中文专利文献摘要的知识表示

郑 红,胡思康
(1. 中国专利技术开发公司,北京 100088;2. 北京理工大学 计算机学院,北京 100081)

摘 要:在分析与领域相关的中文专利文献摘要内容和结构的基础上,提出用三元组语义网络表示知识以及知识间的语义关系,并用一阶谓词逻辑分析语义三元组的语义。推导出专利文献知识融合将要面临的问题,包括句法分析后获取的描述性知识的不一致、信息的补足、冗余的发现和模糊信息的处理等。研究成果为后续中文专利文献知识融合分析和推理奠定了基础。

关键词:知识表示;中文专利文献;三元组语义网络;一阶谓词逻辑;本体

中图分类号:TP182 **文献标志码**:A **文章编号**:1672-3767(2015)05-0104-05

Knowledge Representation of Abstracts of Chinese Patent Documents

Zheng Hong[1], Hu Sikang[2]
(1. China Patent Development Corporation, Beijing 100088, China;
2. School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100081, China)

Abstract: Based on the analysis of the contents and structures of domain-related Chinese patent documents, a triple semantic network was put forward to represent knowledge and semantic relations between knowledge, and the first-order predicate logic was used to analyze the semantics of semantic triples. Problems confronting knowledge fusion of Chinese patent documents were deduced, including the inconsistency of descriptive knowledge obtained by syntactic parsing, the supplementation of information, the detection of redundant information and the processing of fuzzy information. The research results will lay foundations for the subsequent knowledge fusion analysis and reasoning of Chinese patent documents.

Key words: knowledge representation; Chinese patent documents; triple semantic network; first-order predicate logic; ontology

知识表示是用给定的知识结构、按照一定原则来组织和表示知识。知识工程借助本体的概念来分析和解决知识表示、知识推理、知识共享等过程中出现的问题。本体表示知识的方法形式不一,常见的包括产生式规则、语义网络、框架、黑板模型以及谓词逻辑等方法。随着互联网的不断发展,许多与 Web 相关的本体知识表示和描述语言相继出现。罗伯特[1]提出概念图(concept graphs)和与之对应的描述逻辑语言。概念图能表达概念和分类关系,并基于一阶谓词逻辑进行推理。资源描述框架[2]被拉斯莱和斯威克用来描述网页资源。曹存根等[3]提出框架和一阶逻辑语言相结合的网页数学知识表示。陆汝钤等[4]采用 Agent 表示基本概念,知识间关系用 OntoNet 定义并进行概念推理。郑红[5]实现了在面向 Agent 知识库上的查询。陆汝钤等[6]还提出形式本体是领域知识共享和重用的基础。

中文专利文献信息具有冗余性、模糊性和不一致性,因而对专利文献采用概念—规则—语义网络的三级分层知识获取和构造体系。知识表示的另一层语义是对知识表示的理解。一阶谓词逻辑借助语义网络进行

收稿日期:2015-08-27

作者简介:郑 红(1972—),女,山西太原人,副研究员,博士,主要从事知识表示、自然语言理解、智能检索等方面的研究.
E-mail: zhenghong@sipo.gov.cn

语义分析，对于专利文献知识获取结果的融合有较强的理论基础和推理能力。因此，本研究基于领域专利文献内容和结构的分析，提出三元组语义网络表示知识以及知识间的语义关系，并用一阶谓词逻辑描述的语义网来分析语义三元组的语义。

1　专利文献文本知识表示

1.1　语义网络

语义网络作为一种知识表示工具，可用于刻画对事物的认识、表达知识及知识间的语义。最简单的语义网络是一个三元组：(节点1，语义弧，节点2)，可用图1表示。其中，A，B 分别代表两个节点，弧 R_{AB} 表示 A 与 B 间的语义联系。

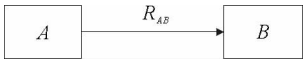

图1　语义网络的基本单元

Fig. 1　Basic unit of semantic net

定义1　三元组语义网络是一个有向图 G，它由节点 A、节点 B 和语义弧 R 构成。其中，节点 A 和 B 或者是领域概念，或者是另一个三元组语义网络 G'。

不难看出，把多个基本网元通过语义弧进行关联，可把独立的基本网元扩展为具有复杂语义表示能力的语义网络。

1.2　三元组语义网络的表示能力

中文专利文献知识通过三元组语义网络来表示，包括单一概念、概念属性和程度信息。

定义2　中文专利文献摘要知识表示的三元组语义网络由如下形式定义：

$$\text{main}:(DC:L0:L1:\cdots:LN,A,B):L0(SC_0,A_0,B_0):L1(SC_1,A_1,B_1):\cdots:LN(SC_N,A_N,B_N)$$

