Journal of Computer Applications
计算机应用,2013,33(3):776-779
ISSN 1001-9081
CODEN JYIIDU
2013-03-01
http://www.joca.cn

文章编号:1001-9081(2013)03-0776-04
doi:10.3724/SP.J.1087.2013.00776

基于 BP 神经网络的 Deep Web 实体识别方法

徐红艳,党晓婉,冯　勇*,李军平
(辽宁大学 信息学院,沈阳 110036)
(*通信作者电子邮箱 fyxuhy@163.com)

摘　要:针对现有实体识别方法自动化水平不高、适应性差等不足,提出一种基于反向传播(BP)神经网络的 Deep Web 实体识别方法。该方法将实体分块后利用反向传播神经网络的自主学习特性,将语义块相似度值作为反向传播神经网络的输入,通过训练得到正确的实体识别模型,从而实现对异构数据源的自动化实体识别。实验结果表明,所提方法的应用不仅能够减少实体识别中的人工干预,而且能够提高实体识别的效率和准确率。

关键词:Deep Web;反向传播神经网络;实体识别;相似度;语义块
中图分类号:TP311.1;TP391.1　**文献标志码**:A

Method of Deep Web entities identification based on BP neural network

XU Hongyan, DANG Xiaowan, FENG Yong*, LI Junping
(*School of Information, Liaoning University, Shenyang Liaoning* 110036, *China*)

Abstract: To solve the problems such as low level automation and poor adaptability of current entity recognition methods, a Deep Web entity recognition method based on Back Propagation (BP) neural network was proposed in this paper. The method divided the entities into blocks first, then used the similarity of semantic blocks as the input of BP neural network, lastly obtained a correct entity recognition model by training which was based on the autonomic learning ability of BP neural network. It can achieve entity recognition automation in heterogeneous data sources. The experimental results show that the application of the method can not only reduce manual interventions, but also improve the efficiency and the accuracy rate of entity recognition.

Key words: Deep Web; Back Propagation (BP) neural network; entities identification; similarity; semantic block

0　引言

随着 Web 技术的发展及其应用的普及,互联网上的信息资源激增,目前可访问的 Web 数据源数量已经超过 250 万个[1]。Web 数据源按其蕴含信息的深度可分为 Surface Web 和 Deep Web[2],通常将传统搜索引擎无法检索到的 Web 数据源称为 Deep Web。在 Deep Web 相关研究中,实体识别是开展模式匹配、数据集成等工作的首要前提[3],因此如何准确、高效地识别出 Deep Web 中的相同实体已经成为 Deep Web 领域的研究热点之一。

在 Deep Web 的实体识别领域,相关学者已开展了较为深入的研究,取得了一些具有代表性的研究成果:Chaudhuri 等[4]提出了一种高效的重复记录模糊检测算法,通过采用特定索引、排序等优化措施有效地搜索出与当前元组最相似的 K 个关联元组进行实体识别;崔晓军等[5]提出了基于距离的自适应 Web 记录匹配方法,该方法提高了实体识别效率;Shen 等[6]提出了一种组合策略,该策略根据现有的训练样本自动选择合适的相似度计算方法提高了实体匹配的准确性。但现有实体识别方法仍普遍存在需要人工干预、对异构 Web 数据源适应性差等不足。

针对现有实体识别方法存在的不足,本文运用反向传播(Back Propagation,BP)神经网络技术对 Deep Web 实体识别开展研究。首先对 BP 神经网络理论进行概述;然后提出了一种基于 BP 神经网络的 Deep Web 实体识别方法,该方法包括三个步骤:实体分块、语义块相似度计算和实体识别模型训练;最后通过实验验证了所给方法切实可行。

1　BP 神经网络概述

人工神经网络(Artificial Neural Network,ANN)是一种模仿生物神经网络的结构和功能、建立在自主学习基础之上的数学模型,通过对大量复杂的数据进行分析,可以完成极为复杂的模式抽取或趋势分析[7]。它的基本功能包括:1)联想记忆功能;2)分类和识别功能;3)优化计算功能;4)非线性映射功能。BP 神经网络属于有监督的学习方法,是一种由非线性变换单元组成的前馈式全连接多层神经网络。一般情况下,BP 神经网络由输入层、隐藏层和输出层组成[7]。

