多 Agent系统计算的意愿理论*

毛新军 王怀民 陈火旺

(国防科技大学计算机系 长沙 410073)

摘 要 意愿是规范和描述多 Agent系统的一个重要抽象认知概念。本文提出了多 Agent系统计算的意愿理论以支持多 Agent系统计算的理论研究。基于多 Agent系统计算的动作模型理论，我们给出了意愿概念新的语义定义，获取和描述了它的一些重要逻辑属性。

关键词 Agent 多 Agent系统，意愿

分类号 TP301

The Intention Theory of Multi-Agent System Computing

Mao Xinjun Wang Huaimin Cheng Huowang

(Department of Computer, NUDT, Changsha 410073)

Abstract Intention is an important cognitive concept to specify multi-agent system. An intention theory is presented in this paper to support the research on the theory of multi-agent system computing. Based on the action modeling theory of multi-agent system computing, the new formal semantics of the intention is defined and some important properties are obtained.

Keywords agent, multi-agent system, intention

分布计算是目前计算机科学领域中的一项关键性主流技术。分布计算系统由一组相对独立并能自主运作的计算实体组成。当前，随着分布计算技术的不断发展及其应用的不断深入，分布计算系统变得日趋复杂和庞大，人们迫切地需要一种严格、有效的软件工程方法来促进分布计算系统的开发。其中，需求获取是软件开发过程中一项极为重要的工作。

Agent是近年来计算机科学领域中的一个重要概念。Agent概念的自主性、交互性、社会性等特征为准确地研究和刻画分布计算系统提供了合理的概念模型。在 AI领域，人们通常基于意向观点（intentional stance)来研究 Agent概念，即将 Agent视为由一组认知成份所构成的意向系统（intentional system）。意向观点为我们研究和分析 Agent提供了高层的抽象认知概念。其中，意愿是规范和描述多 Agent系统的一个重要抽象认知概念，原因是：(1) Agent的意愿被认为是 Agent计算的起因；(2) 这一概念使得我们可以独立于 Agent的具体实现细节来构造 Agent体系结构，定义 Agent状态，研究 Agent行为的规律型特征；(3) 意愿概念具有更强的约束，且体现了 Agent的某些理性特征。

为了使意愿概念能够有效地用于规范和描述分布计算系统，必须深入地研究和分析意愿概念的性质和含义，严格、形式化地给出它的语义定义，获取和描述它的逻辑属性。本文用 Agent概念来刻画分布计算实体，将分布计算系统视为多 Agent系统，提出了多 Agent系统计算的意愿理论。

1 意愿概念讨论

为了指导意愿理论研究，我们首先讨论一下什么是意愿，Agent的意愿具有哪些性质以及意愿概念与其它概念（如动作，信念等等）间的关系。这一工作对于我们进一步开展意愿理论的研究是极为重要

* 国家自然科学基金资助项目
199?年 9月 15日收稿
第一作者：毛新军，男，1970年生，博士生

的, 它将指导我们形式化意愿概念, 有助于我们获取和描述意愿概念的一些重要逻辑属性。意愿概念的基本内涵是对未来行为的合理选择。选择性是意愿概念的本质属性。直觉地, 意愿概念具有下列属性。

· 内部一致性, Agent的多个意愿间是相互一致的。

· 非冲突性, Agent的多个意愿间不应是相互冲突的。Agent的两个意愿是相互冲突的是指 Agent某个意愿的实现将阻止或妨碍另一个意愿的成功实现。

· 可满足性, 意愿的可满足性是指 Agent的意愿是可实现的, 即存在某一将来状态, 在该状态下意愿被实现。

· 与信念的一致性, Agent的意愿与 Agent的信念是相一致的。如果 Agent认为某个命题是不可实现的, 则 Agent就不应意图去实现该命题。

· 有关意愿概念的一个重要直觉性认识是: Agent的意愿是 Agent计算的起因, 它体现了 Agent的某种选择特征。

· 承诺性, 承诺性是意愿概念的另一个重要特征。意愿的承诺性是指 Agent不会随意地放弃其已有的意愿, 即具有某种意愿的 Agent将在不断变化的环境中持续性地拥有该意愿。

2 多 Agent系统计算的动作模型理论

2.1 形式化语言 L

形式化语言 L是对命题分支时序逻辑 CTL* [2]的扩充。语言 L的公式集由状态公式集 L_t和路径公式集 L_s两部分组成。设 Φ 是原子命题符号集合, A是 Agent符号集合, A_t是原子动作符号集合。这些集合均非空且递归可枚举。为了简化说明, 文中具有下列符号约定: (1) p, q表示原子命题; (2) φ, ψ 表示公式; (3) i, j表示 Agent; (4) a, b表示原子动作。

