大数据时代电子政务中 XML 文档相似性

赵 震[1,2],任永昌[1]
(1. 渤海大学 信息科学与技术学院,辽宁 锦州 121013;
2. 东北大学 计算机科学与工程学院,辽宁 沈阳 110819)

摘 要:XML 作为电子政务应用中的数据交换标准已经被广泛研究。随着大数据时代的到来,对电子政务中 XML 数据的管理也显得越来越重要。在 XML 数据的管理中,XML 文档的相似性是 XML 数据集成、XML 数据分类的关键。为了研究 XML 文档的相似性,针对 XML 文档进行了树形变换,并提取树节点的相应特征,然后分别利用这些特征对节点进行相应的相似性计算,再将得到的相似性利用 ELM(超限学习机)算法进行拟合得到最终的节点相似性。在节点相似性的基础上提出了 XML 文档树的相似性比较算法,从而计算得到 XML 文档的相似性。实验部分在给出具体的评估指标的基础上,在两个不同的数据集上给出使用文中方法所得到的精确度、召回率、*F* -measure 值以及相应时间的对比情况,通过实验验证了所提方法的性能优势。

关键词:XML 文档;相似性;特征提取;拟合;数据集成
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2017)01-0186-04
doi:10.3969/j.issn.1673-629X.2017.01.042

Similarity of XML Documents in E-government in Era of Big Data

ZHAO Zhen[1,2], REN Yong-chang[1]
(1. College of Information Science and Technology, Bohai University, Jinzhou 121013, China;
2. School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China)

Abstract: XML has been widely studied as the standard of data exchange in e-government applications. With the arrival of the era of big data, the management of XML data in e-government is also becoming more and more important. In the management of XML data, the similarity of XML documents is the key of XML data integration and XML data classification. In order to study the XML document similarity, the XML document are transformed into tree, extracting the corresponding characteristics of the nodes of the tree, and then using these characteristics to calculate the similarity of nodes, and then the final node similarity can be obtained by the ELM(Extreme Learning Machine) algorithm. Based on the similarity of nodes, the algorithm of similarity comparison of the XML document tree is given, which can obtain the similarity of XML documents. Based on the given specific evaluation indexes, the accuracy, recall, *F* -measure values and the corresponding time are obtained through experiments in two different data sets using the method proposed. The performance advantages of the proposed method are verified by experiments.

Key words: XML documents; similarity; feature extracting; synthesizing; data integration

0 引 言

近年来,随着电子政务的快速发展,XML 作为电子政务应用中的数据交换标准[1]越来越受到重视。众多学者在此基础上提出了许多基于 XML 的电子政务服务模型[2-4]。随着大数据时代的到来,对电子政务中 XML 数据的管理也显得越来越重要。XML 数据的管理包括数据的存储和集成、数据的交换等。在 XML 数据的管理中,XML 数据的相似性是 XML 数据集成[5]、分类[6]的关键。由于各个部门 XML 的数据源是独立构建的,不同部门应用中的 XML 数据结构是有差异的,首先要对这些数据进行识别,找出它们之间的相似性后再进行数据集成或分类。文中工作有利于

收稿日期:2016-03-28 **修回日期**:2016-07-05 **网络出版时间**:2017-01-04
基金项目:教育部人文社会科学研究青年基金项目(15YJC870028);辽宁省自然科学基金(2015020009);辽宁省哲学社会科学规划基金项目(L15BTQ002);辽宁省社科联 2015 年度辽宁经济社会发展立项课题(2015lslktglx-01)
作者简介:赵 震(1977-),男,博士研究生,讲师,CCF 会员,研究方向为人工智能与语义 Web;任永昌,博士,教授,研究方向为云计算与软件项目管理。

解决政府各部门各类应用间的信息孤岛问题,对实现部门间协同工作十分重要。

XML 数据管理问题是以往各国学者研究的热点[7-10]。提出了一些经典方法,对于解决 XML 数据管理问题十分重要。在 XML 文档的相似性研究中,XML 文档可以表示为树,两者的相似性问题可以转化为两棵树的匹配问题,目前的解决方案主要有:将需要进行匹配的 XML 文档转化为树,利用基于树编辑距离的算法计算文档树的相似性[7-8];借助邻接矩阵来计算对应 XML 文档的相似性[9-10]。

