

Agent 的意图模型*

胡山立¹ 石纯一²

¹(福州大学计算机科学与技术系 福州 350002)

²(清华大学计算机科学与技术系 北京 100084)

E-mail: husl@fzu.edu.cn

摘要 意图是 Agent 的一个不可缺少的意识属性,在决定理性 Agent 的行为时起着重要的作用.已经有了若干种基于正规模态逻辑的意图模型,但它们存在着严重的“逻辑全知”问题.该文阐明意图不是正规模态算子,并提出了另一种意图模型,它不存在“逻辑全知”问题和其他相关问题(例如,副作用问题等).这种意图模型与 Konolige 和 Pollack 的意图模型相比,比较简单、自然,且满足 K 公理和联合一致性原理,实际上,为非正规模态算子基于正规可能世界的语义表示提供了一种新的方法.

关键词 Agent, 意图模型, BDI 理论, 非正规模态逻辑.

中图法分类号 TP18

在多 Agent 系统(multi-agent system)中,如何抽象和模型化 Agent,已成为当前人工智能、软件工程等领域的一项重要重要的研究课题.在 AI 领域,人们通常基于意向立场(intentional stance)来研究 Agent. Agent 的思维状态模型用来研究如何描述 Agent 的意识属性和它们之间的联系及其与感知、规划、行为、协商、合作等的关系.从意向立场出发,一般把信念(belief)、愿望(desire)和意图(intention)当作 Agent 的基本意识属性(简称 BDI).其中意图是抽象和模型化 Agent 的一个不可缺少的意识属性,在决定理性 Agent 的行为时起着重要的作用.意图决定了对未来行动的选择,从而使一个有限自治系统维持信念、目标、规划和行动的自我平衡.近年来,关于意图的形式化研究已成为哲学逻辑、人工智能、计算语言学 and 软件工程共同关心的重要课题.在实用推理、认知建模、自然语言处理以及多 Agent 系统的规划、协商、合作与对抗等方面均有着重要的应用.已经有了若干种基于正规模态逻辑的意图模型,但它们的模型存在着严重的“逻辑全知”问题以及由此带来的副作用等问题.

1 意图不能是正规模态算子

在理性 Agent: BDI 结构的形式化研究中,一般采用正规模态逻辑为工具.由于采用 N 规则,正规模态逻辑在逻辑蕴涵下是封闭的:给定一个正规模态算子 \square ,公式 α, β ,如果 $\square \alpha$ 为真,且 $\alpha \rightarrow \beta$,那么 $\square \beta$ 为真.当 \square 代表信念时,由于 Agent 系统是有限系统,知识和推理能力不完备且资源有限,封闭性对 Agent 是不现实的,只能看成是理想 Agent 的情况.

然而意图不同于信念,逻辑推论的封闭性即使在理想情况下也不应对意图进行假设.例如,一个意图拔牙的患者,不会意图享受拔牙的疼痛,尽管他知道拔牙引起疼痛是不可避免的.某些技巧已被用于避免这些副作用问题^[1,2],但还没有令人满意的结果.其根本原因在于,逻辑全知问题以及由此带来的副作用等问题是正规模态算子的固有性质.

* 本文研究得到国家自然科学基金(Nos. 69773026, 69733920)资助.作者胡山立,1974年生,副教授,主要研究领域为人工智能应用基础,多 Agent 系统.石纯一,1935年生,教授,博士生导师,主要研究领域为人工智能应用基础.

本文通讯联系人:胡山立,福州 350002,福州大学计算机科学与技术系

本文 1999-01-13 收到原稿,1999-08-20 收到修改稿

2 Konolige 和 Pollack 的意图表示理论

针对上述问题, Konolige 和 Pollack^[3] 提出以非正规模态逻辑作为描述工具的意图表示理论(KP 系统), 其语义模型是三元组 $M = \langle W, \Sigma, I \rangle$, 其中 W 是可能世界集, $w \in W$ 是状态的时序序列, 假设所有解释均相对于一个共同的特定时刻 *now* (语句的求值点). 对于每个世界 $w \in W$, 存在一个求值函数, 它确定了在语言 L 中语句的值, 可能性模态算子 \Diamond 的语义是

$$W, w \models \Diamond a \text{ iff } \exists w' \in W \text{ 使 } W, w' \models a.$$

必然性模态算子 $\Box a =_{df} \neg \Diamond \neg a$.

