文章编号:1671-5896(2013)06-0621-06

融合概念格约简的中文领域本体学习方法

侯丽鑫[a],郑山红[a],贺海涛[a],赵 辉[a],韩 冬[b]
(长春工业大学 a. 计算机科学与工程学院; b. 软件职业技术学院,长春 130012)

摘要: 在基于形式概念分析的中文领域本体学习中,为提高概念格构建效率,将概念格约简理论应用于概念格构建中。首先对基于语义依存分析获取的形式背景进行对象和属性约简,然后基于约简的形式背景采用Godin算法构造概念格,最后根据修复定理修复约简概念格,得到完整的概念格。通过有关对萝藦科植物的文本学习,得到一个萝藦科植物领域本体。实验结果表明,引入概念格约简理论,概念格的构建效率提高70%,进而提高了领域本体构建的效率。

关键词: 形式概念分析; 概念格约简; 语义依存分析; 领域本体学习
中图分类号: TP39 **文献标识码:** A

Concept Lattice Reduction Application in Field of Chinese Domain Ontology Learning

HOU Li-xin[a], ZHENG Shan-hong[a], HE Hai-tao[a], ZHAO Hui[a], HAN Dong[b]
(a. College of Computer Science and Engineering; b. College of Software Vocational Technology, Changchun University of Technology, Changchun 130012, China)

Abstract: In order to improve the efficiency of building concept lattice in Chinese domain ontology learning based on formal concept analysis, we applied the concept lattice reduction theory to the process of building concept lattice. The main idea is that we reduce the objects and attributes of the obtained formal context which is based on semantic dependency analysis, adopte the Godin algorithm to construct concept lattice based on formal context reduced, and repaire the concept lattice with the reparation theories. The article takes the asclepiadaceae plants ontology construction as an example to verify this method. The experiment results show that the efficiency of concept lattice construction increase by 70%. It improves the efficiency of ontology construction.

Key words: formal concept analysis; concept lattice reduction; semantic dependency analysis; domain ontology learning

0 引 言

本体学习是指自动或半自动地构建本体,目的是利用知识获取技术降低本体构建的开销,目前本体学习已成为本体领域的研究热点之一。形式概念分析(FCA: Formal Concept Analysis)[1]与本体有许多相似之处[2],近年来,人们开始尝试将FCA应用于本体学习领域。Obitkom等[3]将FCA应用于分布式的领域本体开发环境,通过构建概念格探寻潜在的对象和属性,并将现有的和潜在的实体以可视化的方式自动呈现; 张斌等[4]利用FCA与统计理论进行政务本体学习。概念格是FCA的核心数据结构,在机器学习、模式识别、决策分析等领域应用广泛[5-8]。概念格约简理论[9,10]能更容易发现形式背景中隐含的知

收稿日期: 2013-08-01
基金项目: 吉林省科技厅自然科学基金资助项目(20130101060JC); 吉林省教育厅"十二五"科学技术研究基金资助项目(2014131)
作者简介: 侯丽鑫(1986—),女,山东菏泽人,长春工业大学硕士研究生,主要从事本体、智能系统和语义网研究,(Tel)86-13341593698(E-mail)houlixinmingxuan@126.com; 通讯作者: 郑山红(1970—),女(朝鲜族),长春人,长春工业大学副教授,博士,硕士生导师,主要从事智能系统与语义网研究,(Tel)86-13756476636(E-mail)bioszsh2007@yaoo.cn。

识，也使这些知识的表示变得更简单，提高概念格构造的效率。

笔者在已有工作[11]的基础上，提出了一种融合概念格约简的中文领域本体学习方法。该方法应用语义依存分析技术获取领域形式背景，将概念格约简应用于概念格的构造，通过对象约简和属性约简完成初始概念格的构造，再采用约简概念格修复得到完整概念格，最后根据映射规则完成中文领域本体的生成。通过对萝藦科植物及其药用性的领域文本进行学习，得到萝藦科植物领域本体，提高了本体构建的效率。

1 基于语义依存的形式背景的获取

形式背景是基于 FCA 进行本体学习的基础，笔者将语义依存技术用于从中文领域文本中获取形式背景。下面给出形式背景的定义及从领域文本中获取形式背景的具体算法。

定义 1 形式背景是三元组 $K=(G,M,I)$，其中 G 和 M 分别是对象集和属性集，二元关系 $I\subseteq G\times M$。通常用 $(g,m)\in I$ 或 gIm，表示“对象 g 拥有属性 m”。