文中在节点相似性的基础上提出了 XML 文档树的相似性比较算法,从而计算得到 XML 文档的相似性,并进行了实验验证。

1　XML 文档及树形表示

XML 作为可扩展标记语言,以半结构化的方式描述各种类型的数据。XML 文档中允许使用自定义的标签来更准确地描述数据。下面给出一个 XML 文档片段,如图 1 所示。

```
<collegeCname = "NEU">
<departmentDname = "IST">
<teacher TID = "007102">
<tname> George Frank</tname>
<position>professor</position>
</teacher>
<student SID = "20130425">
<age>26</age>
<email>John@ yahoo. com</email>
</student>
</department>
</college>
```

图 1　XML 文档实例

XML 文档可以用树形结构表示。按照文档对象模型(DOM),一个 XML 文档也可以表示为一个单根的有序标签树,其中的节点对应文档中的元素和属性。文中只比较树的结构相似性,所以省略元素和属性的值。图 1 中文档片段对应的树结构如图 2 所示。

2　树节点的特征相似性

对于 XML 文档树,树节点是最基本的数据项。一个节点可以是 XML 文档中的元素或属性。用 $\mathrm{Sim}_{\mathrm{Node}}(N_1, N_2)$ 表示来自不同文档树节点 N_1 和 N_2 的相似度。

可以充分利用节点的特征来更精确地获得节点的相似性。标签名、节点深度、数据类型是最常见的用于计算节点相似度的特征。也就是说,利用节点的这些特征值计算得到来自不同文档树节点的相似性。根据不同的特征,可以得到不同的相似度。

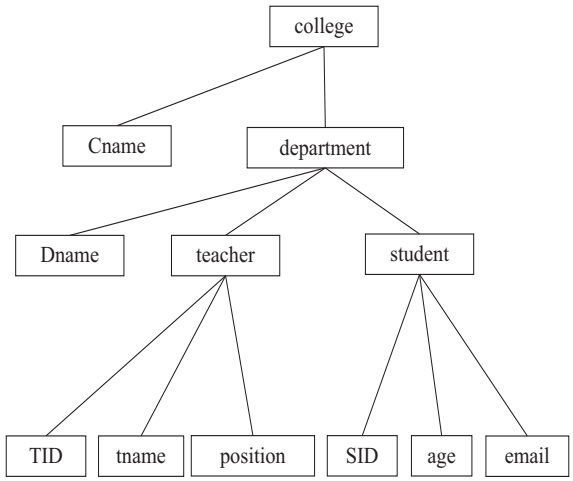

图 2　XML 文档树实例

(1)标签相似性度量。

标签名(Label)是最重要的节点特征。利用字符串匹配来计算标签相似度。当然字符串匹配的方法有很多,这里采用文献[11]中的方法来计算字符串的相似性。那么,节点的相似性可由式(1)得到:

$$\mathrm{Sim}_{\mathrm{Label}}(N_1, N_2) = \frac{1}{1 + \mathrm{editDistance}(L_1, L_2)} \quad (1)$$

其中, $\mathrm{editDistance}(L_1, L_2)$ 是字符串 L_1 转换为 L_2 所需要编辑字符的最小代价。

(2)深度相似性度量。

只用节点标签来度量节点相似性是完全不够的,节点的深度是另外一个重要的考量节点相似性的特征。深度相似性需要考虑节点和它们最近共同祖先节点的深度。那么两个节点的相似性可由式(2)得到:

$$\mathrm{Sim}_{\mathrm{Depth}}(N_1, N_2) = \frac{d_{01} + d_{02}}{d_1 + d_2} \quad (2)$$

其中, d_1 和 d_2 分别是节点 N_1 和 N_2 在相应文档树中的深度; d_{01} 和 d_{02} 分别是 N_1 和 N_2 最近共同祖先 N_0 在相应文档树中的深度。

(3)数据类型相似性度量。

节点的数据类型是另一个用来确定节点相似性的特征。具有相同数据类型的节点具有更大的相似性($\mathrm{Sim}_{\mathrm{DataType}}$)。表 1 说明了不同数据类型节点相似性度量值。

表 1　数据类型相似性列表

Type1	Type2	$\mathrm{Sim}_{\mathrm{DataType}}$
#PCDATA	#PCDATA	1.0
CDATA	CDATA	1.0
#PCDATA	CDATA	0.6
CDATA	NMTOKEN	0.7
ID	IDREF	0.8

