Deep Web数据源发现与分类模型

马 丹,王翰虎,陈 梅,张小平
(贵州大学 计算机科学与信息学院,贵州 贵阳 550025)

摘 要:随着Internet的发展,Web正在不断深入人们的生活。传统搜索引擎只能检索浅层网络(Surface Web),不能直接索引到深层网络(Deep Web)的资源。为了有效利用Deep Web资源,对Deep Web数据源发现并进行领域类别的划分,已成为一个非常迫切的问题。该模型首先抽取Deep Web页面查询接口的特征,构造了一个Deep Web页面过滤器,从而能够发现Deep Web的数据源,其次在对查询接口特征进行分析后,构建了一个基于KNN的分类器,并通过该分类器对新产生的Deep Web数据源进行领域分类。试验结果表明,这种模型的平均分类准确率达到86.9%,具有良好的分类效果。

关键词:深层网络;查询接口;K近邻算法;分类
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2010)07-0065-03

Discovery and Classification Model for Deep Web Sources

MA Dan, WANG Han-hu, CHEN Mei, ZHANG Xiao-ping
(College of Computer Science & Information, Guizhou University, Guiyang 550025, China)

Abstract: With the development of Internet, Web is continuously used in our lives. Traditional search engines are only able to reach surface Web except for Deep Web sources. To make use of Deep Web source efficiently, it's urgent that Deep Web sources are found out and classified. This work was focus on Deep Web classification, and a novel classification model was proposed. Its processing including two steps: at first, the model employed features of query interfaces of Deep Web, to recognize whether the Web page was Deep Web, and then, the specific subject of the Deep Web were be identified in the second step by utilize KNN algorithm. The experiments show that the average correct classification rate is 86.9%, and the detailed results are listed in the end of this paper.

Key words: Deep Web; query interfaces; KNN; classification

0 引 言

Internet上的一些网页是搜索引擎不能索引的,这部分信息用户不可见,称为Deep Web[1]。截至2004年,Deep Web的网站数量已经达到307 000个,其背后的数据库数量已经达到366 000~535 000个。由此看出,Deep Web的信息量远远超过了Surface Web的信息量。为了用户能很好地使用Deep Web的信息,对Deep Web数据进行研究是很有必要的[2,3]。

Internet上Deep Web数据源的个数一直处于动态变化中,每天都有新的生数据源成和死亡,当新的Deep Web数据源加入时,需要做到自动化的分类。对于分类问题,目前的工作主要集中在文本或Web文档的分类研究,对Deep Web查询接口分类的研究工作较少。而Deep Web后台数据库的入口就是查询接口,Internet上的查询接口大多数是Form表单形式,因此,判定网页表单是否是查询接口,并对Deep Web信息分类是至关重要的[4,5]。

1 模型构架

模型首先收集各种领域的Deep Web网页,抽取Deep Web查询接口特征信息。然后,通过过滤器对非Deep Web页面进行过滤。最后,分类器根据KNN算法对输入的待分类表单进行分类,输出Deep Web网页的类别。模型构架如图1所示。

1.1 Deep Web网页表单特征提取及表示

通过对网页表单中的文字信息和表单内部控件信息进行分析,可以发现文字信息是用来对相应的控件进行描述的,而网页表单的内部控件有INPUT控件、SELECT控件和TEXTAREA控件。其中,INPUT控件表示可以输入表单信息,INPUT控件中的submit类

收稿日期:2009-11-11;**修回日期**:2010-02-11
基金项目:贵州省自然科学基金项目(黔科合GY字[2008]3035)
作者简介:马 丹(1977-),女(回族),贵州镇宁人,硕士,讲师,研究领域为Web数据挖掘、多媒体数据库;王翰虎,教授,研究领域为数据挖掘、多媒体数据库、分布式数据库;陈 梅,副教授,研究领域为Web数据挖掘、数据库技术;张小平,研究员,研究领域为数据挖掘、数据库技术,Web信息检索。

