基于语义 Web 的受控自然语言系统推理模型

王缓缓[1,2] 李 虎[3] 石 永[3,4]

(华中科技大学管理学院 武汉 430074)[1] (三峡大学计算机与信息学院 宜昌 443002)[2]
(华中科技大学图像识别与人工智能研究所 武汉 430074)[3] (武汉通信指挥学院 武汉 430070)[4]

摘 要 虽然相关研究组织提供了语义 Web 的一些简化工具,但是对不具备相关背景知识的领域专家来说,语义 Web 的可用性较低。提出了基于语义 Web 的受控自然语言系统推理模型,以解决这个问题。首先给出受控自然语言系统推理模型框架;然后分析受控自然语言的语言处理部分,提出基于 WordNet 的受控自然语言系统的本体词库模型和基于本体词库的受控自然语言解释器,把受控自然语言转换成中间表达语言篇章表述结构;最后通过推理部分把篇章表述结构转换成语义 Web 的本体和规则,通过模板工具映射成 Jess 的事实和规则,根据预定义的语义 Web 的公理和定理对受控自然语言进行推理。试验证明此模型大大提高了知识表示建模的效率,也基本满足简单推理任务,具有实用价值。

关键词 受控自然语言,语义 Web,WordNet,Jess

中图法分类号 TP391.1 **文献标识码** A

Semantic Web Based Reasoning Model for Controlled Natural Language System

WANG Huan-huan[1,2] LI Hu[3] SHI Yong[3,4]

(School of Management, Huazhong University of Science & Technology, Wuhan 430074, China)[1]
(College of Computer and Information Technology, China Three Gorges University, Yichang 443002, China)[2]
(Institute of Pattern Recognition & Artificial Intelligence, Huazhong University of Science & Technology, Wuhan 430074, China)[3]
(Wuhan Communication Command Academy, Wuhan 430070, China)[4]

Abstract Proposed a semantic Web based reasoning model for controlled natural language system which can be used by experts to describe business logic in controlled natural language. The framework of the reasoning model includes two parts, one is language processing part and the other is reasoning part. Firstly, the controlled language sentences were partially made into discourse representation structure through ontology lexicon model based on the WordNet and the controlled natural language interpreter based on ontology lexicon model. Then discourse representation structure was transformed into semantic Web OWL and SWRL through the reasoning part. Finally, semantic Web OWL and SWRL was mapped into Jess facts and rules through template tools, and then reasoning was conducted on controlled natural language according to a predefined axiom of semantic Web. Experiment proves that this mode has greatly enhanced the efficiency of knowledge representation modeling, can basically meet the simple reasoning tasks, and has practical value.

Keywords Controlled natural language, Semantic Web, WordNet, Jess

1 引言

美国 Stanford 大学人工智能研究中心尼尔逊教授将人工智能定义为:"人工智能是关于知识的学科——怎样表示知识以及怎样获得知识并使用知识的科学"[1]。可见知识的表示和获取是人工智能科学的基础研究,是构建强大人工智能系统的关键所在。本体和规则是语义 Web[2] 上知识表示与推理的重要组成部分。目前有大量的语言比如 RDF, OWL, SWRL, RuleML, SPARQL 等被定义出来以满足语义 Web 的研究,另外有 Protégé[3] 等工具用于定义语义 Web 本体库,也有 Pellet[4] 等推理算法用于语义 Web 的推理。然而,知识特别是领域知识由领域专家掌握,但是现实当中领域专家不具有语义 Web 等相关背景知识,也不会使用语义 Web 相关复杂的工具。如果要构建领域内的知识库系统,则需要领域专家通过自然语言的形式描述知识形成文档,我们称之为知识规格说明书,然后知识库构建专家把业务知识转换成计算机所能接受和处理的事实和规则库,整个过程我们称之为知识转换。

