基于主题和表单属性的深层网络数据源分类方法

祝官文,王念滨,王红滨
(哈尔滨工程大学计算机科学与技术学院 黑龙江哈尔滨 150001)

摘 要: 当前深层网络中蕴含着高质量的海量信息并且其数量不断地增长,由于深层网络具有分布、异构、自治等特点,用户高效、快捷地获取自己感兴趣的信息面临巨大挑战.将深层网络数据源按领域分类是解决这一挑战的基础.本文以对航空订票、图书、汽车和房地产领域的 200 多个数据源的统计和分析为基础,充分利用主题和表单属性信息,提出了一种新的深层网络数据源分类方法以及改进的查询接口相似性度量方法,实现深层网络数据源的自动分类.本文还提出了一种查询接口标记策略,以降低随机选择初始中心点所产生的影响.实验结果表明该方法具有较高的分类精度.

关键词: 表单主题和属性; 查询接口标记; 深层网络; 数据源自动分类
中图分类号: TP311.13 **文献标识码:** A **文章编号:** 0372-2112 (2013)02-0260-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2013.02.009

An Improved Method for Deep Web Sources Classification Based on the Theme and Form Attributes

ZHU Guan-wen, WANG Nian-bin, WANG Hong-bin
(*Department of Computer Science and Technology, Harbin Engineering University. Harbin, Heilongjiang* 150001, *China*)

Abstract: Nowadays, Deep web consists of vast amounts of high quality information which is rising rapidly. However, because of its distributed character, heterogeneity, autonomy etc, it is faced with huge challenges for users to obtain the information efficiently and quickly which they are interested in. Deep Web data sources are organized by the domains in the real world, which is the foundation for addressing this challenge. In this paper, based on the statistics and analysis on more than 200 data sources which are from four different fields(i.e., Airfares, Books, Automobiles and Real estates, a novel classification method and an improved similarity measure of query interfaces were proposed to realize the automatic classification of large masses of deep web sources, which make full use of theme information and form attributes. In addition, we present a strategy of tagging query interface to reduce the influence resulted from choosing initial centers randomly. The experimental results indicated that the method is effective and has higher accuracy.

Key words: form theme and attributes; query interface tagging; deep web; automatic classification of sources

1 引言

根据 UIUC[1]统计,深层网络蕴含 45 万个 Web 数据库、125.8 万个查询接口;最近一次统计表明[2]:深层网络数据库增加了 250 万个,而且数量每年还在不断增长.如何有效利用这些海量信息,已成为当前研究的热点.研究 Deep Web 查询接口集成[3]的目的就是建立统一查询接口,从而使用户可以方便、快捷、自动地查询感兴趣的信息.因此,作为 DeepWeb 数据集成的重要组成部分——Deep Web 数据源分类,目前已成为当前 Deep Web 领域[4]的一个研究热点.

目前,面对海量 Deep Web 数据源,一些学者或研究机构主要采用手工方法构建 Deep Web 目录,然而其覆盖率非常低(最高仅为 15.6%)[1].此外,深层网络中数据源每天都处于不断变化中,如果采用手工的方法进行分类,不仅耗时费力、效率很低,而且难于满足用户对分

收稿日期:2012-05-18;修回日期:2012-12-05
基金项目:国家自然科学基金(No.60973028,No.61272185);黑龙江省自然科学基金(No.F201238)

类性能日益增长的要求.因此,对深层网络数据源进行自动化分类的研究具有重要意义.

本文通过对航空订票、汽车、图书销售及房地产领域中 200 多个数据源(如表 1 所示,其中 N1 表示查询接口数,N2 表示包含领域主题特征词的查询接口数,P 表示 N2/N1)统计和分析发现:(1)在查询接口源代码中,绝大多数 title 标记含有内容,而且这部分内容中的有些词往往只出现在某个领域且在一定程度上反映了该查询接口的主题(如图 1 上半部分实线框所示),即所属的相关领域;(2)同一领域查询接口间相似属性的个数往往较多,不同领域接口间相似属性的个数则较少.受此启发,本文提出了一种基于主题和表单属性的数据源分类方法(TAF-SSCC),该方法结合了半监督 K-Means 方法与分类方法,并且利用了表单属性可视化特征及页面主题特征;提出了一种基于领域主题特征词的查询接口标记方法,该方法解决了初始点选择好坏的问题;提出了一种自动构建领域主题特征词词典的方法,同时利用属性同义词词典很好地解决了属性标签间的异名同义问题;本文还提出了一种改进的接口相似性度量方法,用于 Deep Web 数据源的分类.

