

# **Приобретение знаний. Извлечение знаний из данных.**

Курс «Интеллектуальные  
информационные системы»

Лекция 7

# Приобретением знаний

называется выявление знаний из источников и преобразование их в нужную форму, а также перенос в базу знаний интеллектуальной системы.

# Источники знаний:

- Книги, архивные документы, содержимое других баз знаний, т.е. некоторые *объективизированные знания*, приведенные к форме, которая делает их доступными для потребителя;
- ***Экспертные знания***, которые имеются у специалистов, но не зафиксированы во внешних по отношению к ним хранилищах (*субъективные*);
- ***Эмпирические знания (субъективные)***, получающиеся путем наблюдения за окружающей средой (если у ИС есть средства наблюдения)

# Методология приобретения субъективных знаний

Две формы представления:

- модели и формы хранения знаний у человека – эксперта
- Модель, по которой инженер по знаниям (проектировщик ИИС), собираются их описывать

# Схема приобретения знаний



# Когнитивные структуры знаний

- Представление класса понятий через его элементы

*Птица = <чайка, воробей, скворец, ...>*

- Представление понятий класса с помощью базового прототипа

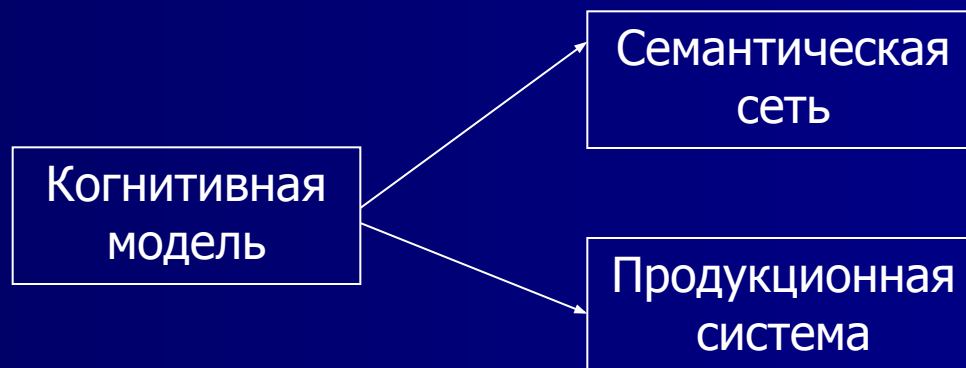
*Птица = <нечто с крыльями, с клювом, летает, ...>*

- Представление с помощью признаков

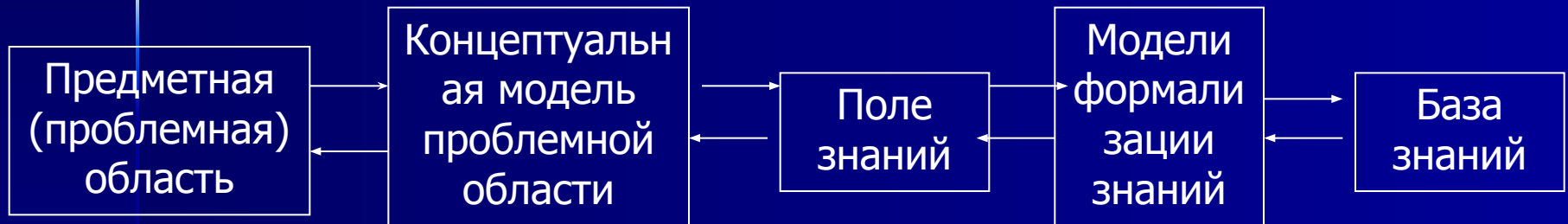
*Птица = <крылья, клюв, две лапы, перья...>*

# Представление КОГНИТИВНОЙ модели

- Отношения между понятиями определяются процедурным способом
- Отношения между оставляющими понятий – декларативным способом.



