基于知识管理的本体自动构建算法研究

郑学伟[1,2]
(1. 辽宁广播电视大学,辽宁 沈阳 110034;
2. 辽宁装备制造职业技术学院,辽宁 沈阳 110161)

摘 要:语义网的研究中,基于领域本体的构建研究方法基本上还处于手工阶段,如何在本体的设计中实现自动构建仍然是目前需要解决的问题,而采用基于图的构建原理,应用 MCL 聚类的本体自动构建算法进行概念提取和关系运算。将领域文本文档映射为文档概念图,在关系运算中采用基于频繁信息子图的 gSpan 算法的任意关系提取算法,得到基于 OWL-DL 描述的领域本体,并通过评价反馈机制进行闭环校正是研究的核心思想。

关键词:领域本体;自动构建;知识管理;gSpan 算法
中图分类号:TP182 **文献标识码**:A **文章编号**:1673-629X(2014)12-0064-05
doi:10.3969/j.issn.1673-629X.2014.12.016

Research on Ontology Automatic Construction Algorithm Based on Knowledge Management

ZHENG Xue-wei[1,2]
(1. Liaoning Radio and TV University, Shenyang 110034, China;
2. Liaoning Vocational and Technical College of Equipment Manufacturing, Shenyang 110161, China)

Abstract: In the research on semantic Web, the construction method based on domain ontology is still basically in the manual stage. In the design of ontology, how to realize the automatic construction is still a problem needs to solve. But use the construction principle based on diagram, and apply the algorithm of MCL clustering for automatic ontology construction to extract concept and operate relation. Mapping the domain text document to document concept map, and in the relational operations, using the arbitrary relations extraction algorithm based on frequent information sub graphs of gSpan algorithm to obtain the domain ontology based on OWL-DL description, and through the evaluation and feedback mechanism to carry out the closed-loop correction is the core idea.

Key words: domain ontology; automatic construction; knowledge management; gSpan algorithm

1 概 述

语义网的实现依赖于三大关键技术:XML(可扩展标记语言)、RDF(描述 Web 资源的标记语言)、Ontology(本体或本体论)[1]。本体(Ontology)的概念起源于西方哲学,现在哲学领域较多翻译为“本体论”[2]。20 世纪 60 年代,人们在人工智能通用问题求解的研究上遇到了障碍,为了解决这一问题,Guarino(1998)在概念的阐述上明确了 AI 中的本体与哲学的本体的区别。本体是基于语言的,但是本体的概念化同语言无关,词汇的概念化其实是一个比信息技术中的本体论更宽泛的概念,更接近于哲学中本体的涵义。未来有望在万维网联盟(World Wide Web Consortium, W3C)的主导下解决在互联网工作时实现对互联网上的语言以及词汇机器自动智能分析的问题,从而实现世界范围内跨语言的知识智能分析和信息共享功能[3]。目前对于语义网中本体的研究主要有以下几个方面:本体的概念化、本体的形式化、本体的标准化和本体的准确性。目前本体的研究按照领域依赖程度可以分为以下几类:顶层(top-level)本体、领域(domain)本体、任务(task)本体、应用(application)本体。领域本体的研究是将研究范围限定在某一个特定的领域范围内,通过小范围的逻辑运算对特定范围内的词汇做出描述,通过对领域本体的描述使对于同一概念无论是计算语言还是人都能共享并且无歧义的理解。

收稿日期:2014-02-26 **修回日期**:2014-05-28 **网络出版时间**:2014-10-23
基金项目:国家开放大学项目(Q0604F-Q);辽宁省现代远程教育学会项目(2012xh1-25,2013XH01-23)
作者简介:郑学伟(1979-),男,辽宁抚顺人,硕士,副教授,研究方向为知识管理、语义网技术。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20141023.1124.037.html

目前构建领域本体的方法按照自动化程度的不同大致可以分为以下三类:自动构建、半自动构建和人工构建。人工构建由于天然具有工作量大、适应性及可移植性差等缺点,难以追踪研究领域的最新变化,自动构建领域本体的算法的思路是通过知识系统的文档追踪领域本体定义与变化的过程。

