Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»

Е.К. Пугачев

Извлечение, структуризация и формализация знаний

**Учебно-методическое пособие
по выполнению домашнего задания по дисциплине «Интеллектуальные технологии и системы»**

**МОСКВА
2022г**

УДК 004.021
ББК 32.973-018.2
И20

Издание доступно в электронном виде по адресу:
https://
Факультет «Информатика и системы управления»
Кафедра «Компьютерные системы и сети»

Рекомендовано
Научно-методическим советом МГТУ им. Н.Э. Баумана
в качестве учебно-методического пособия

Извлечение, структуризация и формализация знаний: учебно-методическое пособие / Е. К. Пугачев. — Москва : Издательство МГТУ им. Н. Э. Баумана, 2022. — , [ ] с. : ил.

ISBN ……..

Представлены краткое описание способов извлечения знаний, классификация видов знаний, способов декомпозиции и формализации предметной области задачи, примеры определения существенной информации и варианты классификации декларативных знаний, способы оценки результатов, варианты заданий, порядок выполнения и требования к защите домашнего задания предусмотренных учебным планом МГТУ им. Н.Э. Баумана.

Для студентов МГТУ им. Н.Э. Баумана, обучающихся по направлению подготовки «Информатика и вычислительная техника».

УДК 004.021
ББК 32.973-018.2

Предисловие

Учебно-методическое пособие предназначено для получения практических навыков в области разработки интеллектуальных систем студентами, обучающимися на кафедре «Компьютерные системы и сети».

Издание составлено в соответствии с самостоятельно устанавливаемым образовательным стандартом (СУОС), основной образовательной программой по направлению подготовки 09.04.01 «Информатика и вычислительная техника» магистров и предназначено для подготовки к выполнению домашних заданий по дисциплине «Интеллектуальные технологии и системы».

Цель учебно-методического пособия — подготовка студентов к выполнению работы на этапе исследования предметной области задачи с целью получения данных, на основе которых может быть создана программная модель экспертной системы.

После выполнения домашних заданий студенты будут:

знать

• основные методы извлечения знаний;
• способы классификации фактов и правил;
• способы формализации видов знаний;
• способы графического представления знаний;
• способы оценки полноты выделенных знаний;

уметь

• выделять существенную информацию в предметной области задачи, учитывая цель создания системы искусственного интеллекта;
• определять качественные и количественные критерии оценки способов представления знаний;
• оценивать полученные результаты, учитывая перспективы создаваемой системы;
• определять главные предикаты и объекты в структурах фактов;
• строить возможные логические цепочки фактов предметной области;

владеть

• методиками построения базовых семантических структур, применительно к классическим способам представления знаний. • методами тестирования программных продуктов;
• способами декомпозиции предметной области;
• практическими навыками анализа методов обработки знаний;

Кроме того, студент получит следующие навыки:

• извлечения существенной информации применительно к конкретной задаче;
• классификации декларативных и процедурных знаний;

• оценки полноты и точности представляемых знаний.

Для выполнения домашнего задания необходимо предварительное освоение следующих дисциплин: основы программирования, объектно ориентированное программирование, технология разработки программных систем.

Одной из важных задач учебно-методического пособия является приобретение студентами способности извлекать знания предметной области, классифицировать, структурировать и формализовать их так, чтобы в дальнейшем было удобно представить в базе знаний экспертной системы и, при этом, могли бы эффективно обрабатываться на ЭВМ.

В домашнем задании 1 рассмотрены способы извлечения знаний предметной области задачи, представленных в текстовой форме. Приведены некоторые принципы классификации знаний, а также способы определения структур основных видов знаний.

В домашнем задании 2 описаны способы декомпозиции предметной области задачи, где особый акцент сделан на логическую, продукционную и инвариантную декомпозиции. Также показаны способы графического представления семантической информации для основных способов представления знаний.

В приложениях представлены примеры оформления домашних заданий.