# Формирование БЗ в ИИС



- Переход от описания некоторой области в поле знаний аналогичен переходу от концептуальной модели БД к ее логической схеме



	<b>Способы извлечения знаний</b>	<b>Способы передачи знаний</b>
<b>документы</b>	<b>Пассивный источник знаний</b>	<b>письменный</b>
<b>специалисты</b>	<b>Активный источник знаний</b>	<b>устный</b>

# Схема приобретения знаний

**Носитель информации →  
Посредник → Модель знания**

**Посредник – человек, обладающий  
специфическими знаниями инженер по  
знаниям или когнитолог**

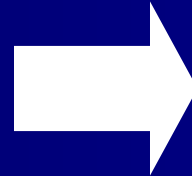
# Причины использования посредника

1. Эксперт владеет субъективными знаниями, которые не всегда можно выразить словами, упускает промежуточные звенья цепочки вывода.
2. Объясняющий в процессе объяснения сам начинает лучше понимать проблему
3. Посреднику, владеющему меньшим объемом знаний о ПО, проще постепенно строить целостную модель Предметной области (ПО)

# **В качестве посредника могут использоваться**

- Инженер по знаниям (когнитолог)
- Специальная программа

По отношению к носителю предметного знания посредник должен обладать метазнанием, к которому относится знание следующих научных областей



- Системный анализ
- Математика
- Модели знания
- Машинное представление моделей знания
- Основы проектирования программных систем
- Психология
- Лингвистика
- Изобразительное искусство
- Музыка

Специалист, обладающий перечисленными знаниями, называется ***системным аналитиком***

# *Приобретением знаний*

называют процесс получения знаний от эксперта или каких-либо других источников и передачи их в ИИС.

Применяют также термины

**извлечение и**

**формирование знаний**

# Правила использования терминов

1. Если при разработке ИИС процесс получения знаний от экспертов происходит без использования компьютерных средств путем непосредственного контакта – это ***извлечение знаний***.
2. Если процесс осуществляется с использованием специальных программных средств – ***приобретение знаний***.
3. Если процесс осуществляется с использованием программ обучения при наличии репрезентативной выборки примеров принятия решений в ПО – ***формирование знаний***.

# Три стратегии получения знаний при разработке ИИС

Этапы разработки ИИС







# Классификация методов извлечения знаний

# Коммуникативные методы.

## *Наблюдение*

- Используется в случаях, когда участие инженера по знаниям (ИЗ) в реальном процессе невозможно. («Чистый» метод) Может потребовать:
  - Техники стенографирования и хронометрирования
  - Серьезного предварительного знакомства с ПО

# Коммуникативные методы. *Анализ протоколов «мыслей вслух»*

- ИЗ протоколирует все слова эксперта

## *Лекции*

- ИЗ ведет конспекты, по ходу лекции задает вопросы.

# Коммуникативные методы.

## *Анкетирование*

Требования к анкете:

1. Она не должна быть монотонной
2. Должна быть приспособлена к языку экспертов
3. Должна быть продумана последовательность вопросов
4. Допускается избыточность вопросов для перепроверки ответов

# Коммуникативные методы.

## *Интервью*

- Серия заранее подготовленных вопросов. На качество интервью влияют:
  1. Язык вопросов (понятность, лаконичность, терминология);
  2. Порядок вопросов (логическая последовательность и немонотонность);
  3. Уместность вопросов (этика, вежливость)

# Коммуникативные методы.

## *Свободный диалог*

- Метод извлечения знаний в форме беседы ИЗ с экспертом, в которой нет жесткого регламентированного плана и вопросника. Следует выбрать правильный темп беседы, не утомляющий эксперта

# Коммуникативные методы.

## *Игры с экспертом*

- *Учитель и ученик* – эксперт выявляет и исправляет ошибки ученика.
- *Медицина – ИЗ* – врач , эксперт - консультант

# Коммуникативные методы.

## *Круглый стол*

- Обсуждение проблем ПО в присутствии привлеченных экспертов, обладающих равными правами. Роль ИЗ – организация обсуждений



# Коммуникативные методы.

## *«Мозговой штурм»*

- Участникам (до 10 чел.) предлагается высказывать любые идеи: чем больше идей, тем лучше. Идеи оцениваются группой экспертов, не участвовавших в их генерации. Эффективен для новых ПО.