在管理过程中,基于知识分析的知识管理系统主要面向的是在系统建设和维护管理过程中主观或者客观产生的各类文档,文档记录了产品从设计理念到调研分析再到可行性分析、详细设计、加工测试、投入使用、维修反馈等一系列环节的信息,从信息采集到知识分析到智能管理是知识管理体系的三个方面,它们之间构成了一条由简单到复杂的序列,在这一系列的知识管理活动中可以通过领域本体的构建有效地解决知识的识别与利用难题。自动构建算法在实现过程中首先把所有文档归纳为构建领域本体的信息源库[4],随着产品或者系统不断的推进,相应的文档也在不断的维护与更新,领域本体就可以通过信息源库中文档资料的更新修正自身[5]。算法具有以下特点:

(1)文档的描述基于图。通过基础的语义分析将信息源库中的各类文档中的词汇提炼出集合成为一个无向图,词频的信息用无向图中对应的顶点和边来描述,文档图中的结构信息可以用各个顶点之间的拓扑关系来描述。

(2)概念提取。先对文档中词汇进行概念性的评价,通过算法将含义相近的词汇汇集成一组,每一组组成一个候选的集合,以避免含义重复。

(3)关系运算。对含义相近的信息集合后进行数据挖掘,引入信息方程,筛选出有意义的领域集合,准确描述出领域本体的信息。

(4)评价反馈。通过对生成的领域本体的实际测试效果对构建本体的信息源库以及产生的领域本体概念进行评价,用过闭环的反馈不断修正自动算法,使之达到一个理想的效果。

2　基本原理与概念

将系统中的文本文档(Word、PDF、txt 等)整合成信息源库。归纳形成的概念图表现为带有标记的无向加权图,概念图标记为 G,集合表示为 $G=(V,E,\alpha,\beta)$。其中 V 是顶点的集合,$E\subseteq V\times V$ 是连接顶点的边的集合,$\beta:E\rightarrow\sum E$ 定义由源数据库中表现出的无向图中顶点到相邻顶点之间的映射关系,$\alpha:V\rightarrow\sum V$ 定义了无向图中从边到相邻边之间的映射。通过以上两种映射关系来分别表示源文档中词汇之间的逻辑关系。基于无向图表示的文档库不仅要反映出词汇出现的频率,更要根据图的结构反映出文档的结构,使这两种信息能在后续的本体构建中被完整地考虑进去,使得文档的源库中的信息能够被准确真实地提炼出来[6]。为解决区别并归纳分析语义相近的概念必须在算法中引入马尔可夫聚类(Markov CLuster,MCL),马尔可夫聚类的基础是随机游走理论,以上信息经过 Markov 聚类运算后得到的结果是双幂等矩阵。关系算法是自动构建算法的关键,目前应用的算法大部分为层次关系算法,基于任意关系运算的很少。文中设计采用的思路是首先从所有文档图集合中分析出所有的频繁子图,采用基于任意关系挖掘的 gSpan 算法,gSpan 算法利用模式增长(pattern-growth)策略,对于整个空间的搜索采用深度优先方式进行遍历搜索,直到发现全部频繁子图为止[7]。引入约束机制来进一步控制子图挖掘过程,在运算的过程中导入文档的实际信息量。

3　领域本体的自动构建算法设计

文中所设计的领域本体自动构建算法由文档预处理、基于图的文档表示、概念提取、关系计算、评价反馈五个部分构成一个闭环控制系统,不同部分之间的数据传递关系及控制流程如图 1 所示。

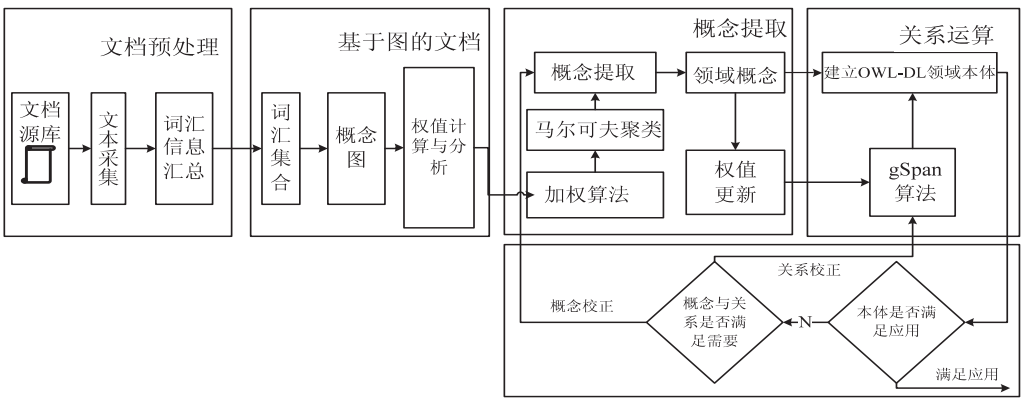

图 1　本体自动构建算法框架图

基本流程如下:先对目标文档库进行预处理,将处理过的词汇采用加权算法进行图形化,在此基础上采

用概念提取和关系提取,概念的分析采用马尔可夫聚类生成,约束条件采用 gSpan 算法,生成的结果由反馈评价系统进行评估,不符合条件的概念和关系返回进行重新演绎生成。

