МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ВИНИТИ РАН)

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. № 8 Москва 2020

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 81’367

Н.И. Сидняев, Ю.И. Бутенко, Е.Е. Болотова

Теории формальных грамматик в методах распознавания неизвестных объектов

Рассматриваются вопросы формирования контекстных грамматик, описывающих как структурную информацию образа, так и взаимодействие образов в сложной сцене. Предлагается использование многоуровневой грамматики, включающей задачу синтаксического анализа последовательности изображений, а также задачу синтаксического анализа объектов различного назначения, когда природа источника не ясна. Показано, что формирование грамматики, описывающей как структурную информацию образа, так и взаимодействие образов, связано с необходимостью разработки алгоритма восстановления грамматики по заданному множеству динамических изображений, представляющих собой обучающую выборку. Излагаются некоторые основные положения, присущие структурным методам описания и распознавания сцены.

Ключевые слова: *распознавание, грамматики, синтаксический метод, оценка, решение, образ, обучение*

DOI: 10.36535/0548-0027-2020-08-1

ВВЕДЕНИЕ

Компьютерное зрение – одно из наиболее востребованных направлений в области интеллектуальных систем, включающих сложные алгоритмы распознавания и комплексные механизмы глубокого машинного обучения. Компьютерное зрение основано на воспроизведении механизмов сложной системы человеческого зрения, позволяющих получать значимую информацию из изображений, видео и других визуальных средств, а также принимать необходимые меры или давать рекомендации на основе полученной информации. До недавнего времени компьютерное зрение работало только в рамках ограниченных возможностей, однако благодаря достижениям в области искусственного интеллекта и инновациям глубокого обучения эта область в последние годы получила активное развитие и во многом превзошла человека в выполнении некоторых задач, связанных с обнаружением и распознаванием объектов [1–3]. Одним из движущих факторов развития компьютерного зрения является объем данных, который используется для обучения и улучшения подобных систем [4–7].

Проблемы, возникшие при решении задач распознавания изображений, привели к разработке различных алгоритмов, одним из которых является структурный, получивший название лингвистического или синтаксического метода [4, 5]. Его особенность заключается в том, что априорными описаниями классов являются структурные описания – формальные конструкции, при которых последовательно проводится принцип учета иерархичности структуры объекта и учета отношений, существующих между отдельными элементами этой иерархии в пределах одних и тех же уровней и между ними. Структурно-ориентированное распознавание применяется для самых разнообразных задач: распознавание символов, речи и отпечатков пальцев, анализ сцены, химических и биологических структур [6].

Во многих случаях апостериорная информация о распознаваемых объектах хранится в записях соответствующих сигналов – электрокардиограмм, энцефалограмм, инфракрасных и акустических сигналов и т.д. Синтаксический метод распознавания позволяет обнаруживать упорядоченность структур в сигналах, порождаемых распознаваемыми объектами, и использовать для описания объектов способы комбинаторной регулярности структур, заключающиеся в том, что, оперируя весьма ограниченным количеством атомарных (непроизводных) элементов и правил комбинирования, можно получать значительное разнообразие описаний объектов с помощью неограниченного (например, рекуррентного) применения к исходным элементам правил комбинирования и результатов соответствующих рекуррентных процедур. Использование комбинаторной регулярности в качестве принципа описания структуры объектов обеспечивает экономное расходование средств [7].

На современном этапе развития интеллектуальных систем стало возможным применить теорию формальных грамматик для распознавания изображений путем описания объекта в виде иерархической структуры [8]. Теория формальных языков и грамматик была разработана в середине 1950-х гг. и применялась для описания алгоритмов обработки естественных языков, однако в последующем стало очевидно, что подобный механизм может быть перенесен и на распознавание изображений, обладающих сложной многоуровневой структурой.

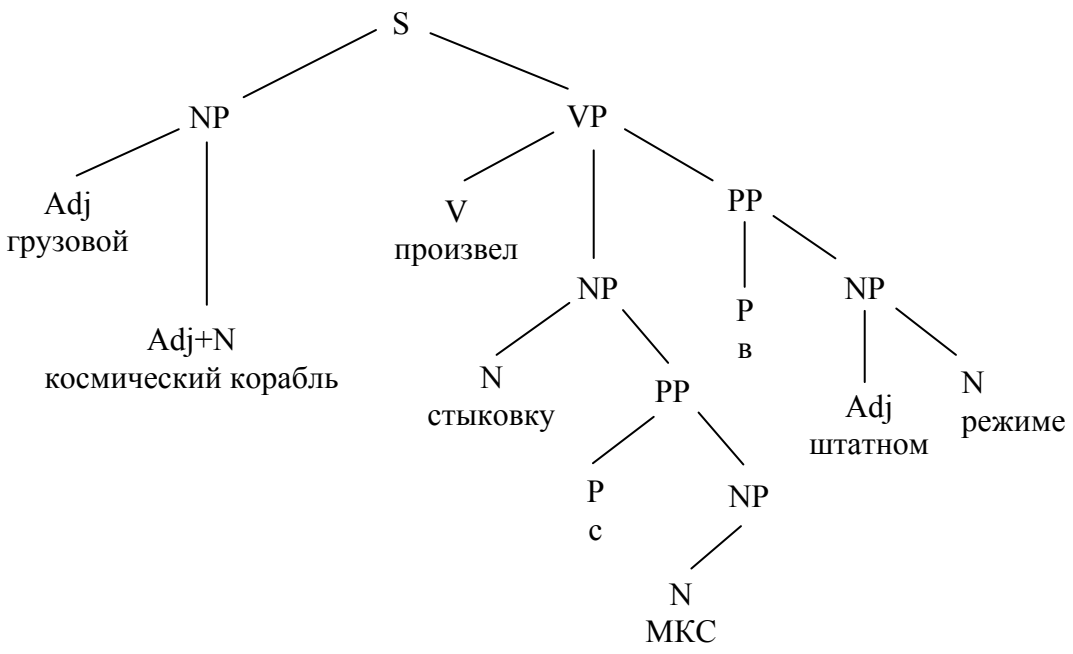

Рис. 1. Синтаксическая структура предложения

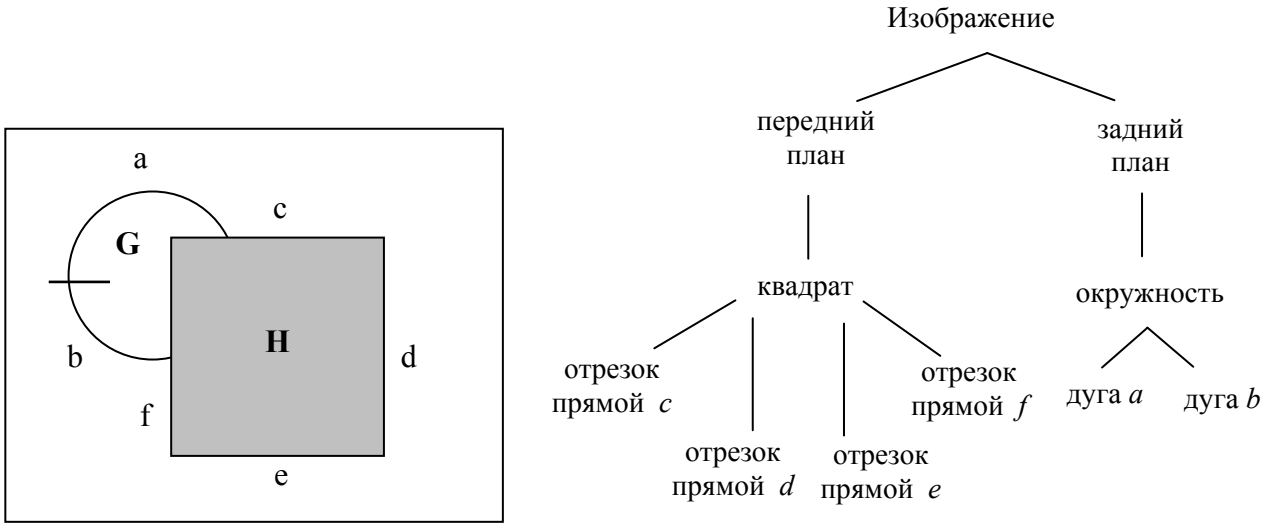

Рис. 2. Синтаксический анализ изображения

В процессе распознавания изображения выделяются такие элементы объекта, как подобразы, которые в свою очередь подразделяются на более простые подобразы, находящиеся в некоторых отношениях с другими выделенными подобразами. Это позволяет представить информацию об объекте в виде иерархической структуры, чаще всего, с помощью графа, в узлах которого находятся подобразы, а дуги обозначают отношения между выделенными подобразами распознаваемого объекта. Таким образом, структурный метод распознавания предельно схож с методом анализа синтаксической структуры предложений естественного языка [9]. Соответствие между синтагматикой и линейной упорядоченностью слов в предложении естественного языка можно представить в виде дерева составляющих. Глубинная структура пропозиционального компонента любого простого предложения представляет собой конструкцию, состоящую из глагольной (VP) и именной (NP) группы, которые находятся в некоторых семантико-синтаксических отношениях. Как показано на рис. 1. дерево составляющих позволяет представить синтаксическую структуру предложения в виде иерархии, составленной из связанных друг с другом непересекающихся единиц.

Схожим образом может быть описано распознаваемое системой изображение. На рис. 2 проиллюстрирована аналогия между синтаксическим разбором предложения на естественном языке и структурным разбором изображения, содержащего геометрические фигуры.

В рамках представленного подхода решение задачи распознавания сводится к использованию лингвистических конструкций, содержащих некоторый словарь признаков распознаваемого объекта и грамматику, по правилам которой описываются элементы этого объекта. После того как каждый подобраз внутри объекта идентифицирован, выполняется его распознавание путем синтаксического анализа предложения, описывающего этот объект в виде древовидной структуры, с целью определить, является ли такое предложение синтаксически правильным по отношению к указанной грамматике. Шаблоны предложений определяются как выстраиваемые последовательности из выделенных подобразов различными способами так же, как и фразы предложения выстраиваются путем конкатенации слов, а слова – путем конкатенации морфем. Синтаксический подход к распознаванию образов дает возможность описать множество подобразов с помощью небольших наборов простых грамматических правил формальной грамматики [6, 9].

ОСНОВНЫЕ ПОНЯТИЯ ФОРМАЛЬНОЙ ГРАММАТИКИ

Поскольку реализация структурных методов распознавания образов основана на аппарате формальной грамматики, рассмотрим некоторые ее основные понятия, необходимые для понимая сущности структурного подхода.

Теория формальных языков в качестве самостоятельной дисциплины, как правило, берет свое начало из работ знаменитого американского лингвиста Ноама Хомского, когда в 1950-х гг. он предпринял попытки дать точную и уникальную характеристику структуры естественных языков. Его целью было определение синтаксиса языков с применением простых и точных математических правил. Позже было обнаружено, что синтаксис языков программирования может быть описан с использованием одной из грамматических моделей Хомского, называемой свободно-контекстной грамматикой [10]. Вскоре после появления современных ЭВМ стало очевидно, что все формы информации – будь то числа, имена, изображения или звуковые волны – могут быть представлены в виде записанных в строку по определенным правилам элементов. Теория формальных грамматик заняла центральное место в прикладных исследованиях в области искусственного интеллекта. Более того, с помощью аппарата формальных грамматик возможно не только описать структуру объек-

та, но и создать синтаксический анализатор для определения того, является ли предложение, описывающее объект, синтаксически верным, т.е. принадлежит ли оно конкретному языку или является грамматически неправильным. Для описания таких распознающих моделей в теории формальных языков используются отдельные формализмы, известные как теория автоматов [11].

В теории формальных языков грамматика G – это набор правил для порождения строк на формальном языке. Правила описывают, как из алфавита языка формировать строки, которые действительны в соответствии с синтаксисом языка. Алфавитом является конечный непустой набор символов. Предполагается, что символы неделимы. Слово (строка) в алфавите Σ представляет собой конечную последовательность элементов. Порождающую грамматику G определяют как производственную систему, состоящую из упорядоченной четверки $G = < T, N, I, P >$, где T – набор терминальных символов алфавита Σ, т. е. исходные элементы словаря, N – конечный набор нетерминальных (вспомогательных) символов, I – начальный символ грамматики и P – конечный набор правил.

Построение грамматики начинается с начального символа I. Применив правила сначала к I, а затем рекурсивно к результату предыдущего преобразования, можно сгенерировать корректное "предложение" формального языка, определенного грамматикой. Грамматика строится с помощью специальных правил подстановки выражений вида «$x \rightarrow y$», что означает «заменить x на y» или «подставить x вместо y», где x и y – цепочки, содержащие любые терминальные или нетерминальные символы. Алгоритм преобразования останавливается, когда выражение больше не содержит нетерминальные символы алфавита. Таким образом, в соответствии с правилами грамматики объект представляется предложением в этом языке (рис. 3). Для дальнейшего описания процесса порождения необходимо введение таких понятий, как непосредственная выводимость, выводимость и язык, порождаемый грамматикой [10, 11].

Если имеются цепочки x и y, которые можно представить в виде $x = z_1 a z_2$ и $y = z_1 b z_2$, где $a \rightarrow b$ – одно из правил грамматики G, то цепочка y непосредственно выводима из цепочки x в грамматике G. Другими словами, цепочка x может быть переработана в цепочку y за один шаг путем применения одной подстановки: x получается из y подстановкой b на место некоторого вхождения цепочки a. Это обозначается как $x \mid = y$ или $x \mid = y$.

Если имеется последовательность цепочек $x_0, x_1, ..., x_n$, в которой каждая следующая цепочка непосредственно выводима из предыдущей, то цепочка x_n выводима из цепочки x_0. Последовательность цепочек $x_0, x_1, ..., x_n$ называется выводом x_n из x_0 в грамматике G. Это означает, что цепочка x_0 перерабатывается в цепочку x_n необязательно за один шаг, а последовательным применением нескольких подстановок. Очевидно, что непосредственная выводимость есть частный случай выводимости. Вывод, начинающийся начальным символом и заканчивающийся цепочкой, состоящей только из терминальных символов, называется полным выводом. Естественно, не всякий вывод, начинающийся начальным символом, является полным; возможны выводы, начинающиеся начальным символом, которые невозможно продолжить до полного вывода, – тупиковые выводы. Невозможность продолжать вывод, хотя он и не является полным, т. е. не кончается цепочкой терминальных символов, определяется тем, что в соответствующей грамматике отсутствует правило, левая часть которого содержалась бы в последней цепочке данного вывода. Считаем, что от «хорошей» грамматики вовсе не требуется, чтобы любой вывод в ней заканчивался правильной терминальной цепочкой: достаточно, чтобы любой полный вывод давал правильную цепочку. В результате построения полных выводов из начального символа в грамматике получается цепочка, состоящая только из терминальных символов. Такую цепочку называют языком, порождаемым грамматикой [9–11].

Важно учитывать, что правильно построенная грамматика не должна содержать бесполезных и недостижимых символов.

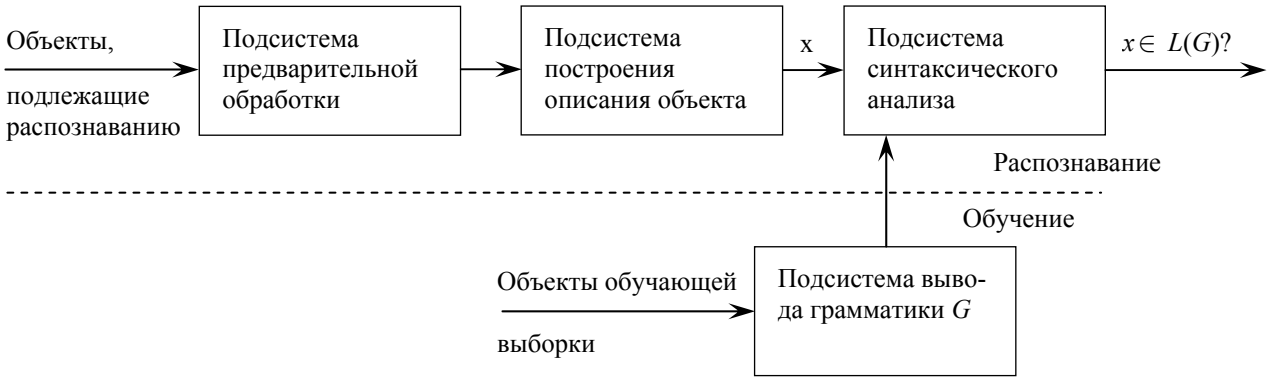

Рис. 3. Грамматика представления объекта

Определение. Символ X называется полезным в грамматике G = {*T*, *N*, *I*, *P*}, если существует вывод вида $A \Rightarrow *\alpha X\beta \Rightarrow * w$, где $w \in N*$.

Отметим, что *X* может быть как переменной, так и терминалом, а выводимая цепочка $\alpha X\beta$ – первой или последней в выводе.

Определение. Если символ *X* не является полезным, то он называется бесполезным. Очевидно, что исключение бесполезных символов не изменяет порождаемого грамматикой языка, поэтому все такие символы можно удалить.