定义 1 形式化语言 L是由下列规则定义的最小封闭集合

(L_1) 如果 $p\in\Phi$, 则 $p\in L_t$;

(L_2) 如果 $i\in A$, ψ, $\varphi\in L_t$且 $i\in A$, 则 $\neg\varphi$, $\psi\wedge\varphi$, $K_i\varphi$, $Intend_i\varphi\in L_t$;

(L_3) $L\subseteq L_s$;

(L_4) 如果 ψ, $\varphi\in L_s$, $i\in A$且 $a\in A_t$, 则 $\neg\varphi$, $\psi\wedge\varphi$, $\psi U\varphi$, $\langle do_i(a)\rangle\varphi$, $[do_i(a)]\varphi\in L_s$;

(L_5) 如果 $\varphi\in L_s$, 则 $A\varphi\in L_t$

2.2 形式化模型和语义

语言 L的一个模型 M是指元偶 $\langle T, <, A, \pi, Y, B, I\rangle$。T是时刻集, T中的每一时刻对应于世界的一个状态 (包括物理系统状态, 系统中各个 Agent的认知状态)。物理系统状态由在该状态下为真的原子命题来表示, Agent的认知状态是指 Agent的信念状态和意愿状态, <是 T上的偏序关系且满足过去线性, <描述了时刻间的先后次序, 任一时刻的过去是确定和线性的, 它的将来可能是分支的。整个形式化模型呈一树形结构。某一时刻 t的一条路径是指始于该时刻, 由 t的将来时刻构成的一条线性分支。不同的路径对应于不同的 Agent动作执行事件与环境事件的组合, 反映了世界的不同发展轨迹。分支时间模型可以帮助我们刻画世界发展的各种可能轨迹以及 Agent可能作出的各种动作选择, 并进一步地形式化意愿概念。形式化模型允许多个 Agent的动作并发、异步地发生, 允许不同的动作具有不同的时间延迟。在任一时刻 Agent通过执行动作来影响和控制世界的发展, 然而这种影响和控制可能是有限的, 世界发展的轨迹还受其它 Agent的动作执行事件以及环境事件的影响, 所有 Agent的动作执行事件以及环境事件共同确定世界的发展。

定义 2 时刻 t的一条路径是指满足以下性质的集合 $S\subseteq T$: (1) $t\in S$; (2) $\forall t_1, t_2\in S$: $(t_1<t_2)\vee(t_2<t_1)\vee(t_1=t_2)$; (3) $\forall t_1, t_2\in S$; $t_3\in T$: $(t_1<t_3<t_2)\Rightarrow(t_3\in S)$; (4) $\forall t_1\in S$; $t_2\in T$: $(t_1<t_2)\Rightarrow(\exists t_3\in S: (t_1<t_3)\wedge\neg(t_3<t_2))$; (5) $\forall t_1\in S$: $(t=t_1)\vee(t<t_1)$

设 S_t表示时刻 t的所有路径集合。S_T是所有路径的集合, 即 $S_T=\bigcup_{t\in T}S_t$。

定义 3 设 $t\leq t'$, 则 $[t, t']=\{t''\mid t\leq t''\leq t'\}$ 为一路径子区间。

A是Agent集合。π: $\Phi \to$ powerset(T), powerset是幂集符号。$\pi(p)$定义了使原子命题p成立的时刻集。Y: $A \times A_T \to$ powerset($T \times T$)。Y定义了动作的发生。$[t, t'] \in Y(i, a)$表示在Agent$_i$在$[t, t']$路径子区间中执行动作a, t是起始时刻, t'是终止时刻。B: $A \to T \times T$。$(t, t') \in B(i)$是指Agent$_i$在时刻t认为时刻t'是可能的。关系B用于定义Agent的信念。I: $A \times T \to$ powerset(S^∞)。$S \in I(i, t)$是指在t时刻Agent$_i$选择了路径S, 因而有$I(i, t) \in S$。函数I用于定义Agent的意愿, 它刻划了在任一时刻Agent对世界发展轨迹的选择。

L_t中公式的可满足语义定义由模型M和时刻t给出。$M \models_t \varphi$表示模型M在时刻t满足公式φ。L_s中公式的可满足语义由模型M, 路径S和时刻t加以定义。$M \models_{s,t} \psi$表示模型M在路径S的时刻t满足公式ψ。

定义**4** (语言L的形式化语义)

