一种有效的深网入口识别方法

吴春明[1,2] 谢德体[2]

(西南大学计算机与信息科学学院 重庆 400715)[1] (西南大学资源环境学院 重庆 400715)[2]

摘 要 深网入口自动识别是深网数据集成的前提和基础。由于表单在设计上具有较大的随意性,使得深网入口缺乏统一的构建标准,难以利用确定性的规则对其进行判断。首先基于统计特征,抽取了部分表单属性作为深网入口与非深网入口的可区分特征,在此基础上,提出了一种利用神经网络进行深网入口自动识别的方法。不同于基于规则的判断方法,神经网络是被训练的,不需要任何先验知识,这种特性使其非常适合于对具有复杂表现形式的深网入口进行判定。实验结果表明了该方法的有效性。

关键词 深网入口,神经网络,特征抽取,机器学习

中图法分类号 TP391 **文献标识码** A

Effective Approach to Deep Web Entries Identification

WU Chun-ming[1,2] XIE De-ti[2]

(College of Computer and Information Science, Southwest University, Chongqing 400715, China)[1]

(College of Resources and Environment, Southwest University, Chongqing 400715, China)[2]

Abstract Automatic identification of deep Web entries is the basis of deep Web data integration. Owing to the subjectivity of form design, deep Web entries lack unified standard and it is difficult to judge whether the form is a deep Web entry by the definite rules. Based on the statistics, this paper first chose several form attributes as the defining features, which can distinguish searchable forms from non-searchable forms. Then, an entry identification algorithm was proposed by using neural network. Unlike previous approaches, neural network can be trained, which is very suitable for entry identification of the deep Web. The experimental results show that our proposed algorithm can be an effective way in automatic identification of the deep Web.

Keywords Deep Web entries, Neural network, Feature extraction, Machine learning

1 引言

深网(Deep Web)是指那些存储在网络数据库中只能通过动态网页技术访问的资源集合,也是互联网上最大、发展最快的新型信息资源[1]。如何自动获取深网数据、建立深网数据集成系统已成为当前研究的热点,也是下一代搜索引擎进行深度搜索必须解决的关键问题[2]。为了自动获取深网中的数据,有几个关键技术必须解决,如入口发现、接口集成、数据库选择、查询结果抽取、结果注释、实体识别和结果合并等[3]。其中,自动准确地发现深网入口是所有后继工作的前提和基础。

深网入口作为Web数据库的唯一入口,一般以HTML表单的形式存在,但网页中通常会包含大量的非深网入口表单,如注册、登录、论坛、购物车、通用搜索引擎等。而各类表单在设计时,往往都带有较大的随意性,通常并没有一个固定的模式,这为在众多表单中自动准确地识别出真正的深网入口带来了极大挑战。针对该问题,已有一些相关研究[5-11],但总体来说,突破性的研究成果相对较少,目前仍然缺乏一种行之有效的实用机制。

本文提出了一种利用神经网络进行深网入口自动识别的方法,实验结果表明该方法具有较高的识别准确率。

2 相关工作

深网入口自动识别技术可分为pre-query和post-query两大类[4]。前者主要利用接口表单及所在页面的可见属性,如表单中控件的类型、数量、表单文本中的某些关键词等相关信息;而后者则采用填写表单并提交查询,再根据返回结果来进行判断的思路,其核心问题在于如何提交一个有效的查询关键字,显然,这种方法的应用前提是对表单语义模式的准确理解,同时更依赖于对查询结果页面中数据实体的有效抽取,文献[5]指出了该方法的困难性,采用这种方案的识别算法也并不多见。

斯坦福大学研究人员设计了一种可以半自动地抽取深网信息的爬虫HiWE[6]。在人工辅助下,HiWE可以向特定领

到稿日期:2011-03-03 返修日期:2011-04-15 本文受中央高校基本科研业务费专项资金(XDJK2010C033),重庆市自然科学基金(CTS2009817)资助。

