [J]. Journal of Southeast University (English Edition)，2008，24 (3)：315-317.

[6] YANG Xiaoqin，JU Shiguang，CAO Qinghuang，et al. Template generation method for Deep Web automatic data extraction [J]. Application Research of Computers，2010，27 (1)：200-203 (in Chinese). [杨晓琴，鞠时光，曹庆皇，等. 面向 Deep Web 数据自动抽取的模板生成方法 [J]. 计算机应用研究，2010，27 (1)：200-203.]

[7] ZHANG Yanchao，LIU Yun，LI Yong，et al. Study of Web information extraction technology based on automatically generated template [J]. Journal of Beijing Jiaotong University，2009，33 (5)：40-45 (in Chinese). [张彦超，刘云，李勇，等. 基于自动生成模板的 Web 信息抽取技术 [J]. 北京交通大学学报，2009，33 (5)：40-45.]

[8] DONG Min，FANG Shu. On Deep Web information extraction [J]. Library and Information Service，2007，51 (10)：25-28 (in Chinese). [董旻，方曙. Deep Web 信息抽取研究 [J]. 图书情报工作，2007，51 (10)：25-28.]

[9] Liu Wei，Meng Xiaofeng，Meng Weiyi. ViDE：A vision-based approach for Deep Web data extraction [J]. IEEE Transactions on Knowledge and Data Engineering，2010，22 (3)：447-460.

[10] BI Lei，SHEN Jie，XU Fayan，et al. Extracting Web business information using domain-specific ontology [J]. Computer Engineering and Design，2008，29 (24)：6393-6396 (in Chinese). [毕蕾，沈洁，徐法艳，等. 领域本体指导的 Web 商品信息抽取 [J]. 计算机工程与设计，2008，29 (24)：6393-6396.]

[11] ZHANG Wenxiu，ZHU Qinghua. Research on construction methods of domain ontology [J]. Library and Information，2011 (1)：16-19 (in Chinese). [张文秀，朱庆华. 领域本体的构建方法研究 [J]. 图书与情报，2011 (1)：16-19.]