ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.2, February 2008, pp.267−274
DOI: 10.3724/SP.J.1001.2008.00267

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

基于网页上下文的Deep Web数据库分类*

马 军+, 宋 玲, 韩晓晖, 闫 波

(山东大学 计算机科学与技术学院,山东 济南 250101)

Classification of Deep Web Databases Based on the Context of Web Pages

MA Jun+, SONG Ling, HAN Xiao-Hui, YAN Po

(School of Computer Science and Technology, Shandong University, Ji'nan 250101, China)

+ Corresponding author: Phn: +86-531-88391528, Fax: +86-531-88392498, E-mail: majun@sdu.edu.cn, http://ir.sdu.edu.cn

Ma J, Song L, Han XH, Yan P. Classification of deep Web databases based on the context of Web pages. *Journal of Software*, 2008,19(2):267−274. http://www.jos.org.cn/1000-9825/19/267.htm

Abstract: New techniques are discussed for enhancing the classification precision of deep Web databases, which include utilizing the content texts of the HTML pages containing the database entry forms as the context and a unification processing for the database attribute labels. An algorithm to find out the content texts in HTML pages is developed based on multiple statistic characteristics of the text blocks in HTML pages. The unification processing for database attributes is to let the attribute labels that are closed semantically be replaced with delegates. The domain and language knowledge found in learning samples is represented in hierarchical fuzzy sets and an algorithm for the unification processing is proposed based on the presentation. Based on the pre-computing a *k*-NN (*k* nearest neighbors) algorithm is given for deep Web database classification, where the semantic distance between two databases is calculated based on both the distance between the content texts of the HTML pages and the distance between database forms embedded in the pages. Various classification experiments are carried out to compare the classification results done by the algorithm with pre-computing and the one without the pre-computing in terms of classification precision, recall and F1 values.

Key words: deep Web; hidden Web; database classification; content text extraction; semantic classification

摘 要: 讨论了提高 Deep Web 数据库分类准确性的若干新技术,其中包括利用 HTML 网页的内容文本作为理解数据库内容的上下文和把数据库表的属性标记词归一的过程.其中对网页中的内容文本的发现算法是基于对网页文本块的多种统计特征.而对数据库属性标记词的归一过程是把同义标记词用代表词进行替代的过程.给出了采用分层模糊集合对给定学习实例所发现的领域和语言知识进行表示和基于这些知识对标记词归一化算法.基于上述预处理,给出了计算 Deep Web 数据库的 *K*-NN(*k* nearest neighbors)分类算法,其中对数据库之间语义距离计算综合了数据库表之间和含有数据库表的网页的内容文本之间的语义距离.分类实验给出算法对未预处理的网页和经过预处理后的网页在数据库分类精度、查全率和综合 F1 等测度上的分类结果比较.

* Supported by the Specialized Research Fund for the Doctoral Program of Higher Education of China under Grant No.20070422107 (高等学校博士学科点专项科研基金); the Key Science-Technology Project of Shandong Province of China under Grant No.2007GG10001002 (山东省科技攻关项目)

Received 2007-08-31; Accepted 2007-11-19

关键词: deep Web;隐式 Web;数据库分类;内容文本抽取;语义分类
中图法分类号: TP393 文献标识码: A

根据BrightPlanet公司在 2001 年的统计结果,Deep Web数据库所存储数据的数量是目前静态网页所表示的数据量的 500 倍以上[1],因此,Deep Web数据库资源的有效利用将成为未来Web信息检索和数据库领域的重要研究之一.目前,研究主要集中在对Deep Web数据库的检索和对Deep Web数据库的组织[2−4]上.对前者的研究主要集中在元搜索引擎设计[5,6]和面向Deep Web的网路爬虫的设计与实现上[7];而对后者的研究则主要是对搜索到的Deep Web数据库资源信息进行分类或聚类[3,7−10].