其中，DC 是领域概念，SC_i 是 DC 的子概念(下位语义或属性)，A_j 是对专利文献进行句法分析后获取的目标动词，$L_i(0\leqslant i\leqslant N)$ 是三元组标号，具有唯一性。

为了增强对知识的描述能力，三元组语义网络的知识表示引入一组专用语言成分，主要包括三种描述。

语义描述　描述专利领域概念各侧面信息。如：专利领域概念(*DCName*)、申请日(*Date*)、下位概念(*SCName*)、属性(*Attribution*)等。

关系描述　描述多个概念之间的相互关系，主要包含三方面：

①领域概念间的关系(*Relation*)，包括弱语义相关性和强语义相关性；

②领域概念的相互影响(*Influence*)，主要是概念描述信息的不一致性；

③领域概念属性间的依赖关系(*Dependence*)，主要是概念间属性的信息补足。

程度描述　描述对领域概念或概念属性的程度或能力。包括情态(*Mode*)、程度(*Adj*、*Adv*)以及对上述词的否定(*Not*)等。

定义3　主三元组指一个语义三元组网络中的第一个三元组。

如定义2中的 $\text{main}:(DC:L0:L1:\cdots:LN,A,B)$，其中，main 是主三元组标记。

定义4　子三元组指一个语义三元组网络中除主三元组以外的三元组，是对主三元组领域概念语义节点的下位语义或属性的描述。

在子三元组：$Li(SC_i,A_i,B_i)$ 中，A_i 语义节点表示 SC_i 和 B_i 语义节点的语义连接。子三元组的各语义节点有以下几种情形：

1)连接弧 A 为空语义节点。在知识获取中，主要有两种情况导致 A 为空语义节点：

①在句法分析向三元组语义网络转换过程中，省略连接弧语义，表明节点 B 是对节点 SC 的直接描述；

②目标动词识别错误或专利文献缺少目标动词，造成连接弧语义为空节点。

2)语义节点 B 为空节点，则节点 A 是节点 SC 的语义描述；

3)语义节点 SC 为空节点，则专利文献句法分析获取的知识句缺少或有缺省概念；

4)语义节点 SC 和 A 均为空节点，则节点 B 作为主三元组或前一子三元组的程度描述的补足；

5)语义节点 SC 和 B 均为空节点，则视前一子三元组为双连接弧语义网络。但这导致语义网络出现歧

义，如图2所示。

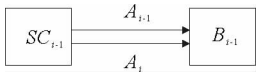

图2 双连接弧语义网络

Fig. 2 Semantic net with double connecting arcs

关于对语义网络歧义的消解分析，将另文探讨。

对上述定义的语义网描述，以下面专利文献摘要自动抽取的文本进行叙述。

本发明公开了一种便捷式电子设备，其包括显示屏、动作感应模块、动作转化模块和输入法模块。

用三元组语义网络表示为：

main:(<*ns*>电子设备:*L*0,包括,: *L*1:*L*2:*L*3:*L*4):*L*0(*,*,<*na*>便携式):*L*1(*,*HAS*,<*na*>显示屏):*L*2(*,*HAS*,<*na*>动作感应模块):*L*3(*,*HAS*,<*na*>动作转化模块):*L*4(*,*HAS*,<*na*>输入法模块)

三元组语义网络的连接关系可以由 n 元关系转换为二元关系的合取表达。即上述三元组语义描述了专利文献的 n 元关系：

ISA(电子设备,便携式) 并且 *HAS*(显示屏,动作感应模块,动作转化模块,输入法模块)