BP 神经网络的学习过程分为两个阶段:信息正向传播阶段和误差反向传播阶段。在信息正向传播阶段,输入样本从输入层进入,经隐藏层处理后传到输出层。若输出层的实际输出与期望输出有较大误差时,网络进入误差反向传播阶段。在误差反向传播阶段主要是进行各神经元阈值和连接权值的修正。反复地进行信息正向传播和误差反向传播,当误差值

收稿日期:2012-09-27;**修回日期**:2012-11-26。　**基金项目**:教育部人文社会科学研究青年基金资助项目(12YJCZH048);辽宁省自然科学基金资助项目(20102083);辽宁"百千万人才工程"培养经费资助项目。

作者简介:徐红艳(1972-),女,辽宁丹东人,副教授,主要研究方向:Web 挖掘、数据管理;　党晓婉(1986-),女,河南洛阳人,硕士研究生,主要研究方向:Web 挖掘、数据管理;　冯勇(1973-),男,辽宁沈阳人,副教授,博士,主要研究方向:社会网络分析、信息管理;　李军平(1987-),女,湖南邵阳人,硕士研究生,主要研究方向:社会网络分析、信息管理。

达到要求,学习过程结束,得到正确的 BP 神经网络模型。

2　基于 BP 神经网络的 Deep Web 实体识别方法

在异构 Web 数据源中,相同实体的表现形式往往不同。Deep Web 实体识别能够从不同数据源中检测出相同实体,为消除数据冗余、相同实体比较等工作奠定基础。将 BP 神经网络融入到 Deep Web 实体识别研究中,可以充分利用神经网络的自主学习特性,减少人工干预、提高识别的准确性。基于 BP 神经网络的 Deep Web 实体识别方法的思想是:首先计算样本集中相同实体各语义块的相似度值,将实体各语义块的相似度值作为输入进行正向信息传播,再通过与期望输出比较,将得到的误差进行反向传播,重复 BP 神经网络学习过程得到正确的实体识别模型,利用此模型可以对任意两个 Deep Web 实体进行识别。下面对该方法的三个主要步骤:实体分块、语义块相似度计算、实体识别模型训练进行详细介绍。

2.1　实体分块

实体分块是对实体按照某个字段或某些字段组合进行分割,具有同一含义的字段形成一个语义块。通过对实体分析可以发现,实体中不仅包含内容信息,同时也包含对实体识别起着干扰作用的元数据信息。这里以图书领域为例,如图 1 中的“作者”“价格”“ISBN”等标签属于元数据信息,这些信息无助于实体识别,反而由于它们的存在影响实体间相似度的计算,因此在进行实体识别前应先将这些元数据信息去除。

在进行训练实体识别模型前,需要对实体的内容信息预处理。首先将内容信息分块出来,如图 1 中矩形框部分,然后将划分出的语义块看作一个文本,将实体的语义块与其他实体进行比较,计算语义块相似度值作为 BP 神经网络的输入。

本文使用基于网页布局标签的语义结构划分方法[8] 对实体进行分块,主要是利用 table、div、span 等常见标签对实体分块,这样不仅提高了块划分的效率,而且提高了块划分的准确性。

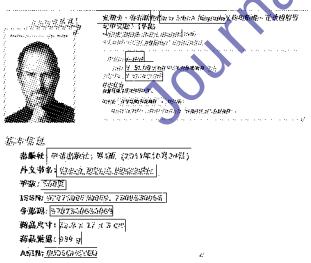

图 1　实体分块示例

在图书领域的 Deep Web 网站中,实体的各语义块的内容大致相似,只有极个别的语义块为某站点独有,例如基本上各网站的实体都包括“作者”“价格”“ISBN”等这些信息,而例如“商品尺寸”等语义块为个别网站的实体所独有。为此,先将所有学习样本实体划分语义块,分割后的语义块已知属于某一属性,统计这些语义块对应的属性值,求并集,形成一个包含所有实体语义块对应属性值的标识集合,记为 $A = \{A_1, A_2, \cdots, A_n\}$,$n$ 代表语义块对应属性值标识的个数。该集合包括了图书领域内所有实体的语义块对应的属性的标识,既有公有属性又有特殊属性,每个语义块可用属性值的标识表示。对数据源中任意一个实体进行语义块划分后,则该实体可以由一个语义块集合 $S = \{S_1, S_2, \cdots, S_n\}$ 表示,其中 S_i 表示语义块 i 的属性值,若在标识集合 A 中未找到对应属性值,则 S_i 为空。