还有很多用于度量节点相似性的特征,用这些特征计算得到节点特征相似性 $S_1,S_2,\cdots,S_N$。但是每一个单一的特征得来的相似性都不足以表示节点的相似性,因此,有必要将这些相似性拟合在一起,从整体上来考虑这些特征,以得到更合理的节点相似性。一般采用权重的方法得到最终的相似性[12-13],但是这种方法得到的结果误差较大。于是利用基于超限学习机的方法得到拟合的节点相似性。

3 超限学习机

超限学习机[14-15]是由黄广斌教授提出的单隐层前馈神经网络。超限学习机的最大优点是提供了非常快的学习速度,其隐藏层的权重和偏移值可以随机指定,并且输出权重可以通过矩阵计算而无需人工调节。

考虑 N 个任意样本 $(\boldsymbol{x}_i,\boldsymbol{t}_i)\in R^{n\times m}$,那么 ELM 可表示为:

$$\sum_{i=1}^{L}\boldsymbol{\beta}_i g(\boldsymbol{W}_i\cdot\boldsymbol{x}_j+\boldsymbol{b}_i)=o_j,j=1,2,\cdots,N \tag{3}$$

其中,L 为隐藏层节点数目;$g()$ 为激活函数;$\boldsymbol{W}_i$ 为输入权重向量;$\boldsymbol{\beta}_i$ 为输出权重向量;$\boldsymbol{b}_i$ 为第 i 个隐藏节点的偏移量。

学习目的是为了达到最小的训练错误,即 $\sum_{j=1}^{L}\|\boldsymbol{o}_j-\boldsymbol{t}_j\|=0$,$\boldsymbol{o}_j$ 是实际输出值。

则存在 $\boldsymbol{W}_i,\boldsymbol{\beta}_i,\boldsymbol{b}_i$,使得

$$\sum_{i=1}^{L}\boldsymbol{\beta}_i g(\boldsymbol{W}_i\cdot\boldsymbol{x}_j+\boldsymbol{b}_i)=t_j,j=1,2,\cdots,N \tag{4}$$

上面的等式可表示为:

$$\boldsymbol{H\beta}=\boldsymbol{T} \tag{5}$$

其中

$$\boldsymbol{H}=\begin{bmatrix}\boldsymbol{h}(\boldsymbol{x}_1)\\ \vdots\\ \boldsymbol{h}(\boldsymbol{x}_n)\end{bmatrix}=\begin{bmatrix}g(\boldsymbol{W}_1\cdot\boldsymbol{x}_1+\boldsymbol{b}_1) & \cdots & g(\boldsymbol{W}_L\cdot\boldsymbol{x}_1+\boldsymbol{b}_L)\\ \vdots & \cdots & \vdots\\ g(\boldsymbol{W}_1\cdot\boldsymbol{x}_n+\boldsymbol{b}_1) & \cdots & g(\boldsymbol{W}_L\cdot\boldsymbol{x}_n+\boldsymbol{b}_L)\end{bmatrix}_{n\times L}$$

$$\boldsymbol{\beta}=[\boldsymbol{\beta}_1^{\mathrm{T}},\cdots,\boldsymbol{\beta}_L^{\mathrm{T}}]_{m\times L}$$

$$\boldsymbol{T}=[\boldsymbol{t}_1^{\mathrm{T}},\cdots,\boldsymbol{t}_L^{\mathrm{T}}]_{m\times L}$$

问题简化为求解线性系统的最小二乘解。则输出权重 $\boldsymbol{\beta}$ 为:

$$\boldsymbol{\beta}=\boldsymbol{H}^{\dagger}T \tag{6}$$

其中,$\boldsymbol{H}^{\dagger}=(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})^{-1}\boldsymbol{H}^{\mathrm{T}}$ 是 $\boldsymbol{H}$ 的伪逆矩阵。

计算得到输出权重 $\boldsymbol{\beta}$ 后,利用它得到:

$$o_t=\boldsymbol{\beta}h(x_t) \tag{7}$$

ELM 算法描述如下:

算法 1:SimNode。

输入:训练集 $D=\{(x_t,y_t)\},t=1,2,\cdots,T$,激活函数 $g(x)$;隐藏节点数 L;(where $L\leqslant T$);