目前,对Deep Web数据库资源信息进行分类的常用的方法有pre-query和post-query[8,10].Post-query的分类过程是通过对数据库提出查询,利用返回结果对数据库进行分类.虽然目前存在可以对数据库进行自动填充并产生输出的软件[8],但因返回的结果毕竟只是数据库的部分内容,而当数据库的记录具有较多属性时,这种方法则难以取得较好的分类效果,因为不同属性的组合方式是属性数的阶乘函数.

Pre-query 基本依赖于描述Deep Web数据库表(forms)的可视特征[8,10],即数据库表中的属性标记(attribute labels)和表中的其他可利用信息.显然,这种方法仅适合数据库的内容可以完全由表的特征表示出来的情形.另外,这种方法的有效性也依赖于软件系统对表特征的提取能力[8].因为涉及到对自然语言理解等困难问题,这种方法的分类精度往往不容易提高.

文献[8]提出用包含表的Web网页的文本内容和表的标记词一起参与对网页的聚类,从而加强对数据库聚类的准确性,并把该方法命名为上下文感知的表聚类法(context-aware form clustering).其核心思想是把包含数据库表的网页中的文本信息作为数据库内容的上下文.然而,文献[8]中对网页和表的处理过于简单,只是把表的标记和HTML的全部词变成文本,然后利用信息检索技术中的词频分析法(term frequency inverse document frequency,简称TFIDF)[11],把表的标记和HTML的全部词的文本用向量表示,然后采用*K*-means算法进行聚类.这样存在的问题是,如果网页中含有噪音信息,如公告、导航、修饰、版权等,则会严重地影响聚类的效果.另外,该方法对数据库的属性标记词缺乏语义分析.因为即使两数据库中的内容相同或相似,但由于数据库系统是面向不同的用户和独立实现的,表的属性标记词也可能采用完全不同的标记词.比如,“Job search”和“employer finder”,这些来自不同数据库系统的标记词语义相同,但计算机可能认为是完全不同的词汇.由于表中的标记词数量比较少,当把语义相同的词作为不同的词来处理时,会严重地影响网页分类的准确性.

本文提出了仅使用包含数据库表格网页的内容文本作为数据库内容描述的上下文的对 Deep Web 数据库进行分类的算法.其主要贡献如下:

1. 给出基于对 HTML 网页中文本块的多种统计特征的对内容文本模块的发现算法.该方法特别适合难以单纯利用网页视觉特征进行判别的网页,如 Deep Web 网页;
2. 基于分类是一种有指导的学习的特点,本文提出从训练样本中找出领域和语言知识,并采用分层模糊集合的对应的有向图来表示这些知识.在此基础上给出把语义相似的数据库属性标记词用代表元替换的算法以及对归一化后的词频进行计算的公式.
3. 基于网页内容文本之间的距离和数据库表之间的距离对 Deep Web 数据库之间的距离进行综合计算,并把这种距离应用在经典的 *k*-NN(*k* nearest neighbors)分类算法框架下,得到对 Deep Web 数据库分类的新算法.

我们在UIUC(University of Illinoise at Urbanan Champain)的实验 数据平台[12]和利用我们垂直网络爬虫所搜集到的数据集合上,进行了大量的实验.比较对经过预处理的网页和未经过预处理网页的Deep Web数据库的分类结果,实验结果表明,经过上述预处理的实验数据的分类结果在分类精度、查全率和综合F1 测度上均好于未经过上述预处理数据的分类结果.

1 基本术语

在基于Web的信息检索的研究中,最常用的对文档的表示方法是采用向量空间模型,即每个文档d_j用一维向量$(w_{1,j},w_{2,j},\ldots,w_{t,j})$表示,其中$t$为标引词的数目,$w_{i,j}$是标引词$k_i$在文档$d_j$的权重.目前比较常用的计算$w_{i,j}$的方法称为词频分析法(TFIDF).

