基于领域特征文本的Deep Web分类研究

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 Deep Web自动分类是建立深网数据集成系统的前提和基础。提出了一种基于领域特征文本的Deep Web分类方法。首先借助本体知识对表达同一语义的不同词汇进行了概念抽象,进而给出了领域相关度的定义,并将其作为特征文本选择的量化标准,避免了人为选取的主观性和不确定性;在接口向量模型构建中,考虑了不同特征文本对于分类作用的差异,提出了一种改进的W-TFIDF权重计算方法;最后采用KNN算法对接口向量进行了分类。对比实验证明,利用所提方法选择的特征文本是准确有效的,新的特征文本权重计算方法能显著地提高分类精度,且在KNN算法中表现出较好的稳定性。

关键词 特征文本,领域分类,向量空间模型,Deep Web

中图法分类号 TP391 **文献标识码** A

Research on Deep Web Classification Based on Domain Feature Text

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 Deep Web classification is the basis of building Deep Web data intergration system. An approach was proposed to classify the Deep Web based on domain feature text. Using the ontology knowledge, the concepts which express the same semantics were firstly extracted from different texts. Then the definition of domain correlation was given as the quantitative criteria for feature text selection, in order to avoid the subjectivity and uncertainty of manual selection. In the process of the interface vector space model construction, an improved weighting method named W-TFIDF was proposed to evaluate the different roles of feature text. At last, a KNN algorithm was used to classify these interface vectors. Comparative experiments indicate that the feature text selected by our method is accurate and effective, and the new weighting method can improve the classification precision significantly and shows good stability in KNN classification.

Keywords Feature text, Domain classification, Vector space model, Deep Web

1 引言

Deep Web代表了Web在线数据库中海量的、高质量的结构化数据信息,其资源容量是Surface Web的400～550倍[1]。为了有效地获取Deep Web中的数据,一个基础工作就是对Deep Web进行自动领域划分,即针对用户查询请求,Deep Web数据集成系统能根据查询主题,自动识别出相关领域的Web在线数据库,并分别在这些数据库上提交查询,再将结果合并后反馈给用户。然而,Deep Web数据隐藏在查询接口(通常为HTML表单)后面,系统只能通过这些查询表单来判断Deep Web所属领域。

目前,对Deep Web进行自动分类的方法主要有pre-query和post-query两种。其中,post-query通过向接口表单提交查询关键词,利用返回结果来对Deep Web进行分类,典型代表如哥伦比亚大学的QProber系统[2],其他工作见文献[3,4]。采用这种方法的核心在于如何确定有效的查询关键字以及对接口表单语义模式的准确理解,文献[5]指出了该方法的困难性,采用这种方法的研究并不多见。更多学者则关注pre-query方法,在这种方法中,将查询接口表单看作Web数据库的一个视图,通过分析接口表单及其所在页面的可视属性来判断和分类后台数据。由于词汇是领域的最直接表征,因此大多数工作集中在基于文本的分类研究上,分别从特征文本选择、接口向量构建、分类器构造等不同角度进行了探讨和改进。在这类研究中,主要面临的挑战有两个:一是如何选择出更具领域代表性的特征文本,从而达到有效降维的目的;另一个是如何确定特征文本的权重,以更好地体现不同特征文本对于分类所起的作用。这两个因素将直接影响查询接口向量的构建,并最终影响分类精度。

到稿日期:2011-05-02 返修日期:2011-07-16 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。

本文结合 Deep Web 分类特点，在特征文本选择和权重计算两个方面进行了深入研究，提出了一种利用领域特征文本对 Deep Web 进行自动分类的方法。通过在 4 个领域共 160 个查询接口上的对比实验，证明了本文所提方法具有较高的分类准确性。

2 领域特征文本抽取

有关文本和 Web 文档的分类已进行了深入地探讨[6-8]，但对于 Deep Web 的分类研究才刚刚起步。类似于传统的文本分类，如何选择有效的特征文本成为 Deep Web 分类首先要解决的问题。在普通的文本分类中，通常存在“特征高维性”和“向量稀疏性”的特点。但 Deep Web 查询接口包含的词汇通常非常有限，文献[9]经过统计，认为每个领域的词汇个数一般稳定在 40 个左右，这使得可以通过使用统计的方法进行特征文本抽取。现在的问题是：1)选择哪些文本作为统计对象；2)如何处理对同一语义的多重表示问题；3)选择特征文本时采用何种度量标准。本节就这 3 个问题展开讨论。