输出:$\boldsymbol{\beta}$。

Begin

步骤 1:随机指定输入权重 $\boldsymbol{W}_i$ 和偏移量 $\boldsymbol{b}_i$;

步骤 2:计算 $\boldsymbol{H}$;

步骤 3:计算 $\boldsymbol{\beta}=\boldsymbol{H}^{\dagger}\boldsymbol{T}$。

Return $\boldsymbol{\beta}$

End

4 文档树的相似性计算

4.1 树节点的相似性

为了得到文档树的相似性,首先要获得文档树中节点的相似度。前文介绍了依据节点特征得到的特征相似性,这一节介绍如何利用超限学习机得到拟合的节点相似性。

用超限学习机拟合节点的相似性如图 3 所示。其中,$S_1,S_2,\cdots,S_n$ 是根据节点特征得到的相互独立的相似度量值;S 是经过 ELM 拟合得到的最终节点相似度。

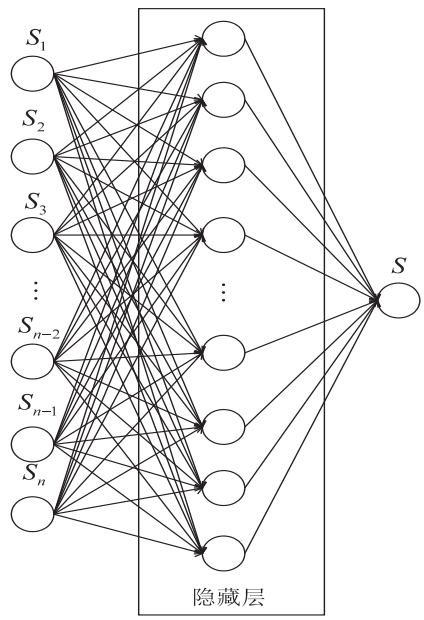

图 3 基于 ELM 的相似性拟合

拟合过程分为训练阶段和预测阶段。这一拟合模型目的是利用训练样本在输入变量($S_1,S_2,\cdots,S_n$)和输出变量(S)间建立一种映射关系。首先随机选择不同文档树中的节点作为训练样本,然后分别计算节点对的特征相似值 $S_1,S_2,\cdots,S_n$,再通过专家确定这些样本节点的最终相似性 S,最后,通过超限学习机算法快速建立预测模型。算法描述如下:

算法 2:SimNode。

输入:$\mathrm{Node}_1,\mathrm{Node}_2$;

输出: Sim_{Node}。

Begin

步骤 1:分别计算特征相似度 $S_1, S_2, \cdots, S_n$;

步骤 2:计算节点相似度 $Sim_{Node} = \boldsymbol{\beta H}, \boldsymbol{\beta}$ 由算法 1 得到。

Return Sim_{Node}

End

4.2　文档树的相似性

给定文档树 D_1 和 D_2,计算文档树的相似性。需要得到节点相似性大于给定阈值(θ)的节点数目。用这一数值与全部节点数目的比值来衡量文档中相似节点所占的比重,据此得出文档的相似性。算法 3 给出了计算文档树的相似性的具体算法。

算法 3:SimDocument。

输入: D_1, D_2;

输出:$Sim_{Document}$。

Begin

步骤 1:遍历 D_1, D_2 中每个节点, $node_i \in D_1$, $node_j \in D_2$;

步骤 2:计算每个节点对的相似度 Sim_{Node}($node_i$, $node_j$);

步骤 3:如果 Sim_{Node}($node_i$, $node_j$)两棵树中相似节点对相似度大于阈值 θ,则相似节点数目 NumSimNode = NumSimNode+1;