对公式 a , 记使 a 为真的可能世界集为 $M_a =_{df} \{w \in W \mid W, w \models a\}$. $\Sigma \subseteq W$ 用来解释信念, 信念模态算子 B 的语义是

$$\langle W, \Sigma, I \rangle \models B(a) \text{ iff } \forall w \in \Sigma, \text{ 有 } W, w \models a, \text{ 即 } \Sigma \subseteq M_a.$$

$I \subseteq P(W)$ ($P(W)$ 标记 W 的幂集) 用来解释意图, 意图模态算子 I 的语义是

$$\langle W, \Sigma, I \rangle \models I(a) \text{ iff } \exists \psi \in I \text{ 使 } \psi = M_a, \text{ 即 } M_a \in I.$$

其实 I 中的每个元素 M_a 对应于一个意图, 其作用相当于一个模态算子, 与正规模态逻辑语义解释不同的是, 在上述解释规则中用 $=$ 而不是 \subseteq , 这使得 KP 系统避免了逻辑蕴涵的副作用. 为了表达意图与信念之间的关系, KP 系统对其语义模型有两条约束:

- (1) 可达性: $\exists w \in \Sigma, \forall M_a \in I, \text{ 有 } w \in M_a$;
- (2) 非平凡性: $\forall M_a \in I, \exists w \in \Sigma \text{ 使 } w \notin M_a$.

满足这两条约束的系统称为可采纳的, 有以下性质:

$$\neg I(a \wedge \neg a); \tag{KP1}$$

$$I(a) \wedge I(b) \rightarrow \Diamond(a \wedge b); \tag{KP2}$$

$$I(a) \rightarrow \neg B(a) \wedge \neg B(\neg a). \tag{KP3}$$

然而, 在 KP 系统中, 用来确定意图的 I 是 W 幂集的子集, 与信念的表示相比不够自然, 而且复杂度高. 例如, 如果有 n 个意图, 确定意图要用 W 的 n 个子集, 而确定信念只要用 W 的一个子集. 另外, KP 系统不满足 K 公理和联合一致性原理.

3 系统定义和直观解释

我们下面提出的意图的语义表示很好地解决了上述问题, 并且满足 KP 系统的所有性质以及人们对意图的要求. 为了突出意图的语义表示, 便于与 KP 系统比较, 也因为对于在 Agent 的思维状态中是否需要愿望或目标作为基本要素还存在争议^[4], 本文暂不涉及目标或愿望. 另外, 必要时也不难增加时间等要素.

定义 1. Agent 的语义模型 M 是一个五元组 $M = \langle W, \Sigma_b, \Sigma_i, \Sigma'_i, V \rangle$, 其中 W 是一个非空集, 可以看成是可能世界集. $\Sigma_b, \Sigma_i, \Sigma'_i$ 均是 W 的非空子集, 且 $\Sigma_i \cap \Sigma'_i = \emptyset$. 赋值 V 的定义同命题逻辑系统, 对每一个世界 $w \in W, V$ 确定了在语言 L 中命题的真值. 由于在讨论中 V 并不重要, 故略去. 增加的模态算子的解释由定义给出.

在语言 L 中引入模态算子 \Diamond, B, I , 分别称为可能算子、信念算子和意图算子.

定义 2. $W, w \models \Diamond a$ iff $\exists w' \in W$ 使 $W, w' \models a$.

$\Box a =_{df} \neg \Diamond \neg a$. 公式 a 的景象(scene) $S_a =_{df} \{w \in W \mid W, w \models a\}$.

定义 3. $\langle W, \Sigma_b, \Sigma_i, \Sigma'_i \rangle \models B(a)$ iff $\forall w' \in \Sigma_b, \text{ 有 } W, w' \models a, \text{ 即 } \Sigma_b \subseteq S_a$.

定义 4. $\langle W, \Sigma_b, \Sigma_i, \Sigma'_i \rangle \models I(a)$ iff $\forall w' \in \Sigma'_i, \text{ 有 } W, w' \models a$ 且

$$\forall w'' \in \Sigma'_i, \text{ 有 } W, w'' \models \neg a.$$

即 $\Sigma'_i \subseteq S_a \subseteq (W - \Sigma'_i)$.