Издание позволит студентам самостоятельно изучить важные разделы дисциплины «Интеллектуальные технологии и системы».

Для успешного освоения материала необходимо внимательно разбирать примеры, целесообразно синтезировать и прорабатывать аналогичные примеры.

Ответы на вопросы позволят обучающемуся самостоятельно оценить степень понимания и усвоения теоретических положений, методик и методов. Выполнение заданий обеспечит приобретение умений и навыков, необходимых для постановки и решения задач проектирования систем искусственного интеллекта.

Для оценки степени понимания и усвоения теоретических положений, методик и методов, приобретения умений и навыков служит информативный отчет по каждому домашнему заданию, где полно и точно представлены результаты и их защита. Домашние задания студенты выполняют самостоятельно, а полученные результаты используют в лабораторных работах, связанных с построением семантических и программных моделей, а также реализации их на языках искусственного интеллекта. Варианты заданий, требования к отчету и защите приведены отдельно по каждому домашнему заданию.

Введение

При создании систем искусственного интеллекта, например, экспертных диагностических систем, необходимо знать методы представления и обработки знаний, а также общие принципы проектирования семантических систем. Важно знать подходы проектирования, как системы в целом, так и отдельных компонентов. Например, чтобы разработать базу знаний, необходимо выбрать наиболее походящую модель представлений, определить все необходимые виды знаний и определить семантические структуры разных уровней. Правильное представление знаний предполагает, впоследствии, гибкую их обработку, например, параллельную, распределенную с возможностью самообучения и др.

Обработка знаний может осуществляться на основе разных алгоритмов, которые, в свою очередь, зависят от способа представления знаний. Например, для эффективной работы механизма логического вывода требуется обеспечить полноту базы знаний, а также наиболее точное представление всех видов знаний.

Удачно принятые решения, связанные с разработкой базы знаний, механизма вывода и других компонентов систем искусственного интеллекта, позволяют экономно использовать ресурсы вычислительной системы, например ресурсы оперативной и внешней памяти, а также процессорное время при принятии решений. Чтобы получить удачные решения, необходимо обладать знаниями, связанными с различными способами конструирования семантических структур различных уровней.

Выбор способов декомпозиций предметной области и грамотного их использования также является важной задачей. Какие способы декомпозиции проводить зависит от модели представления и вида знаний. При этом, важное значение имеет форма представления семантической информации в каждом конкретном случае. Например, если предполагается использовать модель представления с помощью правил, то необходимо провести логическую, продукционную, а в некоторых случаях, и инвариантную декомпозицию предметной области.

Важной задачей для студентов магистров является приобретение навыков, связанных с возможностями проведения качественного исследования предметной области задачи с целью получения исходных данных, на основе которых можно построить адекватную семантическую модель.

Домашнее задание 1

При разработке экспертной системы важной задачей является создать базу знаний, в которой заложена адекватная семантическая информация, описывающая объекты реального мира и их отношения. В рамках выполнения первого задания такими объектами являются предикаты, факты, логические цепочки фактов и др.

Исходными данными для решения вышеуказанной задачи являются описания предметной области на естественном языке. В частности, это может касаться вопросов, связанных с решением задач диагностики, где информация представлена в текстовой форме. Решение задач диагностики является актуальной и может касаться разных предметных областей, например:

- диагностирование заболеваний пациентов;
- определение неисправности автомобиля;
- диагностика компьютеров и др.

При выполнении домашнего задания необходимо ответить на ряд вопросов, связанных с извлечением и формализацией семантической информации представленной в текстовой форме, например,

• какие фразы представляют существенную информацию для решения поставленной задачи?

• какие элементы фраз являются главными?

• какие элементы фраз будут представлять в модели предикаты?

• какие факты связаны между собой?

• какие логические цепочки фактов являются наиболее эффективными?

