基于量子自组织神经网络的Deep Web分类方法研究

张 亮[1] 陆余良[1] 房珊瑶[2]
(解放军电子工程学院网络工程系 合肥 230037)[1] (北方电子设备研究所 北京 100191)[2]

摘 要 针对Deep Web数据源主题分类问题,首先研究了不同位置的特征项对Deep Web接口领域分类的影响,提出一种基于分级权重的特征选择方法RankFW;然后提出一种依赖领域知识的量子自组织特征映射神经网络模型DR-QSOFM及其分类算法,该模型在训练的不同阶段对特征向量和目标向量产生不同程度的依赖,使竞争层中获胜神经元的分布更为集中,簇的区域划分更为明显;最后,在扩展后的TEL-8数据集上进行的实验验证了RankFW和DR-QSOFM的有效性。

关键词 Deep Web接口,特征选择,主题分类,分级权重,领域依赖,量子自组织特征映射
中图法分类号 TP391 **文献标识码** A

Research on Deep Web Classification Approach Based on Quantum Self-organization Feature Mapping Network

ZHANG Liang[1] LU Yu-liang[1] FANG Shan-yao[2]
(Network Engineering Department, Electronic Engineering Institute of PLA, Hefei 230037, China)[1]
(Research Institute of Northern Electronic Equipment, Beijing 100191, China)[2]

Abstract In order to solve the problem of Deep Web data sources classification, this paper firstly researched how features in different position could effect the domain of Deep Web interfaces, and proposed a feature selection method RankFW which is based on Ranked weights. Then, a quantum self-organization feature mapping network model was proposed with a classification algorithm. This model relies on the feature vectors and target vectors incoordinately in different phases of training, making a more centralized distribution of winner neurons in competition layer and more obvious boundaries among clusters. Finally, some experiments were designed and carried out on the expanded TEL-8 dataset to test the validity of RankFW and DR-QSOFM.

Keywords Deep Web interface, Feature selection, Topic classification, Ranked weight, Domain relied, Quantum self-organization feature mapping

1 引言

随着Internet服务的增多,Web信息呈爆炸式增长,截止2004年,连接Internet的数据库已经达到450,000个[1],Deep Web的信息量急剧增加。在国内,截止2006年,中国大陆地区已经拥有24,000个Deep Web站点、28,000个Web数据库和74,000个查询接口,而仅有大约二分之一的Deep Web页面被中文搜索引擎覆盖[2]。Deep Web已经成为网络数据集成领域的热点之一。

Deep Web查询接口主题分类问题是Deep Web数据集成的重要环节之一,其实质是确定与该查询接口对应的数据源中信息所属的领域。查询接口的特征选择以及分类模型的设计是影响主题分类结果的主要因素。

国内外相关人员围绕Deep Web查询接口主题分类开展了一系列的研究。文献[3]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,并使用KNN分类算法和语义距离算法对Deep Web数据源进行分类。文献[4]提出了使用条件最大熵模型对Deep Web查询接口进行分类。其基本思想是为所有已知的因素建立模型,而对于所有未知因素取熵最大的分布为推测的概率分布。最大熵模型最大的优点是能够融合各种各样不同的特征,并且对这些特征不作任何独立性假设。文献[5]提出一种基于朴素贝叶斯的方法来判断表单是否是搜索型的,选择了HTML标签的属性值和控件之间的文本信息作为特征,但该方法忽略了领域的相关性。文献[6,7]使用向量空间模型(VSM)对数据源进行自动分类,将不同的数据源特征的权重以向量空间(矩阵)的形式进行组织,通过计算Cosine值来获取两个特征向量的相似度,并采用概念的信息增益方法进行特征选取。文献[8]提出了一种基于文字语义的相似度算法LSSC,其通过计算将查其询接口按相似度升序排序,并计算接口的平均相似度,据此来获取每个接口与平均相似度的差值,找出最小差值d,从而将接口划入最接近的领域中,但在NQ算法的第二步中,对所有类的所有接口都计算相似度,效率低下;算法终止条件"所有类成员趋于稳定"过于模糊,不易界定;未考虑新成员加

到稿日期:2010-07-28 返修日期:2010-10-20 本文受军队国防科技项目资助。
张 亮(1982—),男,博士生,主要研究方向为Web数据挖掘、网络安全,E-mail:liviocheung@sina.com;陆余良(1964—),男,教授,博士生导师,主要研究方向为计算机网络安全;房珊瑶(1983—),女,硕士,助理工程师,主要研究方向为数据挖掘、网络通信。

入一个类以后对该类的平均相似度的影响，理论上，如果不断地有新成员加入，该类的平均相似度会不停摆动，从而造成其他以其为参考的成员无法判断是否该加入该类。为了解决忽视域相关性的特征提取方法无法适用于 Deep Web 数据源分类的问题，文献[9]描述了一种新方法，即使用多层分类器(两层)对 Deep Web 数据源加以辨识和分类，主要步骤分为 3 步：第 1，通过主题表单爬虫抽取页面；第 2，使用朴素贝叶斯分类器对第一步中所得页面的相关性进行分析，并使用信息增益评估标准来减少特征空间的大小；第 3，根据表单的结构特征(如控件的数量)，使用 C4.5 决策树识别步骤 2 得到的页面相关域的查询接口并进行分类，但该文献仅仅介绍了页面分类及接口分类算法，并未深入研究特征选择的方法。