Определение. Символ *X* называется порождающим в грамматике G = {*T*, *N*, *I*, *P*}, если существует вывод вида $X \Rightarrow * w$, где $w \in N*$.

Заметим, что каждый терминал является порождающим символом, поскольку выводится сам из себя за 0 шагов.

Определение. Символ X называется достижимым в грамматике *G* = {*T*, *N*, *I*, *P*}, если существует вывод вида $A \Rightarrow *\alpha X\beta$ для некоторых α, β.

Полезный символ, как следует из определения, является и порождающим, и достижимым. Если удалить из грамматики сначала непорождающие, а затем недостижимые символы, то останутся только полезные.

Пример. Рассмотрим грамматику *G* = <{*A*, *B*, *C*, *a*, *b*}, {*a*, *b*}, *I*, *P*> с правилами $A \rightarrow BC$; $A \rightarrow a$; $B \rightarrow b$.

Все символы, кроме *C*, являются порождающими, поскольку терминалы *a* и *b* порождают самих себя, *A* порождает *a* и *B* порождает *b*. Удаление *B* приводит к удалению правила $A \rightarrow BC$, т.е. к грамматике *G* = <{*A*, *B, a*, *b*}, {*a*, *b*}, *P*, *A* > с правилами $A \rightarrow a$; $B \rightarrow b$.

По иерархии Н. Хомского [12], формальные грамматики делятся на 4 типа, где каждый последующий тип является подмножеством предыдущего:

грамматика нулевого типа (рекурсивно перечислимая, неограниченная). Такая грамматика включает в себя формальные грамматики всех типов и допускает наличие любых правил вывода. Правила можно записать в виде: $a \rightarrow b$, где на строки *a* и *b* не устанавливаются какие-либо ограничения. Разбор предложения такого языка можно выполнять с помощью машины Тьюринга;

грамматика 1-го типа (контекстно-зависимая грамматика). Для контекстно-зависимых грамматик левая сторона правила может указывать контекст, в котором применяется определенное правило, поэтому один и тот же нетерминальный элемент может быть преобразован по-разному в зависимости от контекста, в котором он появляется. Контекстно-зависимые грамматики имеют правила вида: $xYz \rightarrow xyz$, где подстановка $Y \rightarrow y$ допустима только в контексте *x*, *z*. Разбор предложения такого языка можно выполнять с помощью линейно ограниченной машины;

грамматика 2-го типа (контекстно-свободная грамматика). Подобная грамматика генерируют контекстно-свободные языки, все правила которых принимают форму $A \rightarrow b$, где $A \in T$, $\beta \in (T \cup N)$ *. Подстановки правил внутри грамматики не зависят от контекста. Разбор предложения данного языка можно выполнять с помощью автомата с магазинной (стековой) памятью;

грамматика 3-го типа (регулярная). Правила записываются в виде $A \rightarrow a$, $A \rightarrow a$, *B*, где *A*, $B \in N$, $a \in T$*. Разбор предложения языка третьего типа можно выполнять с использованием конечного автомата.

РЕАЛИЗАЦИЯ ПРОЦЕССА РАСПОЗНАВАНИЯ ОБЪЕКТОВ НА ОСНОВЕ СТРУКТУРНЫХ МЕТОДОВ

Для распознавания неизвестного объекта на основе структурных методов необходимо прежде всего найти его непроизводные элементы и отношения между ними, а затем с помощью синтаксического анализа (грамматического разбора) установить, согласуется ли описание образа с грамматикой, которая, по предположению, могла его породить [13, 14].

Для формирования соответствующей грамматики можно воспользоваться либо априорными сведениями о распознаваемых объектах, либо результатами изучения конечного выборочного множества «типичных» в некотором смысле объектов. В первом случае говорят о задании грамматики на основе эвристических соображений, во втором – о выводе грамматики.

Допустим, что имеем дело с двумя классами объектов (Ω_1 и Ω_2), и объекты, входящие в эти классы, можно описать с помощью признаков, принадлежащих некоторому конечному множеству. Эти признаки можно считать основными символами (терминальными) и обозначать через *T*. Каждый объект может рассматриваться как цепочка или предложение, поскольку он составлен из элементов основного словаря.

Пусть также существует грамматика *G* такая, что порождаемый ею язык состоит из предложений (объектов), принадлежащих исключительно одному из классов, например классу Ω_1. Эта грамматика может быть использована для классификации объектов, так как предъявленный неизвестный объект можно отнести к классу Ω_1, если он является предложением языка *L* (*G*). В противном случае объект приписывается классу Ω_2. В данном случае объект попадает в класс Ω_2 исключительно потому, что не входит в класс Ω_1: если оказывается, что объект не является грамматически правильным предложением в смысле грамматики *G*, то предполагается, что он должен принадлежать классу Ω_2. На самом же деле объект может не относиться и к классу Ω_2. Он может также представлять собой зашумленную или искаженную цепочку, которую в ряде случаев предпочтительно просто изъять из рассмотрения. Для этого необходимо располагать грамматиками G_1 и G_2, порождающими языки *L* (G_1) и *L* (G_2) соответственно. Объект зачисляется в тот класс, в языке которого он оказывается грамматически правильным предложением. Если последнее не выполняется, то очевидно, что данный объект не принадлежит ни одному из двух заданных классов и, следовательно, требуется еще одна грамматика G_3 и т. д. В случае *m* классов рассматриваются *m* грамматик и связанных с ними языков *L* (G_i), *i* = l, 2, …, *m*. Распознаваемый объект относится к классу Ω_2 в том и только в том случае, если он является грамматически правильным предложением языка L (G_i). Если объект оказывается грамматически правильным предложением более чем одного языка, то решение относительно его принадлежности принимается точно таким же образом, как это делается при использовании иных методов распознавания, когда оказывается, что результаты реализации процедуры распознавания не позволяют определенно отнести объект к одному из заданных классов.

Для практического использования структурного метода требуется разрешить следующие основные проблемы:

1) построение адекватного описания распознаваемых объектов;

2) выбор грамматики;

3) реализация процесса распознавания посредством процедур синтаксического анализа;

4) использование процедур обучения для вывода грамматик;

5) применение в рамках структурного подхода процедур из других методов распознавания (например, статистических для учета помех и искажений случайного характера, кластер-анализа и т. д.).

Рассмотрим основные приемы, используемые при практическом применении структурных методов распознавания.

Методы математической морфологии

В процессе распознавания изображений на основе структурных методов важным является этап предварительной обработки анализируемого изображения. На этом этапе предъявленный для распознавания объект подвергается кодированию и фильтрации, восстановлению и улучшению качества изображения. Подобные методы фильтрации рассматриваются в рамках математической морфологии, обеспечивающей эффективный подход к анализу цифровых изображений [15, 16]. В рамках морфологической фильтрации используются не аналитические свойства сигналов, а их геометрические характеристики. Объект кодируется или аппроксимируется таким образом, чтобы дальше с ним было удобно работать. Например, черно-белое изображение можно закодировать с помощью сетки (или матрицы) нулей и единиц. Чтобы повысить эффективность обработки на последующих этапах, часто прибегают к какой-нибудь разновидности «сжатия данных». Затем с помощью методов фильтрации и восстановления устраняются искажения с целью улучшения качества изображения. Предполагается, что по окончании предварительной обработки воспроизводятся образы достаточно хорошего качества.

Выбор непроизводных элементов

Первым шагом в построении структурного описания объекта распознавания является определение набора непроизводных элементов, с помощью которых можно было бы описывать рассматриваемый объект. Этот выбор зависит от характера объекта, вида исходных данных, области приложений и способа практической реализации системы распознавания. В настоящее время не существует некоего универсального решения задачи выбора непроизводных элементов. Обычно при выборе непроизводных элементов пользуются следующими двумя основными критериями:

1) непроизводные элементы должны служить основными элементами образа, чтобы обеспечивалось получение компактного и адекватного описания исходных данных с помощью вполне определенных структурных отношений (например, отношения соединения);

2) непроизводные элементы должны легко поддаваться выделению или распознаванию с помощью известных методов неструктурного характера, так как предполагается, что они представляют собой простые и компактные образы, «внутренняя» структурная информация которых не принимается во внимание.

Так, для речи естественно считать фонемы «хорошим» набором непроизводных элементов, на котором задается отношение соединения. Подобным же образом штрихи (черточки) выбираются в качестве непроизводных элементов для представления рукописных символов.

Однако для изображения вообще нет какого-либо универсального элемента, подобного фонеме или штриху для конкретных объектов – речи или рукописных символов. Иногда для адекватного описания необходимо, чтобы непроизводные элементы содержали информацию, существенную при распознавании объектов или явлений. Если, в частности, при решении задачи распознавания объекта существен его размер (или форма, или местоположение), то непроизводные элементы должны нести информацию о размере (или форме, или местоположении), чтобы объекты, принадлежащие различным классам, оказывались различными независимо от того, какой метод используется для анализа соответствующих им описаний. Это условие, однако, часто приводит к необходимости прибегать к семантической информации при описании непроизводных элементов.

Пример. Допустим, необходимо уметь отличать прямоугольники различных размеров от других фигур.

Выбирается следующий набор непроизводных элементов: a' – 0° – отрезок горизонтальной линии; b' – 90° – отрезок вертикальной линии; c' – 180° – отрезок горизонтальной линии; d' – 270° – отрезок вертикальной линии.

Множество всех возможных прямоугольников (различных размеров) задается с помощью одного предложения – цепочки $a'b'c'd'$ (рис. 4). Если же требуется различение прямоугольников различных размеров, то такое описание оказывается неадекватным.

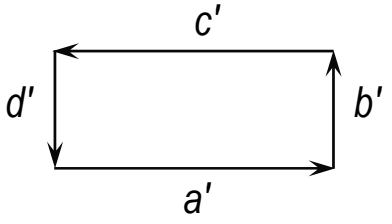

Рис. 4. Пример предложения для описания прямоугольника

В этом случае в качестве непроизводных элементов следует использовать отрезки единичной длины. Множество прямоугольников различных размеров можно описывать с помощью языка:

$$L = \{a^n b^m c^n d^m,\ n, m = 1, 2,...\}.$$

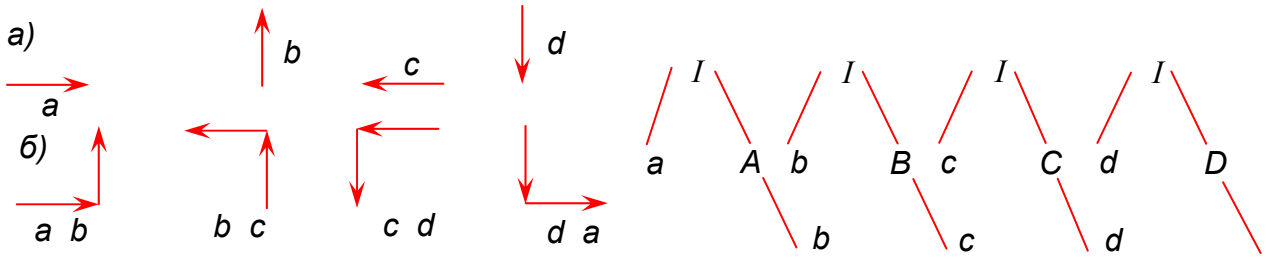

Рис. 5. Пример вывода с использованием дерева

В некоторых случаях второй критерий выбора непроизводных элементов может вступать в противоречие с первым критерием из-за того, что непроизводные элементы, выбранные в соответствии с этим первым критерием, могут плохо поддаваться распознаванию с помощью известных в настоящее время методов. Однако второй критерий допускает выбор достаточно сложных непроизводных элементов, если только они могут быть распознаны. Усложнение непроизводных элементов позволяет упрощать структурные описания, т. е. позволяет пользоваться менее сложными грамматиками. Отыскание компромиссного решения имеет существенное значение при построении реальных систем распознавания, основанных на использовании структурных методов.

Использование грамматик и языков для структурного описания объектов

Грамматику можно использовать для порождения предложений, представляющих собой некий объект, и для грамматического разбора предложений, чтобы установить, является ли структура этих предложений допустимой с точки зрения соответствующей грамматики.

Пример. Рассмотрим грамматику

$G = \{T, N, I, P\}$, где $T = \{a, b, c, d\}$; $N = \{I, A, B, C, D\}$;

$$P: \quad I \to aA \quad A \to b \quad I \to cC \quad C \to d$$

$$I \to bB \quad B \to c \quad I \to dD \quad D \to a\ .$$

В данном случае можно вывести четыре предложения ($\Rightarrow$ – символ вывода):

$$I \Rightarrow aA \Rightarrow ab; \quad I \Rightarrow cC \Rightarrow cd;$$

$$I \Rightarrow bA \Rightarrow bc; \quad I \Rightarrow dD \Rightarrow Da.$$

Если непроизводными элементами считать ориентированные отрезки, изображенные на рис. 5*а*, то с помощью этой грамматики можно получить четыре описания отрезков, образующих прямой угол (рис. 5*б*). Здесь использован прием, позволяющий описывать двумерные объекты с помощью цепочечных грамматик. Этот прием заключается в соединении структур только в особых точках. Один из способов реализации этого требования заключается в том, что в каждой структуре выделяются только две точки. В нашем случае выделенные точки интерпретируются как «головной» и «хвостовой» концы стрелки.

Деревья вывода

В информатике под деревом понимают широко используемый абстрактный тип данных, аналогичный иерархической структуре дерева, с корневым значением и поддеревьями – дочерними элементами с родительским узлом, представленными в виде набора связанных узлов [17, 18]. Каждый узел дерева имеет определенную степень, характеризующую число поддеревьев узла. Узел с нулевой степенью называется листом. Листья представляют собой узлы, из которых не выходит ни одной ветви.

Процесс вывода предложения, описывающего прямоугольник в виде дерева разбора, изображен на рис. 5. Начальный символ I выполняет роль корня дерева, концевые узлы, прочитываемые слева направо, образуют выведенное предложение.

В сущности, любая иерархически упорядоченная схема позволяет представить соответствующий объект в виде дерева. Так, на рис. 6 упорядочение состоит в группировке областей распознаваемого объекта, причем область b находится в области a, в свою очередь находящейся в области r. На рис. 6*б* изображена древовидная структура, естественно вытекающая из этой схемы упорядочения. Каждое дерево вывода грамматики задает одно предложение. Большинство грамматик позволяет получать значительное, а при введении вероятностных правил вывода – и бесконечное, число предложений.

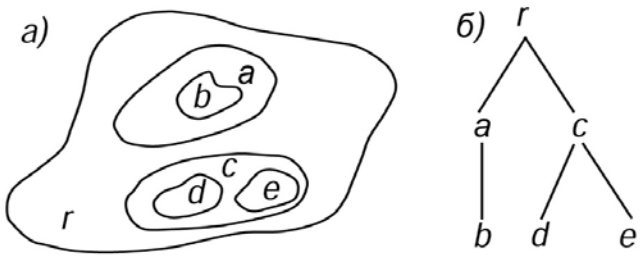

Рис. 6. Упорядоченная схема объекта:
а) – группировка областей распознаваемого объекта;
б) – дерево вывода

В простейшем случае в таких деревьях от каждого узла отходят две ветви: одна соответствует наличию у объекта определенного признака, другая – его отсутствию. Подобные деревья не несут какой-либо информации о процессе вывода решений в указанном выше смысле – они дают иерархическое представление о том, какие признаки (в случае структурного распознавания – какие непроизводные элементы и отношения) характеризуют изучаемый объект. Однако можно сформировать грамматику, описывающую процесс вывода в соответствии со структурой дерева решений (обычно в рамках этого метода строится и используется только дерево).

Процедуры распознавания, связанные с использованием деревьев решений, относят к группе структурных методов, хотя в строгом смысле деревья решений являются инструментом иерархического метода разделения [18], который используется в случаях, когда дерево решений «содержит» много признаков. В каждом узле дерева изучается один признак и в зависимости от результатов этого изучения определяется очередная ветвь дерева. Результат классификации выявляется на нижнем ярусе дерева. Подобная схема распознавания очень удобна для учета априорных сведений об объектах, однако для нее отсутствуют оптимальные процедуры обучения.

РЕАЛИЗАЦИЯ ПРОЦЕССА РАСПОЗНАВАНИЯ ОБЪЕКТОВ С ПОМОЩЬЮ ГРАММАТИЧЕСКОГО РАЗБОРА

Значение формальных грамматик для распознавания объектов заключается, в частности, в том, что они позволяют видеть синтаксическую правильность или неправильность применительно к определенной грамматике представления изучаемого образа с помощью заданных непроизводных элементов и отношений. Подобные элементы и отношения позволяет получить процедура [19], использующая два основных вида грамматического разбора: разбор сверху–вниз и разбор снизу–вверх. Разбор сверху–вниз состоит в последовательных попытках путем использования правил соответствующей грамматики получить заданное терминальное предложение исходя из начального символа. При разборе снизу–вверх необходимо восстанавливать дерево вывода, начиная с элементов основного словаря и применяя инвертированные правила подстановки. Эта процедура начинается с конкретного предложения и заканчивается получением начального символа. Так, непроизводный элемент a, входящий в предложение ab, может быть получен как в результате перехода $D \rightarrow a$, так и в результате перехода $I \rightarrow aA$; элемент D возникает в результате подстановки $I \rightarrow dD$, так что этот вывод не приводит к получению предложения ab. Так как подстановки $I \rightarrow aA$ и $A \rightarrow b$ действительно порождают предложение ab, то это предложение считается грамматически правильным; предложения же ac и ad отклоняются. Описанные схемы грамматического разбора, в принципе, неэффективны, так как предусматривают проведение полного перебора при применении правил подстановки. Часто можно избежать воспроизведения всей последовательности применения правил подстановки, так как можно проверять допустимость промежуточных результатов, выясняя тем самым, обеспечивает ли данная последовательность правил успешный грамматический разбор.