2.1 文本选择

由于 Deep Web 查询接口中已经包含了非常丰富的文本信息，因此大多数研究仅关注于接口本身，充分利用接口表单中的文本对 Deep Web 所属领域进行分类。而少数研究则除了考察接口表单文本外，还综合考虑了接口所在的网页信息。如文献[10]将包含查询接口的网页文本看作是接口表单的上下文，提出了一种上下文感知的表单聚类方法。文献[11]则根据电子商务网站的特点，不仅同时考虑了网页特征词、表单接口属性，而且特别考虑了网页中的价格因素。虽然网页文本能提供更多领域信息，但由于网页中通常会包含广告、导航、版权等大量领域无关文本，因此这种方法势必会增加文本中的噪声，甚至会严重干扰特征项的抽取，增加了算法复杂度，因此本文将关注焦点仅放在查询接口本身。

图 1 展示了一个图书领域的查询接口。通过观察，可以发现特征文本主要分布在以下几个地方：

(1)表单描述信息，如“Used Book Search”，其中的 Book 能清晰表明该接口所属领域。

(2)表单控件的标签文本，如“Author”，“Title”等。

(3)input 标记中的 name 和 value 属性，如提交按钮中的文本(即 value 属性)为“Find the Book”。

(4)控件中的文本，如 select 下拉列表中的选项值。

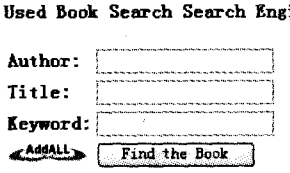

图 1 一个图书领域的查询接口

由此，本文将对以上 4 个位置的文本进行提取。

2.2 概念抽象

在特征文本选择中，另一个需要解决的问题是对同一语义的多重表示，如为了表示作者名字，music 和 book 领域通常会选择使用“musician”和“author”。事实上，由于表单设计的主观性，这种现象不仅出现在不同领域间，即使在同一领域内也广泛存在。这给特征文本的抽取和降维都带来了困难，因此应对这部分词汇进行归纳和合并。语义上，这些文本可以分为同义词、近义词、上下位与包含等几种关系。由于 Deep Web 查询接口中的文本数量有限，因此可以采用基于统计的方法对这些词汇进行概念抽象。

本文以 UIUC 查询接口数据集中的 book，movie，music，airfare 4 个领域作为研究对象[12]，从每个领域中随机选取了近 50 个查询表单，首先对抽取的文本进行去除停用词和词根还原等预处理，之后采用基于本体的概念和方法[13]，分别为这 4 个领域构建本体。图 2 展示了 music 本体的概念层次结构。通过这种概念抽象，将同一领域的相关词汇进行了规约，从而有效减少了特征文本的个数，并且归约后的概念类比单纯的词汇更能反映领域内容，能为后继的分类算法提供语义表示框架。

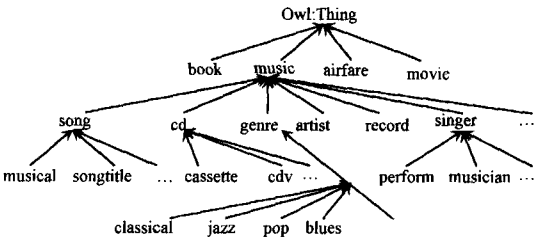

图 2 music 领域本体

2.3 特征文本选择

经过对文本进行概念抽象，查询接口中的文本数量有所减少，但仍存在高维现象。如何选取最少的文本来表征领域，仍是需要解决的问题。在已有研究中，多是基于统计观察而人工确定特征文本集，显然这种方法缺乏量化标准，存在较大主观性。传统的 TFIDF 方法可以有效评估一个特征项对于一个文档的重要程度，但对于分类来说，如果一个文本在一个领域的所有文档中频繁出现，则恰恰说明该文本能很好地代表领域特征，这样的文本应被赋予较高的权重。而利用 TFIDF 方法则会得出完全相反的结果，因此 TFIDF 并不适合作为领域特征文本选择的标准。文献[14]简单地将文本在一个领域样本集合中出现的次数作为特征文本选择的依据，然而某些表单可能会使用大量同一非领域词汇，用此方法将得到不准确的特征项。为此，对单纯的词频方法进行了完善，提出了一种基于领域相关度的特征文本选择量化标准。