综上分析可知，现有的方法大多关注于分类模型的研究，忽视了 Deep Web 接口特征的差异性对分类结果的作用，从而影响了分类进度的进一步提高。针对以上问题，本文首先在第 2 节根据 Deep Web 查询接口所属网页的特点，提出了一种基于分级权重的特征选择方法 RankFW；然后在第 3 节针对主题分类问题设计了依赖领域知识的量子自组织特征映射网络模型 DR-QSOFM；并在第 4 节给出了利用 DR-QSOFM 进行分类的算法；第 5 节设计了 4 个实验，在扩展后的 TEL-8(Ex)数据集上验证了 RankFW 和 DR-QSOFM 的有效性；最后对本文作出总结。

2 特征选择方法 RankFW

对于 Deep Web 查询接口的主题分类而言，仅仅使用表单的结构统计特征无法为分类提供足够的信息量，而查询接口所在网页的上下文信息以及接口中控件的描述信息十分丰富，具有明显的领域性。因此，本文将综合表单上下文信息和控件的描述信息对 Deep Web 查询接口进行主题分类。

图 1 显示了将网页上下文信息和控件描述信息转换为特征向量空间的 3 个步骤：

1. 从包含接口的样本网页中抽取上下文和控件描述信息；

2. 对抽取的特征信息进行分词、去停用词和词根还原等预处理，并对特征维进行约减；

3. 计算特征权重，生成特征向量空间。

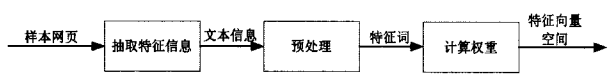


图 1 文本特征选择步骤

2.1 抽取接口的上下文信息和控件描述信息

接口的上下文信息是指接口所在网页中蕴含的与接口相关的文本信息。由于不同站点的设计人员的风格存在差异，接口在网页中所处的位置、接口控件的布局和描述都各不相同，因此接口上下文信息并不处于网页的固定位置。但通过对样本网页的分析发现，与接口相关的上下文信息集中在网页的标题、Meta 内容、导航路径以及与接口在 DOM 结构树中相差不超过 2 层的区域里(如图 2 所示)。本文将抽取这些位置的文本作为接口的上下文特征，同时还将接口控件的描述文本、默认值、可选值和名称作为接口控件的描述信息。

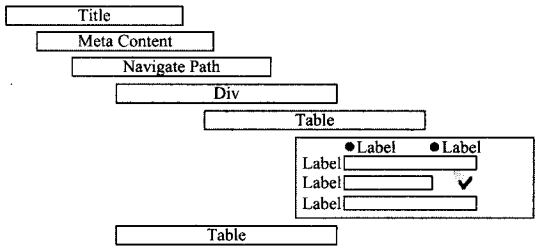


图 2 文本抽取特征抽取位置示意图

2.2 特征信息预处理

在英文中，单词之间以空格为分隔符，具有明显的界限，因此很容易进行分词。而在中文里，表达文本语义的最小单位是字，由字组成词、短语，进而得到句子。但以字为单位并不能准确地划分文本特征的语义，词才是具有独立表达能力和独立意义的有效单位。

本文采用中国科学院计算技术研究所研发的 ICTCLAS 2009[10] 对文本信息进行分词处理，该系统基于层叠隐马尔可夫模型对简繁汉字进行处理，分词精度达到 98.45%。在 ICTCLAS 完成分词和词性标注工作后，本文使用正则表达式选择其中的动词和名词作为特征项。例如，接口的特征信息“China-Pub 网上书店：计算机，经营，外语，建筑，法律”，经过 ICICLAS 分词和词性标注以后，得到：

China-Pub/n 网上/s 书店/n：/w 计算机/n，/w 经营/v，/w 外语/n，/w 建筑/n，/w 法律/n

上述结果采用了北大的二级标注标准。进而，定义正则表达式：[\S]+/(n|v|vn)\b，就可选出所有被标注为“/v”的动词、被标注为“/n”的名词以及具有动名双词性的词。

2.3 计算特征权重

不同的特征对于 Deep Web 查询接口主题分类具有不同的重要性，在抽取特征信息中的动、名词后，需要计算它们的权重。

常用的特征选择方法包括：期望交叉熵，用于衡量特定类别的概率分布和在某个特征出现的条件下类别的概率分布之间的距离；信息增益是一种基于熵的评估标准，用来衡量某个特征出现与否对预测分类的结果影响，表示特征 t 对确定文本类型所提供的信息量的大小；Chi-square 检验是一种基于假设检验的特征选取方法，用于衡量特征项与类别之间的独立无关性的缺乏度，Chi-square 值越大，表示特征与类别的相关度越高；互信息量体现了特征与类别的相关程度，如果特征 t 在类别 c_i 的文本信息中出现的概率较高，而在其他类别的文本信息中出现的概率较低，那么 t 具有较高的互信息量。TF * IDF 的基本思想是：特征 t 在文本中出现得越频繁则越重要；t 在越多的文本中出现则越不重要。

以上 5 种方法中，TF * IDF 得到了广泛应用，并在文本特征选择中表现出了较好的性能[11]。但 TF * IDF 存在一定的局限性：

首先，TF * IDF 考虑特征在整个样本集中出现的频数，在对 IDF 进行计算时，统计的是所有出现过某一特征的文本数，并没有区别特征项在类间文本中出现的概率。其次，TF * IDF忽略了特征在类内的分布情况。再次，TF * IDF 未考虑样本文本的长度对于特征项的影响。文本长度越长，包含的