表 1　包含领域主题特征词的查询接口统计信息

domain	N1	N2	P%	domain theme terms
Airfares	56	54	96.46	airline, flight , ticket, air , airfare, travel
Books	50	38	76.00	book, bookshop , textbook, bookstore , online
Automobiles	53	48	90.57	car, sale, auto, motor, dealer, price , motor, online, vehicle
Real Estate	51	51	100	real estate, home , property, sale , listing, house, property,

```
<!DOCTYPE html>
    <html>
    <head>
  <title>Find Low Airfares δ Desls on All
  Airplane Tickets - Alaska Airlines</tile>
    <meta name="keywords" content="alaska airlines, airplane tickets,
  cheap airfare,hawaii vacations,sheap airfares,discount airfare,alaskan
  airlines,alaskaair,alaskaair com,alaskaairline,alaskaairline,alaska
  airlines goldstreak,alaska airlines arrivals,alaska airlines boarding pass,
  alaska airlines cargo,alaska airlines carry on,alaska horizonairlines,
  alaska airlines mileage partners"/>
    <meta name="description"content="Alaska Arilines offers low
airfare on all airplane tickets,including discount airfare on hawaii vacations
and mexico vacation packages,Book your Alaska Air travel today,"/>
```

图1　查询接口主题信息

2　相关研究工作

深层网络数据源的分类已获得越来越多研究者关注.迄今为止,研究人员提出了许多关于深层网络数据源的分类[5-9]和聚类[10-12]方法.本文对这些方法进行深入研究发现:首先,目前绝大多数的方法都只是针对查询接口的表单属性特征,如:文献[9]对查询接口的属性特征类型进行进一步划分,并采用特征选择过滤器和高斯法分类器对结构化的 Web 数据源进行分类,该方法虽然对属性特征进行了进一步划分,但它过分依赖于查询接口的属性特征,对于那些没有属性标签的简单查询接口(只包含一个文本输出,如图 2 所示),该方法存在局限性;其次,也有部分研究是针对查询接口所在页面内容,如:文献[11]利用 Web 网页表单和表单所在页面内容提出了基于上下文感知的表单聚类方法,但该方法把 HTML 中的所有词组成文本,这导致网页文本内容不可避免地存在大量噪音信息,比如导航、修饰、公告、版权等信息,进而导致聚类效果不佳;第三,在目前存在的方法中,大部分采用分类方法,但这些分类方法有以下缺陷:(1)数据源的覆盖性单一(如:结构化或非结构化);(2)依赖于训练集;(3)需用户事先标记部分数据源.此外,也有部分采用聚类的方法,如文献[11]采用 K-Means 聚类方法.由于聚类方法在初始点选择上带有很大的随机性,因此聚类的效果依赖于初始点的选择.此外,研究表明,尽管在文本领域中关于半监督聚类[13]方法的研究比较多,但是在深层网络数据源分类领域几乎还没有采用半监督的聚类方法.

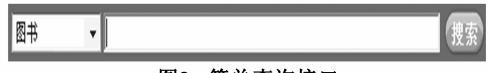


图2　简单查询接口

3　Deep Web 数据源分类框架

Deep Web 数据源分类主要目的是为查询接口集成提供按领域组织的数据源,从而方便用户查询不同领域的信息.图 3 展示了 Deep Web 数据源分类框架.Deep Web 数据源分类框架由四个模块(预处理、标记策略、半监督 K-Means 聚类和后分类)组成.