在查阅语义 Web 及受控自然语言大量文献的基础上,本文提出了基于语义 Web 的受控自然语言系统推理模型。受控自然语言(Controlled Natural Language, CNL)是自然语言的一个子集,以语言学、逻辑学、知识分类理论、心理学和信息学等为理论基础,在一个领域内通过限制 CNL 的词库、语法及意义,达到减少或者消除语言的歧义性和复杂性的目的,从

到稿日期:2010-03-08 返修日期:2010-06-09 本文受国家自然科学基金(60972081)资助。

王缓缓(1978—),女,博士生,讲师,主要研究方向为信息管理、管理系统模拟;李 虎(1978—),男,博士生,主要研究方向为自然语言处理、模式识别等,E-mail:2002_ahu@126.com(通信作者);石 永(1978—),男,博士生,讲师,主要研究方向为人工智能、信息安全等。

而提供了自然语言的应用性。相比目前的语义 Web 工具，本文提供的方法对不具备语义 Web 及相关工具背景知识的领域专家更为实用。除了本文提出的模型外，其他研究机构也有类似的研究工作，比如 ACE[5] 提供了 APE、Race 等 CNL 定义和推理工具；Sydney OWL Syntax[6] 提供了 CNL 到 OWL 的双向映射；Rabbit[7] 提供了通过 CNL 定义本体的方法；Lite Natural Language[8] 映射 CNL 到 DL-Lite。相比上面的实现，本文采用基于 WordNet[9,10] 的本体词库和基于本体词库的 CNL 解释器，通过中间语言篇章表述结构（Discourse Representation Structure，DRS[11]）来表示 CNL 语义信息，然后通过推理部分把篇章表述结构转换成语义 Web 的本体和规则，通过模板工具映射成 Jess 推理机的事实和规则，根据预定义的语义 Web 的公理对受控自然语言进行推理，本文定义的模型具有更好的实用价值。

2 CNL 系统模型框架

本文提出的基于语义 Web 的受控自然语言系统推理模型如图 1 所示。此模型由两个模块组成：语言处理模块和推理模块。这两个模块通过中间表示语言 DRS 关联起来。DRS 基于本体在较高层次表示受控自然语言语句的语义，在此处自然语言语句转换成中间表达式语言逻辑语句，然后转换成本体和规则知识，经过推理引擎运行后产生对应的输出。采用此模型的好处是从逻辑上和物理上分离了语言处理部分和推理处理部分，为此系统移植到其他系统提供了概念上的保障。

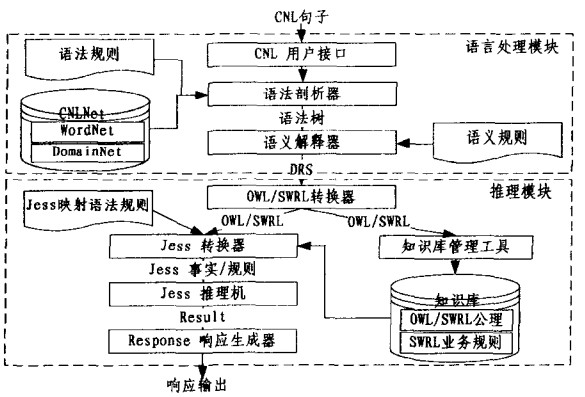

图 1 基于语义 Web 的受控自然语言系统推理模型

语言处理模块主要负责语言本身的处理，包括切词、单词变体处理、句法分析和语义解释等。此模块提供友好的用户接口，用户通过此接口输入查询语句；句法解析器根据句法语法和本体词库解析受控自然语言语句，生成语法树；然后语义解释器根据本体词库和语义语法解释语法树，生成中间表示语言 DRS。DRS 表达式是语言处理模块的最终输出，DRS 包含后面数据库操作部分的所有语义信息。

逻辑推理模块负责进行推理工作。此模块接收语言处理模块的输出 DRS，通过模板技术转换成语义 Web 定义的 OWL 本体和 SWRL 规则，然后通过 Jess 推理引擎进行推理，得到推理结果，并格式化推理结果返回给用户。