特征项越多，从而出现特定特征的概率就越大，频数也越高，因此可能对特征项的 TF 值产生影响。最后，Deep Web 接口的文本特征的特点决定了不同位置上的特征信息具有不同的重要性。对大量样本网页的分析后发现，不同位置的特征表达主题的能力具有以下关系：

标题＞接口描述＞ Meta 内容＞导航＞相邻文本块

而 TF＊IDF 认为样本的特征项具有相同的表达能力，赋予相同的初始权重，忽略了位置差异的影响。

综上所述，以上 4 点局限性限制了 TF＊IDF 在 Deep Web 接口文本特征选择中的作用。为此，本文基于 TF＊IDF 的思想提出一种新的特征权重计算方法 RankFW：

$$w_{jk}=\frac{\sum_{r=1}^{5}\alpha_r\log(tf_{jkr}\times m_{c_ik})*\log(\frac{m_{c_ik}}{m_{\overline{c_ik}}+1})}{\sqrt{\sum_{k=1}^{F}(\sum_{r=1}^{5}\alpha_r\log(tf_{jkr}\times m_{c_ik})*\log(\frac{m_{c_ik}}{m_{\overline{c_ik}}+1}))^2}} \quad (1)$$

式中，α_r 代表 5 种不同位置的特征所具有不同的初始权重，反映了标题、接口描述、META 内容、导航信息和相邻文本块中的特征项对 Deep Web 接口分类的不同重要程度，当相同特征项在不同位置出现时，将其权重进行叠加。m_{c_ik} 表示出现特征项 t_k 的类内样本的数量，用以避免特征分布过分集中于少数类内样本而影响权重计算的情况。$m_{\overline{c_ik}}$ 表示出现特征项 t_k 且不属于 c_i 类的样本数量，引入 $\log(m_{c_ik}/m_{\overline{c_ik}}+1)$ 的目的在于解决 TF＊IDF 忽略样本类间分布的问题。为了消除样本长度对特征权重的影响，分母使用 Cosine 规格化方法对特征向量进行了归一化处理，其中 F 为特征集的大小。

3 DR-QSOFM 模型

3.1 DR-QSOFM 量子神经元模型

不同于普通的自组织特征映射神经网络，DR-QSOFM 将样本的领域信息作为模型输入的一部分，与样本的特征向量一起参与权重向量的调整。

DR-QSOFM 神经元由输入向量、权重向量、量子旋转门、传递函数组成，其中输入向量包含样本的特征向量和目标向量两个部分，模型结构如图 3 所示。

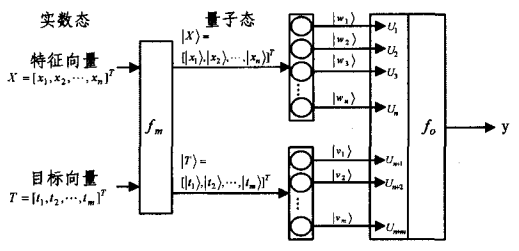


图 3 DR-QSOFM 神经元模型

转换函数 f_m 将处于实数态的特征向量 X 和目标向量 T 转换成对应的量子态输入 $|X\rangle$ 和 $|T\rangle$，f_m 的定义式如下：

$$|p_i\rangle=f_m(p_i)=\cos(\frac{2\pi}{1+e^{-p_i}})|0\rangle+\sin(\frac{2\pi}{1+e^{-p_i}})|1\rangle \quad (2)$$

目标向量 T 对应样本所属领域的向量表达式，如果样本集 S 总共涵盖了 $(a_1,a_2,\cdots,a_m)$ 共 m 个领域，则 T 的维数为 m。假设 $\exists s\in S, s\in a_l$，那么 s 的目标向量满足式(3)。

$$\begin{cases} T_s=[t_1^s,\cdots,t_i^s,\cdots,t_m^s] \\ t_i^s=1 & i=a_l \\ t_i^s=0 & i\neq a_l \end{cases} \quad (3)$$

权重向量同样使用量子态表示，$|W\rangle$ 表示特征向量与量子旋转门 U_i 的连接权重，$|V\rangle$ 表示目标向量与 U_i 的连接权重；量子旋转门 U_i 在量子态特征向量、目标向量和权重向量的作用下进行相位旋转，然后再反作用于 $|W\rangle$，使权重向量的方向朝着输入模式的方向进行调整，不断地向样本的中心位置移动。f_o 表示传递函数，将 DR-QSOFM 神经元的输出映射成一个实数。f_o 的表达式为：

$$\begin{aligned} y=f(\langle W|X\rangle)+f(\langle V|T\rangle)&=|\sum_{i=1}^{n}\langle w_i|x_i\rangle|+|\sum_{i=1}^{m}\langle v_i|t_i\rangle| \\ &=\sum_{i=1}^{n}\frac{\Gamma(|W_i\rangle)\cdot\Gamma(|x_i\rangle)+H(|w_i\rangle)\cdot H(|x_i\rangle)}{\sqrt{(\Gamma^2(|W_i\rangle)+H^2(|w_i\rangle))\cdot(\Gamma^2(|x_i\rangle)+H^2(|x_i\rangle))}}+ \\ &\sum_{i=1}^{n}\frac{\Gamma(|v_i\rangle)\cdot\Gamma(|t_i\rangle)+H(|v_i\rangle)\cdot H(|t_i\rangle)}{\sqrt{(\Gamma^2(|v_i\rangle)+H^2(|v_i\rangle))\cdot(\Gamma^2(|t_i\rangle)+H^2(|t_i\rangle))}} \quad (4) \end{aligned}$$