预处理:针对给定的查询接口,按照是否同时含有

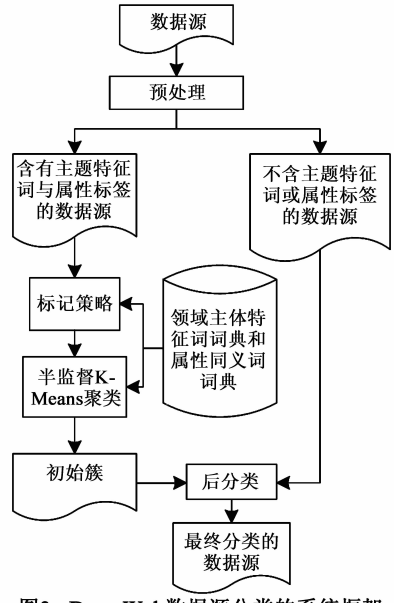


图3　Deep Web数据源分类的系统框架

领域主题特征词和属性标签的标准进行初步分类,其中含领域主题特征词和属性标签的查询接口集为一类,它们用于标记策略模块和半监督 K-Means 聚类模块;其余的为一类,它们用于后分类模块.

标记策略:该模块主要功能是利用领域主题特征词词典(见 4.2)标记部分查询接口,目的是解决半监督 K-Means 聚类的初始中心点选择问题.

半监督 K-Means 聚类:该模块主要是利用标记策略模块产生的结果和人工构建的属性同义词词典(见 4.2)信息对含有领域主题特征词和属性标签的查询接口进行半监督 K-Means 聚类,从而得到初始划分的簇.

后分类:该模块功能是把不含领域主题特征词或属性标签的查询接口划分到初始簇中最相似的簇,从而得到最终的数据源分类结果.

4 Deep Web 数据源分类策略及相关算法

本节首先给出了相关定义,然后对各个主要模块进行了详细描述并且设计了相关算法.

4.1 相关定义

定义 1 (领域主题特征词 DTT):指 title 标记(如图 1 所示)中能够描述某个领域主题的词或词语.

例如,在图 1 上半部分实线框中,"Airline"、"Airfare"、"Ticket"和"Airplane"这些词在航空订票领域中出现的概率明显高于其它领域中出现的概率,并且在其它领域出现的非常少,因此它们可作为该领域的领域主题特征词.

定义 2 (基于领域主题特征词的接口模式 ISDTT):ISDTT 是一个二元组 (T, F). T 为 title 标记所包含的领域主题特征词集合,即 $T=\{t_1, t_2, \ldots, t_n\}$,其中 t_i 表示 title 标记中出现的领域主题特征词;F 为查询接口表单属性集,即 $F=\{A_1, A_2, \ldots, A_m\}$,其中 A_j 表示查询接口表单中的属性标签.

定义 3 (领域主题特征词集 DTTS):$T_{D_k}=\{(t_1, \varphi_1), (t_2, \varphi_2), \ldots (t_n, \varphi_n)\}$,其中,$\varphi_1 > \varphi_2 > \ldots > \varphi_n$,$\varphi_j$ 为领域主题特征词 t_j 隶属于该领域 D_k 的隶属度分数.

定义 4 (领域主题特征词词典 DTTD):$D=\{T_{D_1}, T_{D_2}, \ldots T_{D_n}\}$,其中,$T_{D_k}$ 表示领域 D_k 的领域主题特征词词集.

4.2 词典构建

构建领域主题特征词词典与属性同义词词典的目的在于:形式化描述查询接口模式、标记策略模块进行查询接口标记、辅助度量接口相似性及后分类模块再次分类.图 4 展示了 Deep Web 领域主题特征词词典与属性同义词词典的构建框架.其中语料库为多个领域的查询接口集,比如航空订票、图书销售、汽车等领域.

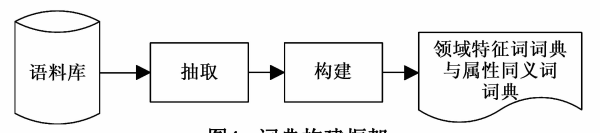


图4 词典构建框架

表 2 部分领域主题特征词词典

domain	< domain theme term, φ >
Airfare	< airline, 1.0000 >, < flight, 0.9175 > < ticket, 0.8126 >, < airfare, 0.6404 >, < travel, 0.6109 >
Books	< book, 1.0000 >, < online, 0.6792 >, < bookshop, 0.6356 >, < textbook, 0.6288 >, < bookstore, 0.4528 >
Automobiles	< car, 1.0000 >, < sale, 0.7105 >, < auto, 0.5882 >, < dealer, 0.5069 >, < price, , 0.4280 >, < motor, 0.3748 >, < online, 0.3733 >, < vehicle, 0.3671 >
Real Estate	< real estate, 1.0000 >, < home, 0.8692 >, < sale, 0.8204 >, < listing, 0.7083 >, < property, 0.6780 >, < house, 0.5089 >