3 CNL 语言处理模型

3.1 基于 WordNet 的本体词库模型

CNL 词库一般按照以下几个原则进行构建：无歧义性，词库中的词汇应该无歧义并且机器可以处理；属于领域内词库，词库中的词汇必须是属于领域范围内的词汇；词汇必须标识，词库中的词汇必须进行标识，包括词性、词义、名词和动词的数、性别等。通常在定义领域词库时，会选用一个通用词库作为基本词库，然后在基本词库上进行扩充，定义领域词库。目前有很多类型的通用词库资源可以选择，比如 WordNet，VerbNet，Brown，Corpus，Oxford Dictionary 等等。本文选用 WordNet 为通用词库，并在 WordNet 基础上扩展领域词库。WordNet 不同于其他传统词库，它是一个面向语义的英语词库，单词按照一定的语义组成一个语义网络。本文按照 WordNet 组织结构定义领域内词库，并利用 WordNet 的同义、反义、上下位等关系与 WordNet 进行关联。这样 CNL 系统能够利用 WordNet 的语义关系识别用户输入的非领域词库的单词，并将其映射到领域词库中的单词，从而提高 CNL 的应用范围。比如，对 CNL 查询语句“Who is the writer of the book ‘The Old Man and the Sea’?”和“Who is the author of the book ‘The Old Man and the Sea’?”应该得到相同的查询结果。我们可以定义 writer 和 author 为同义词，用同一个 Synset Id 表示，在语法分析时通过简单的同义映射得到同义 DRS 输出，从而实现查询结果相同。

基于 WordNet 的 CNL 词库模型（如图 2 所示）由 4 模块组成，即词库、词库解析器、词库通用访问接口和 CNL 系统。词库由 3 部分组成，分别是实义词（名词、动词、形容词、副词）、功能词（冠词、限定词、介词、代词）和变体词（功能词的非规则变体）。其中，实义词又划分为两类：WordNet 实义词和领域实义词。WordNet 实义词直接引用 WordNet 词库，可以通过工具导入 CNL 词库模型。DomainNet 是在 WordNet 的数据结构上根据 CNL 需求进行扩展定义的领域内词库模型。通过一个关联模型连接 WordNet 和 DomainNet。此模型框架有以下特点：

1）采用 WordNet 为基础词典，并以 WordNet 结构定义领域词典。WordNet 是一个面向语义的英语词典，和传统的词典相仿，但是提供了丰富的语义层次关系，包括同义关系、反义关系、整体与部分和上下位关系等等，这些关系把词和词关联起来，形成一个网络。在语义分析中可以充分利用这些关系，将用户输入的单词映射到领域内的概率，以提供单词的识别率，从而可以扩大 CNL 的实用性；

2）提供统一的词库解析算法，独立词库操作为单独的模块，不依赖于具体的 CNL 系统来提高系统的可移植性；

3）定义统一的词库访问接口，对不同的 CNL 系统，具体实现词库访问接口，然后通过接口连接 CNL 系统和词库部分。

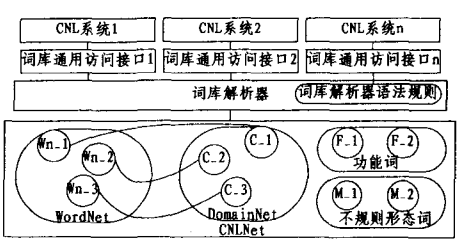

图 2 基于 WordNet 的 CNL 本体词库模型

3.2 CNL 剖析器

在目前自然语言处理中，主要采用两种方式对自然语言

进行剖析，一是基于规则的上下文无关文法(CFG[12,13])，主要代表有 Chomsky 的上下文无关语法；二是基于概率统计模型上下文无关文法的剖析算法，给上下文无关语法的规则加上概率，提出了概率上下文无关语法(PCFG)，其除了给规则加概率之外，还考虑规则的中心词对于规则概率的影响，提出了概率词汇化上下文无关语法(PLCFG)，主要代表有斯坦福的 Stanford Parser[14]。本文由于采用基于 WordNet 的本体领域词库，对领域应用的词汇量已经做了限制。如果采用通用的 PCFG 或 PLCFG 剖析器，就不能发挥词库受限的优势。鉴于上述原因，本文提出了基于概率本体词库化上下文无关语法(POLCFG)。POLCFG 通过两个方面的约束来增强 PLCFG 的剖析算法，一是受限的词库，对每一个已经明确了其词性，降低了单词词性的歧义性，二是通过本体概念关联度，由词与词之间的依赖程度进一步降低单词的歧义性。POCFG 算法可以用一个六元组表示：