式中，$\Gamma(*)$ 表示取量子比特第一状态概率幅的算子，$H(*)$ 表示取量子比特第二状态概率幅的算子。

3.2 DR-QSOFM 神经网络模型

DR-QSOFM 神经网络由输入层和竞争层组成，其中竞争层包含了若干 DR-QSOFM 量子神经元，其示意图如图 4 所示。输入层的每一个结点与竞争层中的每一个神经元输出结点相连接。一个样本的特征向量和目标向量变换成量子态以后，在权重向量的作用下被映射到竞争层中的某一神经元输出结点上，同时对该结点及其一定邻域范围的其他输出结点与输入结点之间的权重进行调整。当所有样本都训练完毕后，样本集的领域信息将蕴含在 DR-QSOFM 的权重向量中，新的测试样本在权重向量的作用下会被映射到与其领域相关的簇中，从而得到该样本所属的领域。

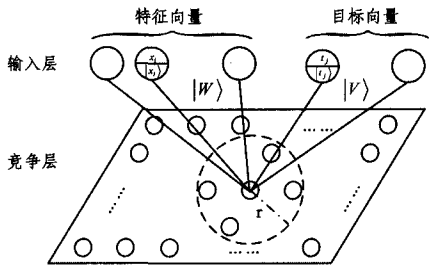


图 4 DR-QSOFM 模型的示意图

下面给出与 DR-QSOFM 的相关定义。

定义 1 任意两个 n 维的量子向量 $|X\rangle=[|x_1\rangle,|x_2\rangle,\cdots,|x_n\rangle]^T$ 和 $|Y\rangle=[|y_1\rangle,|y_2\rangle,\cdots,|y_n\rangle]^T$ 之间的距离定义为：

$$d=|\langle X|Y\rangle|^{-1}=\left(\sum_{i=1}^{n}\frac{\Gamma(|x_i\rangle)\cdot\Gamma(|y_i\rangle)+H(|x_i\rangle)\cdot H(|y_i\rangle)}{\sqrt{(\Gamma^2(|x_i\rangle)+H^2(|x_i\rangle))\cdot(\Gamma^2(|y_i\rangle)+H^2(|y_i\rangle))}}\right)^{-1} \quad (5)$$

定义 2 任意两个量子比特 $|x\rangle$，$|y\rangle$ 之间的夹角是指两者内积与范数乘积比值的反余弦值，计算公式为：

$$\Phi(|x\rangle,|y\rangle)=\arccos\left(\frac{\Gamma(|x_i\rangle)\cdot\Gamma(|y_i\rangle)+H(|x_i\rangle)\cdot H(|y_i\rangle)}{\sqrt{(\Gamma^2(|x_i\rangle)+H^2(|x_i\rangle))\cdot(\Gamma^2(|y_i\rangle)+H^2(|y_i\rangle))}}\right) \quad (6)$$

定义 3 获胜神经元 j_{uin}^s 是指与 s 的特征向量、目标向量之间具有最小距离的神经元，即：

$$j_{uin}^s=\underset{j\in\{1,2,\cdots,k^2\}}{\operatorname{argmin}}\{d_j^s\} \quad (7)$$

根据定义 1，输入样本 s 与竞争层神经元 j 的权重向量的

距离可以表示为：

$$d_j^s=|\langle X_s\bigcup T_s|W_j^s\rangle|^{-1}=$$
$$\left(\sum_{i=1}^{n}\frac{\Gamma(|x_i^s\rangle)\cdot\Gamma(|w_i^j\rangle)+\mathrm{H}(|x_i^s\rangle)\cdot\mathrm{H}(|w_i^j\rangle)}{\sqrt{(\Gamma^2(|x_i^s\rangle)+\mathrm{H}^2(|x_i^s\rangle))\cdot(\Gamma^2(|w_i^j\rangle)+\mathrm{H}^2(|w_i^j\rangle))}}+\right.$$
$$\left.\sum_{i=1}^{m}\frac{\Gamma(|t_i^s\rangle)\cdot\Gamma(|v_i^j\rangle)+\mathrm{H}(|t_i^s\rangle)\cdot\mathrm{H}(|v_i^j\rangle)}{\sqrt{(\Gamma^2(|t_i^s\rangle)+\mathrm{H}^2(|t_i^s\rangle))\cdot(\Gamma^2(|v_i^j\rangle)+H^2(|v_i^j\rangle))}}\right)^{-1}\quad(8)$$

在第 $i+1$ 次迭代中，假设获胜神经元的编号为 j，则连接 j 的权重向量 $|W\rangle$，$|V\rangle$ 的调整公式分别定义为：

$$|W_j(i+1)\rangle=$$
$$\begin{cases}[U_{j1}^w||w_1^j(i)\rangle,\cdots,U_{jn}^w||w_n^j(i)\rangle]^T, & j\in\rho(j_{uin}^i,r(i))\\ |W_j(i)\rangle, & j\notin\rho(j_{uin}^i,r(i))\end{cases}\quad(9)$$

$$|V_j(i+1)\rangle=$$
$$\begin{cases}[U_{j1}^v|v_1^j(i)\rangle,\cdots,U_{jn}^v|v_n^j(i)\rangle]^T, & j\in\rho(j_{uin}^i,r(i))\\ |V_j(i)\rangle, & j\notin\rho(j_{uin}^i,r(i))\end{cases}\quad(10)$$