3.1 文档预处理

文档预处理是整个工作的基础,预处理过程中对文档信息的处理是否准确非常重要,整个环节的主要步骤是:首先将文档中不同的词汇进行语言识别,并分中文英文进行筛选,清除出无意义的停用词后,对不同的词性进行标记,统计出相关的词性的频率和相邻关系,在此基础上计算出图的顶点的初始权值和边的初始权值。

3.2 文档基于图的表示

人在阅读时,对于相邻两个词汇一般是同时映入大脑并自动给出一般情况下的先后顺序,语义分析中在设计出基于文档的图时,可以不采用有向文档图而采用无向文档图。图的基本设定是用文档中分析后的相同的词汇作为图的顶点,一个顶点是一个相同的词汇。两个顶点之间的边标记为两个顶点所代表的词汇的组合,将两个顶点用无向线相连表示在文档中的词义相邻[8]。顶点之间的边用无向线表示,相邻词汇之间的顶点的初始权值用词汇的频率数值来表示,相邻词汇之间无向边的初始权值用词汇的相邻信息来表示。相邻词汇的频率数值也就是代表相对应顶点的词汇在文档中出现的频率,而边则表示为这两个相邻词汇共同出现的次数,先计算出所有顶点初始权值的总数,将每一顶点的初始权值除以这一数值,得到的结果是一个介于 0 到 1 之间的小数,这一数值就作为对应顶点的权值;再计算出所有边的初始权值的总数,将每一边的初始权值除以这一数值,同样得到一个介于 0 到 1 间的小数,这一数值就作为对应边的权值,被除数也可以采用最大权值来表示。

3.3 概念提取

自动提取算法首先对出现的词汇的重要性和信息度进行评估,同时因为引言中提到的问题,消除词汇在不同的环境下产生的歧义,并对其进行分类,产生候选概念。词汇的加权算法采用马尔可夫聚类算法。优点是既可以覆盖全局的信息又可以兼顾顶点的局部信息。马尔可夫聚类算法通过计算图的结构以及图中各个顶点的权值,通过加权运算,一方面消除不同环境下语义的歧义,使词汇的语义清晰,避免语义环境的重复;另一方面也可以将具有含义相近的语义归纳成为一个簇,经过马尔可夫聚类后,消除歧义的顶点被划分为多个的簇,每个簇对应一个候选的概念,代表此簇的词汇由簇中所有词汇的权值中最大的决定。马尔可夫聚类算法的实质是通过计算让所有顶点的权值最后全部归入一个预先定义好的阈值中,在计算的过程中,通过迭代关系使顶点的新旧权值不断更新。由于文档图采用无向标记,对应的权值经过计算后代表的含义由原来出现的次数转换为词汇所代表的含义在文档中所占地位的高低。计算公式如式(1)所示。

$$WS(V_i) = (1 - d) + d \times \sum_{V_j \in \text{Neighbour}(V_i)} \frac{w_{ji}}{\sum_{V_k \in \text{Neighbour}(V_j)} w_{jk}} WS(V_j) \quad (1)$$

式中, $\text{Neighbour}(V_i)$ 表示顶点 V_i 的领域; w_{ji} 表示连接顶点与边的权值;d 为系数调整因子。

由上式可以看出,以上算法计算的是基于文档图的结构,算法不涉及到具体的领域,无需计入额外的领域知识域。由于 MCL 聚类根据图的边信息对顶点进行聚类,边权值由其所连接的两个顶点的权值决定,同时考虑到标准化因素,定义了式(2)来计算边权值。

$$WE(e_{ij}) = \frac{\sqrt{WS(V_i)^2 + WS(V_j)^2}}{|WS(V_i)| + |WS(V_j)|} \quad (2)$$

