Санкт-Петербургский государственный электротехнический университет "ЛЭТИ"
кафедра Вычислительной техники

Магистерская программа
«Семантические технологии и многоагентные системы»

Дисциплина:

«Интеллектуальные агенты и многоагентные системы»

Лекция 3

Логические модели ИА

Преподаватель:
М.Г. Пантелеев

Санкт –Петербург
2021 г.

СПбГЭТУ «ЛЭТИ»
Кафедра Вычислительной техники

Магистерская программа
«Семантические технологии и многоагентные системы»

Дисциплина: «Интеллектуальные агенты и многоагентные системы»

Лекция 3
Формальные (логические) модели ИА

План

- Агенты как интенсиональные системы
- Ограниченность классических логик
- Модальные логики как формальные модели ИА
- Нормальная модальная логика знания
- Формализм Мура
- Теория намерений *P.Cohen* и *H.Levesque*

Формальные модели ИА

• *ИА* принято рассматривать как *интенсиональные системы* и описывать понятиями, отражающими *ментальные состояния* агента: *убеждения*, *намерения* и т. п.

В англоязычной литературе для обобщенного представления таких категорий используется термин “*attitudes*”, т. е. *позиция* агента, его отношение к окружающему миру.

Иногда выделяют две категории:

- *информационная позиция* (information attitudes) – отражает состояние информированности агента об окружающем мире, включает: *убеждения* (belief) и *знания* (knowledge);
- *активная позиция* (pro-attitudes) – отражает активное отношение агента к окружающему миру, включает: *желания* (desire), *намерения* (intention), *обязательства* (commitment) и т. п.

З*нания* – постоянные знания агента о себе и внешнем мире (включая других агентов), которые остается неизменными на протяжении всего жизненного цикла агента.

Убеждения (мнения) – знания агента, которые могут меняться в процессе его функционирования.

Желания – состояния мира (включая самого агента), достижение которых агент считает для себя желательным.

Намерения – действия, которые агент собирается выполнить для осуществления своих желаний или в силу взятых на себя обязательств. Отличие намерений от желаний состоит в том, что, желания могут быть противоречивыми и агент, зная об этом, не ставит себе целью осуществить все желания.

Обязательства – задачи, которые агент берет на себя в рамках кооперации с другими агентами по их просьбе, поручению или в результате реализации переговоров о координации совместных действий (сотрудничестве).

Формальные модели ИА: Ограниченность ЛППП

- Ограниченность классической логики предикатов первого порядка (ЛППП):
  - не позволяет адекватно работать с рассмотренными выше ментальными категориями.

- Пример: высказывание: “**Иван *считает* (*убежден*), что Федор является отцом Петра**”.

- Попытка записать это высказывание средствами ЛППП :

Считает (Иван, ***Отец* (Федор, Петр)**).

- 2-ой аргумент предиката *Считает* сам является формулой ЛППП – нарушен синтаксис ЛППП (аргументом предиката может быть *только терм*!)

- Формальные модели ИА строятся с использованием перечисленных выше категорий и различных неклассических логик, описывающих взаимосвязи между ними.

Модальные логики как формальные модели ИА

- Формальные модели ИА – часто строятся с использованием различных *модальных логик*.
- **Модальная логика** – логика, *содержащая*, помимо стандартных логических связок, переменных и/или предикатов, *модальные операторы* (модальности).
- В соответствии с *типами* используемых *модальностей*, существует множество модальных логик:
  - *алетические*:
    - «необходимо», «возможно»;
  - *временные*:
    - «когда-то в будущем», «всегда в прошлом», «всегда» и т. д.; Пример модального утверждения: «Москва всегда была столицей России»
  - *пространственные*:
    - «здесь», «где-то», «близко» и т. д.;
  - *эпистемические* исследуют высказывания, содержащие теоретико-познавательные понятия: «полагает», «сомневается», «проверяемо» (верифицируемо), «непроверяемо», и т. п.
    - пример закона: «Невозможно полагать что-либо и одновременно сомневаться в этом», «Если субъект убежден в чем-то, неверно, что он убежден и в обратном» и т. п.
  - логики знания («известно, что»), доказуемости («можно доказать, что»), деонтические (изучают логические связи нормативных высказываний), оценок (аксиологические модальности)…