关系描述词 HAS 表示概念与属性间的包含关系，不包括概念间连接关系；同时上述概念属性间亦无依赖关系，则可将其转换为二元关系的合取，即：

$$ISA(\text{电子设备},\text{便携式}) \wedge HAS(\text{显示屏},\text{动作感应模块},\text{动作转化模块},\text{输入法模块}) \xrightarrow{\text{分配律}} ISA(\text{电子设备},\text{便携式}) \wedge HAS(\text{电子设备},\text{显示屏}) \wedge HAS(\text{电子设备},\text{动作感应模块}) \wedge HAS(\text{电子设备},\text{动作转化模块})) \wedge HAS(\text{电子设备},\text{输入法模块}) \xrightarrow{\text{标号限定}} ISA(\text{电子设备},\text{便携式}) \wedge (HAS(\text{电子设备},\text{显示屏}) \wedge HAS(\text{电子设备},\text{动作感应模块}) \wedge HAS(\text{电子设备},\text{动作转化模块}) \wedge HAS(\text{电子设备},\text{输入法模块}))$$

“三元组标号限定”的推导说明“显示屏，动作感应模块，动作转化模块，输入法模块”是对专利领域概念“电子设备”的完整说明，以区别对领域概念的局部描述。

2 三元组语义网络与一阶谓词逻辑

2.1 三元组语义网到一阶谓词逻辑的转换

三元组语义网有丰富的知识表达能力，但同时也导致较复杂的语义推理过程。而一阶谓词逻辑支持目前语义 WEB 的描述逻辑、子属性推理，以及更为复杂的描述结构。因此，本研究考虑把三元组语义分析与一阶谓词相结合。

定义5 假设 φ 是从三元组语义网络到一阶谓词逻辑的转换，当对相同概念 C 的不同三元组语义表示集 TR，$\forall tr \in TR$，如果 $\mathrm{tr}(C)=\varphi(C)$ 成立，则表示 φ 是等价转换。

由于自然语言描述的多样性，穷尽专利文献对同一概念的三元组语义表示是困难的，同时也没有实际意义。因此，对任一个三元组语义网络

TSN:main:(*DC*:*L*0 :*L*1 :…:*LN*,*A*,*B*):*L*0(*SC*$_0$,*A*$_0$,*B*$_0$):*L*1(*SC*$_1$,*A*$_1$,*B*$_1$)…:*LN*(*SC*$_N$,*A*$_N$,*B*$_N$)

按概念属性及程度描述表示，分段为：

TSN':

$$\left.\begin{array}{l}\text{main:}(DC:L0:L1\cdots:LN,A,B):L0(SC_0,A_0,B_0):L1(SC_1,A_1,B_1)\cdots:Li-1(SC_{i-1},A_{i-1},B_{i-1})\\ \text{main:}(DC:L0:L1\cdots:LN,A,B):Li(SC_i,A_i,B_i):Li+1(SC_{i+1},A_{i+1},B_{i+1})\cdots:Lj-1(SC_{j-1},A_{j-1},B_{j-1})\\ \vdots\\ \text{main:}(DC:L0:L1\cdots:LN,A,B):Lk(SC_k,A_k,B_k):Lk+1(SC_{k+1},A_{k+1},B_{k+1})\cdots:LN(SC_N,A_N,B_N)\end{array}\right\} m\text{ 段}$$

根据对专利文献的文本和结构特征分析，能够看出概念 *DC* 在某个侧面上(段)的描述是相对稳定的。

2.2 三元组与一阶谓词的描述能力

针对子三元组语义网络对知识的不同表示语义，有如下假设：

假设1 子三元组(*SC*,*A*,*B*)中，若 *A* 为空语义节点，则缺省连接弧语义为 *ISA*，即(*SC*,*ISA*,*B*)。

假设 2　子三元组(SC,A,B)中，若 B 为空语义节点，则语义网络转变为断言 $A(SC)$，记为 $A(SC,*)$。

假设 3　在子三元组：$L0(SC_0,A_0,B_0):L1(SC_1,A_1,B_1)$中，$SC_1$ 的缺省语义为语义节点 SC_0。

假设 3 说明就近指派原则，以最近概念或属性作为后续子三元组的缺省概念或缺省属性。其对应的语义网络由图 3 所示。

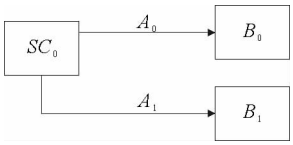

图 3　带缺省语义网络头节点的语义网络

Fig. 3　Semantic net with head node of default semantic net

命题 1　专利文献知识表示的 TSN 能通过 φ 转换为二元关系的一阶谓词逻辑表示。

证明：对任意三元组语义网络 TSN：

$\text{main}:(DC:L0:L1:\cdots:LN,A,B):L0(SC_0,A_0,B_0):L1(SC_1,A_1,B_1):\cdots:LN(SC_N,A_N,B_N)$