$$G=\{N,\sum,P,S,D,O\}$$

式中，D 是给每一个规则指派概率 p 的函数，即

$$p(A\rightarrow\beta)$$

O 表示剖析树的每一个都结点要标上该结点的中心词概率的函数，即

$$P(w_1,tag)=P(w_1,pos)D(w_1,w_2)$$

这个概率函数由两部分组成，一个是单词属于什么词性的概率，表示为：

$$P(w_1,pos)=\frac{1}{count(w_1,pos)}$$

二是本体概念对的关联度，关联度与概念之间的距离成反比，距离越近，表示关联度越强，反之，距离越远，关联度越弱。本体概念对的关联度表示为：

$$D(w_1,w_2)=\frac{\partial}{distance(w_1,w_2)+\partial}$$

式中，∂ 是一个可调节参数。

英语句子首先经过 PCFG 剖析处理后，会得到两个树，一个是 PCFG 语法树结构 T，另外一个中心词依赖树 D，表示词与词之间的依赖关系。两个树合并就得到词汇化语法树 L，语法树 L 上的每一个节点都包含标记、词性和中心词等信息。

词汇化语法树 L 可以用公式 $L=(T,D)$ 表示。这样，将某个非终极符号重写为终极符号 β 符号串的概率。由于规则指派概率和本体概念关联度共同决定 $P(T,O)$，我们假定中心词依赖树概率 $P(D)$ 与 PCFG 语法树结构 $P(T)$ 无依赖，规则指派最终概率可以表示为：

$$P(T,D)=P(T)P(D)$$

后续步骤可以采用 PLCFG 算法继续处理得到词汇化语法树 L 的概率。概率最大的 L 就是句法剖析的最后结果。POCFG 是基于概率统计的标注算法，其把词汇概率和本体概念的相关性引入到剖析算法中，提高了剖析算法的准确率。

3.3 CNL 解释器

受控自然语言的句子经过前面的句法剖析得到一个语法树结构数据，语法树中的每个节点都具有相应的特征信息。我们可以采用一种自顶向下的下推自动机算法结合分类特征形式(Sotred Feature Formalisms)描述语法对语法树进行解释。分类特征形式(Sotred Feature Formalisms)被大量使用在大规模语言处理中，比如 TDL(Krieger and Schafer,1994)，LIFE(Ait-Kaci,1998)，但是这些实现都是针对一定的应用领域，并不适合本文的使用。在前人研究的基础上本文提出一种基于分类特征的语法树解释模型，并通过 ANTLR 工具，根据语法树解释器的语法自动生成此语法的解析器，从而实现该语法，完成对语法树的解释工作。

CNL 句子经过 CNL 剖析器生成语法树，语法树经过语义解释，输出中间表述语言 DRS。DRS 中只包含了句子的语义信息，摒弃了句子中的语法结构信息。在语言中同样语义的句子可以采用不同的形式表示，这些具有同样语义的句子可以得到同样的 DRS 数据。CNL 解释器处理过程可以表示为：

$$AST=SP(s,G)$$

$$DRS=SI(AST,I)$$

式中，s 为用户输入的句子，G 为句法处理文法，SP 表示句法剖析算法，I 表示解释处理文法，SI 表示解释算法，DRS 为输出结果。

CNL 句子映射成 DRS，实义词映射成 DRS 的原子条件，功能词映射成 DRS 的逻辑连接词、DRS 条件参数或共享变量。CNL 句子比对应的 DRS 有更丰富的表达式，且实验证明这样的映射是确实可行的。对名词、不定代词、专有名词、及物动词、系动词、of 结构遵循表 1 的映射规则。