Общность модальных логик

- Модальные понятия разных типов имеют общие формальные свойства
- Независимо от того, к какой группе относятся модальности, они определяются друг через друга по одной и той же схеме:
  - нечто возможно, если противоположное не является необходимым;
  - разрешено, если противоположное не обязательно;
  - допускается, если нет убеждения в противоположном.
  - случайно то, что не является ни необходимым, ни невозможным.
  - безразлично то, что не обязательно и не запрещено.
  - неразрешимо то, что недоказуемо и неопровержимо, и т. п.
- Подобным образом сравнительные модальные понятия разных групп определяются по одной и той же схеме:
  - "первое лучше второго" равносильно "второе хуже первого",
  - "первое раньше второго" равносильно "второе позже первого",
  - "первое причина второго" равносильно "второе следствие первого" и т. д.

Логические модели ИА

- Любой логический формализм характеризуется:
  - языком (синтаксисом)
  - семантикой (моделью)

Соответственно, при разработке логических формализмов для *интенсиональных* понятий приходится решать *два вида проблем*:

- синтаксические (способ записи) и
- семантические (интерпретация)

- Основные подходы к синтаксическим проблемам:
  - использование языка, содержащего *модальные операторы,* применяемые к формулам;
  - использование *метаязыка* – языка первого порядка, содержащего термы, обозначающие формулы некоторого другого *объектного языка*.

Нормальная модальная логика знания

- *Нормальные модальные логики* с *семантикой возможных миров*.

- *Семантика возможных миров* – метод логического анализа *модальных* и *интенсиональных* понятий, основанный на рассмотрении мыслимых положений дел (идеальных альтернатив, описаний состояний). В основе - представление о том, что у настоящего может быть не одно, а несколько направлений развития в будущем (возможные миры).

- *Пропозициональная модальная логика знаний/убеждений* расширяет синтаксис классической пропозициональной логики оператором ***K* – "знать"**.

- Алфавит включает:
  - множество элементарных высказываний Ф = {*p, q, r, . . .* };
  - логические связки: $\neg$, $\wedge$, $\vee$, $\rightarrow$;
  - **модальный оператор *K* ("знать").**

- *Правильно построенная формула* (ППФ):
  - любой элемент множества Ф;
  - отрицание, конъюнкция и дизъюнкция ППФ;
  - выражение вида *K*(*wff*), где (*wff*) – ППФ.

Таким образом, допускается вложенность оператора *K*.

Примеры формул: $K(p \wedge q)$, $K(p \wedge Kq)$.

Нормальная модальная логика знания

- Семантика формул определяется *множеством возможных миров*, соответствующих убеждениям агента

- Пусть W – множество миров (возможных состояний),
- $R \subseteq W \times W$ – бинарное отношение на W, характеризующее, *какие миры* (состояния) *агент считает возможными*

- Если $(w, w') \subseteq R$, то агент, находящийся в мире w, по его убеждению, может находиться в мире w'.

- Семантика формулы задается отношением к миру.
- **Формула $K\phi$ – истинна в мире w, если ϕ истинна во всех мирах w', таких что $(w, w') \subseteq R$.**

- Это определение обладает двумя базовыми свойствами:

  1. Справедлива следующая схема аксиом: $K(\phi \Rightarrow \psi) \Rightarrow (K\phi \Rightarrow K\psi)$
  2. Если справедливо ϕ, то справедливо $K\phi$.

Таким образом, *знания* агента *замкнуты относительно логического следствия.*

Нормальная модальная логика знания

• Свойства этой логики зависят *от ограничений*, налагаемых на *отношение достижимости R*. Наиболее важные из них :

T. $K\phi \Rightarrow \phi$ (аксиома *знания*: *все, что агент знает – истинно*)

D. $K\phi \Rightarrow \neg K \neg \phi$ (аксиома *непротиворечивости знаний*: *если агент знает* ϕ, *он не может знать* $\neg\phi$)

4. $K\phi \Rightarrow KK\phi$ (аксиома *позитивной интроспекции*: *если агент знает* ϕ, *он знает, что он знает* ϕ)

5. $\neg K\phi \Rightarrow K\neg K\phi$ (аксиома *негативной интроспекции*: *агент осведомлен о том, чего он не знает*)

• В зависимости от свойств, которыми требуется наделить агента, могут отбираться разные аксиомы.