定义 1 设有属于类别 $C_i=(A_{i1},A_{i2},\cdots,A_{iq})$ 的 N 个接口表单，A 为样本集中出现的所有词汇，共 P 个，且 A_{ij} ($j\in[1,q]$)在 $M(M\leqslant N)$个文档中出现，共 Q 次，则定义 A_{ij} 在 C_i 中的词频 $TF=Q/P$，定义 A_{ij} 在 C_i 中的分布 $DD=M/N$，定义 A_{ij} 与类 C_i 的领域相关度 DC 为

$$DC(C_i,A_{ij})=TF\times DD=\frac{Q}{P}\times\frac{M}{N} \tag{1}$$

显然，DC 与 A_{ij} 的词频和分布均成正比，这说明如果 A_{ij} 在 C_i 中出现频率越高，分布越广泛，则 A_{ij} 与 C_i 的相关程度越高，其越适合作为该类别的特征项。图 3 展示了 4 个领域的文本相关度计算结果。由图可见，特征文本较好地符合了齐普夫定律，且每个领域的特征文本数量通常在 11 个以内，这为特征文本的确定提供了量化依据。

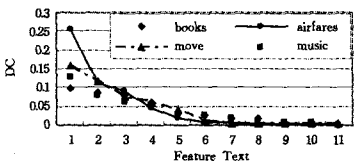

图 3 4 个领域的相关度计算结果

3 接口向量模型

本文采用类似文本分类中使用的向量空间模型(VSM),将 Deep Web 查询接口看作是由一组特征项和相应权重构成的向量集合。

定义 2 设有类别集合 $C=(C_1,C_2,\cdots,C_n)$,其中 C_i 的特征项集合 $T_i=(t_{i1},t_{i2},\cdots,t_{ikm})(i\in[1,n])$,则定义查询接口向量 $IV=\{T,W\}$。

其中,

(1)$T=(T_1,T_2,\cdots,T_n)=(t_{11},\cdots,t_{1k1},\cdots,t_{n1},\cdots,t_{nkm})$为领域特征项集合。由于各领域间的特征文本可能存在部分重合,如 books 和 movies 领域都选用了"title"作为特征文本,因此特征文本数量 $k\leqslant k_1+k_2+\cdots+k_m$。因此,$T$ 可进一步表示为 $T(t_i)_{1\times k}=(t_1,t_2,\cdots,t_k)$。

(2)$W=(w_1,w_2,\cdots,w_k)$为特征项权重,则由 N 个查询接口构成的权重向量集合为

$$W(w_{ij})_{N\times k}=\begin{bmatrix} w_{11} & \cdots & w_{1k} \\ \cdots & \ddots & \cdots \\ w_{N1} & \cdots & w_{Nk} \end{bmatrix}$$

式中,W_{ij} 为第 i 个查询接口中第 j 个特征项的权重。

TFIDF 是在文本分类中被广泛使用的权重计算方法。但我们注意到,不同特征文本对于分类的贡献是不同的,如某些特征文本只会出现在一个领域中(如 ISBN 仅会出现在 book 领域),而某些特征文本则可能同时出现在多个不同领域(如 title 分别出现在 book,movie 和 music 3 个领域)。显然,ISBN 比 title 具有更强的领域区分特性,应被赋予更高的权重值。而 TFIDF 方法则不能对这种特性进行有效区分。为此,本文首先为特征文本进行了词汇特性上的分类。

定义 3 设有领域类别集合 $C=(C_1,C_2,\cdots,C_n)$和特征项集合 $T=(t_1,t_2,\cdots,t_k)$,如果

(1)$t_i\in C_j\wedge t_i\notin C\backslash C_j(j\in[1,n])$,称 t_i 为领域特有词汇,记为 Ss;

(2)$t_i\in(C_j\cap C_m)\wedge t_i\notin C\backslash(C_j\cup C_m)(j,m\in[1,n],j\neq m)$,称 t_i 为领域通用词汇,记为 Sc;

(3)$t_i\in C_1\cap C_2\cdots\cap C_n$,称 t_i 为领域噪声词汇,记为 Sn。

在对特征文本进行词性分类的基础上,本文对 TFIDF 权重计算方法进行了修正,提出了一种加权的 TFIDF 权重计算方法,即通过增加调解系数来区分不同特征文本在领域分类中的作用,记为 W-TFIDF。

$$w_{ij}=\alpha\times tf_{ij}\times idf_i \tag{2}$$

式中,tf_{ij} 代表特征文本 t_i 在文档 d_j 中的词频(term frequency),idf_i 代表特征文本 t_i 在文档集中的逆文档词频(inverse document frequency),α 为调节因子。根据对比实验,本文给出了如下经验公式:

$$\alpha=\begin{cases}1, & t_i\in Ss\\ 1-p/n, & t_i\in Sc\\ 0, & t_i\in Sn\end{cases} \tag{3}$$