定义 1[11]. 设N表示检索系统中文档的总数,n_i表示包含标引词k_i的文献数目,$freq_{i,j}$表示含标引词k_i在文档d_j中的初始频率,则文献d_j中标引词k_i的标准化频率为$tf_{i,j}$,定义为

$$tf_{i,j}=\frac{freq_{i,j}}{\max_l freq_{l,j}} \tag{1}$$

其中,$\max_l freq_{l,j}$表示文献d_j中的最大词频.另一个指标称为标引词k_i的倒文档频度(inverse document frequency),通常由下式计算:

$$idf_i=\log\left(\frac{N}{n_i}\right) \tag{2}$$

则$w_{i,j}$由下面的公式进行计算:

$$w_{i,j}=tf_{i,j}\times idf_i \tag{3}$$

若$\boldsymbol{d}_i,\boldsymbol{d}_j$为两个向量,则两向量之间的距离可以用下面的公式计算:

$$\cos(\boldsymbol{d}_i,\boldsymbol{d}_j)=\frac{\boldsymbol{d}_i\cdot\boldsymbol{d}_j}{\|\boldsymbol{d}_i\|\times\|\boldsymbol{d}_j\|} \tag{4}$$

定义 2. 设U为一个由对象组成的论域,则在U上的一个模糊集合A定义为一个隶属函数$\mu_A(x):U\to[0,1]$,$x\in U$.它把 U 中的元素映射到[0,1]中的一个实数,记为 $A=\int_{x\in U}\mu_A(x)/x$.当 A 可枚举时,直接记为 $A=\{\mu_A(x_1)/x_1,\mu_A(x_2)/x_2,\ldots,\mu_A(x_n)/x_n\}$.

我们扩展文献[13]给出的分层模糊集合的定义如下:

定义 3. 称 A 为具有分层的模糊集合,若 A 是论域 U 上的一个有限模糊集合,并且 A 中的元素之间存在“kind-of”或“语义近似”等关系.这种语义关系可以用权函数表示出来.

定义 4. 由分层的模糊集合 A 中的关系所定义的有向标记赋权图 $G(V,E,W,L)$称为分层模糊集合的图表示.其中,V 是图的顶点集合,$\forall v\in V,L(v)$表示 v 的标记;$e=(u,v)$为图的边,表示顶点 u,v 之间存在语义关系,而 $W(e)(0\le W(e)\le 1)$,则表示这种关系的强度.

图 1 给出一个模糊分层集合的表示.在图 1 中,边 $e=(w,v)$表示“kind-of”关系.e 的权重 $w(e)$表示语义近似程度.如 w(camera phone, digital camera)=0.7;w(camera phone, digital products)=1.

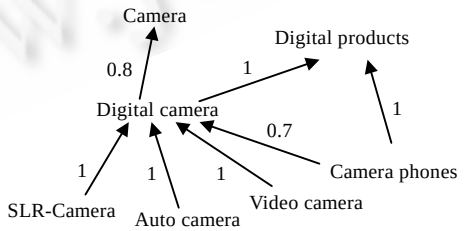

Fig.1 Presentation of digital products by hierachical fuzzy set

图1 利用分层模糊集合表示的数码产品

因为模糊集合 A 中的关系是偏序关系,所以图 G 中无有向回路,因此 G 一定存在顶点 v,v 无出边.称 v 为代表元.G 中的代表元可能多于 1 个.若 v 为代表元,w 为任意顶点,并且在图中有从 w 到 v 的有向路径 P,则称 v 为 w 的祖先.定义路径 P 的权重如下:

$$w(P)=\min_{e\in P}\{w(e)\} \tag{5}$$

并称 w 可归一到 v,其含义是可以用 v 替代 w.在本文中,若 w 定义为代表元,则不归一;否则,归一到距 w 最近的祖先中的代表元.

分层模糊集合与采用分类法(taxonomy)和本体(ontology)有许多不同.领域本体中的最小单位是名词表示的概念.同义词和近义词来自语言学.但分层模糊集合的最小单位是短语,表示查询和数据库属性的检索词.标记词之间的相似性来自语言学、领域知识、常识知识和功能等价等更广泛的关系.比如,手机和相机本是不同的两种物体,但因为手机有拍照功能了,所以手机也被认为是一种相机,这是功能等价性关系.分层模糊集合对应的图不是树,而是有向赋权图,表示同一问题的不同用词的形式.