形式背景是采用 FCA 方法进行本体学习的基础。笔者借鉴 AIFB 研究机构在 IST-Dot Kom 项目中应用 Philipp Cimiano 方法[12]的思想，利用基于汉语的依存语法分析器[13]，给定文本集中的一个句子作为输入，产生一棵标注依存关系、语义角色的语法分析树，从中抽取句子的主干。将主语作为对象，将对应出现的宾语作为描述该对象的属性，这两部分匹配后作为一个对象-属性对放入形式背景中。具体算法如下。

输入：中文领域文本。

输出：对象属性对构成形式背景。

Step 1 初始化 HashMap，用于存储对象-属性对；

Step 2 基于 CRF(Conditional Random Fields)模型，对文本集中的每句子 i(i 是句子的编号)进行分词，得到 wordList；

Step 3 对句子 i 中的每个词 j(j 为词在句中的编号)，采用 GParser(全称为 Graph-based Parser)输出该词在句子中的依存关系；

Step 4 判断词 j 的依存关系的类型，若依存关系的类型为“SBV(Subject-Verb)”，则词 j 为一个对象；若依存关系的类型为“VOB(Verb-Object)”，则词 j 为对象对应的属性；

Step 5 以 HashMap 中所有的对象 G 及属性 M 构成形式背景 $K(G,M,I)$。

以“牛皮消是一种草本植物。”为例，采用 ltp-service 的语义依存分析结果如表 1 所示。

表 1 示例的语义依存分析结果

Tab. 1 The semantic dependency analysis results of the example

词	词性标注结果	依存分析的父节点号	依存句法分析的依存关系
白薇	n	1	SBV
是	v	-1	HED
一	m	3	QUN
种	q	4	ATT
草本植物	n	1	VOB
。	wp	-2	WP

从表 1 可以看出，“白薇”的依存句法分析的依存关系为“SBV”，即主谓关系；“是”为核心(HED: Head)，即“是”为谓语；“草本植物”的依存句法分析的依存关系为“VOB”，即动宾关系。因此可以得出下面的结果

name = “白薇”； value = “草本植物”； hashmap.put(name,value)；

最终得到一组对象属性对是(“白薇”，“草本植物”)。

基于提出的基于语义依存的形式背景获取方法，通过对多篇介绍萝藦科植物以及它们的药用性的领域文本进行学习，得到的部分形式背景如表 2 所示。

表 2　萝藦科植物形式背景

Tab. 2　The formal context of asclepiadaceae plants

I(关系)	草本植物	藤本植物	止咳	祛风湿	圆锥花絮	聚伞花序	伞形花序	总状花序
牛皮消	×					×		
白薇	×		×					
徐长卿	×			×	×			
一枝香	×			×	×			
萝藦	×							×
夜来香		×				×	×	

2　基于概念格约简的概念格构造

概念格是 FCA 的核心，体现概念间的层次(泛化)关系。它是概念格构造算法[14,15]作用于形式背景的结果。为了提高概念格构造效率，笔者引入概念格约简理论。为方便讨论，给出如下定义。

定义 2　对于 $A\subseteq G$ 和 $B\subseteq M$，可定义如下两种映射

$$f(A)=\{m\in M\mid \forall g\in A,\exists(g,m)\in I\}$$

$$g(B)=\{g\in G\mid \forall m\in B,\exists(g,m)\in I\}$$

定义 3　若 $A\subseteq G,B\subseteq M$，满足 $f(A)=B$ 且 $g(B)=A$，则 $C=(A,B)$ 是 $K=(G,M,I)$ 的一个形式概念。A 称为 C 的外延，记作 $\mathrm{Ext}(C)$，B 称为 C 的内涵，记作 $\mathrm{Int}(C)$。

定义 4　形式背景的概念格记作 $L(K)$，$L(K)$ 是一个偏序集，由形式背景中存在层次包含关系的概念形成，即 $(A_1,B_1)\prec(A_2,B_2)\Leftrightarrow A_1\subseteq A_2$(且 $B_2\subseteq B_1$)。此时 (A_2,B_2) 是超概念，(A_1,B_1) 是子概念。