式中, V_i 和 V_j 分别表示边的两个顶点。随后采用 MCL 聚类算法根据顶点和边的新权值进行顶点聚类。算法的流程可参考文献[9]。经过上述聚类 MCL 算法后的候选概念是一个词,将邻接的顶点连接起来后表示的候选概念将合并成一个新的概念。概念提取算法代码如下所示:

```
Set<term>totalGraph;
for each graph G_i in fileTermMap;
Summary Statistics dist_stats;
while 当前迭代次数≤max_iterations{
for each node d_i in Graph totalGraph{
  Double nodeweight=0;
  Double allweight=calculate the sum of weights of edges that connet with d_i;
  Nodeweight * =DAMPING_FACTOR;
  Nodeweight+=1.0-DAMPING_FACTOR;
  Using Nodeweight to Update the weight of node d_i;
    }
  Double standard_error=calculate the standard error of dis_stats;
if standard_error<STANDARD_ERROR_THRESHOLD break;
    }
Upda all edge weights of totalGraph with regard to the new weights of nodes;//MCL 聚类算法
Matrix T_1=the adjacency Markov matrix of totalGraphwithloop;
While(true){
T_2k=Exp(T_2k-1);
T_2k+1=(T_2k);
if(T_2k+1 is idempotent) break;
}
Concept candidates{CC_1, CC_2, …, CC_n}=Interpret T_2k+1 as
```

clustering according to term weights;

Marking{$CC_1, CC_2, \cdots, CC_n$} in the totalGraphwithLoop;

Concept{$C_1, C_2, \cdots, C_k$} = {$CS_1, CS_2, \cdots, CS_p$} {$CM_1, CM_2, \cdots, CM_q$};//最终的输出

3.4　关系运算

关系运算的关键在于基于图的概念提取后关系的提取,频繁子图是关系提取的关键。关系运算要解决的问题是给定文档图的数据库中频繁信息子图,并将频繁信息子图映射为领域关系。具体的运算采用基于信息函数的 gSpan 算法在上一环节生成的数据概念集合中对频繁子图进行挖掘。计算公式如公式(3) ~ (5)所示。

$$I(g) = \sum_{v \in V(g)} i_v(v) + \sum_{e \in E(g)} i_e(e) \tag{3}$$

其中

$$i_v(v) = -\log 2\left\{ \sum_{d \in D} \frac{w_v(v)}{\sum_{v \in d} w_v(v)} \right\} \tag{4}$$

$$i_e(e) = -\log 2\left\{ \sum_{d \in D} \frac{w_e(e)}{\sum_{e \in d} w_e(e)} \right\} \tag{5}$$

在运算中抛弃所有都没有对领域概念进行标注的顶点的频繁子图,同时在运算中认定主体是领域本体相关的动词。关系运算中,被挖掘出的子图中的顶点如果不在表单中,则可以对表单中的概念集合进行补充,这可以看作对前期运算的一个反馈校正机制。具体关系提取算法的代码略。经过上述运算得到的频繁信息子图具体表现为三元组{概念 1,核心关系,概念 2}。核心关系是一个动词,概念是名词,一组关系表现为一个动词同两个名词的组合,运算的步骤是先找到动词标注的顶点,以这一顶点为中心搜寻相连接的顶点,如果找到两个名词,则关系确认,如果有一个仍是动词,则进行持续迭代,一直到确认关系为止,同理适用于两个动词相连。在整个运算过程中,由于核心关系是动词,所以在基于图的分析中,决定最后关系数目的是图中动词顶点的数目。

3.5　评价反馈

在文中所设计的闭环处理系统中,前四个环节都涉及到算法对基于词汇的文档图的处理,每个环节运算对最终结果的影响都较大,判断经运算生成的领域本体是否使用并纠正错误运算对结果的影响,在前四个环节的基础上增加了评价反馈环节,利用那部分与领域相关,但是并没有参与到构建领域本体的文档来测试构建的领域本体是否适用[10]。如果出现前后不一致的情况,根据具体情况分析是概念图还是关系运算出现问题进行纠正,或者在原有基础上添加资料文档完善领域本体。反馈后重新进行运算,直至没有错误产生。