2 网页中内容文本的提取

对网页中内容文本的提取已有许多研究,如基于模板去除噪音来提取内容文本[14]和基于视觉的分块方法(vision based page segmentation,简称VIPS)[15],以及利用网页的可视布局信息对页面进行分块[16].但由于包含Deep Web表的网页属于不同的应用系统,设计风格差异很大.另外,从位置等可视区域,目前许多方法会认定表是网页的内容文本.但我们所需要的是除了表以外的文本中的关于表描述的文本,而内容文本所在的位置难以确定,一般并不在网页的中心位置.而且,一般网页的文字比较少,哪个文本块属于内容文本是相对的,因此需要新的内容文本的发现算法.

基于我们对已有文献的研究和实际网页的分析,采用统计和决策树的方法得到下面的内容文本的特征,表 1 列出内容文本的六大特征.进一步地,通过统计的方法得到文本块对应各个特征的属性值,对连续分布的属性值进行离散化,然后利用贝叶斯公式来计算各个属性值或属性值区域对文本块为主题内容块的概率.

Table 1 Statistical features of content texts

表 1 内容文本的统计特征

No.	Feature	Meaning explanation
1	Special pairs of HTML tags	The content text blocks often contain some pairs of tags, e.g. 〈P〉 〈BR〉. Whether a text block has the pairs of tags can be used to find content texts.
2	Font and size of letters in text blocks	Some letters in content text blocks are often bigger than the letters in non-content text blocks and have special fonts, e.g. the letters of the titles of content texts.
3	Punctuations	Content text blocks often have some special punctuations, e.g. comma and full stop.
4	The area and location of a text block	In most cases the intersection area of a content text block with one of a content text block in a given template is bigger.
5	The pixel number in a text block	Usally the content text block in a HTML page has bigger size in terms of number of pixels among all text blocks.
6	The letter number in a text block	Usally the content text block in a HTML page has bigger size in terms of number of letters among all text blocks.

由表 1 可以看出,文本块中出现的满足表 1 中的特征越多,它是内容文本的可能性就越大,因此,为了将这个可能性准确地表示出来,我们基于统计来求得特征中不同值的概率.

下面以特殊标签特征为例给出如何计算基于这些统计特征确定文本为内容文本的概率.用 T,F 分别表示内容文本和噪音块;R=1,0 分别表示特殊标签出现与否.量化过程可叙述如下:我们取上百个网页作为样本集合 S.采用 VIPS 算法对每个网页进行分块,并对属于内容的文本块进行人工标记.$P(K)$表示文本块 K 为内容文本的先验概率,$P(R)$是特殊符号对 R 出现的概率,则当 K 含有特殊符号对 R 时为内容文本的概率为

$$P(K=T \mid R=1)=\frac{P(K=T)P(R=1 \mid K=T)}{P(R=1)} \tag{6}$$

设集合 C 为 S 中所有内容块的集合,则

$$P(K=T)=\frac{|C|}{|S|};\quad P(R=1 \mid K=1)=\frac{C\text{ 中包含特殊符号对的文本块数}}{|C|} \tag{7}$$

$$P(R=1)=\frac{S\text{ 中包含特殊符号对的文本块数}}{|S|} \tag{8}$$

设P_1,P_2,P_3和P_5为 5 个特征的概率,假定相互独立,P为综合概率值,则

$$P=\min\left\{\sum_i^5 P_i,1\right\} \tag{9}$$

若 $P>\alpha$,α为指定的阈值,则认为 K 为文本块.

对其他特征的概率值的计算可采用类似的方法.表 2 给出针对不同特征的内容文本判断的概率计算方法.