4.2.1 领域主题特征词词典(DTTD)构建

在文本领域中,针对术语的自动化抽取的研究已经非常成熟,大体上分为三类:基于规则的方法、基于统计的方法和基于规则与统计相结合的方法.由于基于规则的方法在构造规则库时费时费力并且覆盖面窄,而基于统计的方法则依赖于语料库的规模,因此本文采用规则与统计相结合的方法自动化抽取领域主题特征词并构建 DTTD.

本文借鉴隶属度思想[14]自动构建 DTTD,其具体步骤如下:①先进行预处理(如:去停用词、标点符号及地名等)并利用语言规则(如:领域主题特征词一般为名词或名称短语——针对英文语料库)获取候选领域主题特征词;②统计词频及查询接口频率;③进行领域隶属度分析;④按隶属度分数降序排序;⑤选择隶属度大于阈值 α 的词,从而得到领域主题特征词集并按照领域来构建 DTTD,其中表 2 为自动构建的 DTTD(φ 表示隶属于该领域的隶属度分数).

4.2.2 属性同义词词典(ASD)构建

研究表明[11]:对于单个领域,尽管查询接口数不断激增,但其属性个数却保持在一个相对稳定的水平.因此,本文进行人工建立属性同义词词典具有一定的可行性.此外,构建的 ASD 可用于下一小节中的查询接口相似性计算.

本文构建 ASD 的步骤主要如下:①利用 Hoa Nguyen[15]提出的 LABELEX 法自动化抽取深层网络查询接口的属性标签;②结合 word-net[16]进行人工构建属性同义词词典.

4.3 查询接口相似性计算

该部分主要介绍了基于领域主题特征词和表单属性的查询接口相似性计算方法并给出了权重公式.

4.3.1 相似性度量

由定义 2 可知,ISDTT 包含领域主题特征词和表单属性两部分,因此本文提出了一种改进的查询接口相似性度量方法 SC-DTTA,公式如下:

$$sim(ISDTT_1, ISDTT_2) = \lambda_1 sim(T_1, T_2) + \lambda_2 sim(F_1, F_2) \tag{1}$$

其中 $\lambda_1 + \lambda_2 = 1$,针对航空订票、图书、汽车和房地产领域,实验表明:λ_1 和 λ_2 值分别取 0.2 和 0.8 效果最佳,$sim(T_1, T_2)$表示查询接口 $ISDTT_1$ 和查询接口 $ISDTT_2$ 的领域主题特征词间的相似性,$sim(F_1, F_2)$表示查询接口 $ISDTT_1$ 和查询接口 $ISDTT_2$ 表单属性间的相似性.

$$sim(T_1, T_2) = \frac{N(T_1 \cap T_2)}{N(T_1 \cup T_2)} + C \tag{2}$$

其中 $N(T_1 \cap T_2)$表示查询接口 $ISDTT_1$ 和查询接口 $ISDTT_2$ 的领域特征词交集的元素个数,$N(T_1 \cup T_2)$表示查询接口 $ISDTT_1$ 和查询接口 $ISDTT_2$ 的领域特征词并集的个数.而 C 为权重调整因子.

$$sim(F_1, F_2) = \left(\sum_{i=1}^{m} w_i + \sum_{j=1}^{n} w_j\right) * \frac{1}{m+n} \tag{3}$$

其中 m, n 分别为查询接口表单属性集 F_1 和 F_2 中的属性个数,w_i 和 w_j 为属性权重因子.直观上就是,两接口中属性相似的个数越多,则查询接口越相似.