式中,n 为领域类别数,p 为 t_i 在 C 中出现的次数。显然,p 值越大,则 α 值越小,表明该特征文本在领域分类中所起的作用也越小。如果 $p=n$,则 $t_i\in Sn$,$\alpha=0$,表明该特征项对于 Deep Web 分类无任何作用。

以上对权重值的处理较好地分区了不同特征文本对于分类所起的不同作用,这也与人们的主观认识是一致的。

4 分类算法

在对接口表单进行了向量表示后,本文选择了 KNN(K Nearest Neighbors)算法来对查询接口向量进行分类。具体描述如下。

算法 1 FTBC(Feature Text Based Classification)

输入:已分类的训练样本集合及待分类的 Deep Web 查询接口表单

输出:Deep Web 所属领域 $C_1,C_2,\cdots,C_n$

(1)按第 2 节方法依次对各领域文本进行概念抽象,并利用式(1)计算文本的领域相关度,在此基础上确定特征文本;

(2)按第 3 节方法为训练样本及待分类的接口表单构造向量 IV_i 和 IV_j,其中权重计算利用式(2)和式(3);

(3)分别计算 IV_j 到 IV_i 的欧式距离:

$$D(IV_i,IV_j)=\sqrt{\sum_{q=1}^{k}(x_{iq}-x_{jq})^2} \tag{4}$$

(4)选取 D 值最小的前 k 个训练样本,依次统计这些样本所属的类别 $n_i=\sum_{1}^{k}C_i$;

(5)选取 $Max(n_i)$所属类别为 IV_j 所属的领域。

在分类算法的评价标准上,本文采用了信息检索中常用的 3 个评估指标:准确率 P(precision)、召回率 R(recall)和综合 F 测度(F-measure)。设 IC 表示分类算法识别出的属于类别 C_i 的接口数,TC 表示被正确地分到类别 C_i 的接口数,MC 表示属于类别 C_i 的实际接口数,则

$$P=\frac{TC}{IC},R=\frac{TC}{MC},F=\frac{2\times P\times R}{P+R}$$

5 实验与分析

为了验证本文所提方法的有效性,利用 UIUC 的 TEL-8 查询接口数据集进行了相关实验。选择 books,airfares,movies,music 4 个领域作为实验对象,从每个领域中手工各选取 40 个查询接口,共计 160 个。

5.1 分类实验

首先依次对 4 个领域的查询接口文本进行提取和预处理,然后进行概念抽象和相关度计算,并依据相关度计算结果确定各领域的特征项。最终结果如表 1 所列。其中,books 领域 10 个、movies 领域 11 个、music 领域 10 个、airfares 领域 11 个。

表 1 4 个领域的特征文本

领域	特征文本
Books	name,book,title,publisher,category,format,isbn,keyword,price,subject
Movies	category,movie,dvd,cast,title,screen,video,release,format,theater,price
Music	song,genre,artist,title,cd,record,album,singer,label,track
Airfares	time,month,airline,destination,passenger,return,from,date,depart,trip,cabin

基于选定的特征文本,利用本文提出的 W-TFIDF 权重计算方法分别为 160 个查询接口构造了向量,并随机从每个领域中选择了 20 个向量作为训练样本。其余的 80 个接口向量均作为测试样本,并利用 KNN 算法进行了分类。当 K 取

值为 5 时，分类结果如表 2 所列。

表 2　4 个领域的分类结果

领域	P(%)	R(%)	F(%)
Books	100	100	100
Movies	100	89.47	94.4
Music	95.45	100	97.67
airfares	95.45	100	97.67

由表 2 可见，本文所提分类算法在 4 个领域上均达到了较高的准确率、召回率和综合 F 测度。

5.2　对比实验

为了进一步评估本文提出的特征文本选择方法以及 W-TFIDF 权重计算方法的有效性，分别在分类精度和稳定性两个方面与传统的 TF 和 TFIDF 方法进行了对比实验。

5.2.1　精度对比

在向量构造中，分别使用 TF，TFIDF 及 W-TFIDF 方法作为各特征文本的权重值，之后采用 KNN 算法进行分类。当 $K=5$ 时，3 种分类结果的 F-measure 对比如图 4 所示。