用 R 表示领域概念 DC 和各下位概念或属性的语义关系，其中 TSN 领域概念数目为 1，且 $SC_i\neq SC_j$，$0\leqslant i,j\leqslant N$。

由定义 4 可知，TSN 中的子三元组语义节点可能具有空语义。为此，不妨假设最后一个子三元组：$LN(SC_N,A_N,B_N)$具有如下情况：

1)如果无空节点，则记 $P_N=ISA(SC_N,B_N)$；

2)如果 A_N 为空节点，根据假设 1，记 $P_N=ISA(SC_N,B_N)$；

3)如果 B_N 为空节点，根据假设 2，记 $P_N=A_N(SC_N,*)$；

4)如果 SC_N 为空节点，根据假设 3，记 $P_N=A_N(SC_{N-1},B_N)$，SC_{N-1}是对标号为：LN 的子三元组的最近指派；

5)如果 SC_N 和 A_N 均为空节点，则当 $B_{N-1}\neq\neg B_N$ 时，记 $P_{N-1}=A_{N-1}(SC_{N-1},B_{N-1}\cup B_N)$；若 $B_{N-1}=\neg B_N$，则直接略去 P_N。

TSN 的一阶谓词逻辑为：

$$(TSN)\Rightarrow R(DC,SC_0,SC_1,\cdots,SC_N)\wedge A_0(SC_0,B_0)\wedge A_1(SC_1,B_1)\wedge\cdots\wedge P_N\Rightarrow$$
$$R(DC,SC_0)\wedge R(DC,SC_1)\wedge\cdots\wedge R(DC,SC_N)\wedge A_0(SC_0,B_0)\wedge A_1(SC_1,B_1)\wedge\cdots\wedge P_N \quad (1)$$

i)如果子三元组属于 1)、2)和 3)的情形，则

$$\text{式}(1)\Rightarrow(R(DC,SC_0)\wedge A_0(SC_0,B_0))\wedge(R(DC,SC_1)\wedge A_1(SC_1,B_1))\wedge\cdots\wedge(R(DC,SC_N)\wedge P_N)$$

ii)如果子三元组属于 4)的情形，则

$$\text{式}(1)\Rightarrow(R(DC,SC_0)\wedge A_0(SC_0,B_0))\wedge(R(DC,SC_1)\wedge A_1(SC_1,B_1))\wedge\cdots\wedge(R(DC,SC_{N-1})\wedge A_{N-1}(SC_{N-1},B_{N-1}))\wedge(R(DC,SC_{N-1})\wedge P_N)$$

iii)如果子三元组属于 5)的情形，则

$$\text{式}(1)\Rightarrow(R(DC,SC_0)\wedge A_0(SC_0,B_0))\wedge(R(DC,SC_1)\wedge A_1(SC_1,B_1))\wedge\cdots\wedge(R(DC,SC_{N-1})\wedge P_{N-1})$$

三元组语义网络的描述逻辑包括下列符号：可数谓词集合 $PS=\{ps_1,ps_2,\cdots,ps_m\}$，可数变元集合 $VS=\{vs_1,vs_2,\cdots,vs_n\}$，常量集合 $CS=\{Z,DC\}$，Z 是数集，DC 是领域概念集，以及逻辑演算符 $\wedge$（合取）、$\vee$（析取）和 $\neg$（否定）。

命题 2　由 $DS=PS\times VS\times CS$ 构成二元一阶谓词逻辑，其中“$\times$”运算受到 TSN 定义的约束。

证明：

1) 如果 $\exists vs\in VS$ 且 VS 值域为 Z，则 $\exists z\in\mathbf{Z}$，$ps\in PS$，使得 $ps(vs,z)$是对数值属性的描述，且 ps 是谓词 ISA。

2)如果 $\exists vs\in VS$ 且 VS 值域为 DC，则 $\exists dc\in DC$，$ps\in PS$，使得 $ps(dc,vs)$表示领域概念的语义描述。

3) 由 1)和 2)通过逻辑演算符 $\wedge$（合取）、$\vee$（析取）和$\neg$（否定）构成的 $Horn$ 子句 H，同样也使得 $H\in DS$ 成立。