吴春明(1972—),男,博士生,副教授,主要研究方向为农业信息技术、Web信息获取,E-mail:springsun@swu.edu.cn;谢德体(1957—),男,博士,教授,博士生导师,主要研究方向为农业信息技术、土壤学。

域的深网入口表单提交查询，以此来获得深网数据信息。但这些工作都是在已获得深网入口的基础上完成的，关于如何识别深网入口，文中并未讨论。文献[7]描述了一个从 PIW (Publicly Indexable Web)出发来寻找深网入口的爬虫，该爬虫必须由预分类的文档及相关关键词进行初始化，文中仍然未对深网入口的判定方法给出讨论。文献[8]提出用两条根据实际经验总结出来的规则来对深网入口进行判断，一是表单中必须有 text 控件，二是至少出现一组关键词中的一个，如 search、query 等。这种方法虽然可以获取大量深网入口，但它不具备自动学习能力，也不能针对给定的深网入口来进行自动识别，具有一定的局限性。文献[9]采用了一种表单特征自动提取技术，抽取了表单名、控件名、控件属性值及 form 标签的 action 属性共 4 类参数作为区分特征，利用 C4.5 决策树算法来进行接口识别。实验表明，该方法对于 Web 中随机数据集的分类正确率为 85%，且不能把简单深网入口与搜索引擎的查询接口区分开，离实际使用还有较大差距。文献[10]针对上述方案存在的缺陷进行了改进，共抽取了 14 个表单特征作为深网入口的可区分特征，实验表明改进方案对数据集的分类正确率提升至 90.95 %，但文中仍然没有对简单深网入口与搜索引擎二者间的可区分特征进行详细分析。文献[11]提出了一个三分类器框架，依次对表单结构、表单文本和页面文本进行考查，用于自动发现深网入口并自动对深网入口所属领域进行划分，在入口识别部分，仍然使用的是 C4.5 分类算法。

总体而言，现有的识别方法都是试图通过确定性的规则对深网入口进行识别，从实验结果看，准确率都不是很理想，特别是对简单深网入口和搜索引擎之间的可区分特征，到目前仍没有相关文献予以讨论和解决。

3 基于神经网络的深网入口自动识别模型

本节首先介绍本文算法的设计思路，分析深网入口与非深网入口间的内在差异，进而给出一种利用神经网络进行深网入口自动识别的模型。

3.1 问题分析与解决思路

在以往工作中，更多关注的是接口表单具体的细节特征，希望由此来构建一系列用于入口识别的规则集。然而，深网入口设计者在设计表单时，往往是根据 Web 数据库的属性字段，从查询的便捷性、查询结果的准确性以及表单布局的美观性等角度出发，设计上带有较大的随意性，通常并没有一个固定模式。因而，这种基于预定义规则的判定方法不能很好地适应灵活多变的表单对象，难以取得较高的识别精度。

虽然很难找到确定的规则来区分深网入口和非深网入口，但我们都熟知一些基本规律。例如，在数据库查询条件中，一般不需要密码字段，即入口表单中通常不会包含 password 控件；在 Web 数据库查询接口中，一定不会包含用于上传文件的 file 控件；为了方便用户操作和提高查询精度，深网入口设计者更愿意选择使用 check、radio、select 等选择类控件；如果 form 标签的 action 属性中含有 mailto 关键词，则该表单一定是邮件列表，而不是深网入口；搜索引擎中的 text 控件通常看起来更长，能容纳的字符数也更多，即 size 和 maxlength 属性通常会有更大的取值。

以上规律说明，深网入口和非深网入口之间应该存在着一种普遍性的内在区别，这种区别或者体现在表单控件(Ctrl)的选择使用上，或者体现在控件属性(Attr)及其值域(Val)的差异上，又或者体现在某些关键词(Keyword)上。假设这些特征属性都具有一定的影响因子，那么判断一个表单是否为深网入口就可以通过以下公式来描述：

$$IS\_DWI=\sum_{i=1}^{n} a_i * Ctrl_i + b_i * Attr_i + c_i * Val_i + d * Keyword$$