- $M \models_t p$ iff $p \in \pi(p)$, 其中p为原子命题;
- $M \models_t \psi \wedge \varphi$ iff $M \models_t \psi$且$M \models_t \varphi$;
- $M \models_t \neg\varphi$ iff $M \not\models_t \varphi$;
- $M \models_t A\varphi$ iff $\forall S \in S_t$: $M \models_{s,t} \varphi$;
- $M \models_t K\varphi$ iff $\forall t'$: $(t, t') \in B(i) \Rightarrow M \models_{t'} \varphi$;
- $M \models_{s,t} \psi \wedge \varphi$ iff $M \models_{s,t} \psi$且$M \models_{s,t} \varphi$;
- $M \models_{s,t} \neg\varphi$ iff $M \not\models_{s,t} \varphi$;
- $M \models_{s,t} (\psi U \varphi)$ iff $\exists t' \in S$: $t \le t'$且$(M \models_{s,t'} \varphi)$且$(\forall t''$: $t \le t'' < t' \Rightarrow M \models_{s,t''} \psi)$
- $M \models_{s,t} \langle do_i(a) \rangle \varphi$ iff $\exists t_u \in S$: $t < t_u$且$[t, t_u] \in Y(i, a)$且$(\exists t''$: $t < t'' \le t_u$且$M \models_{s,t''} \varphi)$;
- $M \models_{s,t} [do_i(a)] \varphi$ iff $\forall t_u \in S$: $t < t_u$且$[t, t_u] \in Y(i, a) \Rightarrow (\exists t''$: $t < t'' \le t_u$且$M \models_{s,t''} \varphi)$;
- $M \models_{s,t} \varphi$ iff $M \models_t \varphi$, 其中$\varphi \in L_t$

根据上述语义定义, 我们可以派生出其它命题连接词和算子。U是until时序算子。$F\varphi = true U \varphi$是必然算子。A是全称路径算子。$A\varphi$在时刻$t$为真当且仅当对于时刻$t$的任一路径S, φ在路径S上为真。E是A的对偶算子即$E\varphi = \neg A \neg \varphi$。$\langle\ \rangle$和$[\ ]$是动作算子。直觉地$\langle do_i(a) \rangle \varphi$表示Agent$_i$执行动作a且具有结果$\varphi$。$[do_i(a) \varphi]$表示如果Agent$_i$能够完成动作a, 则具有结果$\varphi$。$\langle\ \rangle$和$[\ ]$的语义遵循了标准动态逻辑中的定义, 然而上述语义定义具有更强的灵活性, 即要求φ在动作执行过程中(而不是在动作完成之时)成立。$K\varphi$表示Agent具有信念φ。为了刻划多Agent系统计算, 我们假定关系$B(i)$满足自反性和传递性。

3 Agent的意愿

意愿概念的基本内涵是对未来行为的合理选择。选择性是意愿概念的本质属性。Agent具有意愿φ是指Agent对世界发展轨迹(即路径)的选择, 在这些被选择的世界发展轨迹中, φ将最终成立。

定义**5** $M \models_t Intend\varphi$ iff $M \models_{t,\neg}\varphi$且$(\forall S$: $S \in I(i, t) \Rightarrow M \models_{s,t} F\varphi)$

上述语义定义揭示了意愿概念的最本质特征即选择性。不同于已有的许多方法[1, 3, 4], 我们没有基于可能世界间的可达关系来定义意愿概念的形式化语义, 而是将Agent的意愿视为Agent对世界发展轨迹的选择。在形式化模型M中, 动作、时刻、世界发展轨迹三者是紧密相关的。世界发展轨迹由一系列线性的时刻组成, 每一时刻对应于世界的一个状态。动作和时刻是紧密相关的, 任何动作的执行使得时刻发生改变, 动作发生的过程亦是时刻不断流逝的过程, 同时动作的执行影响世界发展轨迹, Agent通过选择并执行动作在一定程度上影响和控制世界的发展, 但Agent的这种影响和控制是有限的。所有Agent的动作执行事件和环境事件共同确定世界的发展, 因而世界发展轨迹对应于Agent的某个特定的动作执行序列, Agent对世界发展轨迹的选择本质上是对动作执行序列的选择。

上述语义定义还揭示了Agent接收、放弃其意愿的条件。Agent不会接收一个已实现的或者已不可能实现的命题作为其意愿, 同样如果某个命题已成立或者已不可能成立, 则具有该意愿的Agent将放

弃这一意愿. 基于意愿概念的上述语义定义, 我们可以获得意愿概念的一系列逻辑属性.

命题 1 $\models \neg(\mathrm{Intend}_i\varphi \wedge \mathrm{Intend}_i(\neg\varphi))$

上述命题表明 Agent的意愿是一致的. 在任意时刻, Agent不可能既有意愿 φ, 同时又有意愿 $\neg\varphi$. 可根据意愿概念的语义定义来证明该命题.