Цель задания: выделить и классифицировать семантическую информацию с учетом цели, определить связанные цепочки фактов, формализовать извлеченные знания в вид, пригодный для использования при реализации продукционной модели.

Продолжительность работы: 9 часов.

1 Методы извлечения знаний

При выполнении домашних заданий необходимо провести исследование предметной области задачи с целью определения данных и знаний, с учетом которых будут приниматься проектные решения при разработке, например, экспертной системы.

Методы извлечения знаний делят на две группы: коммуникативные и текстологические. Источниками знаний, другими словами, источниками первичной семантической информации могут быть: эксперты предметной области, результаты исследований, учебники, техническая документация и др.

Получение знаний может осуществляться разными способами:

- эксперты, на основе своего опыта, сами определяют, какую информацию и в каком виде передавать разработчикам системы;
- эксперты отвечают на вопросы инженеров по знаниям, задача которых получить качественные результаты, связанные со структурированием знаний в форме, удобной для обработки на вычислительной машине;

- разработчики извлекают знания из литературных источников, где информация может быть представлена в текстовой или графической форме и др.

На рисунке 1 представлена классификация способов извлечения знаний, из которого видно, что отдельное место занимает выделение знаний из литературных источников.

Рисунок 1 – Классификация способов извлечения знаний

В рамках выполнения первого домашнего задания необходимо получить фрагмент описания предметной области на естественном языке. Далее, в соответствии с теорией информационных семантических систем, необходимо выполнить операцию сжатия семантической информации. Данная операция выполняется студентами вручную, поэтому результаты будут субъективными и не отвечать требованиям полноты и точности. Грамотное выполнение операции сжатия позволяет сократить объемы информации, что приводит к меньшей вычислительной сложности и положительно влияет на эффективность. С другой стороны, семантическое сжатие предполагает потери информации, поэтому данная операция сильно влияет процесс проектирования базы знаний.

Чтобы получить вторичную семантическую информацию, которая была бы адекватной первичной семантической информации необходимо учитывать следующие факторы:

- цели и подцели, закладываемые в экспертную систему;
- требования используемой модели представления знаний;
- квалификация разработчика, выполняющего операцию сжатия и др.

На рисунке 2 показано несколько шагов операции сжатия семантической операции, которые нужно выполнить в первом домашнем задании.