表 1 DRS 映射规则

词性	DRS 元素
可数名词	object(Ref,Word,countable,na,QType,QNum)
不定代词	object(Ref,something,dom,na,na,na)
专有名词	object(Ref,Word,named,na,eq,1)
及物动词	predicate(Ref,Word,SubjectRef,ObjectRef)
系动词	predicate(Ref,be,SubjectRef,ObjectRef)
介词 of	relation(OwnedRef,of,OwnerRef)

表中 Ref，SubjectRef，ObjectRef，OwnedRef，OwnerRef 指的是 DRS 指示，Word 是对应的单词，QNum 表示数量，QType 表示比较(eq，geq，leq，greater，less)。比如英语句子“Every man is a human.”可以映射成下面的 DRS：

```
[]
    [A]
    object(A,man,countable,na,eq,1)
    =>
    [B,C]
    object(B,human,countable,na,eq,1)
    predicate(C,be,A,B)
```

4 CNL 推理模型

4.1 语义 Web 推理机模型

推理模型是本文提出的基于语义 Web 的 CNL 系统推理模型的推理部分，由 6 个部分组成：

1)OWL/SWRL 转换器：把 DRS 转换成对应的 OWL 本体或 SWRL 规则；

2)Jess 转换器：把 OWL 本体或 SWRL 规则转换成 Jess 规则引擎能够处理的 Jess Fact 或 Rule；

3)知识管理工具：对知识库中的知识进行维护，包括增加、删除、修改、查找以及一致性检查；

4)Jess 推理机：Jess 规则引擎推理机，本文采用 Jess 规则引擎实现对 OWL 和 SWRL 的推理；

5)Response 生成器：输出产生器，在本模型中定义输出

的接口，在不同的应用中根据业务需求有不同的实现；

6)知识库：存储 CNL 推理模型中的知识，这些知识包括由受控自然语言表述的业务规则、OWL 事实及公理、SWRL 公理及其他预定义的业务知识。

CNL 推理模型的输入是语言处理部分的输出 DRS，输出分为两种，一种是知识库中的业务规则，另一种是推理结果。为了得到上述两种输出，分别对应两种运行模式：训练模式和推理模式。

训练模式是接收语言处理模块的输出 DRS；然后通过 OWL/SWRL 转换器把 DRS 转换成 OWL 事实和 SWRL 规则；再通过 Jess 生成器转换成 Jess 对应的 fact 和 rule；最后通过知识管理工具把 Jess 的 fact 和规则保存到知识库中，以备推理时使用。

推理模式同样是接收语言处理模块的输出 DRS；然后 OWL/SWRL 转换器把 DRS 转换成 OWL 事实和 SWRL 规则；再通过 Jess 转换器转换成 Jess 对应的 fact 和 rule；Jess 推理机根据知识库中的知识和受控自然语言描述的事实和规则进行推理，产生推理结果；最后通过 Response 生成器进行格式化输出。

4.2 语义 Web 知识转换器

受控自然语言经过语言处理部分处理后，输出中间表达语言 DRS，DRS 带有本体信息及本体之间的关系，包含了受控自然语言所表述的所有知识信息。DRS 是一种文集语义描述语言，因此可以对其进行推理。目前没有现成的对 DRS 进行推理的工具，DRS 需转换成其他知识表述形式，然后再进行推理。本文由于采用语义网作为知识表示，因此 DRS 首先得转换成语义网的 OWL 本体和 SWRL 规则。而作为 SWRL 的本体组件，OWL 自身也存在着不够完善的部分，尤其是 OWL 推理实现一直在研究中。当然，以描述逻辑 DL 为逻辑基础的 OWL 可以借助 DL 推理机，但至今 DL 推理机仍不擅长处理大规模数据。本文采用 Jess 规则引擎对语义网的本体和规则进行推理，所以语义网的 OWL 本体和 SWRL 规则在推理前需转换成 Jess 对应的事实和规则。