式中，IS_DWI 代表某个表单是否为深网入口，以概率的形式表现。如某表单计算结果为 0.91，则以 91%的概率认为它是一个深网入口；而另一个表单的计算结果为 0.32，则以 68%的概率认为该表单不是深网入口，阈值通常取 0.5；a、b、c、d 分别为各特征值的影响因子，很明显，这些影响因子与判断结果之间存在着一种非线性关系。现在的问题就是：(1)应该选择表单的哪些属性作为评价的标准。这可以通过对候选特征指标进行统计分析，根据统计结果来进行确定；(2)这些特征属性的影响因子该如何计算。本文提出了利用神经网络进行计算的方法。

3.2 BPNN 识别模型

根据前面分析，对深网入口进行识别的过程即为对表单进行分类的过程，而直接影响分类结果的是各特征属性的影响因子。既然这些因子难以人为事先确定，本文试图选用一种具有自学习能力的机器学习算法，这里选择了神经网络。不同于传统的基于规则的判断方法，神经网络是被训练的、不需要任何先验知识，它通过对给定实例进行主动学习来不断调整输入与输出间的权系数，从而能实现任何复杂的非线性映射，进而可以将其应用于新的实例。这种特性使神经网络非常适用于对具有复杂表现形式的深网入口进行判定。

BP (Back Propagation) 网络是应用最广泛的一种神经网络形式。已经证明，任意函数都可由一个三层网络以任意精度逼近[12]。为此，本文构建了如图 1 所示的三层 BPNN (Back Propagation Neural Network)网络模型。

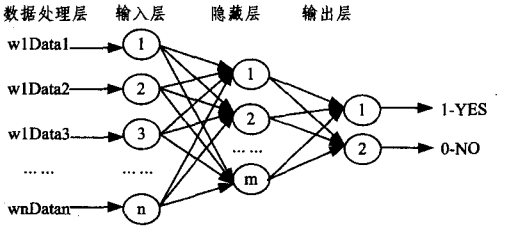

图 1 BPNN 结构

首先，收集一组由深网入口和非深网入口构成的表单数据集 $R(n)$，选择候选表单特征集 $C(m)$，通过对 $R(n)$ 进行统计分析来最终确定用于深网入口判别的特征集 $C(k)$ $(k \leqslant m)$，接下来：

Step1 将所有特征数据进行数值化处理

```
for(i=1;i<=n;i++)
  for(j=1;j<=k;j++)
    {if (data(i,j) is not numerical) then numeric(data(i,
    j))}
```

例如，method 属性的取值可以是 GET 或者 POST，此时应将其数字化为 0 和 1，以便于 BPNN 进行处理。

Step2 为各特征数据赋予不同的权重

```
for(i=1;i≤n;i++)
```

```
for(j=1;j≤k;j++)
  { data (i,j)= w(j) * data (i,j)}
```

基于前面的讨论，某些表单特征对于判断结果具有较高的影响度，如 password 控件、size 属性等，为此，选定的 $C(k)$ 增加了 w 参数，w 代表了一定的权重值，例如，通常用 1 和 0 分别表示表单中是否含有 password 控件。鉴于该控件在入口识别中的重要作用，取 w 为 100，则实际取值变为 100 和 0，以此来增大该特征值的差异度。这样做的目的是为了让网络在学习时对该特征赋予更高的权重。

Step3　输入学习集，训练网络 p 次

```
For (r=1;r≤p;r++){
  for(i=1;i≤n;i++)
    for(j=1;j≤k;j++)
      {Obtain weight(j) of BPNN}
}
```

其中的 weight(j)即为各特征指标的影响因子。

Step4　输入预测表单数据集 $P(q)$，计算各表单的识别概率 IS_DWI

```
for(i=1;i≤q;i++)
  for(j=1;j≤k;j++)
    {IS_DWI(i)+= weight(j) * data(i,j)}
```