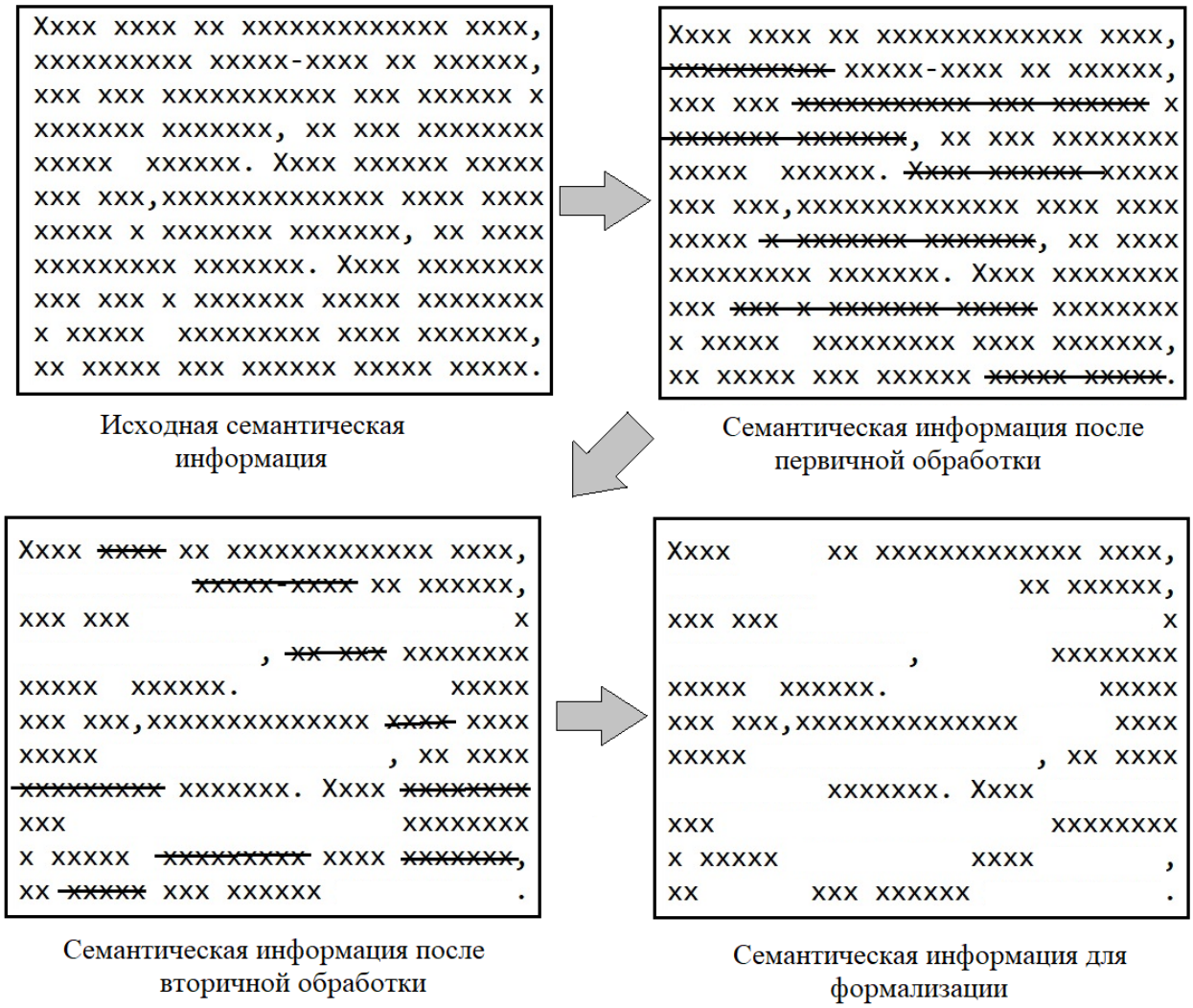

Рисунок 2 – Сжатие семантической информации текстовой формы представления

После первичной обработки текста получаем информацию, в которой, например, присутствует только те данные, которые необходимы во время работы механизма логического вывода для достижения цели экспертной системы.

После вторичной обработки получаем текст, в котором отсутствуют слова или словосочетания, которые, впоследствии, не будут влиять на эффективность работы, и при этом, не будут снижать точность представления семантической информации.

В итоге получаем существенную информацию, которая может быть пригодна для процесса формализации в соответствии со способом представления знаний.

2 Классификация знаний

При выполнении первого домашнего задания, предполагается, что будет использоваться продукционная модель.

В общем виде продукционная система представляется тремя компонентами:

$$S = (F, R, I)$$

где

F – множество фактов предметной области, участвующих в логических цепочках при достижении целей;

R – множество правил, полученных в результате продукционной декомпозиции;

I – интерпретатор, включающий процессы: определения активных фактов, сопоставления, разрешения конфликтов и выполнения правила.

На этапе исследования важно, извлеченные знания, правильно разбить на группы, т.к. это позволит использовать эффективные алгоритмы при поиске решений.

Чтобы провести классификацию семантической информации, необходимо учитывать, что видов или уровней знаний несколько:

- декларативные знания (факты конкретных ситуаций, факты базы знаний и др.);
- процедурные знания (на их основе строится вывод, они показывают, как факты связаны между собой и порядок их вывода);
- управляющие знания (могут представлять некоторый набор стратегий) и др.