信念和意图的语义的直观解释: 信念表示 Agent 对某些可能世界(即 Σ_b)的“偏爱”. 只有在这些可能世界上

均为真的命题,Agent 才是相信的.而意图也表示 Agent 对某些可能世界的“重视”.不过与信念不同的是,这些被重视的可能世界被分成两个不相交的部分 Σ 和 Σ' ,在 Σ 上为真的命题 Agent 认为是可能实现的或已实现的,而在 Σ' 上为假的命题 Agent 认为是当前尚未实现的目的不是必然实现的.显然,只有尚未实现不是必然实现而又可能实现的命题才是理性 Agent 值得去意图的.

定义中信念仍用 W 的一个子集 Σ_0 来表示,意图用 W 的两个不相交的子集 Σ 和 Σ' 来表示,而不是像 KP 系统那样用 W 的幂集的子集来表示(或当有 n 个意图时,用 W 的 n 个子集来表示).由于意图要满足的性质与信念不同,容易理解意图不可能用 W 的一个子集来表示.因此,上述表示应是意图基于可能世界的语义表示的最好结果.当然,还需证明这种表示具有所期望的性质.

为了使意图满足与信念有关的性质,自然必须使意图的表示和信念的表示联系起来. Rao 和 Georgeff 的系统^[2]要求每一个信念世界必存在一个意图子世界. Konoigé 和 Pollack 的 KP 系统^[3]对其语义模型要求满足可达性与非平凡性.对此,我们提出以下更简明的约束条件,称为可实现性与非平凡性.

定义 5. 模型 $M = \langle W, \Sigma_0, \Sigma, \Sigma' \rangle$ 称为可采纳的,当且仅当它是可实现的, $\Sigma_0 \cap \Sigma \neq \emptyset$ 且是非平凡的, $\Sigma_0 \cap \Sigma' \neq \emptyset$.

在以下的讨论中,模型 M 认为是可采纳的.

4 系统的合理性

Bratman^[5]提出有限自治系统必须满足合理性的必要条件:“反对称论题”和“无副作用原理”.反对称论题是说,允许一个 Agent 以某一行动为意图,同时又相信它不能实现是不合理的,即信念-意图一致性;另一方面,一个 Agent 以某一行动为意图,但不相信它能实现是合理的,即信念-意图不完全性.这种一致性与不完全性的不对称性称为反对称论题.形式表示为

$$\forall M \text{ 有 } M \not\models I(a) \wedge B(\neg a), \quad \text{一致性} \quad (A11)$$

$$\exists M \text{ 使 } M \models I(a) \wedge \neg B(a), \quad \text{不完全性} \quad (A12)$$

无副作用原理是说,以 a 为意图的 Agent 不应被迫以 a 的逻辑结论为意图.形式表示为

$$\exists M \text{ 使 } M \not\models (a \rightarrow \beta) \wedge I(a) \wedge \neg I(\beta). \quad (A13)$$

Rao 和 Georgeff^[6]建议增加与副作用相关的非迁移原理作为合理性约束之一,形式表示为

$$\exists M \text{ 使 } M \models B(a) \wedge \neg I(a). \quad (A14)$$

当然,意图还应满足其他一些性质.首先证明,在我们的模型中,意图满足无副作用原理(A13),信念与意图满足反对称论题(A11 和 A12)和非迁移原理(A14).设 α, β 是公式, W 中使 α 为真的可能世界集 $S_\alpha = \{w \in W \mid W, w \models \alpha\}$

命题 1. 意图满足无副作用原理(A13).

证明:构造 $M_1 = \langle W_1, \Sigma_0, \Sigma_1, \Sigma'_1 \rangle$. 其中 $W_1 = \{w1, w2, w3\}$, $\Sigma_0 = \{w1, w2\}$, $\Sigma_1 = \{w1\}$, $\Sigma'_1 = \{w2\}$.

取 $S_\alpha = \{w1\}$, $S_\beta = \{w1, w2\}$, 容易验证 $M_1 \models (a \rightarrow \beta) \wedge I(a) \wedge \neg I(\beta)$. □

命题 2. 信念与意图满足一致性条件(A11).