## *Ролевые игры*

Используются для обучения персонала.

Эксперты сами распределяют роли между собой

# Текстологические методы.

- Данная группа методов объединяет методы извлечения знаний, основанные на изучении специальных текстов в учебниках, монографиях, статьях, методик и других носителей профессиональных знаний

# Схема извлечения знаний

$M_1 \rightarrow \text{Вербализация} \rightarrow \text{Текст} \rightarrow$   
 $\text{Понимание} \rightarrow M_2$

$M_1$  – модель мира автора текста;

$M_2$  – модель, возникающая при чтении  
текста (модель ИЗ)

Модели  $M_1$  и  $M_2$  не могут совпадать в силу  
искажения смысла при вербализации  $M_1$   
и интерпретации  $M_2$

Научный текст строится из следующих компонент:

- a) Наблюдения объективной информации;
- b) Системы научных понятий;
- c) Взгляды и опыт автора;
- d) Общие места
- e) Заимствования (материалы из других источников)

Модель автора:  $M_1 = \langle a, b, c, d, e \rangle$

Модель ИЗ формируется из экстракта  $\langle a, b, c, e \rangle'$  прочитанного текста и индивидуальных свойств ИЗ.

Индивидуальные свойства ИЗ  
характеризуются:

f) Личным опытом

g) Общенаучной эрудицией

h) Предварительными сведениями о  
ПО.

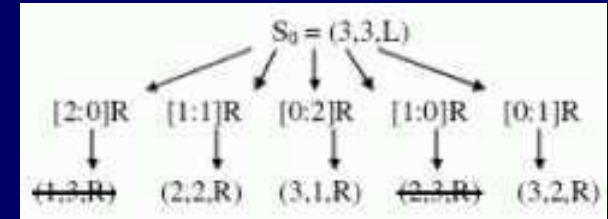
Модель ИЗ имеет вид

$$M_2 = [ \langle a, b, c, e \rangle', \langle f, g, h \rangle ]$$

# Методы поиска решений в пространстве

- **Задача.** Три миссионера и три людоеда находятся на левом берегу реки и им нужно переправиться на правый берег, однако у них имеется только одна лодка, в которую могут сесть лишь 2 человека. Поэтому необходимо определить план, соблюдая который и курсируя несколько раз туда и обратно, можно переправить всех шестерых. Однако если на любом берегу реки число миссионеров будет меньше, чем число людоедов, то миссионеры будут съедены. Решения принимают миссионеры, людоеды их выполняют.

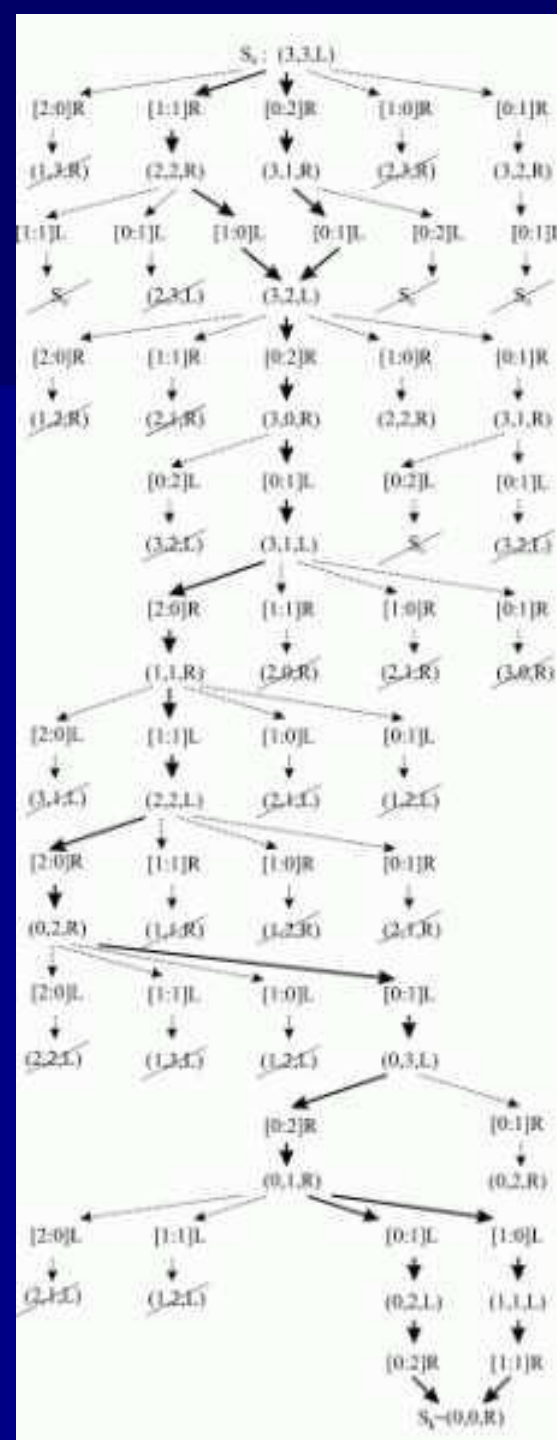
# Основой метода являются следующие этапы:



1. Определяется конечное число состояний, одно из состояний принимается за начальное и одно или несколько состояний определяются как искомое (конечное, или терминальное). Обозначим состояние  $S$  тройкой  $S=(x,y,z)$ , где  $x$  и  $y$  - число миссионеров и людоедов на левом берегу,  $z= \{L,R\}$  - положение лодки на левом (L) или правом (R) берегах. Итак, начальное состояние  $S_0=(3,3, L)$  и конечное (терминальное) состояние  $S_k=(0,0, R)$ .
2. Заданы правила перехода между группами состояний. Введем понятие действия  $M:[u, v]w$ , где  $u$  - число миссионеров в лодке,  $v$  - число людоедов в лодке,  $w$  - направление движения лодки (R или L).
3. Для каждого состояния заданы определенные условия допустимости (оценки) состояний:  $x \geq y$ ;  $3-x \geq 3-y$ ;  $u+v \leq 2$ .
4. После этого из текущего (исходного) состояния строятся переходы в новые состояния, показанные на рисунке. Два новых состояния следует сразу же вычеркнуть, так как они ведут к нарушению условий допустимости (миссионеры будут съедены).
5. При каждом переходе в новое состояние производится оценка на допустимость состояний и если при использовании правила перехода для текущего состояния получается недопустимое состояние, то производится возврат к тому предыдущему состоянию, из которого было достигнуто это текущее состояние. Эта процедура получила название бэктрекинг (bac tracing или BACKTRACK).

# Метод поиска в пространстве состояний

- Теперь мы можем проанализировать полностью алгоритм простейшего поиска решений в проблемном пространстве, описанный группами состояний и переходами между состояниями на рисунке. Решение задачи выделено жирными стрелками. Такой метод поиска  $S_0 \rightarrow S_k$  называется **прямым** методом поиска. Поиск  $S_k \rightarrow S_0$  называют **обратным** поиском. Поиск в двух направлениях одновременно называют **двунаправленным поиском**.





- фундаментальным понятием в методах поиска в ИС является идея рекурсии и процедура BACKTRACK. В качестве примера многоуровневого возвращения рассмотрим задачу размещения на доске 8 x 8 восьми ферзей так, чтобы они не смогли "съесть" друг друга.
- Допустим, мы находимся на шаге размещения ферзя в 6 ряду и видим, что это невозможно. Процедура BACKTRACK пытается переместить ферзя в 5 строке и в 6 строке опять неудача. Только возврат к 4 строке и нахождение в ней нового варианта размещения приведет к решению задачи.