定义 5　设 $L(G,M_1,I_1)$ 和 $L(G,M_2,I_2)$ 是两个概念格。若对于 $\forall(A_2,B_2)\in L(G,M_2,I_2)$，$\exists(A_1,B_1)\in L(G,M_1,I_1)$，使 $A_1=A_2$，则称 $L(G,M_1,I_1)$ 细于 $L(G,M_2,I_2)$，记作 $L(G,M_1,I_1)\leqslant L(G,M_2,I_2)$。若 $L(G,M_1,I_1)\leqslant L(G,M_2,I_2)$，且 $L(G,M_2,I_2)\leqslant L(G,M_1,I_1)$，则称两个概念格同构，记作 $L(G,M_1,I_1)\cong L(G,M_2,I_2)$。

定义 6　在形式背景 $K=(G,M,I)$ 中，若 $\exists A\subseteq M$，$I_A=I\cap(G\times A)$，使 $L(G,A,I_A)\cong L(G,M,I)$，则称 A 是 $K=(G,M,I)$ 的协调集。如果 $\forall a\in A$，有 $L(G,A-\{a\},I_{A-\{a\}})$ 不同构于 $L(G,M,I)$，则称 A 是 $K=(G,M,I)$ 的约简。

2.1　形式背景的对象约简和属性约简

对于形式背景 $K=(G,M,I)$，既可以对对象又可以对属性进行约简。下面给出对象、属性约简定理。为方便讨论给出如下定义。

定义 7　对于形式背景 $K=(G,M,I)$，其中 $G=\{g_1,g_2,\cdots,g_m\}$，$M=\{m_1,m_2,\cdots,m_n\}$，$I\subseteq G\times M$。对于 $\forall g\in G$，$\forall m\in M$，定义 $R(g)=\{m\in M\mid gIm\}$ 为对象的属性空间，$D(m)=\{g\in G\mid gIm\}$ 为属性的对象空间。

基于定义 7，根据各对象的属性空间之间的关系和各属性的对象空间之间的关系，描述如下 4 条形式背景的约简定理。

定理 1(对象的交约简)　在形式背景 (G,M,I) 中:

1) 如果 $\exists g_i,g_j\in G,i,j\in[1,m]$，使 $R(g_i)=R(g_j)$，则对象 g_i 或 g_j 是冗余的，因此，可约简 g_i 或 g_j 所在的行;

2) 如果 $\exists g_i,g_j,g_l\in G,i,j,l\in[1,m]$，使 $R(g_l)\subset R(g_i)$，$R(g_l)\subset R(g_j)$ 且 $R(g_i)\cap R(g_j)=R(g_l)$，则对象 g_l 是冗余的，因此，可约简 g_l 所在的行。

定理 2(对象的全约简)　在形式背景 (G,M,I) 中，如果 $\exists g_i\in G,i\in[1,m]$，使 $R(g_i)=M$，则对象 g_i 是冗余的，则可约简 g_i 所在的行。

定理 3(属性的交约简)　在形式背景 (G,M,I) 中:

1) 如果 $\exists m_i,m_j\in M,i,j\in[1,n]$，使 $D(m_i)=D(m_j)$，则属性 m_i 或 m_j 是冗余的，因此，可约简 m_i 或 m_j 所在的列;

2) 如果 $\exists m_i, m_j, m_l \in M, i, j, l \in [1, n]$，使 $D(m_l) \subset D(m_i)$，$D(m_l) \subset D(m_j)$ 且 $D(m_i) \cap D(m_j) = D(m_l)$，则属性 m_l 是冗余的，因此，可约简 m_l 所在的列。

定理4(属性的全约简) 在形式背景 (G, M, I) 中，如果 $\exists m_i \in M, i \in [1, n]$，使 $D(m_i) = G$，则属性 m_i 是冗余的，因此，可约简 m_i 所在的列。

由于渐进式构造算法可根据形式背景的变化对原有概念格做相应调整，而不用重新构造概念格，从而节省了大量时间。因此渐进式构造算法一直是概念格构造算法的研究热点。笔者采用基于对象的渐进式构造算法——Godin 算法，在概念格构造过程中引入概念格属性约简理论，采用以上约简定理，对已获得的形式背景进行约简；以约简后的形式背景为数据源，采用 Godin 算法构造概念格。