证明:对任意满足可实现性约束的模型 $M = \langle W, \Sigma_0, \Sigma, \Sigma' \rangle$, 假设 $M \models I(a) \wedge B(\neg a)$.

那么,由 $M \models I(a)$ 得出 $\Sigma \subseteq S_\alpha$, 由 $M \models B(\neg a)$ 得出 $\Sigma_0 \subseteq S_{\neg a}$.

显然 $S_\alpha \cap S_{\neg a} = \emptyset$, 故 $\Sigma_0 \cap \Sigma = \emptyset$, 与可实现性约束矛盾. □

命题 3. 信念与意图满足不完全性条件(A12).

证明:构造 $M_1 = \langle W_1, \Sigma_0, \Sigma_1, \Sigma'_1 \rangle$. 其中 $W_1 = \{w1, w2, w3\}$, $\Sigma_0 = \{w1, w2\}$, $\Sigma_1 = \{w1\}$, $\Sigma'_1 = \{w2\}$.

取 $S_\alpha = \{w1\}$, 容易验证 $M_1 \models I(a) \wedge \neg B(a)$. □

命题 4. 信念与意图满足非迁移原理(A14).

证明:构造 $M_1 = \langle W_1, \Sigma_1, \Sigma'_1, \Sigma''_1 \rangle$. 其中 $W_1 = \{w1, w2, w3, w4\}, \Sigma_1 = \{w1, w2\}, \Sigma'_1 = \{w1, w3\}, \Sigma''_1 = \{w2\}$.

取 $S_a = \{w1, w2\}$, 容易验证 $M_1 \models B(a) \wedge \neg I(a)$. □

意图是对将来可能实现的目标的承诺,意图除了要满足无副作用原理外,还应满足:

K 公理: $\forall M$ 有 $M \models I(p \rightarrow q) \rightarrow (I(p) \rightarrow I(q))$; (A15)

一致性原理: $\forall M$ 有 $M \models \neg I(\alpha \wedge \neg \alpha)$; (A16)

联合一致性原理: $\forall M$ 有 $M \models I(\alpha) \wedge I(\beta) \rightarrow I(\alpha \wedge \beta)$; (A17)

联合不完全性原理: $\exists M$ 使 $M \models I(\alpha \vee \beta) \wedge \neg I(\alpha \wedge \beta)$, (A18)

$\exists M$ 使 $M \models \neg I(\alpha \wedge \beta) \wedge \neg I(\alpha) \wedge \neg I(\beta)$. (A19)

信念与意图除了满足反对称论题和非迁移原理外,还应满足

非平凡性原理: $\forall M$ 有 $M \models I(\alpha) \rightarrow \neg B(\alpha)$. (A110)

正规模态逻辑系统和一般非正规模态逻辑系统都有 K 公理,认识论模态逻辑系统也有 K 公理. 意图应该满足 K 公理也是容易理解的:假设理性 Agent 意图 p ,且意图 $p \rightarrow q$,当然它意图 q ,否则它就不应该意图 $p \rightarrow q$ 了.

意图要满足一致性原理 A16 是不言而喻的,理性 Agent 当然不应有互相对立的意图. 意图要满足联合一致性原理 A17 是因为 Agent 意图 α 且意图 β ,当然意图 $\alpha \wedge \beta$. 注意,这和“意图 α 或 β ,不一定要意图 $\alpha \wedge \beta$ ”完全不同,后者的形式化是 A18. 这也和“意图 $\alpha \wedge \beta$,不一定要意图 α ,也不一定要意图 β ”不同,后者的形式化是 A19. 应该注意它们的区别. 我们把 A18 和 A19 称为联合不完全性原理. 信念与意图还应满足非平凡性原理 A110. 这是因为理性 Agent 意图 α ,那么它不相信没有它的行动, α 能自然实现,否则就不必意图 α 了. 在约束条件 $\Sigma_a \subseteq \Sigma$ 下,我们即可证明信念与意图满足不完全性条件 A12(在命题 3 的证明中,改取 $\Sigma'_1 = \{w3\}$ 即可),现在把约束条件加强到非平凡性约束 $S_a \cap \Sigma'_a \neq \emptyset$,即可证明信念与意图满足非平凡性原理 A110.