Table 2 Probability computation based on features

表 2 基于特征的概率计算方法

No.	Features	Probability computation on whether a text block is a content text block
1	Special pairs tags in text blocks	See the discussion above.
2	Font and size of letters in text blocks	Similar with (1), where the probability is calculated based on the average letter size equal to 8, 10, 12, 14, 16 or in (0,8) or (16,+∞) respectively.
3	Punctuations	The probability is calculated based on the numbers of special punctuations in a text block.
4	The area and location of a text block	Calculating the probability based on the size of the intersection area of a text block A with assumed content blocks given by page-templates, e.g. the layout given in Ref.[14]. The support probability is calculated based on Bayesian conditional probability formula and the sample pages.
5	The pixel number in a text block	The probability is calculated based on the pixel number of a text block fall into the interval of (0,10), (10,20), (20,30), (30,+∞) respectively, where the count unit is ten thousand.
6	The letter number in a text block	The probability is calculated based on the letter number of a text block fall into the interval of (0,20), (20,40), [40,+∞] respectively.

把基于上述概率判断内容文本的算法命名为 MFM(multiple feature model),并另外实现以下算法进行比较:

1. IM(information method)[6]:把具有最大信息熵的文本块作为内容文本.
2. PM(position method)[17]:根据文本块的坐标位置.
3. DTM(DOM tree method)[14]:基于文本在HTML的DOM的位置.

实验数据为随机从CWT200G[18]数据集合上抽取的 10 万多个网页.实验采用下面的指标:

噪音去除率 NRR=(正确去除的噪音文本长度)/(总的噪音文本长度) (10)

内容提取率 CER=(正确提取的主题内容文本长度)/(总的主题内容文本长度) (11)

表 3 给出利用 4 个算法在噪音去除率和内容提取率上的实验比较结果.可以看出,PM 虽然噪音去除率较高,但内容提取率却相对较低,所以综合考虑,我们的方法是最优的,有着较高的内容提取率和噪音去除率.

Table 3 Evaluation on found content text

表 3 对找出内容文本的评测

Algorithm	NRR (%)	CER (%)
IM	86.3	93.74
PM	93.5	91.25
DTM	84.1	94.62
MFM	92.8	97.87

3 内容文本和表的表示及相似度计算

设 P 为包含数据库表的 HTML 网页,我们可以比较容易地把 P 分成两部分:表部分,记为 FC(form contents);剩余的部分,记为 PC.把 FC 中关于表格式自身的描述标记去掉,把表中属性的标记词取出.下面我们叙述如何计算归一化后新的词频计算方法.

因为我们研究的是分类,可以假设待分的类别有m个,用C_i 表示第i类,$1\le i\le m$.精选足够数量的属于该类的HTML网页作为学习样本.并选出这些网页中数据库表格中的标记词建立起分层模糊集合.根据模糊集合建立对应的有向图并计算出图的代表元.对图中任意顶点w计算与之最近的代表元v.

对每个网页 p,w 为 p 的 PC 的词或 FC 中的标记,并且 w 在模糊图中可归一到代表元 v.设集合 A 为 p 的 FC(PC)中所以可归一到中心词 v 的标记词(关键词)集合,若 $tf(v)$表示其词频,$v\in A$,则按下面的公式增加代表元 v 的词频数.

$$tf(v)=\sum_{w\in A} w(P(w,v)) \tag{12}$$

其中,P为对应模糊集合的图中连接顶点w到v的路径,并定义$P(v,v)$=1.把FC和PC中的w的出现用v替代,并且把

$tf(w)$用$tf(v)$替代.而idf和权重的计算与第 2 节的内容相同.在归一化后,我们可以在新的PC和FC的基础上计算权重$w_{i,j}$,对每个文档建立起向量空间表示FCV(form content vector)和PCV(page content vector).