2.2　语义块相似度计算

对实体进行块划分之后,由于每个语义块对实体识别的贡献度不同,需要对一个实体的各语义块与待匹配实体进行相似度计算,作为训练神经网络的输入量。即对实体 A 划分 n 个语义块,组成对应语义块集合 $S = \{S_1, S_2, \cdots, S_n\}$,分别将每个语义块与实体 B 进行相似度计算,得到一组相似度值,将它们表示为一组向量,记为 $\boldsymbol{T}$:

$$\boldsymbol{T} = (Sim(S_1, B), Sim(S_2, B), \cdots, Sim(S_n, B))$$

其中:$Sim(S_x, B)$ 表示实体 A 的任一语义块与实体 B 的相似度。

因为 HTML 文档的结构性较差,在进行相似度计算时如果将实体 B 进行内部语义块划分,找到与实体 A 相匹配的语义块后再进行语义块间的相似度计算则需要引入模式匹配的概念,这将增大计算的复杂度。所以,本文采用先将实体 A 进行语义块划分后,再计算其各语义块与实体 B 的相似度的策略。在计算语义块与一个实体的相似度时,根据语义块所属特征属性类型采用不同的计算方法:

1) 非字符串型。对于非字符串类型的属性,如数值型、货币型,可以通过范围距离算法计算它们之间的相似度。将实体 A 的语义块记为非字符串 p_1,将从实体 B 中抽取出的非字符串记为 p_i(从实体 B 中可能会抽取出多个非字符串)。利用式(1),把实体 B 的每个非字符串分别与 p_1 进行相似度计算,并基于 p_i 的视觉特征[9] 将最大的相似度作为对应的神经网络输入。

$$Sim(p_1, p_i) = 1 - \frac{\sqrt{\dfrac{(p_1 - \bar{p})^2 + (p_i - \bar{p})^2}{2}}}{\bar{p}} \tag{1}$$

其中 $\bar{p}$ 为 p_1 和 p_i 的均值。

2) 字符文本型。对于字符文本型的属性,可以通过基于字符串编辑距离的方法[10] 计算语义块与实体 B 中字符文本部分的相似度。基于字符串编辑距离的方法是通过将源串 t_i 转换到目标串 t_j 所需的最少的插入、删除和替换的操作数目。在得到编辑距离的基础上运用式(2) 将其转化为相似度。

$$Sim(t_i, t_j) = \max\left(0, \frac{\min(|t_i|, |t_j|) - D(t_i, t_j)}{\min(|t_i|, |t_j|)}\right)^2 \tag{2}$$

其中:t_i 表示实体 A 中的字符串语义块,t_j 表示实体 B 中的字符文本部分,$D(t_i, t_j)$ 表示编辑距离,三种基本操作分别为:① 将串 t_i 中的一个字符替换 t_j 中的字符;② 将 t_i 中的一个字符删除;③ 在串 t_j 中插入一个字符。

2.3　实体识别模型训练

利用 BP 神经网络对 Deep Web 实体识别时,首先需要通过 BP 神经网络建立实体识别模型,采用 BP 神经网络的优势在于不再需要计算各属性的权重,而通过 BP 神经网络自主学习训练各属性的内在关系来判断 Deep Web 实体是否匹配,这种方法能随着环境的变化调节自身的权值、阈值等参数,有较强的适应能力;然后将样本实体分为两类(一类是匹配实体集,一类是不匹配实体集);接下来将这两类实体分别输入 BP 神经网络进行权值和阈值的训练,得到实体识别模

型;最后将待匹配的实体数据输入训练好的实体识别模型,由此判断实体是否匹配。在实体识别中对 BP 神经网络进行训练的流程如图 2 所示。

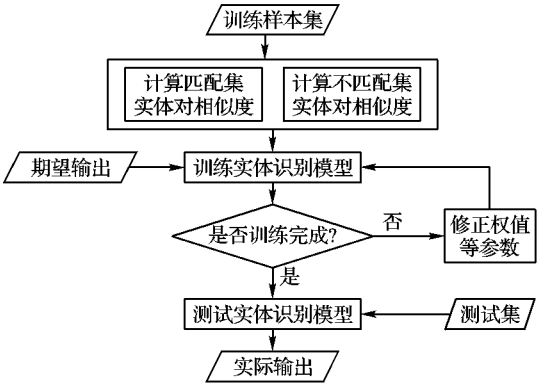

图 2 BP 神经网络训练流程

具体训练步骤如下:

1)建立基于 BP 神经网络的实体识别模型,输入层有 n 个节点,分别对应实体的 n 个语义块和另一实体的相似度,如果没有对应的相似度则输入为 0。输出层为两个神经元,当输出向量为(1,0) 时代表实体对匹配,输出为(0,1) 时代表实体对不匹配,隐藏层神经元个数使用式(3) 确定。

$$k = \mathrm{sqrt}(i + j) + d \tag{3}$$

其中:k 表示隐藏层神经元个数,i 表示输入层神经元个数,j 表示输出层神经元个数,d 为区间[1,10] 内的常数。

2) 将样本实体分为匹配实体集 $M(r)$ 和不匹配实体集 $U(r)$。分别计算样本数据集合 $M(r)$ 和 $U(r)$ 中实体对(A,B) 的实体 A 各分块与实体 B 的相似度 $Sim(A_i,B)$。

3) 将 $M(r)$ 实体集中相似度 $Sim(A_i,B)$ 作为实体识别模型的输入。将语义块的属性值标识与输入节点相对应,对应上的节点输入语义块的相似度值 $Sim(A_i,B)$,没有对应上的节点则输入为 0,期望的输出为(1,0)。利用神经网络的误差反向传播,如果输出层的实际输出与期望输出误差较大时,网络进入误差反向传播阶段,对神经网络权值和阈值进行调节修正,直到神经网络收敛,误差精度满足要求。

4) 将 $U(r)$ 实体集中相似度 $Sim(A_i,B)$ 作为输入,期望的输出为(0,1),与 3) 同理,直到神经网络收敛,误差精度满足要求。

5) 将待测试的实体对相似度 $Sim(A_i, B)$ 作为训练好实体识别模型的输入,得到输出 S。

6) 若 S 的误差范围在 $M(r)$ 实体集的目标模式范围内,则待匹配实体对匹配;若 S 的误差范围在 $U(r)$ 实体集的目标模式范围内,则待匹配实体对不匹配。本文将目标模式范围定义为训练集的上限和下限,以本文的实验为例,经训练 $M(r)$ 实体集的目标模式下限为(0.884,0.116),S 的误差范围在([0.884,1],[0,0.116]) 时则匹配;$U(r)$ 实体集的目标模式上限为(0.373,0.627),S 的误差范围在([0,0.373],[0.627,1]) 时则不匹配。如果误差范围位于([0.373,0.884],[0.116,0.627]) 时则实体既不属于匹配集,也不属于不匹配集,需要专家人工识别。

3 实验与分析

实验的数据来源于 Amazon.com 和 dangdang.com 两个大型购书网站,通过向网站的查询接口提交查询请求继而从两个网站获取类似的实体,将这些实体分为训练集和待测集。首先对所有实体进行语义块划分,本实验中共获取 12 个语义块对应的属性值的标识作为合集,记为 K = {ISBN,书名,作者,市场价,价格,折扣,出版社,平装,条形码,ASIN,重量,尺寸}。K 中的内容只是标识,用以指明输入内容的含义。然后计算训练集中任一实体的各语义块与另一实体相似度,将其作为训练实体识别模型的输入数据,如表 1 所示。

最后,建立基于 BP 神经网络的实体识别模型。假定神经网络的最大迭代次数不限直到误差符合要求为止,目标函数误差为 0.001,学习率为 0.2。神经网络的输入节点为 12 个,输出节点为 2 个,根据式(3) 和文献[11] 所述方法确定隐藏层单元数为 10。基于 BP 神经网络的实体识别模型如图 3 所示。经训练,$M(r)$ 实体集的目标模式下限为(0.884,0.116),$U(r)$ 实体集的目标模式上限为(0.373,0.627)。从测试集中取出一对匹配实体 A 和 B,取出一对不匹配实体 A 和 C 用于测试,将实体 A 进行分块分别与实体 B 和实体 C 进行相似度计算,相似度计算结果如表 2 所示。实体 A 与 B 的测试结果为(0.999,0.001),位于目标模式([0.884,1],[0,0.116]) 的范围内,因此判断实体 A 与实体 B 匹配。而实体 A 与 C 的测