命题 2 $\models \mathrm{Intend}_i\varphi \rightarrow \mathrm{EF}\varphi$

这一命题指出 Agent的意愿是可满足的, 或者说是可实现的, 即存在某一世界发展轨迹, φ 在该世界发展轨迹上必然成立. 为了使命题 2成立, 我们对形式化模型作了下列约束.

$M_{CON}-1$ $\forall i \in T; t \in A: \quad I(i,t) \neq \varnothing$

命题 3 $\models \neg(\mathrm{Intend}_i\varphi \wedge K_i\neg\mathrm{EF}\varphi)$

上述命题指出 Agent的意愿与 Agent的信念是一致的. 在任意时刻, Agent不可能既有意愿 φ 同时又认为 φ 是不可能成立的. 可根据意愿概念的语义定义以及命题 2来证明该命题.

命题 4 $\models \mathrm{Intend}_i\varphi \wedge \mathrm{Intend}_i\psi \rightarrow E(F\varphi \wedge F\psi)$

这一命题揭示了 Agent多个意愿间的非冲突性. 如果 Agent在某一时刻具有多个意愿, 则存在一条路径, 在该路径上 Agent以某种时序关系来实现这些意愿, 或者先实现 φ, 再实现 ψ, 或者先实现 ψ, 再实现 φ, 或者同时实现 φ 和 ψ. 可根据意愿概念的语义定义以及模型约束 $M_{con}-1$来证明该命题.

命题 5 $\models \mathrm{Intend}_i(\varphi\wedge\psi) \wedge \neg\varphi \wedge \neg\psi \rightarrow \mathrm{Intend}_i\varphi \wedge \mathrm{Intend}_i\psi$

然而上述公式的逆并不一定成立. 直觉地, 公式 $\mathrm{Intend}_i(\varphi\wedge\psi)$ 是指 Agent i 意图同时实现 φ 和 ψ, 公式 $\mathrm{Intend}_i\varphi \wedge \mathrm{Intend}_i\psi$ 是指 Agent i 意图以某种时序关系实现 φ 和 ψ, 但不一定同时实现 φ 和 ψ. 可根据意愿概念的语义定义来证明该命题.

命题 6 $\models \mathrm{Intend}_i\varphi \wedge K_i(AG(\varphi\rightarrow\psi) \wedge \neg\psi) \rightarrow \mathrm{Intend}_i\psi$

上述命题揭示了 Agent的目标 – 手段推理, 我们将新派生出的意愿称为 Agent实现其已有意愿的一种手段. 目标和手段是相对而言的, Agent的意愿既可作为实现某个命题的目标, 也可作为实现其它意愿的一种手段. 可根据意愿概念的语义定义来证明该命题.

命题 7 $\models A(\mathrm{Intend}_i\varphi \rightarrow (\mathrm{Intend}_i\varphi U (\varphi \vee \neg\mathrm{EF}\varphi)))$

上述命题刻画了意愿概念的持续性特征. 持续性是意愿概念的一个重要特性, 亦是 Agent成功地实现其意愿的一个重要条件. 意愿的持续性是指在动态、不确定的多 Agent系统中 Agent将尽可能地保持其意愿. 为了确保上述命题成立, 我们对形式化模型作了下列约束:

$M_{CON}-2$ $\forall i \in A; t \in T; S \subseteq S_t: \quad M|=_t \mathrm{Intend}_i\varphi \Rightarrow M|=_{s,t}(\mathrm{Intend}_i\varphi U(\varphi \vee \neg\mathrm{EF}\varphi))$

4 结论

意愿是规范和描述多 Agent系统的一个重要抽象认知概念. 近年来, 意愿概念的理论研究在计算机科学领域引起了人们的注意和重视[1,3,4]. 本文提出了多 Agent系统计算的意愿理论以支持多 Agent系统计算的理论研究. 不同于已有研究, 我们没有基于可能世界间的可达关系来定义意愿概念的形式化语义, 而是将 Agent的意愿视为 Agent对世界发展轨迹的选择. 基于这一语义定义, 本文获取和描述了意愿概念的一些重要逻辑属性.

参考文献

1 Cohen P R, Levesque H J. Intention is Choice with Commitment. Artificial Intelligence, 1990, 42: 213~261

2 Emerson E A. Temporal and Modal Logic. Handbook of Theoretical Computer Science. Edited by J. Van Lecuwen, 1990

3 Haddadi A, Sundermeyer K. Belief-Desire-Intention Agent Architectures. Foundations of Distributed Artificial Intelligence, edited by G. M. P. O'Hare, N. R. Jennings, 1996

4 Singh M P. Multiagent System: A Theoretical Framework for Intentions, Know-how, and Communications. Springer-Verlag, 1994