基于表2萝藦科植物形式背景，概念格构造过程如下。

步骤1 根据表2，列出所有属性的对象空间和对象的属性空间:

D(草本植物) = {牛皮消,白薇,徐长卿,一枝香,萝藦}　D(藤本植物) = {夜来香}

D(止咳) = {白薇}　D(祛风湿) = {徐长卿,一枝香}　D(圆锥花絮) = {徐长卿,一枝香}

D(聚伞花序) = {牛皮消,夜来香}　D(伞形花序) = {夜来香}　D(总状花序) = {萝藦}

R(牛皮消) = {草本植物,聚伞花序}　R(白薇) = {草本植物,止咳}　R(徐长卿) = {草本植物,祛风湿,圆锥花絮}

R(一枝香) = {草本植物,祛风湿,圆锥花絮}　R(萝藦) = {草本植物,总状花序}　R(夜来香) = {藤本植物,聚伞花序,伞形花序}

步骤2 分析 $D(m)$ 之间的关系和 $R(g)$ 之间的关系，得到:

1) D(祛风湿) = D(圆锥花絮);

2) D(藤本植物) = D(伞形花序);

3) R(徐长卿) = R(一枝香)。

步骤3 根据 $D(m)$、$R(g)$ 间的关系结果以及定理1、定理3，约简“圆锥花絮”、“伞形花序”所在的列及“一枝香”所在的行，得到的约简形式背景如表3所示。

表3 约简的萝藦科形式背景

Tab. 3 The reduction formal context of asclepiadaceae plants

I(关系)	草本植物	藤本植物	止咳	祛风湿	聚伞花序	总状花序
牛皮消	×				×	
白薇	×		×			
徐长卿	×			×		
萝藦	×					×
夜来香		×			×	

步骤4 采用 Godin 算法对表3的形式背景构造概念格，其结果如图1所示。

图1的 Hasse 图是一个简化的概念格视图，实际上每个节点都包含一个对象集和一个属性集，每个节点的对象集合由该节点下所有子节点中出现的对象集构成，而每个节点的属性集合则由该节点的所有父节点中出现的属性集构成。

由比较表2和表3结果可知，构造概念格时，运算次数由原来的 $2^6 + 2^8$ 次减少到 $2^5 + 2^6$ 次，即运算效率提高70%。

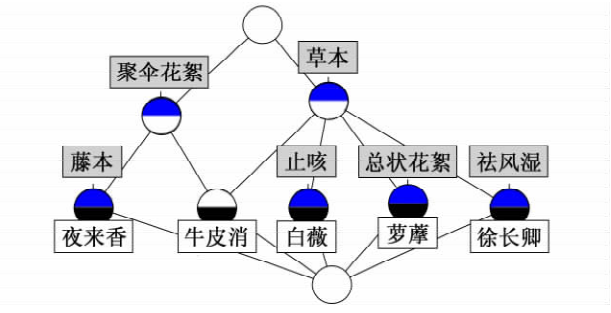

图1 约简概念格

Fig. 1 The reduction concept lattice

2.2 约简概念格的修复

为了保持概念格约简前后的同构性，笔者根据2.1节采用的定理1、定理3得出如下概念格修复定理。

定理5(直接添加约简项法) 对于形式背景 (G, M, I)，如果采用 $R(g_i) = R(g_j)$，$D(m_i) = D(m_j)$ 的方式约简 g_j 所在的行和 m_j 所在的列，则在修复概念格时直接将 g_j, m_j 添加到约简格中所有格节

点中。

根据概念格修复定理，对图 1 的约简概念格进行修复，即将“圆锥花絮”直接添加到约简概念格中内涵含有“祛风湿”的格节点中。将“伞形花序”直接添加到约简概念格中内涵含有“藤本植物”的格节点中。将“一枝香”直接添加到约简概念格中外延含有“徐长卿”的格节点中。修复后完整的概念格如图 2 所示。