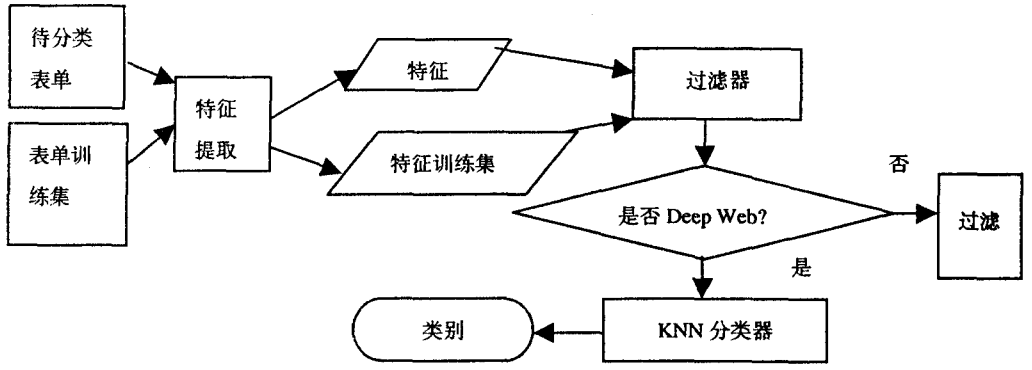


图 1 模型构架

型控件用来提交网页表单内容。每种控件都有一个指定控件名称的 name 属性。SELECT 控件表示下拉列表框或复选框。TEXTAREA 控件表示可以输入多行的文本框。于是网页表单信息可以形式化表示为：

WebFormfeature = {fName, URL,
<name1, type1, dvalue1(s)>;
<name2, type2, dvalue2(s)>;
...
<namei, typei, dvaluei(s)>;
...}

其中，fName 是网页表单标签中的 name 属性值，URL 表示站点信息，namei 为第 i 个控件的 name 属性值，typei 是第 i 个控件的类型，dvaluei(s)是第 i 个控件的缺省值。

1.2 Deep Web 查询接口表单识别

通过对样本进行分析发现，用户注册信息、查询接口，以及搜索引擎都是通过 Form 表单来表示，但是并不是所有的表单都是查询接口，所以网页中的表单是否是查询接口，是至关重要的。我们将表单分为可填写类、可查询类和查询接口类表单。可填写类表单指一些虽然可填写，但不具有查询功能的表单，例如注册表单。可查询表单指有可查询的功能，但是不会和后台的数据库进行交互。而查询接口类表单才是真正 Deep Web 查询接口。

基于以上原因，设计了一个用于过滤非查询接口表单的过滤算法：

步骤 1：页面有 Form 标签吗？如果有，则转步骤 2；否则输出“不是 Deep Web 页面”并转步骤 7。

步骤 2：Form 标签中出现了 INPUT 控件，SELECT 控件和 TEXTAREA 控件中的一种或几种吗？如果是，则转步骤 3；否则输出“不是 Deep Web 页面”并转步骤 7。

步骤 3：若表单中没有具有提交功能的按键(submit)，输出“不是 Deep Web 页面”并转步骤 7；否则转步骤 4。

步骤 4：若表单中含有 password 输入框，就有可能是用户注册类的，则输出“不是 Deep Web 页面”并转步骤 7；否则转步骤 5。

步骤 5：若表单中有 email 值出现，则这种接口很可能是注册表单，输出“不是 Deep Web 页面”并转步骤 7；否则转步骤 6。

步骤 6：查看表单中控件数，控件总数小于三个，输出“不是 Deep Web 页面”并转步骤 7；否则输出“此网页是 Deep Web 页面”。

步骤 7：退出。

经过以上过滤算法[6]得到的才是 Deep Web 页面。

1.3 对过滤后的 Deep Web 页面进行分类

首先收集各种领域的 Deep Web 网页，抽取它们的查询接口特征，接下来的工作就是怎样计算两个查询接口特征的相似性，而查询接口中控件不止一个，于是怎样找到它们之间对应的控件是解决这个问题的关键。下面给出了两控件的相似性计算方法，以及两查询接口表单的相似性计算方法。