Step5　判断表单是否为深网入口

```
for(i=1;i≤q;i++)
  {if (IS_DWI(i)>θ) then Output 1 else Output 0}
```

这里的 θ 代表阈值，取 0.5。当输出 1 时代表该表单是深网入口，输出 0 时则表示该表单不是深网入口。

4　实验与结果

为了验证所提方法的可行性和准确性，本文对 262 个表单数据分别进行了表单特征提取以及 BPNN 训练与预测的实验。实验数据部分来源于 UIUC 的 TEL-8 表单数据集[13]。此外，为了对非深网入口进行统计分析，我们从 Internet 上随机抓取了部分典型的非深网入口表单，涵盖了注册、登录、论坛及搜索引擎等相关领域。人工分类的结果是：深网入口 195 个，非深网入口 67 个。

4.1　表单特征提取

为了确定深网入口与非深网入口的可区分特征，本文首先对表单中各类控件的数量进行了统计，同时参照其他文献的建议，分别统计了 form 标签的 method 和 action 属性、text 控件的 size 和 maxlength 属性，以及表单中是否含有 email、search、mailto 等关键词，并对各特征值进行了对比分析。表 1 展示了各特征指标的平均统计结果及相应的比率。

分析表 1 中的数据，可以得到以下结论：

(1)深网入口往往更多地使用 text、check、selection、radio 等选择类控件，包含有更多的 option 选项，而较少使用 textarea 控件，基本不使用 file、password 控件；

(2)在 hidden、image、button 3 类控件的使用频率上，深网入口与非深网入口基本持平；

(3)非深网入口中 text 控件的 size 和 maxlength 属性值往往较深网入口的大；

(4)深网入口中往往更多地出现如 search、look up、query 等关键词，而非深网入口中往往更多地出现 email 关键词；

(5)深网入口的 action 属性中不会出现 mailto 关键词；

(6)深网入口较非深网入口会更多地使用 GET 方法。

表 1　深网入口与非深网入口特征对比

表单特征	深网入口	非深网入口	比率
password	0	0.418	0:418
text	1.969	1.388	1.42:1
check	5.123	0.806	6.36:1
selection	3.051	0.358	8.52:1
option	134.128	4.746	28.26:1
radio	1.636	0.761	2.51:1
hidden	2.754	2.527	1.09:1
image	0.441	0.343	1.28:1
file	0	0.075	0:746
button	0.282	0.299	1:1.06
textarea	0.005	0.104	1:20.37
HasAction	0.964	1	1:1.04
Method_GET	0.297	0.194	1.53:1
Text _Size	18.596	21.828	1:1.17
Text _Maxlength	37.307	52.419	1:1.41
HasEmail	0.031	0.209	1:6.79
HasSearch	0.877	0.239	3.67:1
HasMailto	0	0.030	0:299

以上结论与人们的常识是一致的(见 3.1 节)。文献[11]也做了类似的工作，得到了与我们类似的结论，而且他们统计出的差异比例值更大、特征值更为明显，这可能与样本数据的选择有关。这说明只要选择出合理的特征值，就可以利用机器学习来自动对表单进行分类。

4.2　BPNN 的训练与预测

基于前面的统计结果，本文选择了差异较大的 14 个表单特征作为区分深网入口与非深网入口的特征指标(见表 2)，并按前述规则对相应的特征数据进行了数值化和适当的缩放处理，最后按照神经网络的要求对这些数据进行了归一化。这些处理后的数据将作为 BPNN 模型的输入。

表 2　选取的表单特征(共 14 个)

种类	表单特征
控件	password、text、check、selection、radio、option、file、textarea
属性	size、maxlength、method
属性值	mailto
关键词	email、search