另外，我们基于 WordNet 定义了领域本体词库 DomainNet，这在推理过程中也需要用到，所以需要把 DomainNet 转换成 OWL 存储到知识库中，并转换成 Jess 的 fact 事实，以便推理使用。

知识转换过程如图 3 所示。输入的 DRS 及 DomainNet 数据经过 DomainNet2SW 和 DRS2SW 转换器转换后，生成中间数据 OWL 和 SWRL，然后再通过 SW2Jess 转换成对应的 Jess 推理机的 fact 和 rule。

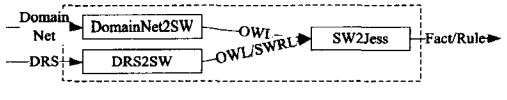

图 3 知识转换过程

结束语 为了对本文提出的推理模型进行实验，本文进行了原型实现，命名为 ORCNLR。本文收集了 25 个简单的可推理的应用案例，这些案例基本覆盖语义 Web 本体和规则对应的各种语义形式。按照案例所表示知识的复杂程度(自然语言描述所需句子的数量)划分为 5 组，每组 5 个测试案例，可以表示为 $C=\{CG_1, CG_2, CG_3, CG_4, CG_5\}$，其中每一组 $CG_i=\{c_{i1}, c_{i2}, c_{i3}, c_{i4}, c_{i5}\}$。每个实验案例分别采用 ORCNLR，ACE 和 Protégé 进行知识库实现。另外，测试用户分 2 组进行，每组有 4 个用户。其中第一组测试用户表示为 $G_1=\{u_{11}, u_{12}, u_{13}, u_{14}\}$，这组用户没有知识表示理论背景，也不会使用 Protégé 工具；第二组用户表示为 $G_2=\{u_{21}, u_{22}, u_{23}, u_{24}\}$，这组用户具有知识表示理论背景，会熟练使用 Protégé 工具。每个测试人员分别使用上述 3 个工具对测试案例进行实现，统计每组测试用户实现案例的正确性率 r 和所需平均时间 t。其中正确率 r 的计算公式如下：

$$r_{ij}=\frac{\text{测试正确数}}{G_i\text{ 用户数}\times CG_j\text{ 测试案例数}}\rightarrow[0..1]$$

所需平均时间 t 的计算公式如下：

$$t_{ij}=\frac{\sum\text{每个测试案例花费时间}}{G_i\text{ 用户数}\times CG_j\text{ 测试案例数}}$$

在实验之前，需要通过词库构建工具来构建领域本体词库和扩展 ACE 词库，为实验做准备。试验结果如表 2 所列。

表 2 测试案例实验结果

测试案例		G_1			G_2		
		ORCNLR	ACE	Protégé	ORCNLR	ACE	Protégé
CG_1	r(%)	100	100	0	100	100	100
	t(s)	87	154	0	92	146	323
CG_2	r(%)	100	100	0	100	100	100
	t(s)	128	235	0	124	241	466
CG_3	r(%)	100	95	0	95	90	100
	t(s)	224	312	0	218	332	580
CG_4	r(%)	95	80	0	90	80	100
	t(s)	386	503	0	497	494	836
CG_5	r(%)	90	60	0	85	65	95
	t(s)	542	733	0	535	751	1145

从表 2 可以很明显地看出对于没有知识表示背景并且不熟悉 Protégé 工具的测试用户，没有办法通过 Protégé 定义知识库系统的本体和规则，而此类测试可以使用 ORCNLR 和 ACE 进行知识库构建；ORCNLR 和 ACE 对复杂程度不高的案例都具有较高的正确率，但是在复杂程度达到一定程度时，ORCNLR 和 ACE 在正确率方面都有所降低，且 ORCNLR 的降低幅度比 ACE 要低；Protégé 是专门用于语义 Web 的本体和规则开发工具，对于有知识表示背景和能熟练使用 Protégé 工具的用户，其准确率高于 ORCNLR 和 ACE。