下面证明在模型 M (M 是可采纳的)中,意图满足上述诸原理.

命题 5. 意图满足 K 公理 A15.

证明:对任意模型 $M = \langle W, \Sigma_a, \Sigma'_a, \Sigma''_a \rangle$,任意命题 p, q ,因为 M 是非平凡的,所以 $\Sigma'_a \neq \emptyset$. 对任意 $w \in \Sigma'_a, p \rightarrow q$ 和 p 均为假是不可能的,故 $I(p \rightarrow q)$ 和 $I(p)$ 不能同时为真. 这就证明了 $I(p \rightarrow q) \rightarrow (I(p) \rightarrow I(q))$ 对模型 M 有效. □

Konolige 和 Pollack^[3]的 KP 系统,不满足 K 公理 A15. 因为当 $M_e \in I$ 且 $M_{\beta, \alpha} \in I$ 时,不一定有 $M_\alpha \in I$.

命题 6. 意图满足一致性原理 A16.

证明:对任意模型 $M = \langle W, \Sigma_a, \Sigma'_a, \Sigma''_a \rangle$,任意公式 α ,显然 $S_{\alpha \wedge \neg \alpha} = \emptyset$,而 $\Sigma'_a \neq \emptyset$ (由可实现性约束),故 $\Sigma'_a \subseteq S_{\alpha \wedge \neg \alpha} \subseteq (W - \Sigma'_a)$ 不成立,所以 $M \models \neg I(\alpha \wedge \neg \alpha)$. □

命题 7. 意图满足联合一致性原理 A17.

证明:对任意模型 $M = \langle W, \Sigma_a, \Sigma'_a, \Sigma''_a \rangle$,任意公式 α, β ,若 $I(\alpha) \wedge I(\beta)$ 为真,有 $\Sigma'_a \subseteq S_\alpha \subseteq (W - \Sigma'_a)$ 且 $\Sigma'_a \subseteq S_\beta \subseteq (W - \Sigma'_a)$,而 $S_{\alpha \wedge \beta} = S_\alpha \cap S_\beta$,因此, $\Sigma'_a \subseteq S_{\alpha \wedge \beta} \subseteq (W - \Sigma'_a)$,即 $I(\alpha \wedge \beta)$ 为真,所以 $M \models I(\alpha) \wedge I(\beta) \rightarrow I(\alpha \wedge \beta)$. □

Konolige 和 Pollack^[3]的 KP 系统不满足联合一致性原理 A17. 因为当 $M_\alpha \in I$ 且 $M_\beta \in I$ 时,不一定有 $M_{\alpha \wedge \beta} \in I$.

命题 8. 意图满足联合不完全性原理 A18 和 A19.

证明:构造 $M_1 = \langle W_1, \Sigma_1, \Sigma'_1, \Sigma''_1 \rangle$. 其中 $W_1 = \{w1, w2, w3\}, \Sigma_1 = \{w1, w2\}, \Sigma'_1 = \{w1\}, \Sigma''_1 = \{w2\}$.

取 $S_\alpha = \{w1\}, S_\beta = \{w3\}$, 容易验证 $M_1 \models I(\alpha \vee \beta) \wedge \neg I(\alpha \wedge \beta)$.

构造 $M_2 = \langle W_2, \Sigma_2, \Sigma'_2, \Sigma''_2 \rangle$. 其中 $W_2 = \{w1, w2, w3, w4\}, \Sigma_2 = \{w1, w2\}, \Sigma'_2 = \{w1\}, \Sigma''_2 = \{w2, w3\}$,

取 $S_\alpha = \{w1, w2\}, S_\beta = \{w1, w3\}$, 容易验证 $M_2 \models \neg I(\alpha \wedge \beta) \wedge \neg I(\alpha) \wedge \neg I(\beta)$. □

命题 9. 信念与意图满足非平凡性原理 A110.