(1)两控件的相似性计算方法。

假设有两个查询接口表单 A 和 B，则 A 的控件 controli 和 B 的控件 controlj 的相似性的计算方法为：

$$sim(controli, controlj) = a \times sim(name1, name2) + b \times sim(type1, type2) + c \times sim(dvalue1, dvalue2) \quad (1)$$

公式(1)中的 a, b, c 是权值。根据 name, type, dvalue 对分类的贡献度，分别给它们赋予的权值为：a = 0.5, b = 0.3, c = 0.2。

① sim(name1, name2)是比较两个控件 name 属性名的相似性，当 name1 和 name2 是同义词关系、多义词关系、上下位关系、包含关系，则 sim(name1, name2) = 1，否则 sim(name1, name2) = 0。

② sim(type1, type2)是比较两个控件类型的相似性。当 type1 = type2，则 sim(type1, type2) = 1，否则 sim(type1, type2) = 0。

③ sim(dvalue1, dvalue2)表示两个控件默认值的

相似性。若缺省值为两个数字,其相似度[7]为:

$$sim(dvalue1,dvalue2) = 1 - \frac{|dvalue1 - dvalue2|}{MAX(dvalue1,dvalue2)} \quad (2)$$

如果缺省值为离散值,则两者的重复数据的多少反映了其相似程度,相似性计算为:

$$sim(dvalue1,dvalue2) = \frac{(dvalue1 \cap dvalue2)}{(dvalue1 \cup dvalue2)} \quad (3)$$

如果缺省值为范围型集合,则两者的重叠程度反映了相似程度,假设 dvalue1 = {s1, s2…}, dvalue2 = {t1, t2…},则相似性计算为:

$$sim(dvalue1,dvalue2) = \frac{((s1 \cup s2 \cup \cdots) \cap (t1 \cup t2 \cup \cdots))}{((s1 \cup s2 \cup \cdots) \cup (t1 \cup t2 \cup \cdots))} \quad (4)$$

(2)查询接口表单 A 和查询接口表单 B 之间相似性计算方法。

在上面,已经提供了两个控件的相似性计算方法,接下来就是统计两个查询表单之间的相似性。假设 A 和 B 中的控件采用链表组织[8],表头分别为 HeadA, HeadB,且 A 表长度较小。寻找匹配属性算法如下:

```
/* MAX存储相似度最大值;r,s分别指向A,B表中当前比较节点 */
{...
P:=HeadA;
while p->next<>NULL do
{ SUM=0;//用SUM统计查询接口表单A和B相似度
r=p->next;MAX=0;q=HeadB;
while q->next<> NULL do
{
s=q->next;
if MAX<Sim(r,s) MAX=Sim(r,s);
q=s->next:
}//该循环用来遍历表单B中所有控件,找到与A表单中r最相似的控件
SUM=SUM+MAX; //累加当前两表单中的控件相似度
P=r->next;} ...}
```

此算法结束后统计出的 SUM 即是查询接口表单 A 和查询接口表单 B 的相似度。

(3)采用 KNN 算法对未知类别的 Deep Web 网页分类。

有了求两个查询接口表单相似度的算法后,只需采用 KNN 算法[9,10],求与未知类别的 Deep Web 网页相似度[11]最大的前 N 个样本网页,哪个类别所占比例最大,则该 DeepWeb 网页就属于哪一类。

采用 KNN 算法对含有 Deep Web 查询接口的网页进行分类,其算法描述如下:

输入:待分类的 Deep Web 网页,M 个类别的样板网页集合 S。

输出:类别 C1,C2,…,CM.