3.6　实验结果与总体评价

对于整个自动构建算法运算结果的完整性和准确性可以采用查准率—查全率曲线(PR 曲线)来进行验证。查准率反映的是对词汇的区分能力,查全率则测试了算法能否完备地抽取概念,同样的测试标准也可以用来测试关系运算的性能[11]。在具体运算过程中通过加权运算可以完整地评估目标词汇在整个文档中的重要程度,马尔可夫聚类运算则将相似词汇之间的歧义进行了消除[12]。通过一系列的运算最终提高了算法的查准率,为保证算法设计在实践中的可行,特设计实验来与传统构建算法进行比较,实验将对概念提取和关系运算的实际效果以文档语料库的 TREC-9 为标准进行比较分析。本算法记为 GR-AONTO,目标对比算法采用 TextRank[13] 和 SIGNUM[14] 算法,以此比较三种算法随着查全率的增长查准率下降的速度,进而得出不同算法的性能指标。运算公式如下:

$$\text{Precision}(C-P) = \text{Precision}(R-P) = \frac{A}{A+B} \tag{6}$$

$$\text{Recall}(C-P) = \text{Recall}(R-P) = \frac{A}{A+E} \tag{7}$$

实验的结果如图 2 所示。

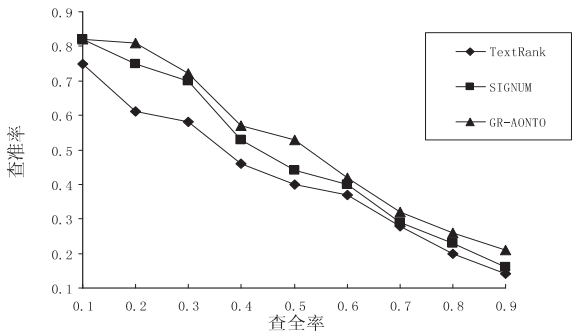

图 2　查准率-查全率曲线

根据图中数据线变化情况可以得知,在查全率上升的过程中,GR 算法的表现最优。通过以上运算可以证明,马尔可夫聚类运算的采用确实可以提高整个算法框架的性能。在整个本体构建的过程中通过概念的提取各关系的运算最终得到符合要求的本体模型,在词汇的分类和概念的提取过程中采用基于图的运算可以有效地消除词汇间的歧义,提高算法的精确度,在关系运算中采用基于任意关系的频繁子图挖掘 gSpan 算法具有重要意义[15]。

4　结束语

在文中所提出的研究中,在手段上采用基于图的运算方式,同时兼顾文档的频率信息和结构信息。主要应用的算法有三个:首先是词汇的加权算法上采用图上随机算法,然后在概念的提取上采用 MCL 聚类算

法将无效的语法歧义进行消除,通过分类筛选出候选概念,最后采用 gSpan 算法对频繁子图中的数据进行挖掘,形成 OWL-DL 格式的领域本体描述。在反馈评价阶段,根据前四个环节运算的结果应用了一种自我调整与修正的机制,整个框架能够完整准确地提取概念和关系,有效地自动构建本体,但是在准确率上,随着图上顶点数量的增加有下降的趋势,需要在以后的研究中进一步设计解决。

参考文献:

[1] Ding Shengchun, Jiang Chaonan. Excavating implicit relation based on SWRL[J]. New Technology of Library and Information Service, 2011, V27(3): 68-72.

[2] 侯 鑫,张旭堂,金天国,等. 面向知识与信息管理的领域本体自动构建算法[J]. 计算机集成制造系统,2011,17(1):159-170.

[3] Mao Yuxin, Chen Huajun, Jiang Xiaohong, t al. Domain knowledge resource management based on sub-ontology[J]. Computer Integrated Manufacturing Systems, 2008, 14(7): 1434-1440.

[4] Gacitua R, Sawyer P, Rayson P, et al. A flexible framework to experiment with ontology learning techniques[J]. Knowledge-Based System, 2008, 21(3): 192-199.

[5] Zhao Jianxun, Zhang Zhenming, Tian Xitian, et al. Ontology & its applications in mechanical engineering[J]. Computer Integrated Manufacturing Systems, 2007, 13(4): 727-737.

[6] 李志国,冯 永,钟 将,等. 基于 Super-P2P 的分布式知识管理模型[J]. 计算机科学,2007,34(7):184-186.

[7] 王永贵. 分布式知识管理中的语义交互式框架与方法研究[D]. 大连:大连理工大学,2008.

[8] 张海霞,吴 江. 基于语义网的知识管理系统框架设计[J]. 计算机技术与发展,2006,16(4):46-48.