• Совокупность всех перечисленных аксиом (KTD45) образует логическую систему S5, называемую ***логикой идеализированных знаний***.

• S5 без T представляет собой *слабую* S5 или KD45.

Формальные модели ИА

- При построении практических агентов ключевым аспектом является *взаимодействие между знаниями и действиями*.
- R.Moore предложил подход к этой проблеме [Moore R.C. A formal theory of knowledge and action// *Readings in Planning*, Morgan Kaufmann, 1990], основанный на следующих идеях:
  - используется модальная логика с семантикой Крипке, действия представляются в стиле динамической логики;
  - показано, как семантика Крипке может быть аксиоматизирована в метаязыке первого порядка;
  - модальные формулы, используя аксиоматизацию, транслируются в метаязык;
  - доказательство модальных теорем сводится к доказательству теорем метаязыка.

Взаимодействие знаний и действий: формализм Мура

Два аспекта взаимодействия знаний и действий в подходе Мура:

1. Для выполнения неких действий агенту необходимы знания, являющиеся *предусловиями* действий.

   - например, чтобы открыть сейф надо знать шифр

2. Агент может *приобретать знания в результате выполнения действий*, поэтому он может выполнять «тестовые» действия для пополнения знаний.

Аксиомы стандартных логических связок:

$\forall w.True\ (w, \ulcorner \neg\phi \urcorner) \Leftrightarrow \neg True\ (w, \ulcorner \phi \urcorner)$;

$\forall w.True\ (w, \ulcorner \phi \wedge \psi \urcorner) \Leftrightarrow True\ (w, \ulcorner \phi \urcorner) \wedge True\ (w, \ulcorner \psi \urcorner)$;

$\forall w.True\ (w, \ulcorner \phi \vee \psi \urcorner) \Leftrightarrow True\ (w, \ulcorner \phi \urcorner) \vee True\ (w, \ulcorner \psi \urcorner)$;

$\forall w.True\ (w, \ulcorner \phi \Rightarrow \psi \urcorner) \Leftrightarrow True\ (w, \ulcorner \phi \urcorner) \Rightarrow True\ (w, \ulcorner \psi \urcorner)$;

$\forall w.True\ (w, \ulcorner \phi \Leftrightarrow \psi \urcorner) \Leftrightarrow True\ (w, \ulcorner \phi \urcorner) \Leftrightarrow True\ (w, \ulcorner \psi \urcorner)$,

где *True* – предикат метаязыка;
w – мир;
ϕ и ψ – формулы модального языка;
$\ulcorner\ \urcorner$ – кавычки для выделения формул модального языка (кавычки Фреге).

Базовая аксиома “знания”

- Базовая аксиома связки “знать” в семантике возможных миров:

$$\forall w.True(w, \ulcorner (Know\ \phi) \urcorner) \Leftrightarrow \forall w'.K(w, w') \Rightarrow True(w', \ulcorner \phi \urcorner),$$

где K – предикат метаязыка, выражающий *знание отношения достижимости миров*

- Свойства знания выражены следующими аксиомами:

$\forall\ w.K(w, w)$; (рефлексивность)

$\forall\ w,w',w''.K(w, w') \wedge K(w', w'') \Rightarrow K(w, w'')$; (транзитивность)

$\forall\ w,w',w''.K(w, w') \wedge K(w'', w') \Rightarrow K(w, w'')$ (евклидовость)

В соответствии с этими аксиомами K является отношением эквивалентности.

Представление действий

$R(a, w, w')$ – предикат метаязыка, введен для представления *действий:*

- *результатом действия* a *в мире* w *может быть мир* w'.