В первом домашнем задании на проработку выносятся первые два вида знаний, т.е. декларативные и процедурные. Например, если необходимо разработать диагностическую медицинскую экспертную систему, то декларативные знания можно разбить на следующие группы:

- факты симптомы заболевания;
- факты, связанные с причинами заболеваний;
- факты диагнозы и др.

Под симптомами обычно понимаю статистически значимые отклонения тех или иных показателей от границ его нормальных величин. Возможно возникновение качественно нового, не свойственного здоровому организму явления.

В свою очередь, факты симптомы заболевания можно разбить на подгруппы:

- диагностические симптомы (свойственны только одному заболеванию);
- специфические симптомы (характерны для группы заболеваний органов одной системы);
- неспецифические симптомы (характерны для многих заболеваний);
- нехарактерные для данного заболевания симптомы.

В зависимости от вида метода исследования выделяют следующие виды:

- субъективные симптомы (выявляются методом расспроса больных, основаны на описании больным своих ощущений, возникающих в ходе развития болезни);
- объективные симптомы (осмотр, пальпация, перкуссия, аускультация, лабораторные и инструментальные исследования).

Также можно разделить в зависимости от времени появления симптома:

- начальные – возникающие на самых ранних стадиях развития болезни (например, боли в горле, боле в сердце, кашель и лихорадка и др.);
- поздние – возникающие в период разгара заболевания или в период его разрешения (например, налеты на миндалинах при ангине, бронхиальное дыхание при пневмонии и др.).

Могут быть использованы и другие принципы разбивки симптомов на группы, например, связанные с болью или нет и др.

Из вышесказанного следует, что требуется обладать глубокими знаниями предметной области задачи.

В рамках выполнения домашнего задания каждому студенту предлагается определенная предметная область и предлагается решить задачу диагностики. При этом, учитывается ограниченное время, отведенное на выполнение задания. Другими словами, предлагается проработать небольшой фрагмент, который касается, например, одного или двух заболеваний. С другой стороны, это связано с реализацией фрагментов двух компонентов экспертной системы: базы знаний и механизма вывода.

3 Формализация знаний

При разработке экспертной системы стоит задача – получить формальную модель предметной области, которая зависит от способа представления знаний.

При формализации знаний используют разные знаковые системы. В зависимости от типа знаковой системы знания могут быть представлены в виде естественных или искусственно-языковых, а также графических и других форм.

Дифференцированными методами формализации знаний являются: естественно-языковое описание, лексикографическое описание, тезаурусное описание, формально-языковое.

Формализация знаний на основе естественно-языковой знаковой системы приводит к формированию естественно-языкового описания. Текст может содержать как декларативную, так и процедурную компоненты знаний. Отдельные образы текста связаны между собой двумя типами отношений: синтагматическим и парадигматическими. Структура и характер отношений между синтагмами отражают все многообразие отношений между элементами знаний.

Для формализации знаний могут быть использованы искусственные знаковые системы, т.е. формальные языки. Могут вводиться абстрактные языковые конструкции, которым ставятся в соответствие объекты и процессы

реального мира, таким образом, осуществляется описание знаний в виде формальных моделей.

Для формализации знаний применяются: формальные грамматики, логические модели (дедуктивные и индуктивные модели, псевдофизические логики и др.), сетевые (простые, иерархические, однородные и неоднородные, функциональные, семантические, фреймовые, сценарии и др.), продукционные.

Инструментарий формально-языковой технологии описания знаний включает теоретические и практические методы определенных разделов математики (алгебры, теории множеств, теории формальных языков и программирования, математической лингвистики и др.), информатики, кибернетики и других научных дисциплин.

При выполнении первого домашнего задания предлагается использовать продукционную модель представления знаний с последующей реализацией ее на языке Пролог.

В общем виде факт можно представить так

$$<Q_{11},Q_{12},\ldots,Q_{1n}> \circ <Q_{21},Q_{22},\ldots,Q_{2m}>$$

где $Q_{11},Q_{12},..,Q_{1n},Q_{21},Q_{22},..,Q_{2m}$ - объекты,

$\circ$ - знак отношения.