式中，$\rho(j_{uin}^i,r(i))$ 表示以获用于神经元 j_{uin}^i 为中心、$r(i)$ 为半径的邻域，U_{ji}^* 表示量子旋转门，用于修正权重向量 $|W\rangle$，$|V\rangle$ 的相位。在不同的训练阶段，特征向量和目标向量的作用是不同的，因此使用量子旋转门进行相位调整时，U_{ji}^w 和 U_{ji}^v 具有不同的周期，其定义如下：

$$U_{ji}^w=\begin{bmatrix}\cos(\eta(i)\theta_{ji}^w) & -\sin(\eta(i)\theta_{ji}^w)\\ \sin(\eta(i)\theta_{ji}^w) & \cos(\eta(i)\theta_{ji}^w)\end{bmatrix}\quad(11)$$

$$U_{ji}^v=\begin{bmatrix}\cos((1-\eta(i))\theta_{ji}^v) & -\sin((1-\eta(i))\theta_{ji}^v)\\ \sin((1-\eta(i))\theta_{ji}^v) & \cos((1-\eta(i))\theta_{ji}^v)\end{bmatrix}\quad(12)$$

式中，θ_{ji}^* 表示调整的相位大小，由特征向量、目标向量与权重向量的夹角大小决定，其计算公式为：

$$\begin{cases}\theta_{ji}^w=-sign\left(\begin{vmatrix}\Gamma(|x_i^s\rangle) & \Gamma(|w_i^j\rangle)\\ \mathrm{H}(|x_i^s\rangle) & \mathrm{H}(|w_i^j\rangle)\end{vmatrix}\right)\cdot\Phi(|x_i^s\rangle,|w_i^j\rangle)\\ \theta_{ji}^v=-sign\left(\begin{vmatrix}\Gamma(|t_i^s\rangle) & \Gamma(|v_i^j\rangle)\\ \mathrm{H}(|t_i^s\rangle) & \mathrm{H}(|v_i^j\rangle)\end{vmatrix}\right)\cdot\Phi(|t_i^s\rangle,|v_i^j\rangle)\end{cases}\quad(13)$$

$\eta(i)$ 表示第 i 次迭代的学习速率，其计算公式为：

$$\eta(i)=\frac{0.5}{i+1}+0.5\quad(14)$$

由式(11)、式(12)可知，在训练开始阶段，$|W\rangle$ 调整相位的周期较大，因此特征向量所对应的权重调整较快；随着训练次数的增加，$|W\rangle$ 调整相位的周期逐渐减小，$|V\rangle$ 则逐渐增大，最终两者调整的频率将趋于一致，使特征向量 $|X\rangle$ 和目标向量 $|T\rangle$ 获得均匀的权重参与获胜神经元的选择。

4 DR-QSOFM 分类算法及分析

上节提出的 DR-QSOFM 分类模型同时将特征向量和目标向量作为输入来训练模型，然后利用该模型对未知样本进行分类。其分类过程如算法 1 所述。

算法 1 DR-QSOFM_Classification

输入：训练样本集 S，测试样本集 N

输出：模式类向量 Y

// step 1：训练模型

1 随机初始化权重向量 $|W\rangle$，$|V\rangle$；

2 设置初始学习速率 η_0、初始邻域半径 r_0、距离阈值 σ；

3 foreach s in S

4 抽取 s 的特征向量 $|X^s\rangle$ 和目标向量 $|T^s\rangle$；

// i_s 表示样本 s 对应的迭代序数

5 更新邻域半径：$r(i_s)=|r_0(1-i_s/|S|)|$；

6 foreach j in C

7 根据式(5)计算 $|X^s\rangle$ 与 $|W_j^s\rangle$ 之间的距离 d_{wj}^s、$|T^s\rangle$ 与 $|V_j^s\rangle$ 之间的距离 d_{vj}^s；

8 $d_j^s=(1-\eta(i_s))d_{wj}^s+\eta(i_s)d_{vj}^s$；

9 end

10 $j_{uin}^s=\underset{j}{\mathrm{argmin}}\{d_j^s\}$；// 获得获胜神经元，并加入 s 对应模式类的获胜神经元集合 C_{a_l}；

11 确定以 j_{uin}^s 为中心，$r(i_s)$ 为半径的邻域 $\rho(j_{uin}^i,r(i_s))$；

12 按照式(9)、式(10)分别调整 $|W\rangle$，$|V\rangle$ 相位；

13 按照式(14)调整 η_i；

14 end

// step 2：样本分类

15 foreach n in N

16 获取 n 对应的获胜神经元 j_{uin}^n；

17 if $j_{uin}^n=j^*,j^*\in\{C_{a_1},C_{a_2},\cdots,C_{a_m}\}$

18 $Y(i_n)$ 等于 j^* 对应的模式类序号；

19 else

20 $diff=\mathrm{argmin}|d_{j_{uin}^n}^n-d_{j^*}^n|,j^*\in\{C_{a_1},C_{a_2},\cdots,C_{a_m}\}$；