表 1 训练样本的相似度

训练项目	实体序号	ISBN	书名	作者	市场价	价格	折扣	出版社	平装	条形码	ASIN	重量	尺寸
$M(r)$ 相似度	1	0.91	0.94	0.82	0.88	0.94	0.78	0.83	0.86	0.71	0.76	0.75	0.00
	2	0.95	0.91	0.78	0.89	0.83	0.74	0.79	0.79	0.88	0.81	0.70	0.70
	3	0.93	0.86	0.81	0.92	0.80	0.71	0.79	0.88	0.71	0.77	0.00	0.67
	4	0.96	0.93	0.95	0.92	0.88	0.81	0.80	0.71	0.73	0.70	0.68	0.00
	5	0.98	0.88	0.94	0.91	0.93	0.87	0.85	0.76	0.79	0.65	0.71	0.60
$U(r)$ 相似度	6	0.26	0.24	0.19	0.24	0.20	0.11	0.08	0.10	0.12	0.06	0.09	0.00
	7	0.22	0.18	0.12	0.23	0.10	0.13	0.08	0.08	0.07	0.06	0.10	0.09
	8	0.17	0.23	0.19	0.20	0.18	0.14	0.09	0.13	0.06	0.15	0.11	0.10
	9	0.21	0.12	0.10	0.14	0.15	0.07	0.09	0.10	0.11	0.09	0.06	0.00
	10	0.26	0.21	0.18	0.11	0.16	0.08	0.09	0.10	0.12	0.06	0.00	0.07

表 2 测试集实体对的相似度

测试项目	ISBN	书名	作者	市场价	价格	折扣	出版社	平装	条形码	ASIN	重量	尺寸
B 与 A 的语义块相似度	0.97	0.92	0.94	0.88	0.90	0.86	0.91	0.87	0.83	0.79	0.70	0.00
C 与 A 的语义块相似度	0.16	0.10	0.13	0.10	0.11	0.12	0.09	0.09	0.08	0.07	0.00	0.05

试结果为(0.03,0.997),位于目标模式([0,0.373],[0.627,1])范围内,因此判断实体 A 与 C 不匹配。

接下来通过增加样本数量来观测 BP 神经网络的误差变化情况,结果如图 4 所示。从图 4 可以看出随着样本数量的增加,模型的准确率会相应提高,但训练样本数量过多会导致计算的时间复杂度升高。本文方法的性能通过 F 值评价指标进行衡量,F 值的计算公式[10] 如下所示:

$$F = \frac{准确率 \times 召回率 \times 2}{准确率 + 召回率} \tag{4}$$

将本文方法与文献[12-13]方法进行比较的结果如图 5 所示。实验结果表明本文方法具有较优的性能。

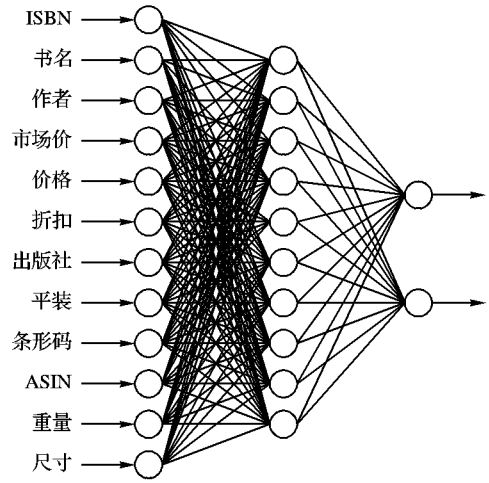

图 3　基于 BP 神经网络的实体识别模型

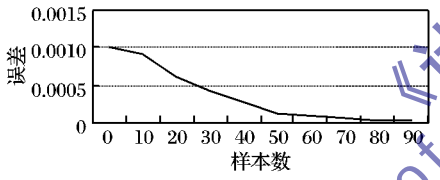

图 4　样本数量与误差变化关系

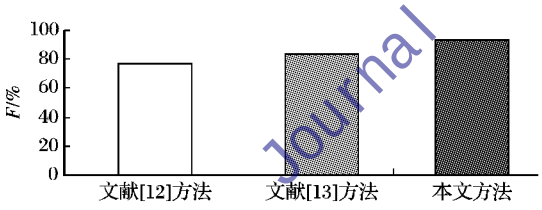

图 5　F 值比较

4　结语

互联网资源的极大丰富对实体识别的有效性和性能提出了较高的要求,本文充分利用 BP 神经网络的优势,提出了一种基于 BP 神经网络的 Deep Web 实体识别方法,该方法无需人工干预,通过对同类实体进行分块,再用每个语义块分别与另一实体计算相似度作为神经网络训练的输入,降低了实体识别的复杂度。再通过 BP 神经网络的自主训练提高了实体识别方法的适应性和自动化水平。实验结果表明,本文方法具有良好的识别准确率和效率。

参考文献:

[1] MADHAVAN J, JEFFERY S R, COHEN S, *et al*. Web-scale data integration: you can only afford to pay as you go [C]// Proceedings of the 3rd Biennial Conference on Innovative Data Systems Research. California, USA: CIDR, 2007: 342-350.