$(Res\ a, \phi)$ – модальный оператор, означающий, что после выполнения *действия* a *станет истинным* ϕ.

$$\forall w.True(w, \ulcorner(Res\ a, \phi)\urcorner) \Leftrightarrow \exists w'.R(a, w, w') \wedge \forall w''.R(a, w, w'') \Rightarrow True(w'', \ulcorner \phi \urcorner)$$

Первый конъюнкт говорит о том, что *действие возможно*, второй – что *необходимым следствием выполнения действия* a *является* ϕ.

Способность (*ability*) определяется через модальный оператор *Can*:

$$\forall w.True\ (w, \ulcorner(Can\ \phi)\urcorner) \Leftrightarrow \exists a.True(w, \ulcorner(Know\ (Res\ a\ \phi))\urcorner)$$

Агент может достигнуть ϕ, если существует некоторое действие a, такое что агент знает, что результатом выполнения a является ϕ.

Формализм Мура

Более слабое определение *способности* учитывает случай, когда *агент выполняет действие, чтобы узнать, как достичь цели*:

$$\forall w.True(w,\ulcorner(Can\ \phi)\urcorner) \Leftrightarrow \exists a.True(w,\ulcorner(Know\ (Res\ a\ \phi))\urcorner) \vee \exists a.True(w,\ulcorner(Know\ (Res\ a\ (Can\ \phi)))\urcorner).$$

Формализм Мура - первая серьезная попытка применить к рациональным агентам аппарат матлогики (включая модальную и динамическую логики)

Недостатки:

1. Трансляция модального языка в язык первого порядка и последующее доказательство теорем в языке первого порядка *неэффективно* по сравнению с использованием встроенной в модальную логику процедуры доказательства.

2. Формулы, получающиеся в результате трансляции, сложны и интуитивно трудно понимаемы. Исходная структура (а значит смысл) при этом теряются.

3. Основан на возможных мирах, следствием чего является проблема логического всезнания. Определение способности неработающее.

Теория намерений *P.Cohen* и *H.Levesque*

(Cohen P., Levesque H. Intention is choice with commitment// *Artificial Intelligence,* 1990, V. 42)

- *Знания* и *убеждения* не полностью характеризуют агента.
- Необходимо также множество выражений, характеризующих *активную позицию* (pro-attitudes) агентов.
- При этом агенту необходимо достигать некоторого рационального баланса между его ментальными категориями.
  - Например, агент не должен быть перегружен обязательствами, но и не должен быть недогружен.
- Одним из подходов к согласованию компонентов когнитивного состояния агента является *теория намерений*, разработанная *P. Cohen* и *H.Levesque*
- Существует два вида намерений: *направленные на настоящее* и *направленные на будущее*

Теория намерений *P.Cohen* и *H.Levesque*

Идентифицировано *7 свойств*, которым должны *удовлетворять намерения*:

1. Намерения *определяют задачи* агента. Агент должен определить способы их решения:

   • *Если агент имеет намерение ϕ, ожидается, что он выделит ресурсы, чтобы решить, как добиться ϕ*

2. Намерения являются «фильтром» для адаптации других намерений, чтобы *избегать конфликта* между ними:

   • *Если агент имеет намерение ϕ, ожидается, что он адаптирует намерение ψ, так чтобы ϕ и ψ не являлись взаимно исключающими*

3. Агент отслеживает успешность своих намерений и *склонен к повторным попыткам*, если его попытки оказываются неудачными :

   • *Если первая попытка агента достичь ϕ оказываются неудачной, то при прочих неизменных условиях, он будет пытаться достичь ϕ с помощью альтернативного плана.*

Теория намерений *P.Cohen* и *H.Levesque*

4. Агент *убежден в осуществимости* своих намерений, т. е. в существовании по крайней мере одного способа претворить их в жизнь.

5. Агент *не считает*, *что* он *не осуществит* свои намерения.

6. Агент *считает*, *что* при определенных условиях он *осуществит* свои намерения.

7. Агент не обязательно имеет в качестве намерений все ожидаемые побочные эффекты своих намерений. Если агент имеет намерение ϕ и убежден, что $\phi \Rightarrow \psi$, он не обязательно имеет также намерение ψ. (Намерения не являются замкнутыми относительно импликации!).