4.3.2 权重

$$C = \begin{cases} \dfrac{N(T_1 \cap T_{D_i})}{N(T_{D_i})} * \dfrac{N(T_2 \cap T_{D_i})}{N(T_{D_i})}, \\ \qquad 若\ T_1 \subseteq T_{D_i} \text{and}\ T_2 \subseteq T_{D_i} \text{and}\ T_1 \cap T_2 = \varphi \\ 0, \qquad 其它 \end{cases} \tag{4}$$

其中,如果查询接口 $ISDTT_1$ 的领域主题特征词 T_1 和查询接口 $ISDTT_2$ 的领域主题特征词 T_2 均属于同一领域的领域主题特征词词集并且 T_1 和 T_2 之间没有公共领域特征词,则 C 的值为查询接口 $ISDTT_1$ 中 T_1 与领域 D_i 中 T_{D_i}(即领域 D_i 的领域主题特征词集)的相似度乘以 T_2 与领域 D_i 中 T_{D_i}的相似度.这里,C 可解决如下问题:虽然两查询接口的领域主题特征词词集没有交集,但从领域主题特征词词典角度看,如果两查询接口的领域主题特征词属于同一领域主题特征词词集,直观上那么它们应该具有一定的相似度(尽管相似度非常低).

$$w_i = \begin{cases} 1, \text{if}\ \exists\ A_{2k} \in F_2\ \text{s.t}\ A_{2k} = A_{1i} \text{or}\ A_{1i} \in C_r \wedge A_{2k} \in C_r \\ 0, \qquad 其它 \end{cases} \tag{5}$$

其中,接口 1 中第 i 个属性 A_{1i}的权重 w_i 为 1,当且仅当存在一个属性 A_{2k},它属于接口 2 中的第 k 个属性,使得 A_{1i}和 A_{2k}名称相同或它们都是 ASD 中的一个概念 C_r(同义词);否则为 0.同理,w_j.

4.4 查询接口标记策略

由于预处理模块对输入的查询接口按照是否含有领域主题特征词和属性标签的标准进行了初步分类,其中一类为含领域主题特征词和属性标签的查询接口,记为 $Interfaces_{TF}$.因此本模块则对这部分查询接口进行标记,并把已标记的查询接口作为聚类的初始中心点.即尽可能标记出 K 个不同领域的查询接口,这样半监督 K-Means 聚类模块利用这些已标记的查询接口指导初始中心点的选择.其过程描述如算法 LabelInterface 所示.

具体查询接口标记策略如下:首先从已构建的 DTTD 中的每个领域的领域特征词词集 T_{D_i}中选择最能代表该领域 D_i 的领域主题特征词词集,即从中选出隶属度分数排名最靠前的三个词,记为 Tset,共选出 K 个 Tset(line 1);然后对于已选出的每个 Tset,从包含该 Tset 的查询接口中随机选择一个查询接口并标记它(line 3);接着把已标记的接口放入已标记接口集 LabeledInterfaces 中(line 4);直到已标记接口集 LabeledInterfaces 中元素个数为 K 时终止.

算法 1　LabelInterface($Interfaces_{TF}$, K, DTTD)

/* 输入:K 需要标记的接口数,DTTD 为领域主题特征词词典

输出:已标记接口集合 */

1: TsetS = SelectRepresentativeTerms(DTTD, K)

2: repeat

3: for each Tset which belong to TsetS, select a interface I to label it randomly from interfaces whose domain theme terms contain all terms of Tset

4: LabeledInterfaces = AssignInterface(I)

5: until number of LabeledInterfaces is K

6: return LabeledInterfaces

4.5 半监督 K-Means 聚类

本模块对那些含有主题领域特征词和属性标签的查询接口采用半监督 K-Means 聚类算法获得初始簇,以便后分类模块再次分类,其过程描述如算法 TAFSSC 所示.