步骤 4: $Sim_{Dcoument} = \dfrac{NumSimNode}{Min(|D_1|, |D_2|)}$。

Return $Sim_{Document}$

End

5　实　验

下面通过实验进一步评估文中提出的 XML 文档相似性计算方法的性能。评估相似性比较的性能主要考虑两方面:有效性和效率。

评估有效性主要有两个指标:精确度和召回率。下面简单介绍它们的定义。

精确度表示正确匹配的程度,召回率表示匹配的完整性,分别为:

$$P = \frac{A}{A + B} \tag{8}$$

$$R = \frac{A}{A + C} \tag{9}$$

其中, A 为正确匹配的 XML 文档数量; B 为错误匹配的 XML 文档数量; C 为没有被识别出的正确匹配的 XML 文档数量。

两者的调和平均值可以用 F – measure 来表示。

$$F - measure = \frac{2 \times P \times R}{P + R} \tag{10}$$

为保证数据的真实性,选用的数据集为 DBLP 和 SigmodRecord。同时,需要将数据集分割为 0.1 M 到 2 M 的数据,以便对比算法响应时间。

图 4 显示了在 DBLP 和 SigmodRecord 数据集上使用文中方法所得到的精确度、召回率、F – measure 值的对比情况。

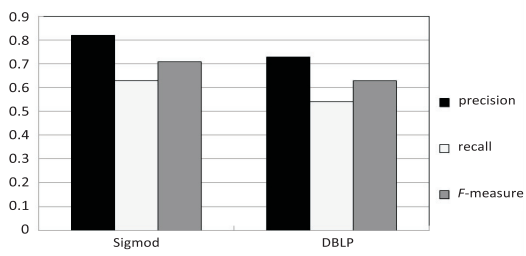

图 4　DBLP 和 SigmodRecord 数据集匹配有效性对比

从图中可以看出,SigmodRecord 数据集上的有效性要优于 DBLP 数据集,这是因为 DBLP 数据集的结构比 SigmodRecord 复杂。

图 5 显示了在 DBLP 和 SigmodRecord 数据集上执行文中算法所得到的响应时间的对比情况。

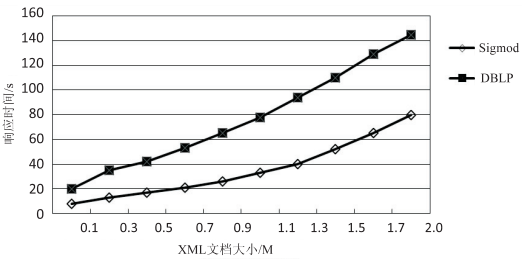

图 5　DBLP 和 SigmodRecord 数据集响应时间对比

从图中可以看出,SigmodRecord 数据集上的响应时间远小于 DBLP 数据集,由此可以看出 DBLP 数据集结构比较复杂。

6　结束语

在大数据的背景下,研究了电子政务中 XML 数据的相似性。首先将 XML 文档转换为对应的 XML 文档树,然后根据抽取的 XML 树节点的特征,计算对应的特征相似性,再使用基于 ELM 的算法得到 XML 节点的相似性,并给出了 XML 文档树的相似性比较算法,从而得到 XML 文档的相似性。通过实验验证了所提方法的正确性和有效性。

参考文献:

[1]　赵慧勤,赵慧玲. 电子政务数据交换标准—XML 语言[J]. 山西大同大学学报:社会科学版,2003,17(3):76-78.

[2]　钟福金,辜丽川,张友华. 基于语义 Web 服务的电子政务模型研究[J]. 微电子学与计算机,2010,27(3):144-147.

[3]　陈　桦,麻风梅,韩艳艳. 基于 XML 的异构数据集成模式

(下转第 194 页)

知:当压缩比取 0.50 时相对于仅进行 QR 优化和仅进行梯度下降优化的观测矩阵,采取所提优化方法用于图像重构的峰值信噪比分别提高了 1.5 dB 和 3.1 dB,而相对未优化的观测矩阵,其提升幅度更大;尤其当压缩比较小时,提出的优化方法在信号重构方面具有更加明显的提升效果。另外,该方法在稳定性方面也有较大优势。文中的研究工作还有许多待改进的地方,例如进一步减小观测矩阵与稀疏矩阵间的相关性,采用更少的观测数目以及减少重构时间得到更精确的重构效果等。

参考文献:

[1] Donoho D. Compressed sensing[J]. IEEE Transactions on Information Theory,2006,52(4):1289-1306.

[2] Candes E. Compressive sampling[C]//Proceedings of the international congress of mathematicians. Madrid,Spin:[s. n.],2006:1433-1452.

[3] Candes E J,Tao T. Near optimal signal recovery form random projections:universal encoding strategies? [J]. IEEE Transactions on Information Theory,2006,52(12):5406-5425.