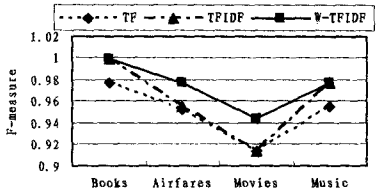

图 4　3 种权重计算方法的综合 F 测度对比

由图 4 可见：1）利用所提方法选出的特征文本进行向量构建，3 种方法均达到了较高的 F 测度，4 个领域的平均 F-measure 值均达到 91.4%以上，说明所提出的特征文本选择方法是准确有效的；2）W-TFIDF 方法较传统的 TF、TFIDF 各评价指标均有所提升，说明本文提出的权重计算方法更为准确，能显著提高分类精度。

5.2.2　稳定性对比

在 KNN 算法中，由于 K 的取值并没有固定标准，导致不同 K 值下的分类结果可能存在较大不确定性。为了判断 W-TFIDF 方法对于分类结果的稳定性影响，依次选取 K 值从 1 至 17，并与 TF、TFIDF 方法进行了对比。结果如图 5 所示。

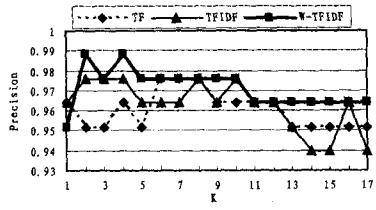

图 5　不同 K 值下 3 种方法的分类精度对比

由图 5 可见，W-TFIDF 方法不仅在分类精度上总体高于 TF 及 TFIDF 方法，而且当 K 取 2～17 时，始终保持了较高的稳定性。这再次证明利用 W-TFIDF 作为特征文本的重值是适宜的。

结束语　本文提出了一种基于领域特征文本的 Deep Web 分类方法，主要贡献如下：1）在深入分析领域词汇的基础上，给出了一种基于本体概念的语义抽象方法，增强了特征文本对领域的表征能力；2）提出了一种领域相关度的评价方法，并将其作为特征文本选择的量化标准，避免了传统的人工选择的主观性和不确定性；3）在接口向量模型构建中，考虑了不同特征文本的词性特征，提出了一种新的权重计算方法 W-TFIDF，实验证明其能有效提高 Deep Web 分类精度。

后续工作中，拟综合考虑查询接口所在网页的文本，研究如何进行领域相关文本的有效提取，以进一步提高 Deep Web 分类算法的准确性。

参 考 文 献

[1] Bergman M K. The deep web: Surfacing hidden value[EB/OL]. http://www.brightplanet.com/, 2001

[2] Gravano L, Ipeirotis P G, SahamiM. QProber: A system for automatic classification of hidden-web databases[J]. ACM TO IS, 2003, 21(1): 1-41

[3] Hedley Y L, YounasM, James A. The categorisation of hidden Web databases through concept specificity and coverage[C]// 19th Int'l Conf. on AINA. Washington: IEEE Computer Society, 2005: 671-676

[4] Gong Z G, Zhang J B, Liu Q. Hidden-Web database exploration [C]// 6th Int'l Conf. on ISDA. Washington: IEEE Computer Society, 2006: 838-843

[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] Lim C S, Lee K J, Kim G C. Multiple sets of features for automatic genre classification of web documents[J]. Information Processing & Management, 2005, 41(5): 1263-1276

[7] Yang Li, Zuo Chun, Wang Yu-guo. K-nearest neighbor classification based on semantic distance[J]. Journal of Software, 2005, 16(12): 2054-2062

[8] Fortuna B, Grobelnik M, Mladeni D. Background knowledge for ontology construction[C]// Proc. of the 15th Int'l Conf. on WWW. New York: ACM Press, 2006: 949-950

[9] He B, Chang K C C. Statistical Schema Matching Across Web Query Interfaces[C]// Proceedings of the 22th ACM SIGMOD Int'l Conf. on Management of Data. San Diego, 2003: 217-228

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

[11] Peng Q, Meng W, He H, et al. WISE-cluster: clustering e-commerce search engines automatically[C]// Proc. of the 6th ACM Int'l Workshop on WIDM. Washington: IEEE Computer Society, 2004: 104-111

[12] UIUC Web integration repository[EB/OL]. http://metaquerier.cs.uiuc.edu/repository/, 2003

[13] 张爱琦，左万利，王英，等. 基于多个领域本体的文本层次被定义聚类方法[J]. 计算机科学，2010，37(3)：199-204

[14] He B, Tao T, Chang K C C. Clustering Structured Web Sources: A Schema-based Model-differentiation Approach [C]// Proc of the 9th Int'l Conf. on Extending Database Technology. Heraklion: Springer-Verlag, 2004: 22-31