网页P_1,P_2表示的数据库间的相似度由下面的公式计算,其中,$\boldsymbol{PCV}_1$,$\boldsymbol{FCV}_1$和$\boldsymbol{PCV}_2$,$\boldsymbol{FCV}_2$分别对应于P_1和P_2的内容文本向量和数据库表向量.k_1和k_2为两个权重调节参数.

$$sim(P_1,P_2)=\frac{k_1\cos(\boldsymbol{PCV}_1,\boldsymbol{PCV}_2)+k_2\cos(\boldsymbol{FCV}_1,\boldsymbol{FCV}_2)}{k_1+k_2} \tag{13}$$

4 分类算法

我们采用 k-NN 算法对含有 Deep Web 数据库表的网页进行分类,其形式化描述如下:

算法. CDC(context-aware database classification).

输入:待分类的网页,参数k_1,k_2.根据类别数和样板个数确定合适的正整数k.

输出:类别C_1,C_2,…,C_M.

步骤 0:初始化,C_1,C_2,…,C_M为人工选出的样本网页,$S=C_1\cup C_2\cup\ldots\cup C_M$.

步骤 1:输入新网页 P,按第 3 节和第 4 节的方法分析该网页,建立对应的 PCV 和 FCV 向量.

步骤 2:$\forall P'\in S$,根据公式(13)计算 $sim(P,P')$;

步骤 3:选取前 k 个具有最大 sim 值的网页组成集合 A.

for i=1 to M **do**

$$score(C_i\mid p)=\sum_{p'\in A}sim(p,p')I(p',C_i)$$

其中函数$I(p',C_i)$=1 若$p'\in C_i$;否则,$I(p',C_i)$=0;

步骤 4:把p放入第C_j类,若$score(C_j|p)\geq score(C_i|p)$,$1\leq i,j\leq M$;

步骤 5:若还有未确定的网页,则转向步骤 1;否则输出C_1,C_2,…,C_M,结束.

我们采用以下常用的标准来评价我们的分类方法.设$n_{i,j}$表示属于C_i的网页分配到类C_j的网页数目,n_j表示类C_j的网页数目,则定义分类精度:

准确率
$$precision(i,j)=\frac{n_{i,j}}{n_j} \tag{14}$$

查全率
$$recall(i,j)=\frac{n_{i,j}}{n_i} \tag{15}$$

准确率越高,表明分类器在该类上出错的概率越小;而查全率越高则表明分类器在该类上可能漏掉的网页越少.F1 标准则综合了精度和查全率,将两者赋予同样的重要性来考虑.F1 的计算由下面的公式决定:

$$F(i,j)=\frac{2\times recall(i,j)\times precision(i,j)}{recall(i,j)+precision(i,j)} \tag{16}$$

5 实验分析

我们的实验数据集合主要来自 UIUC 中以及我们自己的垂直检索系统所爬取的近千个网页.为了表示得简单,给出类别编号,见表 4.

Table 4 Numbering the classes

表 4 类别编号

Class No.	Airfares C_1	Automobiles C_2	Books C_3	CarRentals C_4	Hotels C_5	Jobs C_6	Movies C_7	MusicRecords C_8

首先分别检查单独使用未归一化的网页分类情况.表 5 和表 6 分别给出分类的测试值,其中R和P分别表示查全率和查准率.显然,因为F1 值是综合评价指标,有 6 种利用FCV的分类比使用PCV的分类结果要好,有 2 种利用FCV的分类比使用PCV的分类结果要差.因此,应该给予FCV之间的相似度更大的权值.在原理上,参数k_1,k_2可以是任何正数.下面我们限定k_1,k_2取值为{0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9}.通过对k_1,k_2的 64 种配对方式测试分类

后的F1 的取值变化,得到当k_1=0.4,k_2=0.6 时,对多数的类别而言,都可以得到较好的分类效果这一结论.因此,我们取k_1=0.4,k_2=0.6 来测试算法CDC对经过预处理和未经过预处理的网页的分类比较.