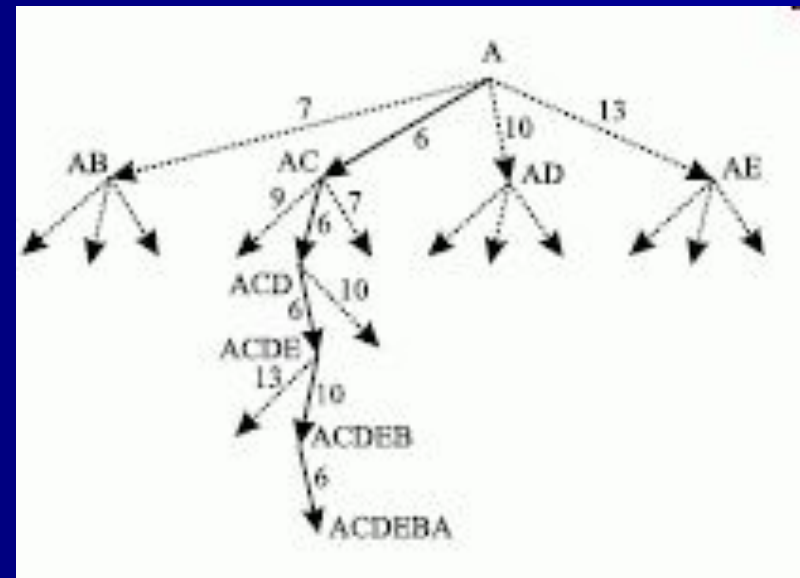
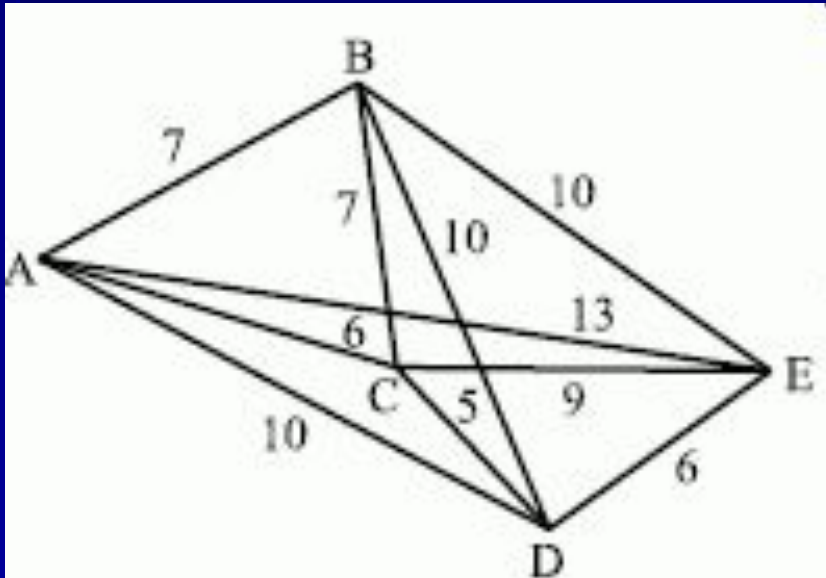
X							
		X					
				X			
	X						
			X				

# Алгоритмы эвристического поиска

- В рассмотренных примерах поиска решений число состояний невелико, поэтому перебор всех возможных состояний не вызвал затруднений. Однако при значительном числе состояний время поиска возрастает экспоненциально, и в этом случае могут помочь алгоритмы эвристического поиска, которые обладают высокой вероятностью правильного выбора решения. Рассмотрим некоторые из этих алгоритмов.

# Алгоритм наискорейшего спуска по дереву решений

- Пример построения более узкого дерева рассмотрим на примере задачи о коммивояжере. Торговец должен побывать в каждом из 5 городов, обозначенных на карте
- Задача состоит в том, чтобы, начиная с города А, найти минимальный путь, проходящий через все остальные города только один раз и приводящий обратно в А. Идея метода исключительно проста - из каждого города идем в ближайший, где мы еще не были.



# Алгоритм оценочных (штрафных) функций

- Умело подобранные оценочные функции (в некоторых источниках - штрафные функции) могут значительно сократить полный перебор и привести к решению достаточно быстро в сложных задачах. В нашей задаче о людоедах и миссионерах в качестве самой простой целевой функции при выборе очередного состояния можно взять число людоедов и миссионеров, находящихся "не на месте" по сравнению с их расположением в описании целевого состояния. Например, значение этой функции  $f = x + y$  для исходного состояния  $f_0 = 6$ , а значение для целевого состояния  $f_1 = 0$ .