试验证明本文提出的受控自然语言推理模型大大提高了知识表示建模的效率，准确率高于同类系统 ACE 的方法，基本满足简单推理任务。没有正确推理的案例是由于在描述事实和规则时漏掉了一些条件，因此导致推理失败。由于 CNL 句子表达方式的局限性，对复杂逻辑推理的案例正确率偏低。

参 考 文 献

[1] Nilsson N J. Artificial Intelligence: A New Synthesis[M]. San Francisco: Morgan Kaufmann, 1998: 1-17

[2] Berners-Lee T. Semantic Web Road Map [OL]. http://www.|w3.org/ DesignIssues/Semantic.html. 1998

[3] Noy N F, Fergerson R W, Musen M A. The Knowledge Model of Protégé-2000: Combining Interoperability and Flexibility[C]// Proceedings of the 12th EKAW Conference. Berlin: Springer-Verlag, 2000: 17-32

[4] Sirin E, Parsia B, Grau B C, et al. Pellet: A practical OWL-DL reasoner[J]. Journal of Web Semantics, 2006, 5(2): 51-53

(下转第 240 页)

如某个孤立的事件、某个突然出现的情况等，实际应用时，可通过判断社区的大小来确定。若某剩余节点数不符合社区要求，就不认为它们构成一个社区，只有规模达到一定程度，才认为其构成社区。

4 示例

如图3所示，销售网络的最大熵为0.529，网络的总熵为7.48，最小熵为0.113，熵变化的平均值为0.356，与最大熵的差为0.463，图中边(ab,bf,fa,fg,gb,ga,bc,cj,jb,ce,eb,eh,hl,lk,ki,ih,im,cd,db,da,ca,fk)及其对应的熵分别为(0.36,0.113,0.152,0.113,0.152,0.332,0.442,0.529,0.006,0.152,0.292,0.113,0.152,0.113,0.113,0.152,0.006,0.387,0.113,0.113,0.006,0.006)，其值的变化及趋势如图2所示。GN算法发现的社区包含节点$abcdefg$,$hijklm$。而CDBE算法发现的社区为$abcjdg$,$efhlkmi$，如图3、图4所示，可见两种算法发现的社区有差异，CDBE着重于节点关联关系，而GE算法着重于节点连接的介数，在实际的运行过程中，把节点熵与最大节点熵的差作为熵的变化，如果该熵变化小于所有节点熵的平均值，则认为该节点属于社区，否则不属于该社区，本示例只运行算法一次，得到两个社区，如果把没有标识的节点继续细分，可以得到多个社区，当然应该考虑节点数是否符合社区要求，如果节点数很少，不认为构成一个社区。而CDBE算法结合商品销售的实际情况，每个社区内成员关联性很强，可以通过熵来度量，传统KDD算法一次运行其支持度、满意度是固定的。另外，假设节点中节点数为n，K-N算法的运行时间复杂度为$O(n^3)$，K-L算法的运行时间复杂度为$O(n^2)$，层次聚类法的时间复杂度为$O(n^2\mathrm{Log}^n)$，GN算法的时间复杂度为$O(m^2n)$。本文算法的时间复杂度最好的情况为$O(n^2)$。