[4] Baraniuk R G. A lecture on compressive sensing[J]. IEEE Signal Processing Magazine,2007,24(4):118-121.

[5] 徐 静,王彩云. 压缩感知测量矩阵优化混合方法[J]. 深圳大学学报:理工版,2014,31(1):58-62.

[6] 傅迎华. 可压缩传感重构算法与近似 QR 分解[J]. 计算机应用,2008,28(9):2300-2302.

[7] 彭玉楼,何怡刚,林 斌. 基于奇异值分解的压缩感知噪声信号重构算法[J]. 仪器仪表学报,2012,33(12):2655-2660.

[8] 郑 晓,薄 华,孙 强. QR 分解与特征值优化观测矩阵的算法研究[J]. 智能系统学报,2015,10(1):149-155.

[9] Donoho D L,Stark P B. Uncertainty principles and signal recovery[J]. SIAM Journal on Mathematical Analysis,1989,49(3):906-931.

[10] 赵瑞珍,秦 周,胡绍海. 一种特征值分解的测量矩阵优化方法[J]. 信号处理,2012,28(5):653-658.

[11] Duarte-Cavajalino J M,Sapiro G. Learning to sense sparse signals:simultaneous sensing matrix and sparsifying dictionary optimization[J]. IEEE Transactions on Image Processing,2009,18(7):1395-1408.

[12] Nhat V D M,Vo D,Challa S,et al. Efficient projection for compressed sensing[C]//Proceedings of the computer and information science. [s. l.]:IEEE,2008:322-327.

[13] Elad M. Optimized projections for compressed sensing[J]. IEEE Transactions on Signal Processing,2008,55(12):5695-5702.

[14] Abolghasemin V,Ferdowsi S,Makkiabadi B,et al. A robust approach for optimization of the measurement matrix in compressed sensing[C]//International workshop on cognitive information processing. Elba Island:IEEE Press,2010:388-392.

[15] Abolghasemin V,Ferdowsi S,Makkiabadi B,et al. On optimization of the measurement matrix for compressive sensing [C]//Signal processing conference. [s. l.]:IEEE,2010:427-431.

(上接第 189 页)

的研究[J]. 微电子学与计算机,2009,26(1):137-139.

[4] 李冬睿. 基于 XML 与 Web Service 的电子政务数据交换模型的设计与实现[D]. 桂林:广西师范大学,2008.

[5] Thomo A,Venkatesh S. Rewriting of visibly pushdown languages for xml data integration[C]//Proceedings of the 17th ACM conference on information and knowledge management. Napa Valley,California,USA:ACM,2008:521-530.

[6] Algergawy A,Mesiti M,Nayak R,et al. XML data clustering:an overview[J]. ACM Computing Surveys,2011,43(4):25-41.

[7] Nierman A,Jagadish H V. Evaluating structural similarity in XML documents[C]//Proceedings of the ACM SIGMOD international workshop on the web and databases. [s. l.]:ACM,2002:61-66.

[8] Tekli J,Chbeir R. A novel XML document structure comparison framework based-on sub-tree commonalities and label semantics[J]. Journal of Web Semantics,2012,11(3):14-40.

[9] Zhang X,Yang T,Fan B Q,et al. A novel method for measuring structure and semantic similarity of XML documents based on extended adjacency matrix[C]//Proceedings of international conference on service science. [s. l.]:[s. n.],2012:1452-1461.

[10] Chowdhury I J,Nayak R. A novel method for finding similarities between unordered trees using matrix data model[M]. Berlin:Springer,2013:421-430.

[11] Lin Dekang. An information-theoretic definition of similarity [C]//Proceedings of the international conference on machine learning. Madison,Wisconsin,USA:[s. n.],1998:296-304.

[12] Algergawy A,Nayak R,Saake G. Element similarity measures in XML schema matching[J]. Information Sciences,2010,180(24):4975-4998.

[13] Tekli J,Chbeir R. Minimizing user effort in XML grammar matching[J]. Information Sciences,2012,210(10):1-40.

[14] Huang Guangbin,Zhu Qinyu,Siew C K. Extreme learning machine:theory and applications[J]. Neurocomputing,2006,70(1-3):489-501.

[15] Huang Guangbin. An insight into extreme learning machines:random neurons,random features and kernels[J]. Cognitive Computation,2014,6(3):376-390.