步骤 1:初始化,F1,F2,…,FM 为人工选出的样本网页;

步骤 2:输入新网页 P,抽取该网页的查询接口特征;

步骤 3:计算 sim(P,P');其中 P'∈S

步骤 4:选取前 K 个具有最大 sim 值的网页;

步骤 5:统计这 K 个网页集合中,哪个类别的网页数最多,则新网页 P 就属于哪一类。

2　实验结果

在实验中,共收集了 804 个 Deep Web 网页,把它们分为 4 类:book, airfares, movies, music。各类网页所占比例基本一致。Deep Web 查询接口分类的评价指标可以采用分类准确率 P(precision)。TP 表示被正确分类的网页数,FP 表示被错误分类的网页数 ,那么 P = TP/(TP + FP)。在 KNN 算法中,当 K 取 100 时,实验结果如表 1 所示。

表 1　分类结果

领域	测试网页数	被正确分类网页数	分类准确率 P(%)
Airfares	207	201	97.3
Movies	179	135	75.6
Musics	186	153	82.1
Books	232	215	92.6

从实验结果来看,该模型的平均分类准确率达到 86.9%,具有良好的分类效果。对于 Airfare 和 Books 领域来说,由于和其它领域的 Deep Web 查询接口的控件相似性较低,所以分类过程中有很高的准确率;Movies 和 Musics 的 Deep Web 查询接口的控件相似性较高,所以比较起来准确率相对低一点。

3　结束语

在文中的研究工作中,提出了一种基于查询接口的 Deep Web 数据源发现及分类模型,该模型通过提取 Deep Web 查询接口特征,从而进行 Deep Web 数据源的过滤,有效消除了一些非 Deep Web 网页表单对分类模型的影响;进一步使用 KNN 算法对 Deep Web 数据源分类。在下一步工作中,将研究 Deep Web 中多数据源数据集成问题[12],在现有的工作基础上实现多 Deep Web 数据源集成。

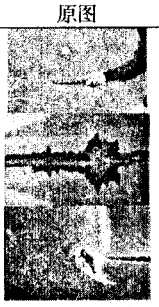
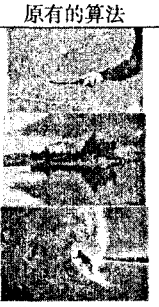
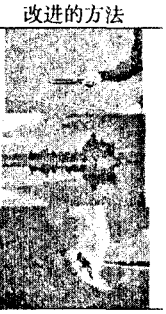
参考文献:

[1] Raghavan S, Garcia - Molina H. Crawling the Hidden Web [C]//Proceedings of the 27th International Conference on Very Large Data Bases. Roma: [s. n.], 2001: 129 - 138.

(下转第 71 页)

其进行编号,"宽 * 高"列展示了图片的大小属性;"相邻区域(P * Q)"列是求解 Histon 直方图时的相邻区域大小,经过大量实验表明,相邻区域 5 * 5 为比较好的选择;"门限值"列显示了隶属矩阵的门限值,原有的和改进的粗糙集直方图求解过程中门限值的差别是比较大的,经过大量的试探,确定了如表 1 所示的门限值,这些门限值是比较合理的,对于门限值的动态自适应调整有待进一步研究。根据以上数据分别对 3 张图片进行求解,得到其原有的粗糙集直方图和改进的粗糙集直方图,如表 2 所示,在粗糙集直方图的基础上,利用单峰子集法对图像进行分割,得到实验结果,如表 3 所示,从结果不难看出,文中所述方法较之原方法更符合视觉感知,贴近原图。

表 3　原有的和改进的方法分割结果对比

图片	原图	原有的算法	改进的方法
1			
2			
3			

6　结束语

文中在深入研究基于粗糙集的彩色图像分割的基础上,提出了针对粗糙集直方图的改进,进而改善了彩色图像分割效果。引入 YUV 空间来计算像素点间的颜色差异,进而得到更精确的 R、G、B 粗糙集直方图,以求符合知觉差异。经过大量实验验证,文中所述方法较之原方法更符合视觉感知,贴近原图。然而,文中未对求粗糙集直方图过程中门限值的自适应调整、直方图的去噪性等问题做深入研究,这些有待于进一步的探索。

参考文献:

[1] 边肇祺,张学工. 模式识别[M]. 第 2 版. 北京:清华大学出版社,2000.