在样本选择上，首先从数据集中选出 51 个表单作为预测数据，其中深网入口 24 个，非深网入口 27 个，这其中特别选取了部分简单深网入口与通用搜索引擎表单，以检验该方法的通用性。余下的 211 个表单全部作为训练样本。

在 BPNN 模型中，通过对比实验，确定的隐层节点数为 6，$alpha$ 值为 0.05。在对网络进行充分训练后，对预测数据进行判断，结果如表 3 所列，51 个预测数据中被正确识别出 48 个，准确率为 94.12%，神经网络的平均精度为 92.71%。

表 3　BPNN 对预测样本的判断结果

类别	实际数	预测数	正确率	精度
深网入口	24	23	95.83%	93.44%
非深网入口	27	25	92.59%	90.98%

4.3　结果分析

在 BPNN 学习中，当训练结果平方误差为 9.37 时，学习效果如图 2 所示。其中的小菱形代表神经网络对表单样本的实际拟合值，目标值分别为 0 和 1，代表非深网入口和深网入口。

(下转第 230 页)

如果 xRy, yRz，那么 $x\in C(\{y\}), y\in C(\{z\})$，所以 $\{y\}\subseteq C(\{z\})$。由于 $C(\{y\})$ 是包含集合 $\{y\}$ 的最小闭集，因此 $C(\{y\})\subseteq C(\{z\})$，故 $x\in C(\{z\})$，即 xRz。这就证明了 R 是传递关系。

综上所述，当 U 不限制是有限集时，U 上的所有拓扑的集合与 U 上所有满足自反、传递关系的集合之间是一一对应的。

结束语 本文对加拿大学者 Y. Y. Yao 所做的工作进行了改进与扩展。Yao 研究了当论域为有限集时满足自反、传递关系的粗糙集模型的拓扑结构。本文则讨论了当论域不限制是有限集时满足自反、传递关系的广义粗糙集空间中的上、下近似算子的拓扑结构；证明了 U 上满足自反、传递关系的集合与 U 上所有的拓扑的集合是同势的，并指出了该拓扑空间的基为 $\{R_s(x); x\in U\}$。对于满足自反、对称关系的粗糙集模型以及模糊粗糙集模型的拓扑结构，我们将另文讨论。

参 考 文 献

[1] Pawlak Z. Rough Sets[J]. International Joural of Computer and Information Science, 1982, 11: 341-356

[2] Pawlak Z. Rough sets: Theoretical Aspects of Reasoning About Data[M]. Boston: Kluwer Academic Publishers, 1991

[3] 陈德刚，张文修. 粗糙集和拓扑空间[J]. 西安交通大学学报，2001(12)

[4] Yao Y Y. Two views of the theory of rough sets in finite universes[J]. International Journal of Approximate Reasoning, 1996, 15(4): 291-317

[5] Qin K, Pei Z. On the topological properties of fuzzy rough sets [J]. Fuzzy Sets and Systems, 2005, 151(3): 601-613

[6] Morsi N N, Yakout M M. Axiomatics for fuzzy rough sets[J]. Fuzzy Sets and Systems, 1998, 100(1-3): 327-342

[7] Lashin E F, Kozae A M, Abo K A, et al. Rough set theory for topological spaces [J]. International Journal of Approximate Reasoning, 2005, 40(1/2): 35-43

[8] Kuncheva L I. Fuzzy rough sets: Application to feature selection [J]. Fuzzy Sets and Systems, 1992, 51(2): 147-153

[9] 秦克云，乔全喜. 粗糙集的拓扑结构[J]. 计算机科学，2007(12): 161-162

[10] 秦克云，裴峥，杜卫锋. 粗糙近似算子的拓扑性质[J]. 系统工程学报，2006(1): 81-85

[11] 张文修，等. 粗糙集理论与方法[M]. 北京：科学出版社，2001

(上接第 201 页)