算法 2　TAFSSC($Interface_{TF}$, K, DTTD)

/* 输入:$Interfaces_{TF}$指那些含有领域主题特征词和属性标签的查询接口集,K 为簇的个数,DTTD 为领域特征词词典

输出:簇 */

1: Centroids = LabelInterfaces($Interfaces_{TF}$, K, DTTD)

2:repeat
3:Clusters = AssignClosestCentroid(Interfaces$_{TF}$, Centroids)
4:Centroids = UpdateCentroid(Clusters)
5:untril Centroids don't change
6:return Clusters

半监督 K-Means 聚类策略具体如下:首先利用查询接口标记算法标记初始的 K 个接口,并把它们作为初始中心点(step 1);然后将 Interfaces$_{TF}$中剩余未标记接口中的每个接口指派给与中心点最近的簇(step 3),接着更新簇中心点(step 4);最后直到簇中心点几乎不再变化.其中簇 i 中心点计算公式为:

$$I_{p,i} = \text{argmax} \sum_{\forall I_p \in C_i \wedge I_k \in C_i \wedge p \neq k} sim(I_p, I_k) \qquad (6)$$

其中,$I_{p,i}$表示第 i 个簇的中心点,C_i 表示第 i 个簇.

4.6 后分类

本模块则是对预处理模块中那些不属于 Interfaces$_{TF}$的查询接口进行再次分类,从而生成最终分类,其描述过程如算法 PostClassification 所示.

后分类具体方法如下:首先计算不属于 Interfaces$_{TF}$集合的 I 接口数(step 1);接着对 I 中每个接口采取如下操作:如果该查询接口表单属性集为空,那么利用公式(2)来计算它与初始簇之间的相似性(step 3),如果该查询接口领域主题特征词集为空,就利用领域公式(3)来计算它与初始簇之间的相似性(step 4);然后把该查询接口指派给最近的簇(step 5);之后对簇进行更新(step 6);接着把计数 K 减 1(step 7),直到 K 的值为 0.

算法 3 PostClassification(InitClusters, I)
/* 输入:InitClusters 为半监督 K-Means 聚类生成的初始簇,I 为那些领域主题特征词为空或属性标签为空的查询接口集合
输出:InitClusters */
1:K = getNumber(I)
/* compute number of interface set I */
2:repeat
3:if $F = \Phi$,
Compute $sim(I_j, InitClusters)$ by using formula(2)
4:if $T = \Phi$,
Compute $sim(I_j, InitClusters)$ by using fomula(3)
5:Assign interface I_j to closest cluster
6:InitClusters = Update(InitClusters)
7:K--
8:until $K = 0$
9:return InitClusters

5 实验及分析

为评估算法执行情况,本文从航空订票、图书、汽车和房地产四个领域中收集了 210 个数据源,这些数据源一部分来自于 UIUC 知识库[17],其余的通过 Web 爬虫获取.其中 T(主题特征词集合)和 F(表单属性集合)均不为空的数据源用于半监督 K-Means 聚类,T 为空或 F 为空的数据源用于后分类,不同领域数据源数量分布如表 3 所示.

表 3 深层网络数据集

	Airfares	Books	Automobiles	Real Estate	Total
$T! = \Phi$ 且 $F! = \Phi$	54	34	45	48	181
$T = \Phi$ 或 $F = \Phi$	2	16	8	3	29

深层网络数据源分类的目的是尽可能地按领域划分数据源,从而为查询接口集成提供"纯净"的数据源.本文利用准确率和召回率衡量簇的好坏,其准确率和召回率公式如下所示:

$$\text{Precision}(i,j) = \frac{n_{ij}}{n_j}, \qquad \text{Recall}(i,j) = \frac{n_{ij}}{n_i} \qquad (7)$$

其中,n_{ij}表示簇 C_j 中属于领域 i 的查询接口数,n_i 表示领域 i 的查询接口数,n_j 表示簇 C_j 中的查询接口数.

图 5 展示了本文对 T 和 F 均不为空的 181 个数据源采用改进的 K-Means 聚类效果.可以明显看出,C_1 簇中绝大部分数据源属于航空订票领域,C_2 簇中绝大部分数据源属于图书销售领域,C_3 簇中绝大部分数据源属于汽车领域,C_4 簇中绝大部分数据源属于房地产领域,这反映了簇内数据源间的同质性非常高.