- Эвристические процедуры поиска на графе стремятся к тому, чтобы минимизировать некоторую комбинацию стоимости пути к цели и стоимости поиска. Для задачи о людоедах введем оценочную функцию:
  - $f(n) = d(n) + w(n)$   
где  $d(n)$  - глубина вершины  $n$  на дереве поиска и  $w(n)$  - число находящихся не на нужном месте миссионеров и людоедов. Эвристика заключается в выборе минимального значения  $f(n)$ .  
Определяющим в эвристических процедурах является выбор оценочной функции.

- Рассмотрим вопрос о сравнительных характеристиках оценочных целевых функций на примере функций для игры в "8" ("пятнашки"). Игра в "8" заключается в нахождении минимального числа перестановок при переходе из исходного состояния в конечное (терминальное, целевое).

2	8	3
1	6	4
7	*	5

1	2	3
8	*	4
7	6	5

Рассмотрим две оценочные функции:

$$h_1(n) \& = Q(n)$$

$$h_2(n) \& = P(n) + 3S(n),$$

где  $Q(n)$  - число фишек не на месте;  $P(n)$  - сумма расстояний каждой фишки от места в ее целевой вершине;

$S(n)$  - учет последовательности нецентральных фишек (штраф +2 если за фишкой стоит не та, которая должна быть в правильной последовательности; штраф +1 за фишку в центре; штраф 0 в остальных случаях).

# Сравнение этих оценочных функций приведено в таблице

Оценочная функция $h$	Стоимость (длина) пути $L$	Число вершин, открытых при нахождении пути $N$	Трудоёмкость вычислений, необходимых для подсчета $h S$	Примечания
$h_1 S_0$ $S_1$	5 >18	13 $100-8!(=40320)$	8	Поиск в ширину
$h_2 S_0$ $S_1$	5 18	11 43	$8*2+8+1+1$	Поиск в глубину



# На основе сравнения ЭТИХ ДВУХ оценочных функций можно сделать выводы.

- Основу алгоритма поиска составляет выбор пути с минимальной оценочной функцией.
- Поиск в ширину, который дает функция  $h_1$ , гарантирует, что какой-либо путь к цели будет найден. При поиске в ширину вершины раскрываются в том же порядке, в котором они порождаются.
- Поиск в глубину управляется эвристической компонентой  $3S(n)$  в функции  $h_2$  и при удачном выборе оценочной функции позволяет найти решение по кратчайшему пути (по минимальному числу раскрываемых вершин). Поиск в глубину тем и характеризуется, что в нем первой раскрывается та вершина, которая была построена самой последней.
- Эффективность поиска возрастает, если при небольших глубинах он направляется в основном в глубь эвристической компонентой, а при возрастании глубины он больше похож на поиск вширь, чтобы гарантировать, что какой-либо путь к цели будет найден. Эффективность поиска можно определить как  $E=K/L*N*S$ , где  $K$  и  $S$  (трудоемкость) - зависят от оценочной функции,  $L$  - длина пути,  $N$  - число вершин, открытых при нахождении пути. Если договориться, что для оптимального пути  $E=1$ , то  $K=L^0*N^0*S^0$ .



- Развивая метод минимакса, назначим вероятности для выполняемых действий в задаче о миссионерах и людоедах:

$$P([2 : 0]R) = 0,8; \quad P([1 : 1]R) = 0,5;$$

$$P([0 : 2]R) = 0,9;$$

$$P([1 : 0]R) = 0,3; \quad P([0 : 1]R) = 0,3;$$

- Интуитивно понятно, что посылать одного людоеда или одного миссионера менее эффективно, чем двух человек, особенно на начальных этапах. На каждом уровне мы будем выбирать состояние по критерию  $P_i$ . Даже такой простой подход позволит нам избежать части тупиковых состояний в процессе поиска и сократить время по сравнению с полным перебором. Кстати, этот подход достаточно распространен в экспертных продукционных системах.