从图 2 可以看出，大多数表单都达到了较为理想的拟合效果（$IS\_DWI<0.2$ 或 $IS\_DWI>0.8$），少部分表单以相对粗糙的概率对表单做出了正确识别（$0.2<IS\_DWI<0.5$ 或 $0.5<IS\_DWI<0.8$）。这说明：(1)本文选取的区分指标是合理的；(2)本文提出的机器学习方法可以根据这些特征指标对表单进行准确分类，特别是在简单深网入口与搜索引擎的区分上，达到了令人满意的效果。

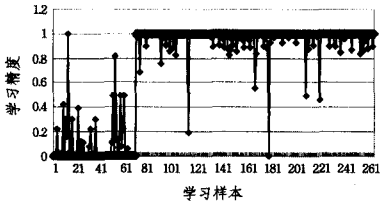

图 2 学习结果图

但从图中也可看出，有极个别表单被错误分类，通过对这些表单进行分析，发现它们都采用了较为特殊的设计，如在 amazon 网站中，允许用户在一个 textarea 中输入关于图书的搜索信息，而这与论坛表单的特征非常类似，致使系统做出了错误判断。如何对这部分表单进行识别，进一步完善该算法的动态适应性，还需要做进一步研究。

结束语 不同于以往的基于规则的判定方法，本文从抽取深网入口与非深网入口的可区分特征入手，采用机器学习的思路，提出了一种利用神经网络进行深网入口自动识别的方法，实验证明了该方法的有效性。在下一步研究中，我们拟结合考虑表单及所在页面的语义信息，以进一步提高该识别方法的准确率。

参 考 文 献

[1] Ghanem T M, Aref W G. Databases Deepen the Web [J]. IEEE Computer, 2004, 73(1): 116-117

[2] Bergman MK. The deep Web: Surfacing hidden value. Technical Report, BrightPlanet[EB/OL]. 2001. http:// www. brightplanet. com/pdf/deepwebwhitepaper. pdf

[3] 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] He B, Tao T, Chang K C C. Organizing structured Web sources by query schemas: A clustering approach [C]// Proc. of the 13th Conf. on Information and Knowlege Management. Washington: ACM Press, 2004: 22-31

[5] Wu P, Wen J R, Liu H, et al. Query selection techniques for efficient crawling of structured Web sources [C] // Proc. of the 22nd Int'l Conf. on Data Engineering. Atlanta: IEEE Computer Society, 2006: 47-56

[6] Raghavans, Garcia-Molina H. Crawling the hidden Web [C]// Proc. of the 27th Int'l Conf. on VLDB. Italy: Rome, 2001: 129-138

[7] Bergholz A, Chidlovskii B. Crawling for domain-specific hidden Web resources[C]// Proc. of the Int'l Conf. on Web Information Systems Engineering. Roma: IEEE Computer Society, 2003: 125-133

[8] Lage J P, da Silva A S, Golgher P B, et al. Automatic generation of agents for collecting hidden Web pages for data extraction [J]. Data & Knowledge Engineering, 2004, 49(2): 177-196

[9] Cope J, Craswell N, Hawking D. Automated discovery of search interfaces on the Web [C]// Proc. of 14th Conf. on Database technologies. Australian Computer Society, 2003: 181-189

[10] Barbosa L, Freire J. Combining classifiers to identify online databases[C]// Proc. of the 16th Int'l Conf. on WWW. New York: ACM Press, 2007: 431-440

[11] Wang Hui, Liu Xian-wei, Zuo Wan-li. Using classifiers to find domain specific online databases automatically[J]. Journal of Software, 2008, 19(2): 246-256

[12] Anthony M. Probabilistic Analysis of Learning in Artificial Neural Networks: The PAC model and its variants[J]. Neural Computing Surveys, 1997, 1: 1-47

[13] UIUC Web integration repository[EB/OL]. 2010. http://metaquerier. cs. uiuc. edu/ repository/

[14] 黄勤，龚海清，刘金亨，等. 基于改进的遗传神经网络入侵检测系统[J]. 重庆理工大学学报：自然 科学版，2010, 24(2): 83-86