Повышение эффективности грамматического разбора можно обеспечить с помощью синтаксических правил, устанавливающих допустимые или запрещающих какие-то специфические отношения между объектами.

Пример. Рассмотрим грамматику $G = \{T, N, I, P\}$, где $T = \{a_1, a_2\}$, $N = \{I, O_1, O_2\}$, P: $I \rightarrow$ выше (a_1, O_2),$O_2 \rightarrow$ выше (O_1, a_1),$O_1 \rightarrow$ слева (a_2, a_1),

способную порождать квадраты (непроизвольные элементы, представляющие собой горизонтальный и вертикальный отрезки прямой фиксированной длины); отношения выше (x, y) и слева (x, y) читаются соответственно «x расположен над y» и «x расположен слева от y», при этом часть x обязательно находится непосредственно над отрезком y и часть отрезка x во втором случае находится непосредственно слева от отрезка y.

Последовательное применение правил грамматики приводит к получению структур, построение которых начинается с горизонтального отрезка, затем вводится еще один горизонтальный отрезок, расположенный над первым, после чего между ними помещаются два вертикальных отрезка (см. рис. 2). Разнообразие конструкций обеспечивается наложением ограничений на позиционные отношения выше и слева.

На этом же примере покажем порядок выполнения синтаксически ориентированного грамматического разбора. Пусть требуется установить, принадлежит ли некоторая конструкция к классу объектов, порождаемых данной грамматикой. При разборе снизу–вверх в первую очередь делается попытка обнаружить объект O_1 с помощью вхождения в объект непроизводного элемента a_2 слева от непроизводного элемента a_2. В положительном случае разбор продолжается, в отрицательном – исследуемый объект отклоняется. Поскольку процедура разбора снизу–вверх начинается с терминального предложения, на первом шаге рассматриваются только правила подстановки, приводящие исключительно к непроизводным элементам. На следующем шаге разбора необходимо получить объект O_2, состоящий из объекта O_1, расположенного над непроизводным элементом a_1. При положительном результате предпринимается попытка вывести начальный символ I, отыскивая непроизводный элемент a_1, расположенный над объектом O_2. При положительном результате образ считается допустимым, при отрицательном – отклоняется.

Вывод грамматик

Рассмотрим на конкретном примере способ получения грамматики, пригодной для описания объектов в рамках структурного подхода к распознаванию, с помощью процедуры обучения. В сущности, искомая грамматика формируется на основании изучения конечной выборки предложений некоторого языка, причем иногда при этом учитывается и конечная выборка предложений, не относящихся к данному языку. Элементы первой выборки должны обладать

структурной полнотой в том смысле, что каждое правило искомой грамматики используется при порождении по крайней мере одного предложения выборки.

На рис. 6 представлена схема вывода цепочечных грамматик. Задача заключается в том, что множество выборочных цепочек $\{x_i\}$ подвергается обработке с помощью адаптивного обучающего алгоритма (*алгоритмом вывода*). На выходе блока в результате процесса обучения воспроизводится грамматика G, согласованная с заданными цепочками. Процедура, позволяющая решить указанную задачу в общем случае, до сих пор неизвестна; существует, однако, множество частных алгоритмов. Типичным подходом можно считать следующий: сначала строятся правила подстановки, порождающие именно те цепочки, которые были заданы; затем посредством сращивания элементов основного словаря – более простая рекурсивная грамматика, обеспечивающая порождение бесконечного числа цепочек. Процедура состоит из трех этапов:

1) формирование правил подстановки;

2) преобразование правил подстановки в грамматику;

3) упрощение грамматики.

На примере двумерной грамматики продемонстрируем её вывод, поскольку, как указывалось выше, именно этот тип грамматик представляет в настоящее время наибольший интерес с прикладной точки зрения. Предполагается, что заданы конечный набор отношений, конечный список элементов основного словаря, т. е. непроизводных элементов, и некоторое множество описаний объектов $\{S_1, ..., S_n\}$. Каждое описание содержит список входящих в него непроизводных элементов, каждый из которых может быть снабжен какой-либо дополнительной информацией. Задача состоит в определении «хорошего» набора правил грамматики, записанных через заданные отношения и согласующихся с заданным набором объектов. Таким образом, следует сформировать такую грамматику, чтобы при введении в грамматический анализатор любого из заданных объектов и разборе его в соответствии с полученной грамматикой на выходе анализатора воспроизводилось одно или несколько структурных описаний объекта.

Первый шаг процесса вывода выполняется отдельно и независимо для каждого «входного» объекта. В результате этого должна выявиться структура, задаваемая выбранной системой отношений – в данном случае предикатов, характеризующих как свойства непроизводных элементов, так и отношения между ними. На этом же шаге для каждого элемента объектов проверяется справедливость их предикатов. При положительном исходе такой проверки формируется новый элемент и информация о соответствующих импликантах и истинности предиката фиксируется в виде некоторого правила грамматики. Затем выполняется следующий цикл, в котором предикаты проверяются на элементах, построенных из новых элементов и элементов входных объектов; процедура повторяется до тех пор, пока построение новых элементов оказывается невозможным. В каждом очередном цикле, естественно, проверяются лишь те описания, которые содержат хотя бы один элемент, сформированный в предыдущем цикле [20].

По окончании первого шага проверяется, какие из элементов включают в качестве своих импликант все непроизводные элементы соответствующего входного объекта. Все не прошедшие проверку описания и описания, входящие в качестве импликант в описания, прошедшие проверку, из дальнейшего рассмотрения исключаются. Соответственно исключаются и все правила грамматики, не использовавшиеся при построении оставленных описаний. Полученные таким образом описания являются некоей разновидностью деревьев структурных описаний, которые сначала подвергаются «чистке» (в частности, отождествляются деревья, оказывающиеся эквивалентными в результате замены конкретных непроизводных элементов их типами).

Для каждого входного объекта таким образом строится одна или несколько грамматик, соответствующих различным структурным описаниям, получаемым для входного объекта. Высшему уровню описания – описанию объекта в целом соответствуют лишь те элементы вспомогательного словаря, которые встречаются более чем в одной из построенных грамматик.

Второй шаг процесса вывода состоит в следующем. Строится объединение грамматик, полученных на первом шаге для каждого из объектов. К каждой из этих грамматик применяется специальная процедура преобразования и согласно некоторому критерию из полученных модифицированных грамматик выбирается наилучшая. Предпочтение, естественно, отдается коротким грамматикам, поскольку они действительно являются единой грамматикой для заданной обучающей выборки, а не просто набором частных случаев. Однако целесообразно, чтобы правила подстановки были как можно конкретнее в том, что касается возможных типов непроизводных элементов и отношений. Очевидно, что окончательный критерий выбора грамматики должен базироваться на какой-то комбинации двух этих эвристических приемов.

Преобразование грамматики

Для эффективного распознавания неизвестных объектов с помощью формальной грамматики, её необходимо сократить, т. е. избавиться от избыточных правил. Для этого применяется процедура преобразования грамматики, которая включает несколько основных этапов:

1) устраняются все избыточные вхождения правил, если какое-то правило входит в преобразуемую грамматику G несколько раз;

2) отыскивается такая пара элементов вспомогательного словаря, что отождествление этих элементов (замена одного другим во всей грамматике) приведет к появлению избыточных вхождений правил. При наличии различных вариантов выбора таких пар следует выбирать пару, приводящую к наибольшему числу избыточных вхождений. После отождествления правил следует вернуться к первому этапу;

3) отыскивается такая пара N, n, включающая элемент вспомогательного словаря N и непроизводный элемент n, что включение в грамматику G правила $N \rightarrow n$ (если оно в нее не входит) и выборочная замена N на n в грамматике Г приводят после устранения множественности вхождений правил к сокращению

их числа. При наличии различных вариантов выбора таких пар следует выбирать пару, обеспечивающую наибольшее сокращение числа правил в грамматике. Затем следует вернуться к первому этапу;

4) вводится порог по числу отождествляемых на этапах 1–3 правил: замена производится, если число вхождений или отождествляемых правил не меньше значения порога. Это делается для того, чтобы не снижать существенно разделяющую силу грамматики;

5) вводится порог по числу правил, оказывающихся идентичными во всем (исключение составляет тип непроизводного элемента): если число таких правил больше значения порога или равно ему, то соответствующие правила заменяются одним, содержащим непроизводный элемент произвольного типа. Затем следует вернуться к первому этапу.

Если дальнейшие упрощения невозможны, то процедура прекращается. В результате выводится одна грамматика. После того как такая процедура выполняется для каждой из сформированных на первом этапе грамматик, необходимо выбрать «наилучшую» грамматику. В качестве критерия оптимальности грамматики может быть использовано, например, отношение разделяющей силы грамматики к квадрату числа ее правил, а разделяющая сила грамматики определяется как сумма разделяющих сил каждого из ее правил, которые определяются как сумма входящих в него непроизводных элементов и термов, представляющих собой отношение или свойство.

В качестве примера рассмотрим ситуацию, когда предъявляется только один входной объект, представленный на рис. 7. Здесь непроизводные элементы принадлежат к следующим типам: окружность, точка, квадрат, отрезок прямой. Предположим, заданы следующие предикаты: слева (x, y) (выполняется, если элемент x находится слева от элемента y); выше (x y) (выполняется, если элемент x находится выше элемента y); внутри (x, y) (выполняется, если элемент x находится внутри элемента y). В данном случае входной объект представляется перечнем, включающим пять непроизводных элементов: точка (два), окружность, квадрат и отрезок прямой. Вначале выделяются все случаи, когда элементы входного списка удовлетворяют одному из заданных предикатов. Каждый выполненный предикат образует новый элемент. Далее проверяется, какие новые элементы можно построить из исходных вновь образованных, затем эта процедура повторяется с очередными построенными элементами и так до тех пор, пока возможности построения новых элементов не окажутся исчерпанными. После этого удаляются все элементы, не являющиеся компонентами некоторого элемента, содержащего все непроизводные элементы входного объекта. Практически это выглядит следующим образом. Непроизводные элементы снабжены номерами 1–5. Пусть, например, обнаружен элемент выше (4, 5) – соответствующий предикат «выше (4, 5)» выполняется. Выполненный предикат образует новый элемент, компонентами которого служат непроизводные элементы 4 и 5.

Каждое из полученных таким образом структурных описаний непосредственно переводится в грамматику (при этом символы $G1$, $G2$, $G3$, используются по мере необходимости для обозначения элементов вспомогательного словаря, а элементы основного словаря заменяются непроизводным элементом определенного типа).

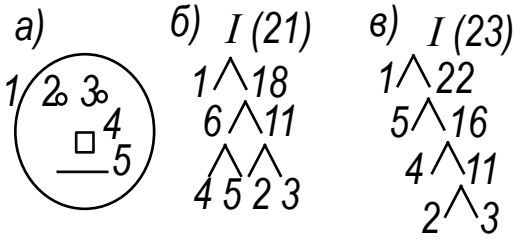

Рис. 7. Представление предикатов с использованием грамматик.

Итак, две построенные грамматики имеют следующий вид:

грамматика 1) I→(x, y): $G1$(x), окружность (y): внутри (x, y): $G1$→(x, y): $G2$(x), отрезок прямой (y): выше (x, y): $G2$→(x, y): $G3$ (x), квадрат (y): выше (x, y); $G3$→(x, y): точка (x), точка (y): слева (x, y);

грамматика 2) I→(x, y): $G4$(x), окружность (x): внутри (x, y); $G4$→(x, y); $G5$ (x), $G6$ (y): выше (x, y); $G5$→(x, y): точка (x), точка (y): слева (x, y); $G6$→(x, y): квадрат (x), отрезок прямой (y); выше (x,y).

Структурный подход к распознаванию объектов не располагает еще стройной и строгой математической теорией. Основные области приложения структурного распознавания в настоящее время – обработка и анализ сигналов различной природы, двумерных и трехмерных изображений. Практика построения систем распознавания свидетельствует о том, что структурные методы целесообразно использовать совместно с другими методами распознавания объектов и явлений. Степень применимости структурных методов будет определяться развитием их математической теории и наличием ориентированных на него нейронных сетей.

Можно отметить новые нормальные формы контекстно-свободных грамматик, в которых все правила имеют одну из двух форм:

$$U \rightarrow u;$$

$$U \rightarrow TN,$$

где U, N, T – нетерминальные символы, u – терминальный символ. Эта форма называется нормальной формой Хомского.

ЗАКЛЮЧЕНИЕ

Для распознавания многоуровневых и плохо структурируемых объектов необходимо выбрать наилучшую грамматику. Распознавание неизвестного объекта на основе структурных методов требует прежде всего нахождения его непроизводных элементов и отношений между ними, а затем с помощью синтаксического анализа (грамматического разбора) установления, согласуется ли описание образа с

грамматикой, которая, по предположению, могла его породить. Для формирования соответствующей грамматики можно воспользоваться априорными сведениями о распознаваемых объектах или результатами изучения конечного выборочного множества понятных в некотором смысле объектов. Каждый объект рассматривается как цепочка или предложение, поскольку он составлен из элементов основного словаря. Представленные грамматики могут быть использованы для классификации объектов. Предъявленный неизвестный объект можно отнести к определенному классу, если он является предложением языка. В противном случае объект приписывается другому классу. Как правило, объект зачисляется в тот класс, в языке которого он оказывается грамматически правильным предложением. Если это условие не выполняется, то очевидно, что данный объект не принадлежит ни одному из заданных классов и, следовательно, требуется еще одна грамматика и т. д. Распознаваемый объект относится к интересующему классу в том и только в том случае, если он является грамматически правильным предложением языка. Если объект оказывается грамматически правильным предложением более чем одного языка, решение относительно его принадлежности принимается точно таким же образом, как это делается при иных методах распознавания, когда оказывается, что результаты реализации процедуры распознавания не позволяют определенно отнести объект к одному из заданных классов.

В качестве критерия оптимальности грамматики может быть использовано отношение разделяющей силы. Слова могут объединяться в более сложные конструкции – предложения. Здесь язык рассматривается как множество предложений. Предложения строятся из слов и более простых предложений по правилам синтаксиса. Синтаксис языка представляет собой описание правильных предложений. Алфавит, лексика и синтаксис полностью определяют набор допустимых конструкций языка и внутренние взаимоотношения между конструкциями. Набор правил синтаксиса образует грамматику языка. Правила синтаксиса могут описывать либо процедуру получения правильных предложений, либо процедуру распознавания «правильности» предложений (т. е. их принадлежности данному языку). В первом случае грамматика будет порождающей, во втором – распознающей. Критерием выбора метода распознавания объекта может служить простота определения меры близости, сложности списания границ классов и образов, разрешающая способность и т. д. При этом весьма важно знать индивидуальные физические особенности распознаваемых объектов, информативность выбранных признаков, количество и качество априорной и текущей информации, возможность введения адаптационных (весовых) коэффициентов и др.