Table 5 Classification based on FCV only

表5 单独使用FCV的分类结果

Class	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
R (%)	93.33	86.49	76.92	33.33	62.50	72.22	66.67	77.78
P (%)	73.68	88.89	90.91	100.00	83.33	38.24	74.07	84.00
F1	0.82	0.88	0.83	0.50	0.71	0.50	0.70	0.81

Table 6 Classification based on PCV only

表6 单独使用PCV的分类结果

Class	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
R (%)	53.33	74.36	61.54	22.22	68.75	90.00	46.43	71.43
P (%)	42.11	78.38	38.10	50.00	84.62	62.07	86.67	90.91
F1	0.47	0.76	0.47	0.31	0.76	0.73	0.6	0.8

实验中我们选择 300 个网页做样板集合,建立分层模糊集合及对应的有向图,对所有的待分类网页中的数据库表的属性标记进行归一操作.对归一化后的网页建立 PCV 和 FCV,重新运行算法 CDC.表 7 给出了算法的分类测度值,其中 $R(R')$和 $P(P')$分别表示针对数据库标记词未进行归一化(进行了归一化)的查全率和查准率.

Table 7 Classification results by algorithm CDC when it uses database attribute label unification and does not

表7 算法CDC采用数据库标记词归一化处理和不采用归一化处理时的分类结果

Class	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
R (%)	86.67	90.00	76.92	77.13	43.75	85.12	87.10	71.43
R′ (%)	100.00	94.60	100.00	100.00	62.50	83.33	76.67	92.59
P (%)	72.22	73.47	88.95	83.00	77.78	73.91	79.41	86.96
P′ (%)	93.75	81.40	76.47	90.00	100.00	93.75	95.83	100.00
F1	0.78	0.80	0.82	0.79	0.56	0.79	0.83	0.78
F1′	0.97	0.87	0.87	0.95	0.77	0.88	0.85	0.96

显然,对 FC 归一化后,分类算法的 F1 的取值均优于未归一化的分类.归一化的分类中有 3 类的 F1 值超过 0.9,F1 值超过 0.85 的有 7 类.有 3 类的查全率达到 100%,有 2 类的查准率达到 100%.显然,仅仅对 Deep Web 数据库表的标记词建立模糊图进行归一化,就可以大幅度提高对 Deep Web 数据库资源的分类精度.

6 结 语

本文给出了对 Deep Web 数据库资源更精细的分类方法.对数据库之间的语义距离计算是基于 HTML 网页的内容文本和嵌入的数据库表之间的向量空间距离的线性组合得到的.这里所谓的内容文本是指那些可能对数据库内容有语义描述的文本块.为了得到这些文本块,本文给出利用文本块的统计特征的内容文本发现算法及相应的概率计算方法,并将得到的内容文本作为对数据库内容描述的上下文,参与对数据库的分类.这种内容文本的识别方法的特点是对文本块所位于的 HTML 网页中的位置具有较少的依赖性,适合一般 DeepWeb 数据库表都处于网页的中心位置以及这类网页中的描述文字比较少的特点.

我们观察到由于同一领域数据库的标记词使用不同的同义或近义词的原因,可能造成对数据库分类精度的下降,因此提出利用模糊分层集合以及对应的图,把同义和近义的标记词进行归一,并依据新的词频计算方法产生对表的向量表示.实验表明,归一化后的算法执行结果在分类精度、查全率和综合 F1 测度上都比较好.这种方法也有可能可以进一步扩展到对跨语言的分层模糊集合,使得算法对使用不同语言的 Deep Web 数据库资源也可以进行分类.下一步的工作是如何能够自动地发现并建立分层模糊集合与对应的有向图.给出对 Deep Web 检索的 PageRank 计算求解,并把我们的面向主题的垂直网络爬虫与本文的网页分类算法结合起来,形成完整的对 Deep Web 资源的检索系统.

致谢 我们感谢审稿老师提出的修改建议.

References:

[1] Brightpanet's investigation. 2001. http://www.brightplanet.com/news/prs/deep-Web-500-times-larger.html

[2] Chang KCC, He B, Zhang Z. Toward large-scale, integration: building a MetaQuerier over databases on the Web. In: Weikum G, ed. Proc. of the Conf. on Innovative Data Systems Research. Asilomar: IEEE Computer Society, 2005. 44−55.