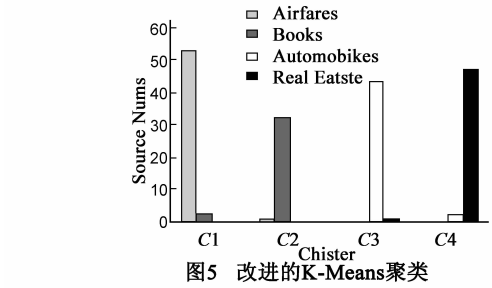
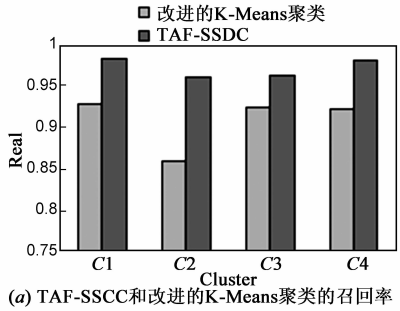


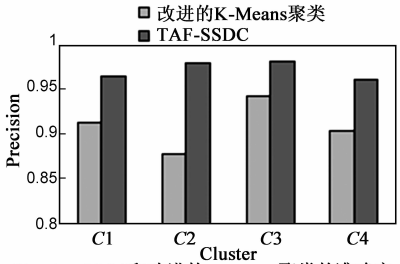
图5 改进的K-Means聚类

图 6(a)、图 6(b)分别展示了 TAF-SSCC 方法与单纯改进的半监督 K-Means 聚类方法的召回率与准确率的对比结果.单纯改进的 K-Means 聚类方法的召回率分别为 92.86%、86.00%、92.45%、92.16%,准确率分别为 91.23%、87.76%、94.23%、90.38%,平均召回率和平均准确率分别为 90.87%、90.90%;而 TAF-SSCC 方法的召回率分别为 98.21%、96.00%、96.23%、98.04%,准确率分别为 96.49%、97.96%、98.08%、96.15%,平均召回率和平均准确率分别为 97.17%、97.12%,其召回率和准

确率都明显提高了,尤其是图书销售领域(即对应簇 C_2).其主要原因是该领域中存在较多的领域主题特征词为空或者属性为空的数据源.所以采取先聚类后分类的 TAF-SSCC 方法可以明显提高最终数据源分类的效果.



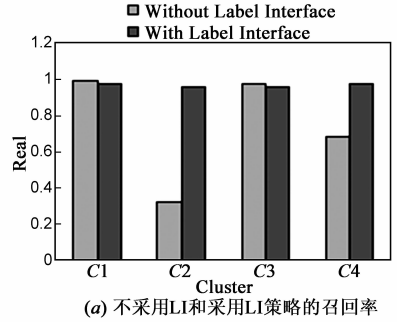
(*a*) TAF-SSCC和改进的K-Means聚类的召回率



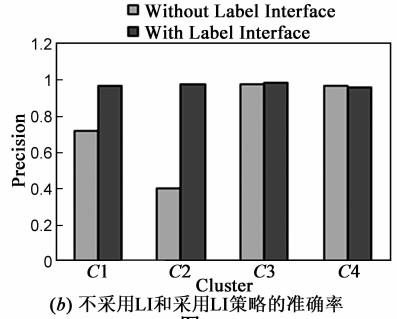
(*b*) TAF-SSCC和改进的K-Means聚类的准确率

图6

图 7(*a*)、图 7(*b*)分别展示了 TAF-SSCC 方法与不采用 LI(查询接口标记策略)方法的召回率与准确率的对比结果.不采用查询接口标记时,其召回率分别为 100%、32.35%、97.78%、68.75%,准确率分别为 72%、40.74%、97.78%、97.04%,平均召回率和平均准确率分别为 80.08%、82.06%;而 TAF-SSCC 方法的召回率分别为 98.21%、96.00%、96.23%、98.04%,准确率分别为 96.49%、97.96%、98.08%、96.15%,平均召回率和准确率分别为 97.17%、97.12%.其召回率和准确率都显著提高了.其主要原因是采用查询接口标记策略来选择初始中心点,极大地降低了初始中心点选择的随机性,使得初始中心点之间的差异性变得尽可能大,而随机选择初始中心点就可能导致初始中心点差异性非常小,进而导致召回率和准确率下降.



(*a*) 不采用LI和采用LI策略的召回率



(*b*) 不采用LI和采用LI策略的准确率

图7

6　结语

本文对 Deep Web 数据源分类开展了深入的研究工作,提出了一种基于主题和表单属性 Deep Web 数据分类方法,即 TAF-SSCC,该方法具有良好的扩展性、鲁棒性及覆盖性;为改进 K-Means 初始中心点选择具有随机性的缺点,本文提出了一种查询接口标记策略;同时基于主题和表单属性提出了一种改进的查询接口数据源相似性度量方法.实验结果表明该方法是切实有效的.