СПИСОК ЛИТЕРАТУРЫ

1. Иванько А.Ф., Иванько М.А., Горчакова Я.В. Методы распознавания образов и задачи логического выделения объектов // Научное обозрение. Технические науки. – 2019. – № 3. – С. 36-40.
2. Ситников В.В., Люминарский В.В., Коробейников А.В. Обзор методов распознавания объектов, используемых в системах машинного зрения // Вестник ИжГТУ им. М.Т. Калашникова. – 2018. – Т. 21, № 4. – С. 222-229.
3. Костылев Д.А., Федотов О.В. Машинное зрение в робототехнических системах // Наука, техника и образование. – 2016. – № 7. – С. 55.
4. Васендина И.С. и др. Разработка методики распознавания образов на изображении на основе структурного подхода // Вестник Брянского государственного технического университета. – 2017. – № 1(54). – С.171-177.
5. Прокофьева Е.В., Прокофьева О.Ю. Кластерный, корреляционный и структурно-лингвистический методы в распознавании образов // Современные исследования в сфере естественных, технических и физико-математических наук. – 2018. – С. 679-685.
6. Хаустов П.А. Алгоритмы распознавания рукописных символов на основе построения структурных моделей // Компьютерная оптика. – 2017. – Т. 41, №. 1.
7. Сидняев Н.И., Бутенко Ю.И., Болотова Е.Е. Экспертная система продукционного типа для создания базы знаний и конструкциях летательных аппаратов // Авиакосмическое приборостроение. – 2019. – №6. – С. 38-52.
8. Фаворская М.Н. К вопросу об использовании формальных грамматик при распознавании объектов в сложных сценах // Решетневские чтения: материалы XIII меж-дунар. науч. конф. Ч. 2. – Красноярск, 2009. – С. 540-541.
9. Singh M., Khan H., Gupta R. Digital Image Processing a Formal Grammar Approach // International Journal for Research in Applied Sciences and Biotechnology (IJRASB). – 2018. – Vol. 5, №2. – С. 3-5.
10. Ходжабекян М.С., Авраменко А.А., Зейтунян В.М. Порождающая грамматика Н. Хомского // Развитие науки и техники: механизм выбора и реализации приоритетов. – 2017. – С. 92.
11. Глушков В.М. Абстрактная теория автоматов // Успехи математических наук. – 1961. – Т. 16, № 5(101). – С. 3-62.
12. Рубцов А.А. Обобщение иерархии Хомского // Дискретные модели в теории управляющих систем. – 2018. – С. 234-237.
13. Гороховатский В.А. Структурный анализ и интеллектуальная обработка данных в компьютерном зрении: монография. – Харьков: Компания СМИТ, 2014. – 316 с.
14. Тупиков В.А., Павлова В.А., Крюков С.Н., Созинова М.В., Шульженко П.К. Лингвистические методы в задачах распознавания изображений // Известия ТулГУ. Технические науки. – 2015. – №11-2. – С. 28-37.
15. Визильтер Ю.В. Обобщенная проективная морфология //Компьютерная оптика. – 2008. – Т. 32. – №. 4.
16. Шашев Д.В., Шидловский С.В. Морфологическая обработка бинарных изображений с ис-

пользованием перестраиваемых вычислительных сред // Автометрия. – 2015. – Т. 51, №. 3. – С. 19-26.
17. Сидняев Н.И., Храпов П.В. Нейросети и нейроматематика: учебное пособие / под ред. Н.И. Сидняева. – М.: Изд-во МГТУ им. Н.Э. Баумана, – 2016. – 83 с.
18. Финогеев А.Г., Четвергова М.В. Разработка и исследование методики распознавания изображений для систем расширенной реальности // Известия Волгоградского государственного технического университета. – 2012. – Т. 15, № 15. – С. 130-136.
19. Фаворская М.Н., Горошкин А.Н. Модель распознавания изображений рукописного текста // Сибирский журнал науки и технологий. – 2008. – № 2(19). – С. 52-57.
20. Соловьев С.Ю. Эквивалентные преобразования контекстно-свободных грамматик // Информационные процессы. – 2010. – Т. 10, № 3. – С. 292-302.

Материал поступил в редакцию 24.05.20.

Сведения об авторах

СИДНЯЕВ Николай Иванович – доктор технических наук, профессор, заведующий кафедрой «Высшая математика» Московского государственного технического университета имени Н.Э. Баумана (национальный исследовательский университет)
e-mail: sidnyaev@yandex.ru

БУТЕНКО Юлия Ивановна – кандидат технических наук, доцент, доцент кафедры «Романо-германские языки» Московского государственного технического университета имени Н.Э. Баумана (национальный исследовательский университет)
e-mail: iuliiabutenko2015@yandex.ru

БОЛОТОВА Елизавета Евгеньевна – студент Московского государственного технического университета имени Н.Э. Баумана (национальный исследовательский университет)
e-mail: bolotovaee@mail.ru

УДК 002::37.016

Н.А. Калинина, В.Р. Милов, А.А. Салтыкова, М.С. Дубов

О формировании информационного обеспечения образовательных программ инженерной направленности

Предложен способ создания информационного обеспечения для достижения целей образовательных программ, заключающихся в формировании у обучающихся компетенций, заданных требованиями образовательного стандарта. С помощью нечеткой сетевой модели определены целевые функции в зависимости от состава информационного обеспечения и порядка изучения информационных фрагментов (элементов). Применение процедуры кластеризации позволяет уменьшить вычислительную сложность при формировании информационного обеспечения. Представленные примеры демонстрируют работоспособность развиваемого подхода.

Ключевые слова: *информационное обеспечение, модель предметной области, цели информационного обеспечения, нечеткая сетевая модель, структура информационного обеспечения*

DOI: 10.36535/0548-0027-2020-08-2

ВВЕДЕНИЕ

Несмотря на многочисленные исследования в области образовательной среды, пространств знаний, организации учебного процесса, включая *e-learning*, можно констатировать отсутствие единого общепринятого подхода к определению состава и структуры информационного обеспечения образовательных программ, а также последовательности применения информационных объектов, обеспечивающих повышение эффективности процессов приобретения знаний. В связи с этим представляется актуальной разработка теоретических основ и процедур анализа, синтеза структуры и последовательности применения учебных материалов, а также способа формирования информационного обеспечения, которые позволили бы достичь поставленной цели образовательных программ с учетом ранее приобретенных знаний.

МОДЕЛЬ ПРЕДСТАВЛЕНИЯ ЗНАНИЙ

Известные модели представления любой предметной области [1] детально описывают иерархическую структуру контента (в нашем случае – учебного курса), используя отношение часть – целое [2]. Последовательность освоения пользователями информационных элементов предметной области, в частности, образовательной программы, поддерживается в сетевых моделях [3], основанных на анализе дидактических связей между объектами.

Использование только сетевых или только иерархических моделей не позволяет всесторонне описать предметную область. Комбинация этих подходов дала возможность построить иерархическую сетевую модель знаний предметной области [4, 5]. В ней информационное множество X, соответствующее рассматриваемой предметной области, разделено на информационные фрагменты первого уровня $C_{1,i}$, которые содержат информационные фрагменты второго более низкого уровня $C_{2,j}$. Информационные фрагменты нижнего уровня состоят из информационных элементов x_n, которые на уровне фрагментов контента считаются неделимыми. Схема иерархической сетевой модели знаний предметной области представлена на рис. 1.

НЕЧЕТКАЯ СЕТЕВАЯ МОДЕЛЬ ФОРМИРОВАНИЯ ЦЕЛЕЙ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

Для решения поставленных задач мы предлагаем использовать нечеткую сетевую модель формирования целей образовательной программы (рис. 2), основанную на частных целевых функциях, отражающих степень достижения *m*-ой цели образовательной программы $J_m(\mathbf{a}) = J_m(a_1,...,a_N)$. Цели образовательной программы заключаются в формировании у обучающихся компетенций, определенных требованиями образовательного стандарта и востребованных работодателями. Компонентами вектора $\mathbf{a} = [a_1,...,a_N]^{\mathrm{T}}$ являются степени освоения информационных элементов $x_n \in X$, где X − множество этих элементов. Дидактические связи, определяющие порядок применения информационных объектов, представляются квадратной $(N \times N)$-матрицей смежности $\mathbf{W}$ и графом, соответствующим этой матрице.

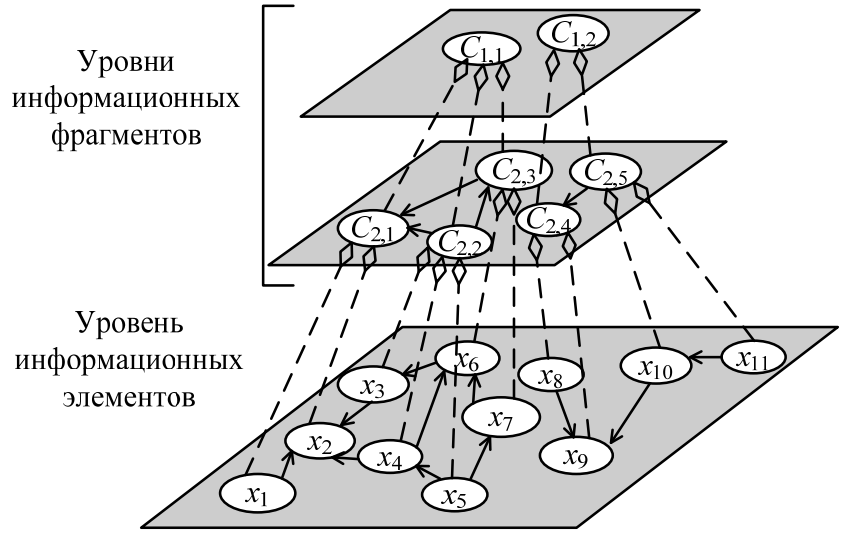

Рис. 1. Иерархическая сетевая модель знаний предметной области

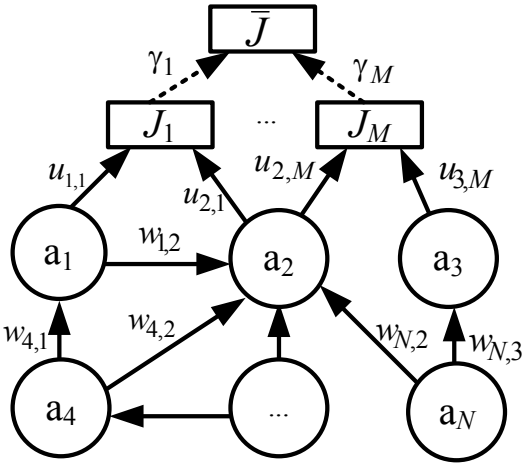

Рис. 2. Нечеткая сетевая модель формирования целей образовательной программы

Связи между информационными элементами и целями ОП описаны $(N \times M)$-матрицей $\boldsymbol{U}$. Так элементы $u_{nm} \in [0;1]$ характеризуют влияние степени освоения n-го информационного элемента на достижение m-й цели программы.

Цель ОП достигается при освоении всех информационных элементов верхнего уровня (рис. 2). Степень достижения целей ОП J_m, $m = \overline{1,M}$ зависит от весов u_{nm}, $n = \overline{1,N}$, $m = \overline{1,M}$, характеризующих влияние степени a_n, $n = \overline{1,N}$ освоения отдельных информационных элементов. Поскольку успешное освоение программы невозможно при пропуске существенных информационных элементов, для формирования частных целевых функций применим мультипликативную свертку:

$$J_m(\mathbf{a}) = \prod_{n=1}^{N} a_n^{u_{nm}}, \quad m = \overline{1,M} \tag{1}$$

Для решения задачи многокритериальной оптимизации воспользуемся методом мультипликативной свертки частных целевых функций:

$$\overline{J}(\mathbf{a}) = \prod_{m=1}^{M} J_m^{\gamma_m}(\mathbf{a}), \tag{2}$$

где $\gamma_m > 0$, $m = \overline{1,M}$ – весовые коэффициенты, определяющие относительные степени важности частных целевых функций. Корректировка этих весовых коэффициентов позволяет учесть специфические требования отдельных работодателей. Подставляя (1) в (2) и логарифмируя, перейдем к аддитивному представлению обобщенной целевой функции:

$$\begin{aligned} \overline{J_L}(\mathbf{a}) &= \ln \overline{J}(\mathbf{a}) = \sum_{m=1}^{M} \gamma_m \ln J_m(\mathbf{a}) = \\ &= \sum_{m=1}^{M} \gamma_m \sum_{n=1}^{N} u_{nm} \ln a_n = \sum_{n=1}^{N} \beta_n \ln a_n \end{aligned} \tag{3}$$

Таким образом, степень достижения целей образовательной программы $\overline{J}(\mathbf{a}) = \exp\left(\overline{J_L}(\mathbf{a})\right)$ зависит от освоенности отдельных информационных элементов a_n, $n = \overline{1,N}$ и от коэффициентов влияния

$$\beta_n = \sum_{m=1}^{M} \gamma_m u_{nm} = \boldsymbol{\gamma}^{\mathrm{T}} \mathbf{u}_n \qquad (4)$$

где

$$\boldsymbol{\gamma} = [\gamma_1, ..., \gamma_N]^{\mathrm{T}}, \ \mathbf{u}_n = [u_{n1}, \ldots, u_{nM}]^{\mathrm{T}}.$$

ПРОЦЕДУРА ФОРМИРОВАНИЯ ИНФОРМАЦИОННОГО ОБЕСПЕЧЕНИЯ

Степени a_n, $n = \overline{1,N}$ освоения отдельных информационных элементов (ИЭ) и, соответственно, достижения цели образовательной программы зависят от состава контента $S \subseteq X$ и порядка освоения ИЭ. Состав контента охарактеризуем вектором $\mathbf{s} = [s_1, ..., s_N]^{\mathrm{T}}$ индикаторных переменных $s_n = \chi_S(x_n)$, $n = \overline{1,N}$, таких, что $s_n = 1$, если $x_n \in S$ и $s_n = 0$ в противном случае. Порядок освоения ИЭ определим вектором $\mathbf{r} = [r_1, ..., r_K]^{\mathrm{T}}$, где $K = \mathrm{card}(S) = \sum s_n$ – мощность множества S, включающего ИЭ, подлежащие изучению. Если k-м по порядку изучается ИЭ $x_n = 0$, то $r_k = n$.

Состав контента и порядок его освоения будем искать из условия максимизации целевой функции (3), характеризующей степень достижения целей образовательной программы. Алгоритм формирования информационного обеспечения включает две вложенные процедуры оптимизации. Внутренняя процедура заключается в определении порядка освоения ИЭ на основе максимизации целевой функции:

$$\hat{\mathbf{r}} = \arg\max_{\mathbf{r}} \tilde{J}_L(\mathbf{r}, \mathbf{s}), \qquad (5)$$

при некотором фиксированном векторе индикаторных переменных $\mathbf{s}$, определяющем состав контента S на текущей итерации алгоритма. Для определения последовательности освоения ИЭ из решения задачи (5), применен алгоритм топологической сортировки на направленном ациклическом графе дидактических связей с обходом в глубину. В задаче (5) целевая функция $\tilde{J}_L(\mathbf{r}, \mathbf{s}) = \overline{J_L}(\tilde{\mathbf{a}})$ вычисляется согласно (3) для значений степени освоения ИЭ, найденных при заданном порядке их применения. Степени освоения ИЭ находятся с помощью итерационной процедуры, основанной на разработанном в [4] нечетком правиле:

$$\tilde{a}_{r_k} = \underset{j \in Pa_{r_k}}{T^*}(\mathbf{a}, \mathbf{w}_{r_k}), \quad k = \overline{1,K}. \qquad (6)$$

Здесь $T^*(\bullet)$ – взвешенная t-норма; Pa_{r_k} – множество предков r_k-го элемента, т.е. множество ИЭ, имеющих дидактические связи, направленные к r_k-му элементу; $\mathbf{r}$ – N-вектор, определяющий порядок освоения контента и являющийся перестановкой; $\mathbf{w} = [w_{1,r_k}, ..., w_{N,r_k}]^{\mathrm{T}}$. Разработанный алгоритм, основанный на последовательном применении нечеткого правила (6), можно трактовать как моделирование процесса «распространения изученности».

Внешняя процедура оптимизации в алгоритме формирования информационного обеспечения заключается в определении состава информационных элементов на основе решения задачи о многомерном рюкзаке. Классическая задача о многомерном рюкзаке состоит в поиске подмножества объектов из условия максимизации суммарной ценности отобранных объектов при наличии нескольких линейных ограничений на так называемые ресурсы. Поскольку изменение состава $\mathbf{s}$ изучаемых информационных элементов приводит также к изменению значений степени их освоения $\hat{\mathbf{a}}$, то в решаемом варианте задачи о рюкзаке целевая функция оказывается нелинейной:

$$\hat{J}_L(\mathbf{s}) = \sum_{n=1}^{N} s_n \beta_n \ln(\hat{a}_n), \qquad (7)$$

где $\hat{a}_n$ – степени освоения информационного элемента, найденные согласно (6), для наилучшего порядка $\hat{\mathbf{r}}$, удовлетворяющего решению задачи (5).

Множество допустимых решений в распространенном случае определяется линейными ограничениями-неравенствами $\mathbf{s}^{\mathrm{T}}\mathbf{t} \le T_A$, $\mathbf{s}^{\mathrm{T}}\mathbf{v} \le V_A$, характеризующими трудоемкость и стоимость применения всех информационных элементов, включенных в состав информационного обеспечения. Здесь $\mathbf{t}$ и $\mathbf{v}$ – N-векторы, компоненты которых представляют значения трудоемкости и стоимости отдельных ИЭ, T_A и V_A – допустимые трудоемкость и стоимость освоения всех отобранных ИЭ. При этом вариант постановки задачи определения состава элементов с единственным ограничением также возможен.

ПРИМЕНЕНИЕ КЛАСТЕРИЗАЦИИ ДЛЯ ФОРМИРОВАНИЯ ИНФОРМАЦИОННОГО ОБЕСПЕЧЕНИЯ

Для снижения вычислительной сложности процедуры определения структуры информационного обеспечения и последовательности применения информационных элементов был предложен подход [4], основанный на выделении кластеров информационного обеспечения.

На рис. 3 представлена процедура определения структуры информационного обеспечения для случая двухуровневой иерархической сетевой модели предметной области

Совокупность информационных фрагментов, состоящих из информационных элементов, а также матриц смежности элементов и фрагментов образуют иерархическую сетевую модель предметной области.