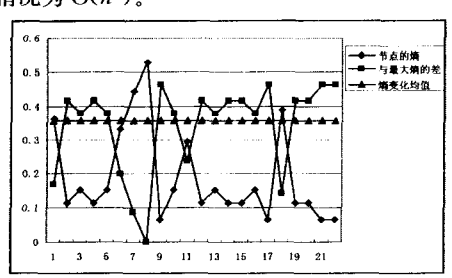

图2 熵，平均熵及熵的变化

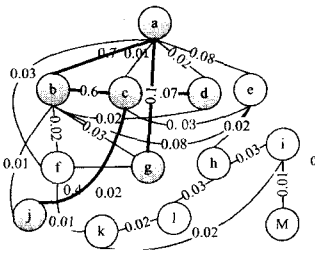

图3 本文发现的社区

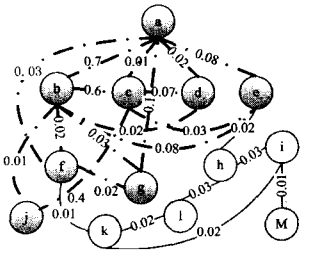

图4 GN算法发现的社区

结束语 根据商品销售记录，构建商品销售记录网络，提出了基于熵的社区划分方法，该方法既注重社区网络的结构，也注重社区网络连接的含义，即社区内熵的变化不会剧烈增加或减少。以前其它算法没有关注到这一点，与基于数据库的关联规则挖掘相比，该方法避免了指定支持度和满意度的不足，通过发现销售商品社区，从而进行推荐。将来的研究将进一步研究信息熵的理论，研究熵信息变化的规律并应用，以求发现更有意义的社区。

参考文献

[1] 丁元竹. 社区研究的理论与方法[M]. 北京：北京大学出版社，1995

[2] 丁连红，时鹏. 网络社区发现[M]. 北京：化学工业出版社，2008

[3] Girval M, Newman M. Community Structure in Social and Biological Network[C]// Proc. natl. acad. Sci. USA, 2002: 8271-8276

[4] Freeman L. A Set of Measure of Centrality Based Upon Betweens[J]. Sociometry, 1997(40): 35-41

[5] Albert R, Barabasi A L. Statistical Mechanisms of Complex Networks[J]. Reviews of Modern Physics, 2002(74)

[6] 唐鹏，张自力. 基于信息熵的多 Agent DDoS 攻击检测[J]. 计算机科学，2008(3)

[7] 唐敏. 关联规则挖掘算法在超市销售分析中的应用[J]. 计算机科学，2006(2)

[8] 范明，孟小峰. 数据挖掘概念与技术[M]. 北京：机械工业出版社，2002：223-243

[9] 高博，周脩，崔英志. Web2.0 网站的特点与社区化模式[J]. 重庆工学院学报：自然科学版，2009，23(6)：102-106

(上接第190页)

[5] Fuchs N E, Kaljurand K, Kuhn T. Attempto Controlled English for Knowledge Representation[J]. Reasoning Web, Fourth International Summer School. Springer, 2008: 104-124

[6] Cregan A, Schwitter R, Meyer T. Sydney OWL Syntax-towards a Controlled Natural Language Syntax for OWL 1.1[C]// 3rd OWL Experiences and Directions Workshop (OWLED 2007). CEUR Proceedings. volume 258, 2007

[7] Hart G, Johnson M, Dolbear C. Rabbit: Developing a Controlled Natural Language for Authoring Ontologies[C]// ESWC 2008. 2008: 348-360

[8] Bernardi R, Calvanese D, Thorne C. Lite Natural Language[C]// IWCS-7. 2007

[9] Miller G A. WordNet: A Lexical Databas[J]. Communication of the ACM, 1995, 38(11): 39-41

[10] Fellbaum C. WordNet: An Electronic Lexical Database [M]. Cambridge, MA: MIT Press, 1998: 5-23

[11] van E J, Kamp H. Representing Discourse in Context[D]. Handbook of Logic and Language, Elsevier, Amsterdam, 1997: 179-237

[12] Pullum G K, Gerald G. Natural languages and context-free languages [J]. Linguistics and Philosophy, 1982, 4(4): 471-504

[13] BalaSundaraRaman L, Ishwar S, Ravindranath S K. Context Free Grammar for Natural Language Constructs-An implementation for Venpa Class of Tamil Poetry[C]// International Forum for Information Technology in Tamil. 2003: 128-136

[14] Klein D, Manning C D. Fast Exact Inference with a Factored Model for Natural Language Parsing[C]// Advances in Neural Information Processing Systems 15. Cambridge, MA: MIT Press, 2003: 3-10