[2] 杜啸晓,杨　新,施鹏飞. 一种新的基于区域和边界的图象分割方法[J]. 计算机图像图形学报,2001(8):755 - 759.

[3] 谭　优,王泽勇. 图像阈值分割算法实用技术研究与比较[J]. 微计算机信息,2007(6):298 - 299.

[4] 彭兴邦,蒋建国 . 一种基于亮度均衡的图像阈值分割技术[J]. 计算机技术与发展,2006,16(11):10 - 12.

[5] Mohabey A, Ray A K. Fusion of Rough Set Theoretic Approximations and FCM for Color Image Segmentation[C]// Systems, Man and Cybernetics, 2000 IEEE International Conference. [s.l.]:[s.n.],2000:1529 - 1534.

[6] Mushrif M M, Ray A K. Color image segmentation: Rough - set theoretic approach[J]. Pattern Recognition Letters, 2008,29:483 - 493.

[7] 张　恒,冯子亮. 一种基于粗糙集合理论的彩色图像分割[J]. 计算机技术与发展,2009,19(2):39 - 42.

[8] 欧阳鑫玉,赵楠楠. 图像分割技术的发展[J]. 鞍山钢铁学院学报,2002,25(5):363 - 368.

[9] 林开颜, 吴军辉, 徐立鸿. 彩色图像分割方法综述[J]. 计算机图像图形学报,2005,10(1):1 - 10.

[10] 纪　滨. 粗糙集理论及进展的研究[J]. 计算机技术与发展,2007,17(3):69 - 72.

[11] 王国胤. Rough 集理论与知识获取[M]. 西安:西安交通大学出版社,1999.

[12] Pawlak Z. Rough sets: Theoretical Aspects of Reasoning about Data[M]. Boston : Kluwer Academic Publisher, 1991.

(上接第 67 页)

[2] Chang K C C, He B, Li C, et al. Structured databases on the web: Observations and implications[J]. SIGMOD Record, 2004,33(3):61 - 70.

[3] Bergman M K. The deep web: surfacing hidden value[J]. In journal of electronic publishing,2002,7(1):8912 - 8914.

[4] 柴春梅,李　翔,林　祥. 基于改进 KNN 算法实现网络媒体信息智能分类[J]. 计算机技术与发展,2009,19(1):1 - 4.

[5] 蒋　云, 赵佳宝. 自动化测试脚本自动生成技术的研究[J]. 计算机技术与发展,2007,17(7):4 - 7.

[6] 李齐会. Deep Web 查询接口的判定技术研究[J]. 计算机与数字工程,2007,37(3):131 - 134.

[7] 姜芳艽,孟小峰,贾琳琳. Deep Web 集成服务的不确定模式匹配[J]. 计算机学报,2008,31(3):1412 - 1421.

[8] 洪　辉,李石君,余　伟,等. 基于语义的中文 Deep Web 查询接口集成[J]. 计算机科学,2008,35(13):61 - 63.

[9] 牛冀平,胡志华,余志超. 可扩展 XML 文本数据自动分析研究与实现[J]. 计算机技术与发展,2006,16(3):8 - 10.

[10] 胡　哲,郑　诚,王艳玲. 语义检索关键技术研究[J]. 计算机技术与发展,2008,18(10):75 - 78.

[11] 梁卓明,陈炬桦. 基于专有名词优先的快速中文分词[J]. 计算机技术与发展,2008,18(3):24 - 27.

[12] 张明宝,马　静. 一种基于知网的中文词义消歧算法[J]. 计算机技术与发展,2009,19(2):9 - 11.