Для кластеризации информационных элементов предложен способ определения коэффициентов близости между ними, основанный на обобщении коэффициентов Серенсена (коэффициенты определяющие бинарную меру сходства) $K = 2\,\mathrm{card}(A \cap B) / (\mathrm{card}(A) + \mathrm{card}(B))$, где $\mathrm{card}(\bullet)$ – мощность соответствующих множеств. Для случая взвешенных дидак-

тических связей, с применением аппарата нечетких множеств, в [5] получены выражения для расчета коэффициентов близости ИЭ по входящим и исходящим связям, учитывающим наличие общих предков и общих потомков, соответственно, на основе которых предложен обобщенный коэффициент близости информационных элементов:

$$K_{ij} = \frac{2\sum_{n=1}^{N} \min\left(w_{ni}, w_{nj}\right) + 2\sum_{n=1}^{N} \min\left(w_{in}, w_{jn}\right)}{\sum_{n=1}^{N}(w_{ni} + w_{nj}) + \sum_{n=1}^{N}(w_{in} + w_{jn})} . \quad (8)$$

Используя формулу (8), сформируем матрицу **D** расстояний между ИЭ с элементами $d_{ij} = 1 - K_{Cij}$.

Для определения весов дидактических связей w_{Lij} между полученными информационными фраг-ментами применяются «внешние» связи между вошедшими в них информационными элементами, т. е. связи от ИЭ одного фрагмента к ИЭ другого фрагмента. Веса дидактических связей между информационными фраг-ментами определены на основе несимметричных коэффициентов близости Серенсена для двух множеств (***A*** и ***B***):

$$K_{AB} = \text{card}(A \cap B)/\text{card}(A) \text{ и } K_{BA} = \text{card}(A \cap B)/\text{card}(B) .$$

Для случая бинарных связей между элементами вес связи между информационными фрагментами ***A*** и ***B*** на иерархическом уровне L определяется как $w_{LAB} = k_{AB}/N_{\text{ИЭ}B}$, где $N_{\text{ИЭ}B}$ – число информационных элементов, составляющих информационный фрагмент ***B***, k_{AB} – число информационных элементов фрагмента ***B***, зависимых от одного или нескольких ИЭ фрагмента ***A***. Для случая взвешенных связей $w_{LAB} = N_{\text{ИЭ}B}^{-1} \sum_{i=1}^{N_{\text{ИЭ}B}} (\max_{j \in A} w_{ABij})$, где $N_{\text{ИЭ}A}$ и $N_{\text{ИЭ}B}$ – количество ИЭ в информационных фрагментах ***A*** и ***B*** соответственно, $\mathbf{W}_{AB}$ – матрица связей от ИЭ фрагмента ***A*** к ИЭ фрагмента ***B*** (рис. 4). Полученные веса дидактических связей w_{Lij} формируют матрицу $\boldsymbol{W}_L$ (см. рис. 3), где L – уровень информационных фрагментов в иерархической модели.

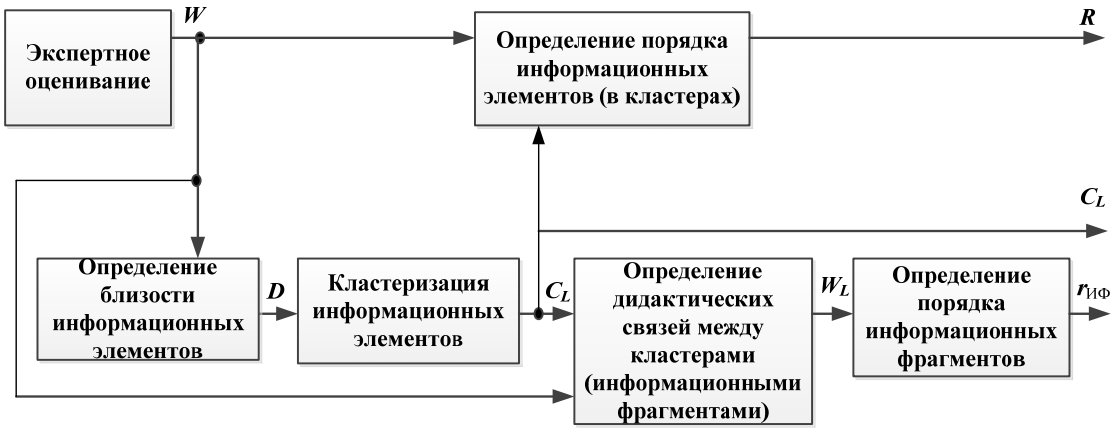

Рис. 3. Схема (процедура) определения структуры информационного обеспечения и последовательности применения информационных элементов: **D** – матрица расстояний между ИЭ, **R** – порядок их изучения, $\mathbf{C}_L$ – кластеры информационных элементов (информационные фрагменты), $\mathbf{W}_L$ – матрица весов дидактических связей между кластерами на уровне L, $\mathbf{r}_{\text{иф}}$ – порядок изучения информационных фрагментов

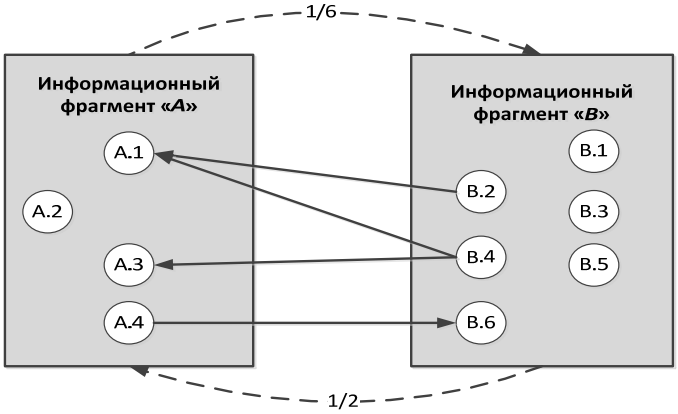

Рис. 4. Определение дидактических связей между информационными фрагментами

Для выделения нескольких уровней в иерархической модели предметной области, полученные на первом этапе информационные фрагменты, объединятся в кластеры, представляющие информационные фрагменты следующего уровня. Кластеризация выполняется на основе расстояний $d_{Lij} = 1 - K_{Lij}$ где коэффициенты близости между информационными фрагментами определяются (8) с использованием весов w_{LAB}.

ПРИМЕРЫ РАЗРАБОТКИ ИНФОРМАЦИОННОГО ОБЕСПЕЧЕНИЯ

В качестве примера разработки информационного обеспечения рассмотрим задачу отбора курсов повышения квалификации в системе корпоративного обучения сотрудников. Задача определения состава и последовательности прохождения таких курсов, заданных графом дидактических связей (рис. 5), решается при наличии ограничений на финансовые и временные ресурсы, которые представляют стоимость и продолжительность программы повышения квалификации соответственно. Заданные цели обучения отражают необходимые знания и компетенции.

В рамках рассматриваемого примера предварительно отобранные тематические курсы повышения квалификации соответствуют $N = 10$ информационным элементам x_n, $n = \overline{1,10}$, которые в совокупности составляют множество X. На первом этапе из множества осваиваемых курсов S исключены информационные элементы x_6 и x_7, направленные на формирование уже имеющихся (базовых) знаний, которые определены с помощью входного тестирования.

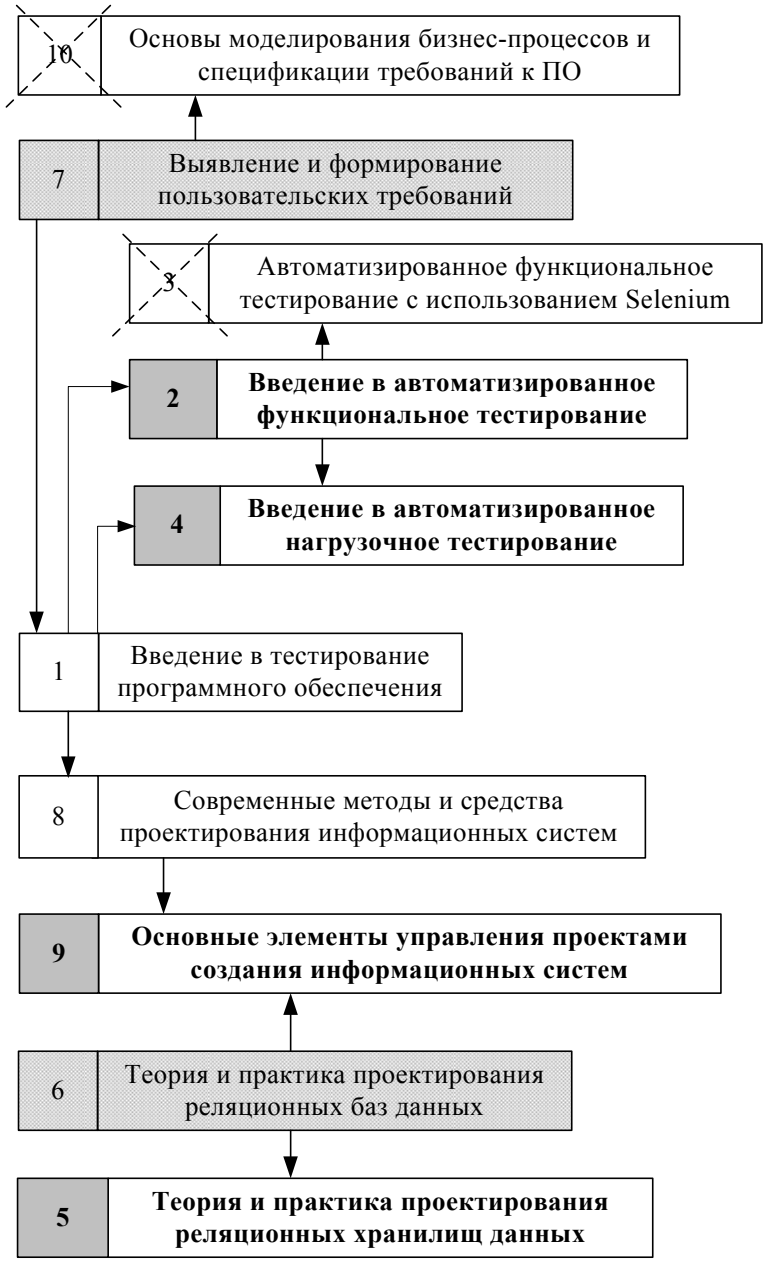

Рис. 5. Структура курсов повышения квалификации

В результате решения задачи о многомерном рюкзаке найдено подмножество S осваиваемых курсов $x_1, x_2, x_4, x_5, x_8, x_9$, которому соответствует вектор индикаторных переменных $\mathbf{s} = [1,1,0,1,1,0,0,1,1,0]^{\mathrm{T}}$. Порядок изучения $\mathbf{r} = [1,8,9,5,2,4]^{\mathrm{T}}$ курсов получен с помощью разработанной процедуры определения последовательности освоения информационных элементов (дисциплин):

- «Ведение в тестирование ПО».
- «Современные методы и средства проектирования информационных систем».
- «Основные элементы управления проектами в проектировании информационных систем».
- «Теория и практика проектирования реляционных хранилищ данных».
- «Введение в автоматизированное функциональное тестирование».
- «Введение в автоматизированное нагрузочное тестирование».

Разработанный подход применялся при определении последовательности изучения материала по дисциплине «Электроника». С помощью алгоритма иерархической агломеративной кластеризации выделены информационные фрагменты, представляющие разделы дисциплины, и установлены связи между полученными кластерами. Для элементов каждого кластера (раздела дисциплины) определена последовательность их изучения с применением алгоритма топологической сортировки графа, а также последовательность изучения разделов.

С помощью предложенного подхода выполнена автоматизированная разработка учебного плана подготовки магистров по направлению «Информационные системы и технологии» в ФГБОУ ВО «Нижегородский государственный технический университет им. Р.Е. Алексеева» (НГТУ). Ограничения в задаче составления учебного плана определяются требованиями образовательных стандартов. Получена структура учебного плана, соответствующая ярусно-параллельной форме графа и удовлетворяющая всем заданным ограничениям.

Ещё одним перспективным направлением применения разработанного подхода является структурирование дидактических материалов предметных областей для проектирования информационно-справочных систем и интерактивных электронных технических инструкций [6], в том числе, при подготовке операторов технологических процессов и других специалистов.

ЗАКЛЮЧЕНИЕ

Сочетание иерархического подхода к представлению информационных элементов и дидактических связей между ними позволило разработать способ формирования информационного обеспечения для образовательных программ инженерной направленности. Предложен алгоритм определения состава информационного обеспечения и порядка изучения информационных элементов для достижения целей образовательных программ.

Снижение вычислительной сложности достигается с помощью процедуры кластеризации для формирования иерархически упорядоченных фрагментов информационного обеспечения, например, материал по темам объединяется в разделы, которые, в свою очередь, составляют дисциплину. Порядок изучения информационных элементов определяется внутри соответствующих фрагментов, которые также упорядочиваются на более высоких иерархических уровнях.

Предложенный подход нашел применение в автоматизированной разработке учебных планов, формировании контента отдельных дисциплин, учебных и справочных ресурсов, а также в поддержке принятия решений при повышении квалификации и др. В результате обеспечиваются объективные решения при определении рационального состава и предпочтительного порядка освоения информационных элементов.

СПИСОК ЛИТЕРАТУРЫ

1. Doignon J.P., Falmagne J.C. Knowledge Spaces and Learning Spaces. – Cornell University Library, 2015. – 51 p. – URL: https://arxiv.org/abs/1511.06757v1
2. Лукашевич Н.В. Отношения часть – целое: теория и практика // Нейрокомпьютеры: разработка, применение. – 2013. – № 1. – С. 7-12.
3. Абрамский М.М., Циммерман А.М., Альмухаметова А.А., Алтынбаева Д.Т. Онтологический подход к проектированию образовательных программ в цифровых средах // Научно-техническая информация. Сер. 2. 2019. – № 3. – С. 14-20; Abramskiy M.M., Tsimmerman A.M., Almukhametova A.A. and Altynbaeva D.T. An Ontological Approach to Education Program Design in Digital Environments. – 2019. – Vol. 53, № 2. P. 64-70.
4. Калинина Н.А., Милов В.Р., Баранов Д.В. Процедуры поддержки принятия решений при управлении информационным обеспечением // Информационно-измерительные и управляющие системы. – 2017.– № 8. – С. 55-62.
5. Калинина Н.А. Способ построения иерархической сетевой модели предметной области и определения порядка освоения знаний // Материалы XXIII международной научно-технической конференции «Информационные системы и технологии» (ИСТ-2017). – Нижний Новгород: Нижегородский государственный технический университет им. Р. Е. Алексеева, 2017. – С.112-116.
6. Киселев В.Ф., Милов В.Р., Санин М.А. Информационно-обучающая система для подготовки и аттестации оперативно-диспетчерского персонала, обслуживающего магистральные газопроводы // Автоматизация в промышленности. – 2009. – № 7. – С. 48-51.

Материал поступил в редакцию 02.06.2020

Сведения об авторах

КАЛИНИНА Наталья Андреевна – кандидат технических наук, доцент кафедры «Электроника и сети ЭВМ» Федерального государственного бюджетного образовательного учреждения высшего образования (ФГБОУ ВО) «Нижегородский государственный технический университет им. Р.Е. Алексеева»
e-mail: alipovana@mail.ru

МИЛОВ Владимир Ростиславович – доктор технических наук, профессор, заведующий кафедрой «Электроника и сети ЭВМ» ФГБОУ ВО «Нижегородский государственный технический университет им. Р.Е. Алексеева»
e-mail: milov@nntu.ru

САЛТЫКОВА Анна Александровна – старший преподаватель кафедры «Иностранные языки» ФГБОУ ВО «Нижегородский государственный технический университет им. Р.Е. Алексеева»
e-mail: saltykova111@mail.ru

ДУБОВ Максим Сергеевич – старший преподаватель кафедры «Электроника и сети ЭВМ» ФГБОУ ВО «Нижегородский государственный технический университет им. Р.Е. Алексеева»
e-mail: demson@decadalab.ru

УДК 004.896

С.Г. Черный, В.А. Доровской, Б.П. Новак

Концепция построения информационной подсистемы АСУ промышленным производством

На современном этапе развития достаточно актуален вопрос цифровизации процессов промышленного производства. Практика разработки и внедрения АСУ приводит к выводу о необходимости применения ориентированного системного подхода на всех стадиях исследовательских и проектных работ. Показано, что проектируемые системы могут быть приняты как элементы банков структуры данных, обеспечивающих наращивание и корректировку без изменения алгоритмов обработки и поиска данных, а также для разработки форм документов и структур массивов, что позволяет оптимизировать работу предприятия в период чрезвычайных ситуаций.

Ключевые слова: *концепция, построение информационной подсистемы, автоматизированные системы управления, промышленные производства*

ВВЕДЕНИЕ

Разработка и внедрение автоматизированных систем управления, базирующихся на современных научных достижениях в области вычислительной техники, теории управления, а также экономико-математических методов, и охватывающих сферу организационного управления, являются одним из главных путей повышения эффективности промышленных производств (ПП). Практика разработки и внедрения АСУ ПП приводит к выводу о необходимости применения системного подхода на всех стадиях исследовательских и проектных работ [1-3]. Это, в первую очередь, относится к разработке информационной подсистемы (ИС) как ядра АСУ.