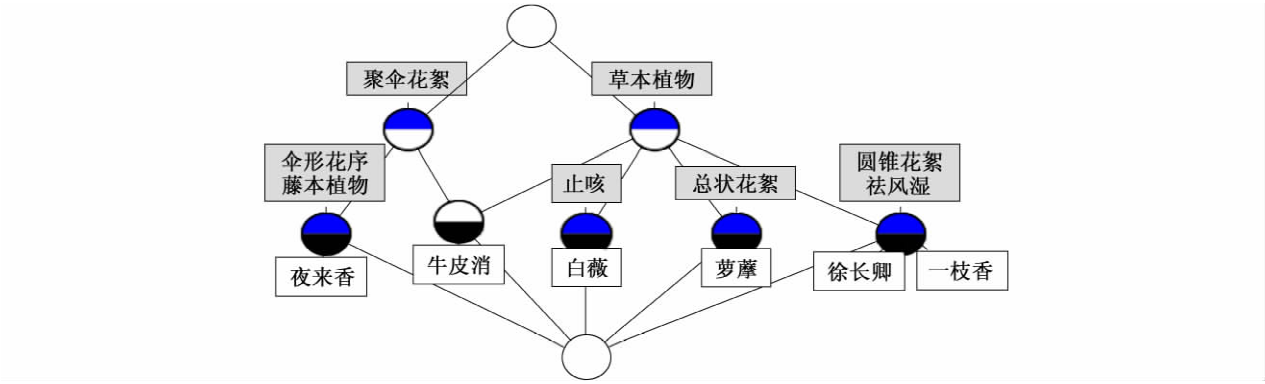

图 2　萝藦科植物完整概念格

Fig. 2　The whole concept lattice of asclepiadaceae plants

3　概念格到中文领域本体的映射

由定义 4 可知，概念格是一种概念聚类的过程，体现了概念间的层次关系，即超概念是子概念的泛化，相应的子概念是超概念的特化，因此，将概念格中的每个概念节点映射成本体中的一个 Class，层次关系映射成本体中的 subClassOf 关系；概念格节点的外延表示该格节点所包含的对象，相当于类的实例，因此映射成 Class 的 Individual；概念格节点的内涵表示该格节点中的对象共同拥有的属性，相当于类的属性，因此将该节点的内涵映射成 Class 的 DatatypeProperty。

采用上述映射规则，实现萝藦科植物概念格到萝藦科植物本体的映射，并使用斯坦福大学开发的本体编辑器 Protégé 中的 OntoGraf 插件对构建的中文领域本体进行图形化描述(见图 3)。

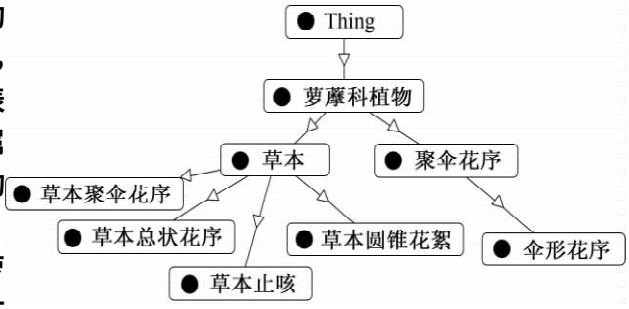

图 3　萝藦科植物本体

Fig. 3　The asclepiadaceae plants ontology

4　结　语

笔者提出一种融合概念格约简的中文领域本体学习方法，并以多篇有关萝藦科植物及其药用性的中文文本为数据源，应用该方法进行中文领域本体学习，最终得到了萝藦科植物本体，并用可视化的方式描述。在学习过程中，采用语义依存分析技术获取形式背景，基于概念格约简理论构造概念格，然后根据映射规则实现概念格到领域本体的映射。将 FCA 应用于本体构建，既能发挥形式概念分析自动客观提取语义的特点，又能使构建的本体在知识重用和知识共享等应用层面上比单纯的分类法具有更大的优势。将概念格约简理论应用于概念格构造，实验验证可提高概念格构建效率。实验结果表明，该本体学习方法能提高本体构建的效率和形式化程度。但仍有需要改进的地方，下一步会设计更明确的映射算法实现概念格到本体的映射，减少人工参与，进一步提高本体学习的自动化程度。

参考文献:

[1] WILLE R. Restructuring Lattice Theory: An Approach Based on Hierarchies of Concept [C] //ICFCA 2009. Berlin: Springer-

Verlag, 2009: 314-339.

[2] 欧阳纯萍, 胡长军, 李扬, 等. 一种基于FCA的面向关系数据库的本体学习方法 [J]. 计算机科学, 2011, 12(12): 167-171.
OUYANG Chun-ping, HU Chang-jun, LI Yang, et al. Approach of Ontology Learning from Relational Database on FCA [J]. Computer Science, 2011, 12(12): 167-171.