[2] 王妍, 宋宝燕, 张佳旸, 等. 基于标签编码的 Deep Web 查询接口识别方法[J]. 计算机应用, 2011, 31(5): 1351-1354.

[3] 刘伟, 肖建国. 多 Web 数据源环境下的重复实体识别方法研究[J]. 计算机科学与探索, 2010, 4(7): 599-607.

[4] CHAUDHURI S, GRANTI V, MOTWANI R. Robust identification of fuzzy duplicates [C]// Proceedings of the 21st International Conference on Data Engineering. Washington, DC: IEEE Computer Society, 2005: 865-876.

[5] 崔晓军, 肖红宇, 丁立新. 基于距离的自适应 Web 数据库记录匹配方法[J]. 武汉大学学报, 2012, 58(1): 89-94.

[6] SHEN W, DEROSE P, VU L, *et al*. Source-aware entity matching: a compositional approach[C]// Proceedings of the 23rd International Conference on Data Engineering. Washington, DC: IEEE Computer Society, 2007: 196-205.

[7] 马锐. 人工神经网络原理[M]. 北京: 机械工业出版社, 2010.

[8] 凌妍妍, 刘伟, 孟小峰, 等. Deep Web 数据集成中的实体识别方法[J]. 计算机研究与发展, 2006, 43(Suupl): 46-53.

[9] 高乐, 张健, 田贤忠. 基于视觉的 Web 页面分块算法的改进与实现[J]. 计算机系统应用, 2009, 18(4): 65-69.

[10] 朱命冬, 申德容, 寇月, 等. 一种应用于 Deep Web 环境下重复记录识别模型[J]. 计算机研究及发展, 2009, 46(Suppl): 14-21.

[11] 沈花玉, 王兆霞, 高成耀, 等. BP 神经网络隐含层单元数的确定[J]. 天津理工大学学报, 2008, 24(5): 13-15.

[12] LI W S. SeEMINT: a tool for identifying attribute correspondences in heterogeneous database using neural networks [J]. Data and Knowledge Engineering, 2000, 33(1): 49-84.

[13] 强保华, 陈凌, 余建桥, 等. 基于 BP 神经网络的属性匹配方法研究[J]. 计算机科学, 2006, 33(1): 249-251.

(上接第 775 页)

[9] 刘小生, 潘燕群. 基于 OWL 的自然灾害领域本体的建立[J]. 黑龙江工程学院学报: 自然科学版, 2008, 22(2): 19-21.

[10] 马朋云. 本体公理推理及其在交通领域中的应用[D]. 大连: 大连交通大学, 2010.

[11] 耿科明, 袁方. Jena 推理机在基于本体的信息检索中的应用[J]. 微型机与应用, 2005, 24(10): 62-64.

[12] KIM J Y, JEONG D W, BAIK D-K. Ontology-based semantic recommendation system in home network environment [J]. IEEE Transactions on Consumer Electronics, 2009, 55(3): 1178-1184.

[13] HOSER B, HOTHO A, JASCHKE R, *et al*. Semantic network analysis of ontologies [C]// Proceedings of ESWC 2006, LNCS 4011. Berlin: Springer, 2006: 514-529.

[14] MESINA M, ROLLER D, LAMPASONA C. Visualisation of semantic networks and ontologies using AutoCAD [C]// Proceedings of CDVE 2004, LNCS 3190. Berlin: Springer, 2004: 21-29.

[15] KNAPPE R, BULSKOV H, ANDREASEN T. Perspectives on ontology-based querying: research articles[J]. International Journal of Intelligent Systems, 2007: 22(7): 739-761.

[16] 吕金丽, 余雪丽. 课程知识本体建模及推理[J]. 计算机工程, 2011 (4): 61-63.

[17] 李春. 基于本体的文本信息检索技术研究与实现[D]. 南京: 南京航空航天大学, 2009.

[18] 杨涛. 基于本体的案例推理系统框架研究[D]. 南京: 南京航空航天大学, 2006.