Разработка и конкретная реализация ИС связана с проведением исследований и решением следующих задач:

1) выбор наиболее рационального метода организации информации в системе;

2) создание набора алгоритмов и программ, обслуживающих систему независимо от метода размещения данных;

3) описание метода организации информации на языке, понятном и человеку, и машине.

В последнее время наметился такой подход решения перечисленных задач, при котором ИС является надстройкой над уже имеющейся вычислительной и операционной системами с элементами дистанционного (удаленного) управления. Этот подход объясняется тем, что существующие в машинном математическом обеспечении библиотеки стандартных программ слабо учитывают специфику информационного поиска и описания данных, обрабатываемых в определенной системе управления. **При этом подходе возможно решение принципиальных проблем разработки ИС методом выделения инвариантных блоков, реализация которых позволяет рассматривать структуру ИС как набор однотипных элементов и возможность построения управляемых объектов.** Важной особенностью любой ИС является обеспечение возможности описания метода организации данных и алгоритмов оперирования данными в терминах используемого информационного языка.

Цель настоящей работы – это не перечисление или выбор возможных инвариантных блоков, (этот выбор зависит от характера решаемых задач системой и от свойств управляемых объектов), а построение такой модели блоков ИС и их описания на языке, понятном человеку и машине, которая позволяет создать набор алгоритмов и программ, обслуживающих систему, независящих от содержимого блоков конкретных ИС, и которые позволят осуществлять удаленную работу при возникновении чрезвычайных ситуаций.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Объекты управления систем производственного управления относительно наблюдателя, с точки зрения информационной теории, представляются информационным отображением. Элементарное информационное отображение (так называемая наблюдаемая динамическая переменная) – это некоторая характеристика определенного объекта, следовательно, ИС есть, прежде всего, отображение объектов [2, 4].

Информационные системы АСУ аналитически можно представить как совокупность

$$I = \left\{ Q(t), U, S\left(\bar{Q}, \bar{U}\right), F(S), \Theta(Q) \middle| \varphi \right\},$$

где $Q(t) = \{q_1, q_2, ..., q_m\}$ – множество наблюдаемых динамических переменных (НДП) объектов управления системы;

$U = \{A, B, ...\}$ – семейство конечных множеств имен признаков (свойств), однозначно определяющих наблюдаемые динамические переменные (логический уровень задания информации); при этом каждое имя признака (свойства) НДП является в свою очередь, множеством значений отдельного признака (свойства) $A = \{a_1, a_2, ..., a_n\}$, а любой элемент множества A содержит информацию, которая может быть записана в поле, размером (длиной) ℓ_j единиц информации;

$S\left(\bar{Q}, \bar{U}\right)$ – множество моделей информационных структур ИС, которые являются функцией, зависящей от двух других параметров: 1) $\bar{Q}$ – мощность множества Q НДП объектов управления системы и 2) $\bar{U}$ – мощность множества признаков (свойств) НДП с учетом множеств, образованных отдельным признаком;

$F(S)$ – алгоритм поиска, однозначно отображающий Q в U;

$\Theta(Q)$ – операторы преобразования моделей информационных структур, определяемые способом их описания и структурой описания данных;

φ – коэффициент нагруженности системы.

Зависимость алгоритма $F(S)$ и состава операторов $\Theta(Q)$ информационной системы от выбранной структуры описания данных метода организации данных в моделях информационных структур – очевидна. Таким образом, чтобы разработать эффективную ИС, необходимо создать метод организации данных в информационных структурах на основе единого языка описания наблюдаемых динамических переменных, а также осуществить четкую идентификацию модели информационных структур (МИС), рассматриваемых как новый подход к организации и разработке информационной системы управления [5, 6].

Основой предлагаемой в настоящей статье МИС являются обобщенные понятия языка экономических показателей и декомпозиция информационного отображения объектов управления.

Модель информационных структур представлена в виде множества $\{M\}$ размером $m \times n$, где: m – строки, n – столбцы.

Каждый вектор-столбец и вектор-строка МИС определяются количеством элементов однородных признаков (свойств) наблюдаемой динамической переменной, объединяемой в этой МИС. Результаты разработки информационной системы целесообразно представлять в виде конструкций форм документов и структур массивов информации, определенных на основе модели информационных услуг.

Результаты проведенных исследований показывают, что модели информационных структур могут быть приняты как элементы банков данных, обеспечивающих наращивание и корректировку без изменения алгоритмов обработки и поиска данных.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ ИССЛЕДОВАНИЙ

Синтез информационной структуры ПП в условиях функционирования АСУ

Под информационной структурой промышленных предприятий понимается совокупность всех сведений, получаемых предприятием извне, передаваемых внутри производства и выходящих за его пределы.

Если рассматривать производство как множество элементов, имеющих общую цель, то, очевидно, что качество управления (в любом смысле) зависит: от полномочий органов управления и разбиения хозяйственного комплекса на различные отделы или подразделения, а также от информации, используемой в процессе решения задач и принятия решений. Все эти обстоятельства требуют решения задачи синтеза рациональной информационной структуры производства в условиях функционирования АСУ.

Такая задача по существу имеет следующий аспект: производство состоит из участников $i \in I = \{1, 2, ..., N\}$, каждый из которых получает информацию в виде вектора $\{x_{1i}, x_{2i}, ..., x_{ni}\} = Y_i(x)$ и управляет вектором $\{k_{1i}, k_{2i}, ..., k_{li}\} = Z_i(k)$.

Общая целевая функция всех участников $I = I(\gamma_1, \gamma_2, ..., \gamma_N)$, где γ_i – правило управления, применяемое участником i $Z_i(k) = \gamma_i\{Y_i(x)\}$. Следовательно, получаем задачу оптимизации информационной структуры системы управления при ограничениях, определяемых набором функций и требуемым режимом их выполнения. Если принимается во внимание каждый из информационных векторов в качестве информационной структуры, то возникает проблема её оптимизации при увеличении объема информации.

Исходные предположения, лежащие в основе модели информационных структур, сводятся к следующему:

1) функция управления, лежащая в основании того или иного органа управления, справедлива для всех систем управления, принадлежащих к достаточно широкому классу (идея однородности или равноправия);

2) наблюдаемая динамическая переменная зависит, по крайней мере, от двух других, а вообще, от нескольких переменных (идея бинарных, тернарных и т.д. отношений);

3) функция управления в своём наиболее первичном виде – это связь между наблюдаемыми динамическими переменными;

4) во множество наблюдаемых переменных и во множество объектов системы управления можно ввести понятия близости, непрерывности, окрестности (методы теории классификаций и компараторной идентификации).

В общем виде задача создания информационной структуры АСУ сводится к определению метода группировки численных значений различных наблюдаемых динамических переменных объекта управления, удовлетворяющих ограничениям по описанию данных – на уровне пользователей и используемых носителей информации, и позволяет отличать описания данных в зависимости от таких уровней, как: информационная структура, логическая структура данных и структура физических записей.

Для лучшего понимания формирования математической модели информационной структуры, рассмотрим множество X наблюдаемых динамических переменных, объекта управления, численные значения которых образуют множество N. Данному множеству X присуще множество $U = \{A, B, C, D, E, R\}$ свойств объекта управления, каждое их которых, представляет множество значений отдельного свойства; обозначим их *a, b, c, d, e, r,* где:

$$a \subseteq A,\ b \subseteq B,\ c \subseteq C,\ d \subseteq D,\ e \subseteq E,\ r \subseteq R,$$
$$a = \{a^1, a^2, ..., a^m\},\ b = \{b^1, b^2, ..., b^m\},$$
$$c = \{c^1, c^2, ..., c^m\},\ d = \{d^1, d^2, ..., d^m\},$$
$$e = \{e^1, e^2, ..., e^m\},\ r = \{r^1, r^2, ..., r^m\}.$$

В нашем случае информационная структура должна содержать $N = m \cdot k \cdot \ell \cdot p \cdot q \cdot h$ данных.

При таком рассмотрении структуры становится очевидна необходимость организовывать модель идентификации в виде совокупности матриц $S = f[Y, Z]$, где Y и Z – соответственно множество строк и столбцов матрицы (предполагается, что $Y > Z$) представляющие собой выбранные множества двух из имен свойств объекта управления. Для этого случая, выбрав в качестве вектора–строки свойство *B*, а вектора–столбца свойство *A* матрицы *S*, получим модель информационной структуры.

Количество матриц в информационной структуре определяется множествами имен свойств объекта управления для данной совокупности НДП и определяется по формуле:

$$\Psi = C_{\ell+p+q+h}^{U-2} - C_{\ell+p}^{U-2} - C_{p+q}^{U-2} - C_{\ell}^{U-2} - C_{p}^{U-2} - C_{q}^{U-2} - C_{h}^{U-2} = \ell \cdot p \cdot q \cdot h. \quad (1)$$

В данном случае $U = 6$.

Порядковый номер любого элемента в информационной структуре определяется формулой:

$$Q_U = \ell + (p-1) + (q-1)p + (h-1)p \cdot q. \quad (2)$$

Порядковый номер матрицы в информационной структуре вычисляется по формуле:

$$\Theta_N - m_i \cdot k_i \cdot \ell_i \cdot p_i \cdot q_i \cdot h_i \quad (3)$$

В формулах (1–3) m, k, ℓ, p, q, h определяют место имени свойств объекта управления в модели информационной структуры.

Пример реализации разрабатываемой системы, авторы публикации интегрируют на модели ИС судового комплекса.

Системная технология KONGSBERG позволяет создать распределенную и открытую систему, использующую стандартизированную коммуникационную сеть, облегчающая интеграцию других систем судов и свободный поток информации из всех подсистем, а информация о судне доступна на многофункциональных рабочих станциях. Общая базовая технология и пользовательский интерфейс обеспечивают безопасную и эффективную рабочую среду с последовательной работой и повышенной надежностью (https://www.kongsberg.com/).

Наиболее важными и критичными узлами этой системы судового комплекса являются системы измерения оборотов главного двигателя, управлением говернёром (серво приводом), а также интерфейс подключения локальной панели управления. Эти узлы и компоненты относятся к Fault-Tolerant Control категории по следующим причинам [5, 6]:

- системы обеспечивают аварийное управление главным двигателем с локального поста, при отказе общей системы (находятся в замкнутой CAN-шине до первой развязки PSS-контроллеров;
- испытывают максимальные механические нагрузки и тяжелые условия окружающей среды (вибрации главного двигателя, высокие температуры машинного отделения и непосредственное расположение на корпусе или подушках двигателя);
- являются необслуживаемыми компонентами системы;
- требуют ежегодного обслуживания специалистами.

При обрыве CAN-шины система выдает ошибку коммуникации всех модулей системы и приводит к повышенной нагрузке на backup-линию связи. Приведем список ошибок модулей связи:

COMM ERR BETW. ACP/ECR – ESU CAN A FAIL
COMM ERR BETW. ACP/ECR – RPMU1 CAN A FAIL
COMM ERR BETW. ACP/ECR – RPMU2 CAN A FAIL
COMM ERR BETW. ACP/ECR – RAI CAN A FAIL
COMM ERR BETW. ACP/ECR – DGU CAN A FAIL
COMM ERR BETW. ACP/ECR – MEI CAN A FAIL
COMM ERR BETW. ACP/ECR – C2LOC CAN A FAIL
COMM ERR BETW. ACP/ECR – LTU BR CAN A FAIL
COMM ERR BETW. ACP/ECR – LTU ECR CAN A FAIL
COMM ERR BETW. ACP/ECR – PBT CAN A FAIL
COMM ERR BETW. ACP/ECR – MPP ECR CAN A FAIL
COMM ERR BETW. ACP/ECR – ACP-BR CAN A FAIL
COMM ERR BETW. ACP/ECR – RDO CAN A FAIL
COMM ERR BETW. ACP/ECR – DGU SIO CAN A FAIL

В режиме удаленной работы инженер может перезапустить систему и устранить необходимые неполадки. В некоторых случаях, будет необходимо вмешательство бортовой службы поддержки, но все же не вызывая специальную службу базового ремонта.

ВЫВОДЫ

В результате наших исследований было установлено, что использование модели информационных структур для разработки форм документов и структур массивов позволяет переходить к инженерным методам разработки информационного обеспечения АСУ. Предложены концепция построения динамической информационной системы управления промышленным предприятием и методология синтеза его рациональной информационной структуры в условиях функционирования АСУ. Показано, что модели информационных структур могут быть приняты как элементы банков данных, обеспечивающих процесс наращивания и корректировки без изменения алгоритмов их обработки и поиска данных в них, а также для разработки форм документов и структур массивов.

СПИСОК ЛИТЕРАТУРЫ

1. Авраменко В.П. Управление производством в условиях неопределенности: Монография. – Киев: НВК ВО, 1992. – 48 с.
2. Баранов В.В., Колянов Г.Н., Попов Ю.И. Автоматизация управления предприятием. – М.: ИНФА-М, 2000. – 239 с.
3. Четвериков В.Н., Воробьев Г.Н., Казаков Г.И. и др. Автоматизированные системы управления предприятиями. – М.: Высшая школа, 1979. – 303 с.
4. Алтунин А.Е., Семухин М.В. Модели и алгоритмы принятия решений в нечетких условиях: Монография. − Тюмень: Изд-во Тюменского гос. ун-та, 2000. – 352 с.
5. Черный С.Г., Доровской В.А. Элементы информационной технологии оптических систем идентификации необитаемых подводных аппаратов // Научно-техническая информация. Сер.2. – 2020. – № 3. – С. 11-16; Chernyi S.G., Dorovskoy V.A. Information Technology Elements for Optical Systems of Identification of Autonomous Underwater Vehicles // *Autom. Doc. Math. Linguist.* . – 2020. – Vol. 54, № 2. – P. 73–78. https://doi.org/10.3103/S0005105520020028.
6. Zhilenkov A., Chernyi S. Models and algorithms of the positioning and trajectory stabilisation system with elements of structural analysis for robotic applications // International Journal of Embedded Systems. – 2019. – Vol.11, № 6. – P. 806-814. DOI: 10.1504/ijes.2019.104005

Материал поступил в редакцию 29.03.20.

Сведения об авторах

ЧЕРНЫЙ Сергей Григорьевич – кандидат технических наук, доцент, заведующий кафедрой электрооборудования судов и автоматизации производства Керченского государственного морского технологического университета, г. Керчь; доцент кафедры комплексного обеспечения информационной безопасности, Государственный университет морского и речного флота имени адмирала С.О. Макарова, Санкт-Петербург
e-mail: sergiiblack@gmail.com

ДОРОВСКОЙ Владимир Алексеевич – доктор технических наук, профессор кафедры электрооборудования судов и автоматизации производства Керченского государственного морского технологического университета, г. Керчь
e-mail: dora1943@mail.ru

НОВАК Богда Петрович – аспирант кафедры электрооборудования судов и автоматизации производства Керченского государственного морского технологического университета, г. Керчь (Fly electromechanic m/v Blumenau (Германия)).
e-mail: bogdan.dsl94@gmail.com

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 004.93: 033.212

А.А. Егоров, М.А. Егорова, Т.В. Демидова, Т.Г. Орлова

Статистический метод автоматизации исследования иероглифических надписей Цзягувэнь (甲骨文)

Описан статистический метод исследования иероглифических надписей на черепашьих панцирях (Цзягувэнь / 甲骨文 / Jiǎgǔwén, XIV–XI века до н.э.), являющихся древнейшими образцами китайского искусства и культуры. Приведены некоторые результаты, показывающие возможность автоматизации статистической обработки текстов типа Цзягувэнь, в частности, – подсчет числа иероглифов. Установлено, что поверхности панциря без иероглифов и с иероглифами описываются разными статистиками. Полученные данные позволили решить актуальную обратную задачу распознавания и определить число древних иероглифов (ключей), нанесенных на поверхность панциря черепахи

Ключевые слова: *иероглифическая надпись, Цзягувэнь, поверхность, статистические характеристики и параметры, китайские ключи, автоматизация, математическая лингвистика, компьютерное моделирование*

DOI: 10.36535/0548-0027-2020-08-3

ВВЕДЕНИЕ

Письмена на черепашьих панцирях Цзягувэнь (甲骨文 / Jiǎgǔwén, XIV–XI вв. до н.э.) относятся к древнекитайским образцам искусства и культуры. Они являются уникальными в своем роде. Цзягувэнь – иероглифические надписи, фиксирующие результаты гаданий или предсказаний [1-4]. Эти объекты имеют большую историческую, культурную и научную ценность.

В настоящей работе кратко описан новый статистический метод исследования иероглифических надписей на черепашьих панцирях, основанный на фотометрировании изучаемой поверхности и определении по полученным данным ее статистических характеристик и параметров, как без иероглифов, так и с ними. Предлагаемый метод может стать хорошим дополнением к используемым сейчас традиционным методикам исследования, в основном визуального.