21 if $diff\leqslant\sigma$

22 $Y(i_n)$ 等于 j^* 对应的模式类序号；

23 else

24 $Y(i_n)=$unknown；

25 end

26 end

27 end

DR-QSOFM 分类算法分为两个阶段，第一阶段(1－14句)是利用训练样本对模型进行训练。在这个阶段中，首先设置模型的训练参数，并随机初始化权重向量，更新邻域半径；然后抽取每一个样本的特征向量和目标向量，并转换为相应的量子态，6－11 句计算样本对应的获胜神经元，将其加入与样本对应模式的获胜神经元集合中，并确定邻域半径；最后，根据样本输入和原始权重调整权重向量的相位。第二阶段(15－27 句)完成未知样本的分类。首先计算未知样本 n 所对应的获胜神经元 j_{uin}^n，如果神经元与获胜神经元集合中的某一神经元 j^* 重合，则赋予 n 与 j^* 相同的模式类别(17－18句)；否则，计算 j_{uin}^n 与获胜神经元集合中的每一个神经元的距离之差，选择其中最小差值 $diff$，如果 $diff$ 的绝对值不大于距离阈值 σ，则将获得最小差值的获胜神经元的模式类别赋予 n，否则将其类别标注为未知。算法中，第 8 句考虑了特征向量和目标向量在初始训练阶段所具有的不同作用，利用 d_{wj}^s，d_{vj}^s 计算出综合距离 d_j^s，从而影响训练过程中获胜神经元的选择结果。另外，从算法 1 可以看出，阈值 σ 的取值将直接影响到算法的收敛性，因此，在 5.3 节将设计实验以检验 σ 值对算法收敛性的影响。

5 实验及结果分析

5.1 文本特征选取

本实验的目的是检验特征权重计算方法 RankFW 的有效性，并将特征选择的结果与 TF * IDF 进行对比。实验测试数据为扩展后的 TEL-8 数据集。TEL-8 数据集源自 UIUC

大学，共包含 8 个领域的 477 个查询表单。但该数据集并未包含非查询表单样本，且全部为英文样本。因此，本文利用主题爬虫收集大量含有表单的中文网页补充到 TEL-8 数据集中，得到扩展后的测试集 TEL-8(Ex)，如图 5 所示。

Domain	TEL-8 数据集			中文数据集		
	DS	SF	NSF	DS	SF	NSF
Airfares	47	49	0	195	206	383
Hotel	39	39	0	177	219	286
CarRental	25	25	0	16	54	22
Book	65	67	0	36	87	60
Movie	73	78	0	112	134	71
Music	65	70	0	60	126	93
Job	49	52	0	184	265	157
Automobile	84	97	0	98	118	141
	449	477	0	878	1209	1213

图 5　TEL-8(Ex)测试集的样本构成

实验中，RankFW 的初始权重赋值为：$\alpha_1=0.7$，$\alpha_2=0.6$，$\alpha_3=0.4$，$\alpha_4=0.3$，$\alpha_5=0.1$。图 6 列举了 RankFW 和 TF * IDF 在 Airfares 领域中得到的权重最大的 10 个特征。

Airfares								
RankFW					TF * IDF			
特征	权重	tf_k	$m_{c,k}$	$m_{\bar{c},k}$	特征	权重	tf_k	df_k
机票	0.372803	616	197	104	飞机票	970.1799	781	259
飞机票	0.361181	781	162	97	机票	700.8264	616	301
航空	0.346383	537	188	113	机场	612.8637	579	339
航线	0.302733	384	152	107	天空	610.9477	537	301
航班	0.298850	574	113	81	公司	520.6157	332	168
预订	0.291717	491	122	96	旅游	486.7451	235	92
查询	0.243359	493	93	85	酒店	396.9078	233	142
单程	0.232336	352	141	110	国际	363.7740	307	281
机型	0.214398	277	149	121	提供	342.6812	195	133
特价	0.189814	384	155	141	民航	334.4857	178	115

图 6　RankFW 和 TF * IDF 在 Airfares 的特征选择结果对比

RankFW 与 TF * IDF 相比，前者选择的 10 个特征更能体现 Airfares 领域的特点。从特征权重的降序排列来看，由于对不同位置出现的特征赋予了不同的初始权重，RankFW 所选的前几个特征都是出现在〈TITLE〉位置中，如“机票”、“飞机票”、“航空”等。查询接口本身所包含的描述文本特征也被 RankFW 赋予较大的权重，如“单程”、“机型”、“往返”和“转机”等。与查询接口相邻的部分文本也被选入特征，如“国际”、“机场”和“旅游”等。对于 TF * IDF 而言，由于只考虑了特征词频和文档频率，TF * IDF 所选择的特征高度依赖于特征在网页中出现的次数，导致了一些非领域相关的特征项被赋予了很高的权重，如“公司”、“酒店”和“提供”等。通过对样本的分析发现，这些特征通常出现在网页的正文内容中，与查询接口的位置相差 2 层以上，因此与领域的相关性并不大，只是特征词频较高，从而导致其获得了较高的权重。由此可见，使用 RankFW 选择的特征具有更高的领域相关性，减少了干扰特征出现的概率。需要说明的是，非类内样本对 RankFW 的影响是较为明显的，当出现特征的类内样本数小于类间样本数(即 $m_{c_i,k}<m_{\bar{c}_i,k}+1$)时，特征将获得负权重。权重为负的特征在排序过程中自然靠后，表明 RankFW 能够有效解决 TF * IDF忽略类间样本差异的问题。