[3] OBITKOM, SNÁELV, SMID J. Ontology Design with Formal Concept Analysis [EB/OL]. [2010-10-08]. http://ftp.information.rwth-aachen.de/Publications/CEUR-WS/Vol-110/paper12.pdf.

[4] 张斌, 刘增良, 余达太, 等. 基于形式概念分析与统计理论的本体构建模型 [J]. 计算机应用研究, 2011, 28(1): 111-113.
ZHANG Bin, LIU Zeng-liang, YU Da-tai, et al. Ontology Construction Model Based on Formal Concept Analysis and Statistical Theory [J]. Application Research of Computers, 2011, 28(1): 111-113.

[5] PENG Qiang-qiang, DU Ya-jun, HAI Yu-feng, et al. Topic Specific Crawling on the Web with Concept Context Graph Based on FCA [C]//Proc of Int Conf on Management and Service Science. Piscataway, NJ: IEEE, 2009: 1-4.

[6] SHYNG J Y, SHIEH H M, TZENG G H. An Integration Method Combining Rough Set Theory with Formal Concept Analysis for Personal Investment Portfolios [J]. Knowledge-Based System, 2010, 23(6): 586-597.

[7] SNASEL V, HORAK Z, KOCIBOVA J, et al. A Analyzing Social Networks Using FCA: Complexity Aspects [C]//Proc of 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence(WI) and Intelligent Agent Technologies(IAT). Piscataway, NJ: IEEE, 2009: 38-41.

[8] LI Yang, YANG Xu. Decision Making with Uncertainty Information Based on Lattice-Valued Fuzzy Concept Lattice [J]. International Journal of General Systems, 2010, 39(3): 235-253.

[9] 张文修, 魏玲, 祁建军. 概念格的属性约简理论与方法 [J]. 中国科学E辑: 信息科学, 2005, 35(6): 628-639.
ZHANG Wen-xiu, WEI Ling, QI Jian-jun. The Theory and Method of Concept Lattice Attributes Reduction [J]. Science in China Series E: Technological Sciences, 2005, 35(6): 628-639.

[10] 杨丽, 徐扬. 基于形式背景的概念格约简及其修复 [J]. 计算机工程, 2008, 34(9): 22-24.
YANG Li, XU Yang. Concept Lattice Reduction and Reparation Based on Formal Context [J]. Computer Engineering, 2008, 34(9): 22-24.

[11] 侯丽鑫, 郑山红, 赵辉, 等. 基于P-集合和FCA的中文领域本体学习方法 [J]. 吉林大学学报: 理学版, 2013, 51(4): 659-665.
HOU Li-xin, ZHENG Shan-hong, ZHAO Hui, et al. Research on Chinese Domain Ontology Learning Based on P-Sets and Formal Concept Analysis [J]. Journal of Jilin University: Science Edition, 2013, 51(4): 659-665.

[12] 黄美丽, 刘宗田. 基于形式概念分析的领域本体构建方法研究 [J]. 计算机科学, 2006, 33(1): 210-212.
HUANG Mei-li, LIU Zong-tian. Research on Domain Ontology Building Methods Based on Formal Concept Analysis [J]. Computer Science, 2006, 33(1): 210-212.

[13] CHE Wan-xiang, LI Zheng-hua, LIU Ting. LTP: A Chinese Language Technology Platform [C]//Proceedings of the Coling 2010: Demonstrations. Beijing, China: [s.n.], 2010: 13-16.

[14] BAIXERIES J, SZATHMARY L, VALTCHEV P, et al. Yet a Faster Algorithm for Building the Hasse Diagram of a Concept Lattice [C]//ICFCA 2009. Berlin, Germany: Springer-Verlag, 2009: 162-177.

[15] 蒋义勇, 张继福, 张素兰. 基于链表结构的概念格渐进式构造 [J]. 计算机工程与应用, 2007, 43(11): 178-180.
JIANG Yi-yong, ZHANG Ji-fu, ZHANG Su-lan. Incremental Construction of Concept Lattice Based on Linked List Structure [J]. Computer Engineering and Applications, 2007, 43(11): 178-180.

(责任编辑: 刘俏亮)