Главные достоинства нашего метода – бесконтактность, информативность и достаточно высокая чувствительность по статистическим параметрам профиля поверхности, что позволяет говорить о его перспективности в исследовании таких древнейших образцов искусства, культуры и науки, как Цзягувэнь.

МЕТОД ИССЛЕДОВАНИЯ. ОБЪЕКТ, РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Статистика и оценки параметров для профиля поверхности без иероглифов и с иероглифами

Используемый нами метод исследования позволяет найти статистические характеристики и параметры профилей исследуемого образца – панциря черепахи. Важнейшей статистической характеристикой неровной поверхности является автокорреляционная функция, характеризующая взаимосвязь между анализируемой функцией и её сдвинутой копией от величины сдвига аргумента [5, 6].

Метод основан на фотометрировании профилей поверхности, при котором осуществляется выборка уровней яркости в дискретном множестве точек с последующим преобразованием значений яркости в цифровой вид [5–8], отражающий форму профиля поверхности. Полученные численные данные подвергаются цифровой обработке с целью определения статистических характеристик и параметров, относящихся к свойствам исследуемой поверхности. В итоге может быть найден вид аппроксимирующих автокорреляционных функций, а также определены соответствующие статистические параметры неровности профилей по-

верхности панциря: среднеквадратичное отклонение и радиус корреляции [5–7, 9]. Этот метод исследования позволил решить важную обратную задачу: идентифицировать иероглифы на исследуемых участках профиля поверхности панциря черепахи и определить их общее число (около 112-117). Подробное описание метода выходит за рамки этой статьи и будет дано в одной из наших следующих работ.

Важнейшими статистическими параметрами неровностей профиля поверхности являются: среднеквадратичное отклонение σ и радиус (интервал) корреляции r. *Среднеквадратичное отклонение* характеризует величину рассеивания значений случайной величины относительно её математического ожидания, т.е. среднего значения. Понятно, что большее значение среднеквадратического отклонения показывает больший разброс в представленном множестве значений, а меньшее, соответственно, демонстрирует, что значения в этом множестве сгруппированы вокруг среднего. *Радиус корреляции* определяет характерный размер «частиц» рассеивателей на поверхности исследуемого объекта. Эти «частицы» не следует понимать буквально – термин используется для указания области, в пределах которой имеет место корреляция, т.е. области «искажения» поверхности. В нашем случае такими характерными «искажениями» являются иероглифические знаки.

Приведем формулы, определяющие вид двух наиболее известных в статистике автокорреляционных функций: экспоненциальной и гауссовой [5, 6, 7, 9].

Экспоненциальная автокорреляционная функция описывается формулой:

$$R(y) = \sigma^2 \exp\left[-|y/r|\right]. \qquad (1)$$

Гауссова автокорреляционная функция имеет вид:

$$R(y) = \sigma^2 \exp\left[-(y/r)^2\right]. \qquad (2)$$

В формулах (1) и (2) переменная y – координата, вдоль которой производится фотометрирование профиля поверхности изучаемого образца. В наших исследованиях сканирование производилось вдоль обеих координат исследуемой плоской поверхности образца (см. рис. 1) – горизонтальной x и вертикальной y, но результаты приведены только для случая фотометрирования поверхности вдоль координаты y.

В статье [4] было проведено оригинальное сравнение некоторых типовых иероглифов и ключей из стандартных СМС-сообщений с иероглифами и ключами, которые можно найти на изображениях панцирей черепах (см. рис. 1). Эти письмена считаются древнейшими текстами и образцами китайской письменности Цзягувэнь (Jiǎgǔwén / 甲骨文 – «письмена на черепашьих панцирях и костях», относятся к XIV–XI вв. до н.э.) [1-4]. Несомненно, древняя письменность Цзягувэнь играла важную роль в распространении и становлении китайского языка.

Нами было выявлено около 10 «символов» (иероглифов и ключей), которые есть в современных СМС сообщениях и на ряде черепашьих панцирей, т.е. около 10% от типовой 100-символьной таблицы черт и ключей китайских иероглифов (пример некоторой условной аналогии с 100-словным списком Сводеша). При этом на черепашьем панцире с рис. 1 нами выявлено 8 иероглифов и ключей (подробнее см. [4]). Проведенное исследование показало сложность визуального исследования Цзягувэнь, поскольку многие образцы плохо сохранились, частично разрушены, а исследуемая поверхность, где находятся иероглифические тексты, не может быть подвергнута исследованиям с помощью контактных методов, например, с помощью профилографов. С другой стороны, многие бесконтактные методы исследования также не могут быть использованы в силу известных ограничений их применения [7].

На рис. 1 показана фотография панциря черепахи[1]. Предметом исследования является поверхность панциря (с нанесенными на неё иероглифами или без них). На рис. 2 приведены вместе экспоненциальная и гауссова автокорреляционные функции, характеризующие, соответственно, профили неровности поверхности вдоль вертикальных линий, отмеченных цифрами «1» и «2» на рис. 1.

Для экспоненциальной функции корреляции (см. рис. 2) определены статистические параметры – среднеквадратичное отклонение σ и радиус корреляции r неровностей профиля поверхности панциря: $r \approx 27$ мм (погрешность определения не более 5%), $\sigma \approx 10$ мкм (погрешность определения не более 3%).

Оценка радиуса корреляции по спаду нормированной автокорреляционной функции в e (≈ 2,72) раз: $r \approx 9$ мм. Оценка радиуса корреляции по второй производной: $r \approx 35$ мм. Среднее по этим двум последним радиусам корреляции получается: (9 мм + 35 мм)/2 = 22 мм. Отличие вышеопределенного $r \approx 27$ мм от среднего значения 22 мм составляет 1,2 раза, т.е. 20%. Эти важнейшие оценки статистических параметров профиля поверхности панциря будут полезны в дальнейших расчетах, поскольку они содержат значимую информацию и позволяют численно охарактеризовать поверхность панциря без иероглифов.

Полученные данные позволяют сделать выводы для профиля поверхности в центре панциря, где нет иероглифов:

а) размер неровностей в плоскости (латеральный) исследуемого объекта изменяется примерно в интервале от 1–2 мм (близко к среднему значению шага неровностей S исходного профиля: S = 3 мм) до 20–30 мм;

б) размер неровностей по высоте h (вертикальный) изменяется примерно в интервале от 5 мкм (шероховатость) до 2 мм (уединенные неровности типа канавок).

Для гауссовой функции корреляции (см. рис. 2) определены соответствующие статистические параметры: среднеквадратичное отклонение σ и радиус корреляции r неровностей профиля поверхности панциря: $\sigma \approx 6{,}9$ мкм (погрешность определения не более 3%), $r \approx 2{,}9$ мм (погрешность определения не более 5%).

[1] Примерные размеры исследуемой поверхности панциря: ширина 160 мм, высота 190 мм.

Рис. 1. Пример Цзягувэнь (甲骨文 – «письмена на черепашьих панцирях») – иероглифические надписи, фиксирующие результаты гаданий или предсказаний. Вертикальные линии в центре:
1 – соответствует центральной части исследуемой поверхности панциря, где нет иероглифов;
2 – соответствует исследуемой части поверхности панциря, где есть столбец из иероглифов (древних китайских ключей)

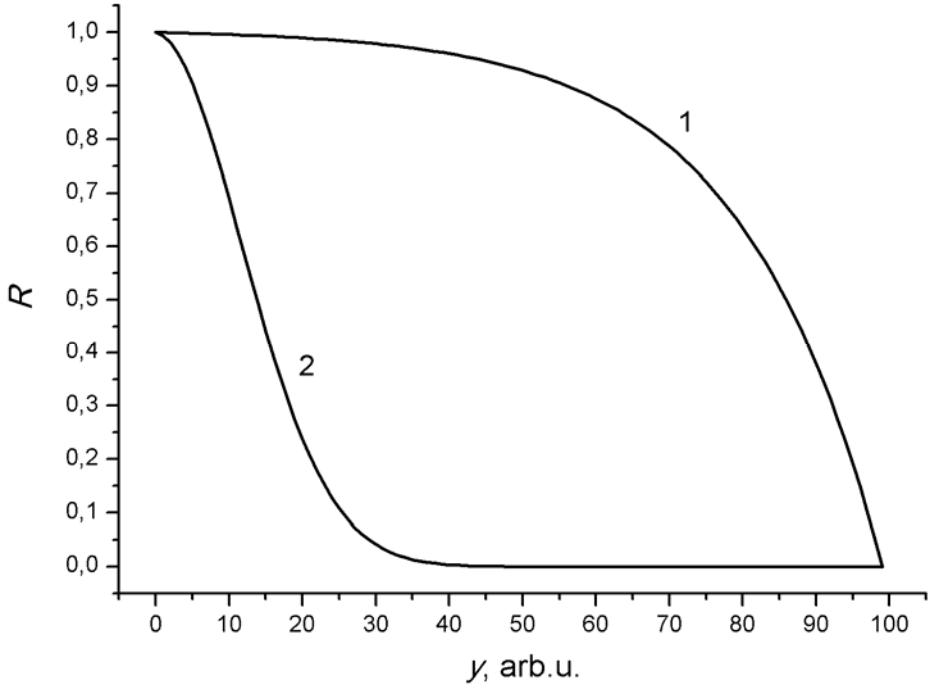

Рис. 2. Экспоненциальная (1) и гауссова (2) нормированные функции корреляции.
Одно деление по оси *y* равно примерно 100 мкм

Отметим, что радиус корреляции профиля поверхности с иероглифами уменьшился в 9,3 раза, а среднеквадратичная высота – в 1,4 раза соответственно по сравнению с параметрами, полученными для профиля поверхности без иероглифов.

Оценка радиуса корреляции по спаду нормированной автокорреляционной функции в e раз: $r \approx 1{,}6$ мм. Оценка радиуса корреляции по второй производной: $r \approx 1{,}7$ мм. Среднее по этим двум последним радиусам корреляции получается: (1,6 мм + 1,7 мм)/2 = 1,65 мм. Здесь радиус корреляции профиля поверхности с иероглифами уменьшился в 16 раз по сравнению с параметрами, полученными для профиля поверхности без иероглифов.

Сделаем некоторые предварительные выводы из полученных результатов:

- свободная от иероглифов поверхность панциря черепахи лучше описывается экспоненциальной автокорреляционной функцией;
- поверхность панциря с иероглифами лучше описывается гауссовой автокорреляционной функцией, при этом происходит одновременное уменьшение, как радиуса корреляции, так и среднеквадратичной высоты неровностей профиля.

Оценка числа иероглифов (ключей)

Для определения числа ключей (древних иероглифов) $\Im$ разделим длину интервала сканирования L профиля поверхности на радиусы (интервалы) корреляции r, характеризующие статистику этих профилей. Действительно, *радиус корреляции*, как было сказано выше, определяет характерный размер «частиц» рассеивателей на поверхности исследуемого объекта. Этот термин используется для обозначения области поверхности «искаженной» («возмущенной») неровностью.

В рассматриваемом случае это искажение поверхности создают, в основном, ключи, поэтому предложенная нами достаточно простая оценка величины $\Im$ вполне справедлива. Примерная средняя длина линии сканирования профиля поверхности панциря: L = 18 см. При этом в расчетах использовались как участки меньшей, так и большей длины для нахождения средних оценок.

Какой радиус корреляции надо брать в расчетах? Если руководствоваться формальными соображениями, то нужен радиус корреляции, определенный для профилей поверхности панциря с иероглифами. Однако надо также учесть определенное влияние свободной от иероглифов, но неровной поверхности, поскольку ее роль в формировании итогового профиля тоже присутствует (она является некоторой случайной помехой, на фоне которой регистрируется информация – иероглифическая надпись).

В качестве начального шага можно предложить следующее: там, где визуально наблюдается достаточно плотное расположение ключей, вклад свободной от иероглифов поверхности будем учитывать в меньшей степени. Поскольку мы используем статистический подход, то для простоты найдем среднее значение радиуса корреляции по двум оценкам радиусов корреляции: для профилей поверхности панциря без ключей (иероглифов) и с ключами[2].

Теперь можем рассчитать число ключей (древних иероглифов) на линии сканирования, содержащей иероглифы, по формуле:

$$\Im = L/r\,. \qquad (3)$$

Найдем среднее арифметическое значение радиуса корреляции r с учетом вышесказанного:

$$r = [(2{,}9 + 27)/2] \text{ мм} = 14{,}95 \approx 15 \text{ мм}.$$

Определим примерное число ключей в соответствии с формулой (3): $\Im$ = 180/15 = 12. Итак, примерное число ключей (древних иероглифов) $\Im$ = 12. По нашим оценкам на вертикальной линии сканирования находится 12 иероглифов (см. рис. 1: вертикальная линия, отмеченная цифрой «2»). Таким образом, получено точное значение ключей.

Если при расчетах по формуле (3) взять не среднее значение L, а размеры использованных линий сканирования панциря по вертикали в разных сечениях (при смещении по горизонтали), то получим, что $\Im$ может варьироваться примерно от 11 до 13.

Как видно из полученных данных, погрешность нахождения числа ключей $\Im$ будет составлять примерно 9%. На этом этапе исследований такую погрешность можно считать вполне удовлетворительной [10]. Тем более, что одной из наших целей была демонстрация возможностей предложенного в работе метода исследования, в первую очередь, с точки зрения возможности определения числа ключей. Как видно, эта цель была успешно достигнута.

АНАЛИЗ РЕЗУЛЬТАТОВ. ВЫВОДЫ

Нами было отмечено, что свободная от иероглифов поверхность лучше описывается экспоненциальной автокорреляционной функцией, а поверхность панциря с иероглифами – гауссовой автокорреляционной функцией. При этом происходит одновременное уменьшение, как радиуса корреляции r, так и среднеквадратичной высоты σ неровностей профиля.

Можно дать следующее объяснение такого поведения неровностей профиля поверхности. Поверхность, особенно в месте расположения иероглифов, подвергалась дополнительной обработке, например полировке, для удобства их нанесения, что привело к некоторому изменению статистики неровностей, и это было обнаружено в процессе исследования.

Уменьшение радиуса корреляции и среднеквадратичной высоты неровностей профиля поверхности панциря с иероглифами можно объяснить некоторым упорядочиванием структуры поверхности, а именно – появлением определенных кластеров (групп ключей, объединенных в сложные иероглифы). Эти кластеры

[2] Для этого заключения выполнена проверка гипотез об *аддитивном* и *мультипликативном* характере формирования статистики поверхности в месте нанесения иероглифов. Было использовано предварительное разложение в ряд формул (1) и (2) с оставлением двух первых членов для получения приближенной результирующей автокорреляционной функции.

выстроены преимущественно вдоль определенных направлений, главным образом вдоль вертикали. Что приводит к некоторому снижению рассеяния света на случайных нерегулярностях структуры поверхности панциря, несмотря на то, что в профиле появились отдельные элементы типа уединенных канавок, чья неровность возросла по сравнению с исходным профилем поверхности. В итоге этот процесс и отображается в данных фотометрирования поверхности панциря вдоль выбранных направлений.

Сделаем выводы о полученных статистических параметрах неровностей поверхности панциря черепахи:

- по порядку величины все радиусы корреляции (для одних и тех же автокорреляционных функций) находятся в хорошем соответствии друг с другом;
- отличие радиусов корреляции (для одних и тех же автокорреляционных функций) зависит, в первую очередь, от метода их определения;
- отличие среднеквадратичных высот обусловлено, в основном, наличием или отсутствием иероглифов.

Отметим, что использование в перспективе цифровой обработки больших массивов экспериментальных данных, например, с помощью таких известных методов обработки сигналов как стандартное быстрое преобразование Фурье или вэйвлет преобразование, может позволить получить новые интересные результаты. Однако это требует дополнительных затрат времени и ресурсов.

Описанный в статье метод исследования основан на фотометрировании профилей поверхности с последующим преобразованием данных в цифровой вид (т.е. дискретное множество точек интенсивности света, рассеянного поверхностью исследуемого объекта), отражающий форму профиля поверхности. Расчеты производились на компьютере, оснащенном процессором *Intel Pentium* 4. Полное время обработки одной диаграммы рассеяния света (около 120 точек) составляет не менее 20-30 минут. При вариации параметров подгоночных автокорреляционных функций с целью получения улучшенных аппроксимаций искомых статистических характеристик затраты времени возрастают примерно на порядок. После определенного совершенствования метод позволит автоматизировать процесс статистической обработки древнейших иероглифических надписей. Очевидно, что это приведет к существенному снижению затрат времени на проведение всего комплекса исследований.

Полученные нами результаты позволяют утверждать, что описанный в работе способ может быть использован как перспективный статистический метод исследования иероглифических надписей и определения числа ключей (древних иероглифов) Цзягувэнь (*Jiǎgǔwén* / 甲骨文, XIV–XI вв. до н.э.).