证明:对任意满足非平凡性约束的模型 $M = \langle W, \Sigma_a, \Sigma'_a, \Sigma''_a \rangle$ 和任意公式 α ,若 $I(\alpha)$ 为真,即 $\Sigma'_a \subseteq S_\alpha \subseteq (W -$

Σ'_i), 因为 M 满足非平凡性约束 $\Sigma_i \cap \Sigma'_i \neq \emptyset$, 从而 $\exists w', w' \in \Sigma_i \cap \Sigma'_i$, 即

$$w' \in \Sigma_i, \quad (1)$$

且

$$w' \in \Sigma'_i, \quad (2)$$

由式(2)得

$$w' \in S_{\alpha}. \quad (3)$$

由式(1), (3)得 $\Sigma_i \not\subseteq S_{\alpha}$, 即 $B(\alpha)$ 为假, $\neg B(\alpha)$ 为真, 所以 $M \models I(\alpha) \rightarrow \neg B(\alpha)$. \square

命题 10. 对于意图, 不存在逻辑全知问题.

证明: 因为 $\Sigma'_i \neq \emptyset$, 所以, 有效的公式不能被意图. \square

5 系统比较与结论

本文讨论了意图的语义表示并提出了一种意图模型, 这种模型有直观的基于可能世界的语义解释, 能满足对意图性质的要求(A11~A110). 与 Cohen 和 Levesque 的工作^[1]以及 Rao 和 Georgeff 的工作^[2]相比, 不存在“逻辑全知”问题以及副作用等问题. 与 Konolige 和 Pollack 的工作^[3]相比, 意图的表示简单、自然, 有基于可能世界的语义解释, 且满足 K 公理和联合一致性原理(A17).

我们提出的这一方法还具有一般性, 不仅适用于 Agent BDI 结构中的信念、愿望或目标等形式表示, 而且为非正规模态逻辑找到了一种基于正规可能世界的语义表示^[7].

参考文献

- 1 Cohen P R, Levesque H J. Intention is choice with commitment. *Artificial Intelligence*, 1990, 42(2~3): 213~261
- 2 Rao A S, Georgeff M P. Modeling rational agents within a BDI architecture. In: Allen J, Fikes R, Sandewall E eds. *Principles of Knowledge Representation and Reasoning. Proceedings of the 2nd International Conference (KR-91)*. San Mateo, CA: Morgan Kaufmann Publishers, Inc., 1991. 473~484
- 3 Konolige K, Pollack M E. A representationalist theory of intention. In: Bajcsy R ed. *Proceedings of the 13th International Joint Conference on Artificial Intelligence*. San Mateo, CA: Morgan Kaufmann Publishers, Inc., 1993. 390~395
- 4 Singh M P. *Multi-Agent System: A Theoretical Framework for Intentions, Know How, and Communications*. Berlin: Springer Verlag KG, 1994
- 5 Bratman M E. *Intentions, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press, 1987
- 6 Rao A S, Georgeff M P. Asymmetry thesis and side effect problems in linear-time and branching-time intention logic. In: Mylopoulos J, Reiter R eds. *Proceedings of the 12th International Joint Conference on Artificial Intelligence*. San Mateo, CA: Morgan Kaufmann Publishers, Inc., 1991. 498~504
- 7 Hu Shan-li, Shi Chun-yi. A semantic interpretation for agent's non-normal modal operators. *Computer Research and Development*, 1999, 36(10): 1153~1157
(胡山立, 石纯一. 适用于 Agent 非正规模态算子的一种语义解释. *计算机研究与发展*, 1999, 36(10): 1153~1157)

An Intention Model for Agent

HU Shan-li¹ SHI Chun-yi²

¹Department of Computer Science and Technology Fuzhou University Fuzhou 350002

²Department of Computer Science and Technology Tsinghua University Beijing 100084

Abstract Intentions, an integral part of the mental state of an agent, play an important role in determining the behavior of rational agents. There are several models of intention based on normal modal logic. But these

theories suffer from the omniscience problem seriously. In this paper, the authors argue that intention is not a normal modal operator, and present another intention model. It doesn't have the logical omniscience problem and other related problems such as side-effect problem, etc. Compared with Konolige and Pollack's model of intention, this model not only is simpler and more natural, but also satisfies the K-axiom and the Joint Consistency. Actually it gives a new method for semantic representation of non-normal modal operators based on normal possible worlds.

Key words Agent, intention model, BDI theory, non-normal modal logic.