未来将进一步研究属性同义词词典的自动构建方法,从而提高本文数据源分类方法的自动化程度.

参考文献

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

[2] Madhavan J, Cohen S, et al. Web scale data integration: you can afford to pay as you go[A]. Proceedings of CIDR'07[C]. United States: CIDR, 2007. 342 – 350.

[3] 刘伟,孟小峰,等. Deep Web 数据集成研究综述[J].计算机学报,2007,30(9):1475 – 1489.
Liu Wei, Meng Xiaofeng, et al. A survey of deep web data integration[J]. Chinese Journal of Computers, 2007, 30(9): 1475 – 1489. (in Chinese)

[4] 申德荣,刘丽楠,等.一种面向 Deep Web 数据源的重复记录识别模型[J].电子学报,2010,38(2):275 – 281.
Sheng Derong, Liu Linan, et al. A duplicate records identification model for deep web data sources[J]. Acta Electronica Sinica, 2010, 38(2): 275 – 281. (in Chinese)

[5] Wu C M, Qiang B H, et al. Deep web classification based on domain feature text[J]. International Journal of Advancements in Computing Technology, 2011, 3(6): 267 – 275.

[6] Feng Y, Zhou Q W. Attribute decentralization algorithm-based deep web sources classification[J]. Advances in Information Sciences and Service Sciences, 2012, 4(1): 423 – 431.

[7] Noor U, Rashid Z, et al. TODWEB: training-less ontology based deep web source classification[A]. ACM International Conference Proceeding Series[C]. United States: ACM, 2011. 190 – 197.

[8] 马军,宋玲,等.基于网页上下文的 Deep Web 数据库分类

[J].软件学报,2008,19(2):267 – 274.
Ma J, Song L, et al. Classification for deep web databases based on the context of web pages[J]. Journal of Software, 2008, 19(2):267 – 274. (in Chinese)
[9] Le H Q, Conrad S. Classifying structured web sources using aggressive feature selection [A]. WEBIST 2009 [C]. United States: ISA, 2009. 618 – 625.
[10] Barbosa L, Freire J, et al. Organizing hidden-web databases by clustering visible web documents[A]. Proceedings of International Conference on Data Engineering[C]. United States: IEEE, 2007. 326 – 335.
[11] He B, Tao T, et al. Organizing structured web sources by query schemas: a clustering approach [A]. Proceeding of CIKM'04[C]. United States: ACM, 2004. 22 – 31.
[12] Zhao P P, Huang L, et al. Organizing structured deep web by clustering query interfaces link graph [A]. Lecture Notes in Computer Science[C]. Germany: Springer, 2008. 683 – 690.
[13] Gao Y, Qi H, et al. Semi-supervised k-means clustering for multi-type relational data[A]. Proceedings of ICMLC'08[C]. United States: IEEE, 2008. 326 – 330.
[14] 于娟,党延忠.领域特征词的提取方法研究[J].情报学报,2009,28(3):368 – 373.
Yu Juan, Dang Yanzhong. Domain feature and its extracting approach[J]. Journal of the China Society for Scientific and Technical Information, 2009, 28(3):368 – 373. (in Chinese)
[15] Nguyen H, Nguyen T, et al. Learning to extract form labels [A]. Proceedings of the VLDB Endowment[C]. New York: Springer, 2008. 684 – 694.
[16] Miller G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11):39 – 41.
[17] CS, UIUC. The UIUC Web integration repository [OL]. http://metaquerier.cs.uiuc.edu/repository/, 2003.

作者简介

祝官文 男,1986 年生于江西南昌.哈尔滨工程大学博士研究生,主要研究方向为海量数据集成技术及数据空间.
E-mail: zhgwen8761851@126.com

王念滨 男,1967 年生于四川达州.哈尔滨工程大学教授,博士生导师,主要研究方向为海量数据处理、数据集成技术,数据空间,并行计算等.
E-mail: wangnianbin@hrbeu.edu.cn