[9] Dongen S. A cluster algorithm for graphs[R]. New York, NY, USA: ACM, 2000.

[10] 刘 莉,何中市,邢欣来,等. 基于语义角色的中文时间表达式识别[J]. 计算机应用研究,2011,28(7):2543-2545.

[11] 刘成山,赵捧未. 语义对等网环境下的数字图书馆原型[J]. 情报杂志,2010,29(6):110-112.

[12] 孙 艳,周学广,付 伟. 基于依存关联分析的情感词扩展[J]. 北京邮电大学学报,2012,35(5):90-93.

[13] Mihalcea R, Tarau P. TextRank: bringing order into texts[C]//Proceedings of the empirical methods in natural language processing. Berlin, Germany: Springer, 2006: 404-411.

[14] Ngomo A C N. SIGNUM: a graph algorithm for terminology extraction[J]. Lecture Notes in Computer Science, 2008, 4919: 85-95.

[15] 邬 桐,周雅倩,黄萱菁,等. 自动构建时间基元规则库的中文时间表达式识别[J]. 中文信息学报,2010,24(4):3-10.

(上接第 63 页)

通过表 2 可以看出,英文论文的提取正确率要高于中文论文的提取情况,主要原因是中文论文中的格式过于多样化,依靠对标题提取规则有时并不能正确地提取到标题,而是提取到页眉的内容,一旦标题提取不正确,作者信息和单位也会提取失败。很多学术论文中并没有收稿日期,尽管程序已经利用页眉或页脚中的日期信息,但是仍然有很多论文甚至连页眉/页脚信息也没有,造成出版日期提取失败。所以提取的规则或者方法还有待进一步完善。

参考文献:

[1] Adobe Systems Inc. PDF reference, Adobe portable document format version 1.4.3nd[EB/OL]. 2001. http://www.adobe.com/suppon/down-loads/product.jsp? product=44& platform=Windows(Accessed Mar. 8, 2005).

[2] Lovegrove W S, Brailsford D F. Document analysis of PDF files: methods, results and implications[J]. Electronic Publishing Origination Dissemination and Design, 1995, 8(2/3): 207-220.

[3] 陈云榕,刘立柱,丁志鸿. PDF 文件中关键信息的提取与组织方法研究[J]. 计算机工程与设计,2007,28(7):1688-1690.

[4] 李贵林,李建中,杨 艳. 用 Plug-in 实现对 PDF 文件的信息提取[J]. 计算机应用,2003,23(2):110-112.

[5] 赵 耀. 基于 PDF 文档的数字化学习资源建设[J]. 临沂师范学院学报,2011,33(6):125-128.

[6] 龙 珑,邓 伟,覃 晓. 绿色网络 PDF 提取系统[J]. 计算机技术与发展,2014,24(1):204-207.

[7] 张秀秀,马建霞. PDF 科技论文语义元数据的自动抽取研究[J]. 现代图书情报技术,2009(2):102-106.

[8] 李兰友,陈 立,谢雪莲. 面向 Web 的 PDF 文档构建技术[J]. 计算机与现代化,2013(12):184-187.

[9] Yuan Fang, Liu Bo, Yu Ge. A study on information extraction from PDF files[C]//Processdings of the 4th international conference on advance in machine learning and cybernetics. Berlin: Spinger-Verlag, 2005: 258-267.

[10] 李 强,刘时进. PDF 阅读器的设计与实现[J]. 计算机工程与设计,2013,31(7):1635-1638.

[11] 李朝光,张 铭,邓志鸿,等. 论文元数据信息的自动抽取[J]. 计算机工程与应用,2002,38(21):189-191.

[12] Chao Hui, Fan Jian. Layout content extraction for PDF documents[C]//Proceedings of document analysis systems. Berlin: Spinger-Verlag, 2004: 213-224.

[13] 宋艳娟,张文德. 基于 XML 的 PDF 文档信息抽取系统的研究[J]. 现代图书情报技术,2005(9):10-13.

[14] 杨道良. 面向对象的中文 PDF 阅读器的设计与实现[J]. 计算机应用,1999,19(6):1-4.

基于知识管理的本体自动构建算法研究

作者： 郑学伟， ZHENG Xue-wei
作者单位： 辽宁广播电视大学，辽宁 沈阳110034；辽宁装备制造职业技术学院，辽宁 沈阳110161
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2014(12)