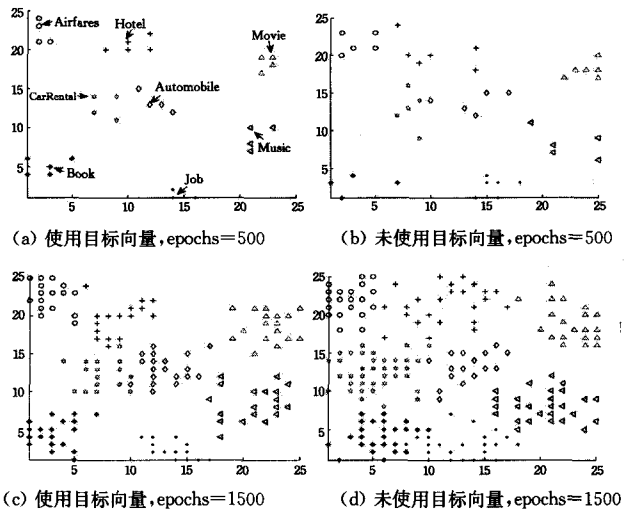
RankFW 和 TF * IDF 在 TEL-8(Ex)的其他 7 个领域上选择特征的情况类似，本文不再赘述，仅在图 7 中直接列举两种方法选择的前 10 个特征，其权重从左至右依次递减。

Hotel									
RFW	酒店	预订	宾馆	饭店	人	房	住	公寓	提供	预付
TF * IDF	酒店	订	预订	游	服务	信息	全国	手机	姓名	房
CarRental									
RFW	车	租车	汽车	租赁	包车	提供	出租	车型	轿车	品牌
TF * IDF	汽车	租车	车	公司	服务	商务	提供	包车	旅游	电话
Book									
RFW	图书	书	图书馆	查询	书名	作者	出版社	更新	搜索	简介
TR * IDF	书	研究	大学	搜索	开发	出版	技术	购买	国际	电子
Movie									
RFW	电影	电影票	票	票房	购买	分级	演员	年份	分类	点评
TF * IDF	动员	资料库	上映	榜	娱乐	类型	主演	查询	明显	分类
Music									
RFW	专卖	音乐	专辑	歌曲	唱片	明星	价格	奏鸣曲	外语	正版
TF * IDF	网	价格	音乐	歌曲	女声	四重奏	限量	新口	经典	舞
Job									
RFW	人才	招聘	求职	简历	工作	搜索	专业	兼职	应聘	岗位
TF * IDF	人	人才	信息	市场	提供	工作	人力	新闻	资讯	能力
Automobile									
RFW	车	汽车	报价	经销	品牌	座位	价格	车型	二手车	口碑
TF * IDF	汽车	车	经销	上牌	培训	企业	销售商	市场	动力	公司

图 7　TEL-8(Ex)特征选择结果对比

5.2　DR-QSOFM 的特征映射

本小节将在两种情况下对 TEL-8(Ex)数据集进行实验，第一种不使用目标向量，即单独使用样本的特征向量训练 DR-QSOFM 模型，模型具有 20-625 结构；第二种同时使用特征向量和目标向量作为模型输入，模型的结构为 28-625。两次实验中，DR-QSOFM 的参数设置为：初始学习速率 $\eta_0=1$，初始邻域半径 $r_0=2$，距离阈值 $\sigma=2$，竞争层神经元呈栅格型排列，样本的输入顺序一致。图 8 分别显示了两种实验中训练样本在竞争层的映射结果。如图 8(a)所示，当迭代次数为 500 次，DR-QSOFM 同时使用目标向量训练模型时，由于在训练初期，目标向量对权重和距离的调节起主要作用，因此初始获胜神经元以目标向量对应的获胜元为中心分布，位置相对集中，而图 8(b)中，初始获胜神经元的分布较散；当迭代次数达到 1500 次时，图 8(c)所示的各个簇区域划分仍然十分清晰，相对而言，在图 8(d)中，各个类别区域的边界出现大量交叉的情况，在 Airfares 与 Hotel，CarRental 与 Automobile，Job 和 Book 的边界处尤为明显。簇区域交叉、边界模糊的情况将直接影响到 Deep Web 查询接口主题分类的精度。由实验结果可知，将目标向量用于 DR-QSOFM 的训练有助于提高 Deep Web 数据源主题分类的精度。



(a) 使用目标向量，epochs=500　(b) 未使用目标向量，epochs=500

(c) 使用目标向量，epochs=1500　(d) 未使用目标向量，epochs=1500

图 8　目标向量对竞争层的影响

5.3 距离阈值 σ 对分类误差的影响

对 Deep Web 查询接口主题进行分类时，如果测试样本取得的获胜神经元并不在获胜神经元集合中，那么 DR-QSOFM_Classification 算法将根据距离阈值 σ 来计算与该测试样本获胜神经元距离最小的神经元编号。从上一个实验的结果来看，距离阈值 σ 的选择对于分类结果具有一定的影响，尤其当获胜神经元远离类别簇中心时，σ 选择不当可能会造成从其他类别簇中误选最近的获胜神经元，从而导致分类结果错误，影响精度。本实验使用 MSE 作为误差函数，其中 t_k 表示样本的实际领域类别，a_k 表示由算法 1 输出的分类结果，目标误差设为 0.01。实验分别在 $\sigma=1$，$\sigma=2$，$\sigma=3$ 三种情况下进行，图 9 显示了相应的误差曲线。从图中可以看出，当 $\sigma=1$ 时，在训练初期，误差迅速下降，此时各个类别簇之间的边界十分明显，分类精度并未受到影响，迭代次数超过 500 次时，误差减小的速率变缓，各个类别的获胜神经元开始出现交叉，最终在迭代次数达到 1421 次时误差减小到目标值，算法收敛。$\sigma=3$ 对应的曲线表明，在迭代过程中，由于距离阈值过大，造成在训练的过程中获胜神经元的误选，因此误差曲线比较粗糙，最终 MSE 误差为 0.0182766，没能达到目标误差值，算法未收敛。当 $\sigma=2$ 时，误差曲线相对平滑，并且在第 961 次迭代时达到了目标误差值。以上结果表明，距离阈值的理想取值为 2，且算法收敛的概率较高。