一个三元组语义网络用于专利领域概念描述有以下推论。

推论 1 专利领域文献集 TS 中，$i \neq j$，$\exists ts_i, ts_j \in TS$；$m \neq n$，$\exists ps_m, ps_n \in PS$；$a \neq b$，$\exists vs_a, vs_b \in VS$，$\exists dc \in DC$。如果 $ps_m(dc, vs_a)|_{ts_i} = \neg ps_n(dc, vs_b)|_{ts_j}$ 成立，则 TS 中存在对领域概念描述出现矛盾。

推论 2 专利领域文献集 TS 中，$i \neq j$，$\exists ts_i, ts_j \in TS$；$m \neq n$，$\exists ps_m, ps_n \in PS$；$a \neq b$，$\exists vs_a, vs_b \in VS$，$\exists dc \in DC$。如果 $ps_m(dc, vs_a)|_{ts_i} \neq ps_n(dc, vs_b)|_{ts_j}$ 且 $ps_m(dc, vs_a)|_{ts_i} \neq \neg ps_n(dc, vs_b)|_{ts_j}$ 成立，则 TS 中存在对领域概念描述信息的补足。

推论 3 专利领域文献集 TS 中，$i \neq j$，$\exists ts_i, ts_j \in TS$；$\exists ps \in PS$；$\exists vs \in VS$，$\exists dc \in DC$。如果 $ps(dc, vs)|_{ts_i} = ps(dc, vs)|_{ts_j}$ 成立，则 TS 中存在对领域概念的冗余描述。

推论 4 专利领域文献集 TS 中，$i \neq j$，$\exists ts_i, ts_j \in TS$；$\exists ps \in PS$；$\exists dc \in DC$，$\exists z \in \mathbf{Z}$，有 $ps(dc, vs_a)|_{ts_i}$ 和 $ps(vs_a, z)|_{ts_j}$ 存在，则 TS 中存在对领域概念的模糊描述。

3 结束语

探讨专利文献摘要知识的一阶谓词演算，目的是推导专利文献知识融合将要面临的问题，包括句法分析后获取的描述性知识的不一致、信息的补足、冗余的发现和模糊信息的处理等。本研究在深入分析中文专利文献文本内容和结构的基础上，提出用三元组语义网络表示知识以及知识间的语义关系，并阐述了语义三元组到一阶谓词逻辑的转换，以此说明三元组语义网络的表示能力，为后续中文专利文献知识融合分析和推理奠定基础。

参考文献：

[1]Kent R E. Conceptual knowledge markup language: The central core[C/OL]//12th Workshop on Knowledge Acquisition, Modeling and Management. Banff, Alberta, Canada, 1999. [1999-10-26][2015-03-23]http://arxiv. org/ftp/arxiv/papers/1109/1109. 1525. pdf.

[2]Lassila O, Swick R. Resource description frameword (RDF) model and syntax specification[C/OL]//W3C Recommendation. [2004-02-10][2015-03-23]http://citeseer. ist. psu. edu/viewdoc/download;jsessionid=4FEC48E562F96F162CB750A35668401B? doi=10. 1. 1. 44. 6030&rep=rep1&type=pdf.

[3]曹存根，眭跃飞，孙瑜，等. 国家知识基础设施中的数学知识表示[J]. 软件学报，2006，17(8)：1731-1742.
Cao Cungen, Sui Yuefei, Sun Yu, et al. Representation of mathematical knowledge in national knowledge infrastructure[J]. Journal of Software, 2006, 17(8): 1731-1742.

[4]陆汝钤，石纯一，张松懋，等. 面向 Agent 的常识知识库[J]. 中国科学：E 辑，2000，30(5)：453-463.
Lu Ruqian, Shi Chunyi, Zhang Songmao, et al. Agent-oriented commonsense knowledge base[J]. Science in China: Series E, 2000, 43(6): 641-652.

[5]郑红. 面向 Agent 知识库的研究[D]. 贵阳：贵州大学，2000：12-40.

[6]陆汝钤，金芝. 形式本体：领域知识共享和复用的基础[J]. 计算机科学技术学报，2002，17(5)：535-548.
Lu Ruqian, Jin Zhi. Formal ontology: Foundation of domain knowledge sharing and reusing[J]. Journal of Computer Science and Technology, 2002, 17(5): 535-548.

（责任编辑：吕文红）