Ниже приведен пример структуры представления декларативных знаний на языке Пролог , в частности фактов заключений разных уровней:

<Имя предиката>(<№ уровня>,<№ факта>,<Заключение на ЕЯ>).

В структуре факта предусмотрено наличие как управляющей информации (учетной), так и информации непосредственно касающейся семантики факта.

После выделения фактов из текста на естественном языке, необходимо определить их конструкцию. Конструкция факта предполагает выделение предикатов и объектов. Вариантов конкретных конструкций фактов может много, поэтому определить наиболее правильную конструкцию является важной задачей.

Рассмотрим пример определения структуры факта. Пусть дана фраза на естественном языке в текстовой форме: «Высокая температура тела», которая представляет собой факт-симптом, однако, может быть и фактом-заключением.

Чтобы получить возможность гибкой обработки данного факта необходимо правильно определить элементы его структуры. Возникает два вопроса:

- Какое слово будет представлять предикат?
- Какое слово будет первым объектом?
- Какое слово будет вторым объектом?

От решения данных вопросов будет зависеть, насколько эффективно в дальнейшем будут использованы возможности инвариантности, например, при реализации на языке Пролог.

Смысл каждого слова может быть таким: «высокая» - является характеристикой параметра; «температура» - является параметром объекта; «тело» - является названием объекта.

В таблице 1 представлены варианты структур фактов, которые можно объявить на языке Пролог.

Таблица 1 – Варианты структур факта

№	**Предикат**	**Объект 1**	**Объект 2**
1	высокая	температура	тело
2		тело	температура
3	температура	высокая	тело
4		тело	высокая
5	тело	температура	высокая
6		высокая	температура

Чтобы принять правильное решение можно воспользоваться «частотным» принципом и определить наиболее правильный вариант предиката. Порядок объектов имеет вторичное значение и мало влияет качество будущей программы. Необходимо определить какое слово в качестве предиката дает возможность задать большее количество фактов в данной предметной области. Это имеет большое значение с точки зрения унификации программного кода интеллектуальной системы.

После определения структур фактов необходимо связать данные факты в логические цепочки в соответствии с семантикой предметной области задачи. В результате, получим структуры, которые будут представлять процедурный вид знаний. Представить цепочки логических фактов и правил можно следующим образом:

$R_1 : F_{d1} \leftarrow (F_1 \wedge F_4 \wedge F_7 \wedge F_8 \wedge F_9 \wedge F_{11}) \vee (F_2 \wedge F_3 \wedge F_{14}) \vee (F_1 \wedge F_4 \wedge F_{10})$

$R_2 : F_{d2} \leftarrow F_1 \wedge F_5 \wedge F_6 \wedge F_{12} \wedge \neg F_{13}$

$R_3 : F_{d3} \leftarrow F_1 \wedge F_{14} \wedge F_{15} \wedge F_{16}$

где $F_1 \ldots F_{16}$ - факты условных частей правил,

$F_{d1} \ \ldots \ F_{d3}$ - факты заключений (диагнозов),

$R_1 \ \ldots \ R_3$ - правила, реализующие логические цепочки.

Домашнее задание 2

На этапе исследования предметной области важно получить качественные исходные данные, на основе которых будут приниматься решения на этапе проектирования интеллектуальной системы.

Один из основных процессов – это декомпозиция предметной области. Способов декомпозиций существует несколько, но не все используются при разработке семантических систем. Особое внимание при выполнении второго домашнего задания уделено: логической, продукционной и инвариантной декомпозициям.

Также в рамках данного задания необходимо получить результаты, связанные с выбором и обоснованием методов представления знаний и способов их обработки.

Кроме вышесказанного, могут быть получены результаты обоснования выбора средств разработки системы искусственного интеллекта.