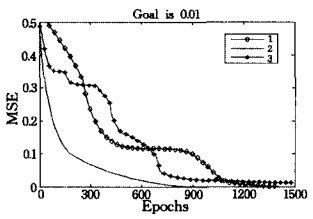


图 9 σ 不同取值的误差曲线

5.4 DR-QSOFM 的分类性能

本实验目的在于检验 DR-QSOFM 的分类能力，采用召回率和精度作为性能评估标准。模型网络结构为 28-625，竞争层神经元采用栅格排列，初始学习速率为 1，初始邻域半径为 3，距离阈值为 2。从 TEL-8(Ex)8 个领域中分别随机抽出三分之一(共 562 个)的的 SF 样本作为测试样本，其余三分之二(共 1124 个)的 SF 样本作为训练样本。实验结果与 SOFM，SVM 进行比较，SOFM 采用和 DR-QSOFM 相同的网络结构，SVM 采用默认设置。实验进行 3 次，取 3 种方法所得的分类精度和召回率的平均值作为最终结果，如表 1 所列。

表 1 DR-QSOFM 与 SOFM，SVM 的分类性能对比

	DR-QSOFM		SOFM		SVM	
Domain	R(%)	P(%)	R(%)	P(%)	R(%)	P(%)
Airfares	95.33	96.63	90.20	88.33	92.13	92.33
Hotel	94.41	96.07	91.13	89.24	89.46	87.94
CarRent	92.67	91.57	88.37	86.73	90.61	87.43
Book	96.52	95.18	89.52	92.84	91.59	90.50
Movie	91.04	92.40	87.44	89.17	88.32	89.25
Music	93.83	90.92	90.12	87.53	89.69	86.13
Job	96.56	97.63	92.45	89.20	90.16	86.63
Auto	91.39	92.43	87.51	88.46	88.42	89.71
Avg	93.97	94.10	89.59	88.94	90.05	88.74

使用 DR-QSOFM 对 8 个领域的样本分类时，平均查全率为 93.97%，平均精度为 94.10%，其中在 Airfares，Book 和 Job 领域的查全率、精度都在 95%以上，在 CarRental，Music 领域相对较低，但也达到 90%以上。分析样本可知，CarRental 和 Automobile 领域的特征空间重合区域较多，因此对分类结果产生了一定的影响。而 Music 领域的样本通常源于综合性信息网站，这些网站集图书、音像制品、玩具和家电等商品信息于一体，查询表单中包含了一些非 Music 领域的特征，从而导致了分类精度下降。与 SOFM 相比，DR-QSOFM 的平均查全率和平均精度分别提高了 4.38%，5.16%；与 SVM 相比，两值分别提高了 3.92%，5.36%。以上结果表明，DR-QSOFM 综合选择 5 种不同位置的特征和样本的领域知识参与 Deep Web 查询接口主题的分类过程，对分类精度和召回率产生了积极的影响，与 SOFM 和 SVM 相比其分类效果得到一定的提高。

结束语 实验结果表明，提出的 RankFW 权重计算方法综合考虑了不同位置的文本在 Deep Web 接口主题分类中的不同影响，所选文本特征更加符合领域特点；DR-QSOFM 在适当的距离阈值下能够使特征在竞争层上的映射更加集中，簇边界更加明显。在 TEL-8(Ex)数据集上的实验表明，DR-QSOFM 与 SOFM，SVM 相比在查全率和精度上具有一定的优势。

参 考 文 献

[1] 申德荣，聂铁铮，余恩运，等. 支持 Web 深层数据库网格的部分关键技术的研究[J]. 计算机科学，2007，34(8)：123-125

[2] 赵朋朋，崔志明，高岭，等. 关于中国 Deep Web 的规模、分布和结构[J]. 小型微型计算机系统，2007，10：1799-1802

[3] 马军，宋玲，韩晓晖，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

[4] 黄健斌. 基于条件概率图模型的 Deep Web 数据抽取与集成研究[D]. 西安：西安电子科技大学，2007

[5] Gao Ling，Zhao peng-peng，Cui Zhi-ming. Automatic Judgement of Deep Web Query Interfaces[J]. Compute technology and development，2007，17(15)：148-151

[6] Xu He-xiang，Zhang Cheng-hong，Hao Xiu-lan，et al. A Machine Learning Approach Classification of Deep Web Sources[C]// Fourth International Conference on Fuzzy Systems and Knowledge Discovery. 2007，4：561-565

[7] Xu He-xiang，Hao Xiu-lan，Wang Shu-yun，et al. A Method of Deep Web Classfication[C]// Proceedings of the Sixth International Conference on Machine Learning and Cybernetics. 2007：4009-4114

[8] Lin Pei-guang，Du Yi-bing，Tan Xiao-hua，et al. Research on Automatic Classification for Deep Web Query Interfaces[C]// International Symposiums on Information Processing. 2008：313-317

[9] Li Zhi-tao，Liu Quan，Cui Zhi-ming，et al. A Method to Automatically Discover and Classify Deep Web Data Source Using Multi-Classifier[C]// 2009 World Congress on Compter Science and Information Engineering. Vol. 3，2009：736-740

[10] ICTCLAS Org[EB/OL]. http：//ictclas. org，2010-6-29

[11] Sebastiani F. Machine learning in automated text categorization [J]. ACM Computing Surveys，2002，34(1)：1-47