- Последняя проблема известна как *проблема дантиста*: агент убежден, что визит к дантисту влечет боль и, тем не менее, намерен посетить его, не имея намерения испытать боль.

Многомодальная логика *P.Cohen* и *H.Levesque*

P.Cohen и *H.Levesque* использовали *многомодальную логику* со следующими основными конструкциями:

(*Bel x* ϕ) – *x* убежден в ϕ ;

(*Goal x* ϕ) – *x* имеет цель ϕ ;

(*Happens a*) – следующим будет действие *a*;

(*Done a*) – действие *a* только что произошло.

Используется семантика возможных миров. Каждый мир представляет собой бесконечно длинную линейную последовательность состояний. Каждый агент располагает:

1. *Отношением достижимости убеждений – B*. Каждой паре «агент, момент времени» сопоставляется множество достижимых миров убеждений.
   - Это отношение является транзитивным и евклидовым отношением порядка и порождает *логику убеждений KD45*

2. *Отношением достижимости целей – G*. Каждой паре «агент, момент времени» сопоставляется множество достижимых миров целей.
   - Это отношение порядка порождает *логику целей KD*

Многомодальная логика *P.Cohen* и *H.Levesque*

На убеждения и цели накладывается ограничение $G \subseteq B$. Это ограничение дает следующую, связывающую модальности, аксиому:

$$\models (Bel\ i\ \phi) \Rightarrow (Goal\ i\ \phi).$$

Кроме того, оно *выражает свойство реалистичности* – агенты принимают неизбежное.

Аксиома:

$$\models (Goal\ i\ \phi) \Rightarrow \Diamond\neg(Goal\ i\ \phi)$$

отражает два свойства:

- агенты *не придерживаются целей вечно*;
- агенты не откладывают работу по достижению целей на неограниченный срок.

Многомодальная логика *P.Cohen* и *H.Levesque*

Для описания структуры последовательности событий вводятся операторы:

α; α ' – за α следует α'

α? – действие-проверка

Определяются операторы темпоральной логики:

"$\Box$" – "*всегда*";

"$\Diamond$" – "*иногда*";

Later – "позднее";

(*Before p q*) – оператор предшествования.

Операторы выражаются через друг друга:

$\Diamond \alpha \equiv \exists x.(Happens\ x; \alpha?)$;

$\Box \alpha \equiv \neg\Diamond \neg\alpha$;

$(Later\ p) \equiv \neg p \vee \Diamond p$.

Представление устойчивой цели и намерения

Описание *устойчивой* (*persistent*) *цели*:

$$(P\text{–}Goal\ x\ \ p) \equiv (Goal\ x\ (Later\ p)) \wedge (Bel\ x\ \neg p) \wedge$$
$$[Before\ ((Bel\ x\ p) \vee (Bel\ x\ \Box\neg p))\quad \neg(Goal\ x\ (Later\ p))]$$

Таким образом, агент имеет *устойчивую цель p*, если:

1. Он имеет цель, чтобы *p со временем стало истинным* и убежден, что в данный момент *p* не является истинным.

2. Прежде чем агент отбросит цель, должно быть выполнено одно из следующих условий:
   - агент убежден, что цель достигнута *или*
   - агент убежден, что цель никогда не будет достигнута.

Представление намерения:

$$(Intend\ x\ \alpha) \equiv (P\text{–}Goal\ \ x\ [Done\ x\ (Bel\ x\ (Happens\ \alpha))?;\ \alpha]\ )$$

Таким образом, агент *имеет намерение* сделать α , если он имеет *устойчивую цель* убедиться в том, что он готов сделать α и затем сделать α.

Недостатки теории намерений *P.Cohen* и *H.Levesque*

- не вводится адекватное понятие способностей (компетенций);
- отсутствует адекватное представление намерений для выполнения составных действий;
- требуется, чтобы агент точно знал, что он собирается делать – допущение исчерпывающе продуманных намерений;
- не допускаются множественные намерения.

Логические модели ИА позволяют формально описать связи между ментальными категориями, однако их *практическое применение при проектировании агентных систем в настоящее время ограничено.*