[3] He H, Meng W, Yu CT, Wu Z. Automatic integration of Web search interfaces with WISE-integrator. VLDB Journal, 2004,13(3): 256−273.

[4] He H, Meng W, Yu C, Wu Z. Wise-Integrator: An automatic integrator of Web search interfaces for e-commerce. In: Lockemann P, ed. Proc. of the Int'l Conf. on very Large Data Bases. Berlin: IEEE Computer Society, 2003. 357−368.

[5] Gravano L, Garcia-Molina H, Tomasic A. Gloss: Textsource discovery over the Internet. ACM Trans. on Database Systems, 1999, 24(2):229−246..

[6] Yi L, Liu B. Web page cleaning for Web mining through feature weighting. In: Cohn AG, ed. Proc. of the 18th Int'l Joint Conf. on Artificial Intelligence (IJCAI 2003). Acapulco: Kluwier Academic Publisher, 2003. 64−75.

[7] Bergholz A, Chidlovskii B. Crawling for domain-specific hidden Web resources. In: Spaccapietra S, ed. Proc. of the 4th Int'l Conf. on Web Information Systems Engineering. Rome: IEEE Computer Society, 2003. 125−133.

[8] Barbosa L, Freire J, Silva A. Organizing hidden-Web databases by clustering visible Web documents. In: Doqac A, ed. Proc. of IEEE the 23rd Int'l Conf. on Data Engineering. Istanbul: IEEE Computer Society, 2007. 326−335.

[9] Gravano L, Ipeirotis PG, Sahami M. QProber: A system for automatic classification of hidden-Web databases. ACM TOIS, 2003, 21(1):1−41.

[10] He B, Tao T, Chang KCC. Organizing structured Web sources by query schemas: A clustering approach. In: Gravano L, ed. Proc. of ACM the 13th Conf. on Information and Knowlege Management. Washington: ACM Press, 2004. 22−31.

[11] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval. Boston: Addison Wesley, 1999. 27−30.

[12] The UIUC Web integration repository. 2007. http://metaquerier.cs.uiuc.edu/repository/datasets/tel-8/index.html

[13] Thomopolos S, Buche P, Haemmerle O. Fuzzy sets defined on a hierarchical domain. IEEE Trans. on Knowledge and Data Engineering, 2006,16(10):1395−1409.

[14] Wang J, Loehovsky F. Data-Rich section extraction from HTML pages. In: Cham TS, ed. Proc. of the 3rd Int'l Conf. on Web Information Systems Engineering. Singapore: IEEE Computer Society Press, 2002. 1−10.

[15] Cai D, Yu SP, Wen JR, Ma WY. VIPS: A vision-based page segmentation algorithm. Technical Report, MSR-TR-2003-79, Redmond: Microsoft Research Corporation, 2003. 1−79.

[16] Song RH, Liu HF, Wen JR, Ma WY. Learning important models for Web page blocks based on layout and content analysis. SIGKDD Explorations, 2004,6(2):14−23.

[17] Feng HM, Liu B, Liu YM. Framework of Web page analysis and content extraction with coordinate trees. Journal of Tsinghua University, 2005,45(S1):1767−1771 (in Chinese with English abstract).

[18] CWT200G. 2007. http://www.cwirf.org/SharedRes/DataSet/cwt.html

附中文参考文献:

[17] 封化民,刘飚,刘艳敏.含有位置坐标树的 Web 页面分析和内容提取框架.清华大学学报,2005,45(S1):1767−1771.

马军(1956－),男,山东汶上人,博士,教授,博士生导师,CCF 高级会员,主要研究领域为信息检索,并行计算,算法分析与设计.

韩晓晖(1983－),男,博士生,主要研究领域为信息检索.

宋玲(1969－),女,博士生,副教授,主要研究领域为信息检索.

闫泼(1985－),女,硕士生,主要研究领域为信息检索.