По нашему мнению полученные результаты будут полезны в математической лингвистике, лексикостатистике и социолингвистике, особенно при статистическом анализе подобных иероглифических надписей. Особый интерес может представлять изучение связи между информацией, содержащейся в Цзягувэнь, и социально-культурными условиями развития языка и общества. Не только в тот период, когда Цзягувэнь создавались и имели достаточно широкое применение, но и в последующие времена, поскольку многие из сохранившихся образцов использовались, например, теми же иероглифистами-писцами как для подражания старым стилям написания иероглифических знаков, так и для их улучшения и последующего совершенствования. Именно этот процесс и способствовал тому, что древнейшие иероглифические надписи со временем развились во множество, в основном, не пиктографических знаков, включающих все основные типы китайских иероглифов, которые используются и сейчас.

Важно подчеркнуть, что в итоге внешняя форма китайских иероглифов в III в. до н.э. и позднее претерпела сильные изменения, вызванные, главным образом, изменением техники письма. В I в. н.э. впервые появился тот стиль письма, который сохранился до настоящего времени («*кайшу*» – образцовое или уставное письмо *kǎishū* / 楷书). Он отличается от существовавших ранее древних форм тем, что все знаки в этом стиле строятся из небольшого числа одних и тех же основных графических элементов (горизонтальная черта, вертикальная черта, крюк, точка и т.д.), которые окончательно потеряли всякое сходство с рисунками, от которых произошли.

ЗАКЛЮЧЕНИЕ

В настоящей работе с помощью статистического метода исследованы письмена на черепашьих панцирях (Цзягувэнь) и продемонстрирована возможность достаточно точного определения по цифровым данным фотометрирования поверхности панциря числа ключей (древних иероглифов). Погрешность метода на данном этапе исследований не превышает в среднем 10%. Основные достоинства метода: бесконтактность, информативность, не деструктивность, достаточно высокая разрешающая способность.

Статистический метод позволяет автоматизировать процесс статистической обработки древнекитайских текстов, что дает возможность сделать вывод о его перспективности в исследовании образцов Цзягувэнь, которые часто существуют в единичных экземплярах и нередко имеют плохую сохранность, когда зрительная идентификация иероглифических знаков затруднена, а отдельные элементы надписей плохо видны или почти не различимы.

Определенная простота и наглядность реализации (от постановки задачи до полученных результатов) позволяют применять описанный метод в междисциплинарных исследованиях, в которых участвуют специалисты из разных предметных областей знания, например, гуманитарных (лингвистика, история, педагогика и др.).

* * *

Публикация подготовлена при частичной поддержке Программы РУДН «5-100». Авторы выражают благодарность коллегам за доброжелательные отзывы и сделанные во время обсуждения замечания, которые помогли доработать статью наилучшим образом.

СПИСОК ЛИТЕРАТУРЫ

1. Keightley D.N. Sources of Shang history: the oracle-bone inscriptions of Bronze Age China. – London: Berkeley, 1985.
2. Oracle bone. – URL: https://en.wikipedia.org/wiki/Oracle_bone
3. Яхонтов С.Е. Древне-Китайский язык. – М.: Наука, 1965.
4. Егорова М.А., Егоров А.А., Соловьева Т.М. Моделирование распространения и видоизменения письменности в пра-Китайских языковых сообществах // Научно-техническая информация. Сер. 2. – 2020. – № 4. – С. 22–35; Egorova M.A., Egorov A.A., Solovieva T.V. Modeling the distribution and modification of writing in proto-chinese language communities // Automatic Documentation and Mathematical Linguistics. – 2020. – Vol. 54, №. 2. – P. 92–104.
5. Левин Б.Р. Теоретические основы статистической радиотехники. – М.: Сов. Радио, 1966.
6. Басс Ф.Г., Фукс И.М. Рассеяние волн на статистически неровной поверхности. – М.: Наука, 1972.
7. Егоров А.А. Методы исследования субмикронных объектов // Сб. научн. трудов РУДН. – М.: РУДН, 2000. – С. 366-371.
8. Pratt W.K. Digital image processing. – NY.: Wiley, 1978.
9. Born M., Wolf E. Principles of optics. – NY.: Pergamon, 1986.
10. Егорова М.А., Егоров А.А. Прародина народов носителей праиндоевропейского языка: математические модели для исследования лингвистической информации // Научно-техническая информация. Сер. 2. – 2019. – № 5. – С. 27–37; Egorova M.A., Egorov A.A. The ancestral homeland of the carriers of the proto-Indo-European language: mathematical models for the study of linguistic information // Automatic Documentation and Mathematical Linguistics. – 2019. – Vol. 53, № 3. – P. 127–137.

Материал поступил в редакцию 24.02.2020

Сведения об авторах

ЕГОРОВ Александр Алексеевич – доктор физико-математических наук, внештатный профессор-консультант, Российский университет дружбы народов (РУДН), Москва.
e-mail: alexandr_egorov@mail.ru

ЕГОРОВА Майя Александровна – кандидат политических наук, доцент кафедры иностранных языков факультета гуманитарных и социальных наук РУДН, Москва
e-mail: Mey1@list.ru

ДЕМИДОВА Татьяна Васильевна – кандидат филологических наук, доцент кафедры иностранных языков факультета гуманитарных и социальных наук РУДН.
e-mail: tademidova52@mail.ru

ОРЛОВА Татьяна Геннадьевна – кандидат филологических наук, доцент кафедры иностранных языков факультета гуманитарных и социальных наук РУДН.
e-mail: orlova_tg@rudn.ru

УДК 81’322.2

В.А. Яцко

Критерии классификации лингвистических технологий*

Предлагаются две группы критериев классификации современных лингвистических технологий: семиотические и технологические. Семиотические критерии включают семантические, синтаксические и прагматические признаки технологий. Семантические соотносятся с обрабатываемой единицей языка; синтаксические – с последовательностью выполнения программ; прагматические – с различными категориями пользователей. Технологические критерии предусматривают идентификацию технологий по видам входных данных, выходных данных, методам и алгоритмам обработки текстовых документов. Рассматриваются некоторые перспективы развития предметной области, связанной с автоматической обработкой и анализом текстов на естественном языке.

Ключевые слова: *лингвистические технологии, критерии классификации, семиотические критерии, технологические критерии, лингвистическое программное обеспечение, виды*

ВВЕДЕНИЕ

Характерная черта развития современных информационных технологий – широкое использование лингвистического программного обеспечения. Отправляя запросы в информационно-поисковые системы, отдавая голосовые команды электронным устройствам, миллионы пользователей во всём мире не подозревают, что в основе функционирования этих устройств лежат сложные процессы обработки текстов на естественном языке, реализация которых стала возможной в результате развития лингвистических технологий. В связи с этим приобретает актуальность задача системного представления таких технологий, что возможно на основе выделения критериев их идентификации и создания соответствующей таксономии. Решение этой задачи имеет непосредственное значение для: 1) определения конфигурационных характеристик создаваемого программного обеспечения. Разработчики должны иметь представление о существующих видах лингвистических программ, алгоритмов, возможностях их сочетания и последовательности применения; 2) обучения и подготовки кадров. Студенты, изучающие лингвистические технологии, должны получать системные знания, отражающие особенности функционирования и применения различных видов программного обеспечения. На основе классификационной схемы следует проводить структурирование учебных курсов, планов и стандартов; 3) развития научных исследований и разработок. Исследования должны быть основаны на развитии предметной области, специфике применения основных методов исследования и обработки текстовых документов.

Адекватная таксономия лингвистических технологий позволяет оценивать состояние и перспективы развития предметной области; 4) организации и стандартизации как научных, так и прикладных разработок, включая стандартизацию терминологии, модификацию номенклатуры научных специальностей, организацию повышения квалификации и переподготовки научно-педагогических, исследовательских, инженерных кадров.

В.И. Вернадский отмечал, что ядро науки составляют математические, логические науки, а также научные факты в их системе, классификации, лежащие в основе научного аппарата [1, с. 428].

Следует констатировать, что в настоящее время отсутствует общепринятая классификационная схема, отражающая структуру и состав предметной области, связанной с автоматической обработкой и анализом документов на естественном языке, что, в конечном счёте, отрицательно сказывается на её развитии. Предлагаемая статья представляет собой попытку восполнить данный пробел, сформулировать подходы к определению критериев идентификации и классификации существующих лингвистических технологий.

СЕМИОТИЧЕСКИЕ КРИТЕРИИ КЛАССИФИКАЦИИ

Выделение семантических, синтаксических и прагматических признаков лингвистических технологий предусматривают семиотические критерии, которые предполагают соотнесение с обрабатываемыми единицами текста. Ранее [2, 3] мы предлагали классификацию с учётом уровня системы языка, к которому относятся

* Исследование поддержано грантом РФФИ № 20-07-00124.

обрабатываемые единицы. В соответствии с этим критерием можно выделить технологии, работающие на графемном, фонетическом, морфологическом, лексическом, синтаксическом, дискурсивном уровнях.

На графемном уровне используются системы оптического распознавания символов, которые позволяют с помощью сканеров и специализированного программного обеспечения оцифровывать бумажные носители информации [4]. Системы этого типа широко применяются с целью создания электронных библиотек и текстовых корпусов.

На фонетическом уровне технологии распознавания звуков обеспечивают преобразование устной речи в письменный текст и лежат в основе различных программ, применяющихся для распознавания речи.

На морфологическом уровне разрабатываются технологии распознавания стемм, лемм, суффиксов и окончаний, благодаря которым идентифицируются одинаковые морфологические конструкции, что необходимо для адекватного взвешивания терминов и распознавания частей речи.

На лексическом уровне применяются технологии лексической декомпозиции и аннотирования, позволяющие получать списки токенов и приписывать им условные символы (теги), которые указывают на их семантические, морфологические, синтаксические и прагматические характеристики.

На синтаксическом уровне программы синтаксической декомпозиции позволяют получать на выходе списки предложений, словосочетаний, н-грам, клауз. Синтаксические парсеры генерируют графы, представляющие иерархическую структуру предложений.

На дискурсивном уровне разрабатываются технологии клаузальной декомпозиции и сегментации текста, которые позволяют проводить разрешение анафоры и кореференции, определять и моделировать тематическую структуру текста с помощью экспертных систем и систем искусственного интеллекта [5]. В табл. 1 представлены виды лингвистических технологий, идентифицируемые по уровням системы языка.

Синтаксический критерий классификации предполагает установление определённой последовательности выполнения программ, указанных в табл. 1. Полагаем, что можно выделить прямую последовательность выполнения алгоритмов, обратную последовательность и их синхронное выполнение. Под прямой последовательностью имеется в виду выполнение сначала алгоритма более низкого уровня, а затем – более высокого. При обратной последовательности сначала выполняется алгоритм более высокого уровня, а затем – более низкого. Лексическая декомпозиция обычно предшествует синтаксической: сначала текст разбивается на токены, а затем – на предложения. Однако возможна и обратная последовательность: сначала распознаются предложения, а затем они разбиваются на токены [2]. В случае, если лексическая и синтаксическая декомпозиции не связаны, они могут выполняться одновременно, синхронно. В сложных системах обычно выделяется модуль предварительной обработки и модуль основной обработки входного документа[1].

Таблица 1

Лингвистические технологии и уровни системы языка

Технологии		Обрабатываемая единица	Уровни системы языка
Алгоритм	**Программа**		
Распознавание символов	OCR	Символ	Графемный
Распознавание звуков	Распознавание речи	Звук	Фонетический
Стемминг	Стеммер	Стемма	Морфологический
Лемматизация	Лематайзер	Лемма	
Токенизация (Лексическая декомпозиция)	Токенайзер	Токен	Лексический
Аннотация	Теггер	Тег	
Взвешивание терминов	Фильтр взвешивания	Стемма/лемма/токен/ н-грам/предложение/текст	Все уровни
Разбивка на н-грамы	Н-грам сплитер	Н-грам	Синтаксический
Фразовая декомпозиция	Чанкер	Фраза	
Парсинг	Парсер	Предложение	
Синтаксическая декомпозиция	Синтаксический сплитер	Предложение	
	Клауз-сплитер	Клауза	
Сегментация	Дискурсивный сплитер	Отрезок текста	Дискурсивный
Разрешение анафоры и кореференции	Резольвер	Отрезок текста	
Моделирование текста	Моделятор	Отрезок текста/текст	

[1] Входные тексты/документы загружаются пользователем в систему/приложение. Выходные тексты/данные выдаются системой/приложением пользователю и являются результатом обработки входного текста (документа)

Сначала выполняются алгоритмы предварительной обработки, а затем – основной. К первым обычно относятся алгоритмы лексической и синтаксической декомпозиции [6]. В зависимости от особенностей лингвистических систем один и тот же алгоритм может применятся и на этапе предварительной обработки, и на этапе основной обработки. В системах автоматической классификации частеречное аннотирование выполняется на этапе предварительной обработки, в то время как в корпусной лингвистике оно выступает в качестве основного алгоритма.

Существенной характеристикой лингвистических технологий являются пользователи, для которых предназначены соответствующие системы и приложения. По этому критерию можно выделить глобальные, специальные и специализированные технологии и системы.

Глобальные технологии предназначены для любых пользователей, независимо от их возраста, социального положения, уровня образования. Типичный пример – документальные информационно-поисковые системы индексного типа такие, как Гугл, Яндекс, принимающие на входе запрос на естественном языке и выдающие на выходе ссылки на веб-ресурсы.

Пользователями специальных технологий являются специалисты в конкретной предметной области.

Технологии корпусной лингвистики используются для поддержки лингвистических исследований, а основными пользователями текстовых корпусов различного типа являются специалисты-лингвисты.

Полагаем, что созданные к настоящему времени текстовые корпуса можно разделить на два вида: предназначенные для исследования языка и предназначенные для обучения языку. Пользователи текстовых корпусов, предназначенных для обучения языку, выступают преподаватели, использующие их в лингво-дидактических целях. Дидактические корпуса (*learner corpora*) представляют собой коллекции текстов (устных и письменных), которые создаются студентами, изучающими иностранные языки. Корпуса этого типа аннотируются как тегами частей речи, так и тегами, указывающими на ошибки разных видов (орфографические, лексические, грамматические), допущенные обучаемыми. Сопоставление распределения лингвистических единиц в исследуемых текстах с их распределением в эталонных текстах, создаваемых носителями языка, позволяет выявить наиболее типичные ошибки и отклонения от стилистической нормы, допускаемые при изучении иностранного языка [7]. Так, анализ, проведённый в [8], позволил выявить, что студенты, изучающие английский язык как иностранный, намного чаще, чем носители языка, используют вводные слова и выражения. Очевидно, что это должно учитываться при обучении английскому языку.

Пользователями исследовательских корпусов являются специалисты в различных областях языкознания. Исследовательские корпуса можно разделить на два основных: национальные корпуса представляют язык на определенном этапе его развития и используются для поддержки синхронных исследований; исторические корпуса применяют специалисты в области сравнительно-исторического языкознания.

По критерию параллельности можно выделить моноязычные и многоязычные корпуса. Параллельные трансляционные корпуса используются специалистами в области машинного перевода и обычно содержат эквивалентные тексты на разных языках. Сопоставительные корпуса содержат тексты предопределённого жанра/жанров и используются для классификации текстовых документов специалистами в области компьютерной лингвистики [9]. Корпуса снабжаются поисковыми системами фактографического типа, которые предусматривают формулирование запросов на некотором искусственном информационно-поисковом языке. Для адекватной работы пользователи должны изучить соответствующий язык, его семантику (значение символов) и синтаксис. К специальным относятся и приложения, предназначенные для статистического анализа текстовых документов – конкордансы, которые выдают информацию о распределении единиц входного текста, загруженного пользователем [10]. Конкордансы генерируют три основных типа выходных данных – контекст ключевого слова, заданного пользователем; список слов с указанием частотностей; список коллокаций, с указанием частотностей. В этом случае обязательно указывается количество уникальных слов и общее количество токенов. Некоторые конкордансы предоставляют пользователю дополнительные функции распознавания и статистического распределения *н*-грам и кластеров, генерации ключевых терминов текста на основе вероятностно-статистического анализа, обработки аннотированных текстов. Эти данные являются исходными для дальнейшей автоматической обработки текстов специалистами в области компьютерной лингвистики.

Специализированные технологии предназначены для поддержки принятия решений их пользователями и предполагают использование интеллектуального анализа, позволяющего выявить информацию, имплицитно присутствующую в тексте [11]. К такой информации могут относиться, например, числовые коэффициенты, указывающие на интенсивность отрицательной или положительной оценки какого-то продукта покупателями. Основываясь на такой информации, менеджмент фирмы, производящей данный продукт, может принять решение о его модификации, продвижении или снятия с продажи.

Отличие специализированных технологий от интеллектуального анализа, проводимого с целью классификации, состоит в том, что они являются онтологически и дискурсивно-ориентированными. Для поддержки функционирования таких систем создаются лингвистические онтологии – многоуровневые таксономии, отражающие содержание соответствующей предметной области, а также формальные грамматики, в которых описываются правила распознавания единиц онтологии в речи (связном тексте). Наряду с интеллектуальным анализом мнений пользователей к этой предметной области относятся системы распознавания экстремистского или террористического контента, а также системы обмена опытом, разрабатываемые в разных предметных областях.

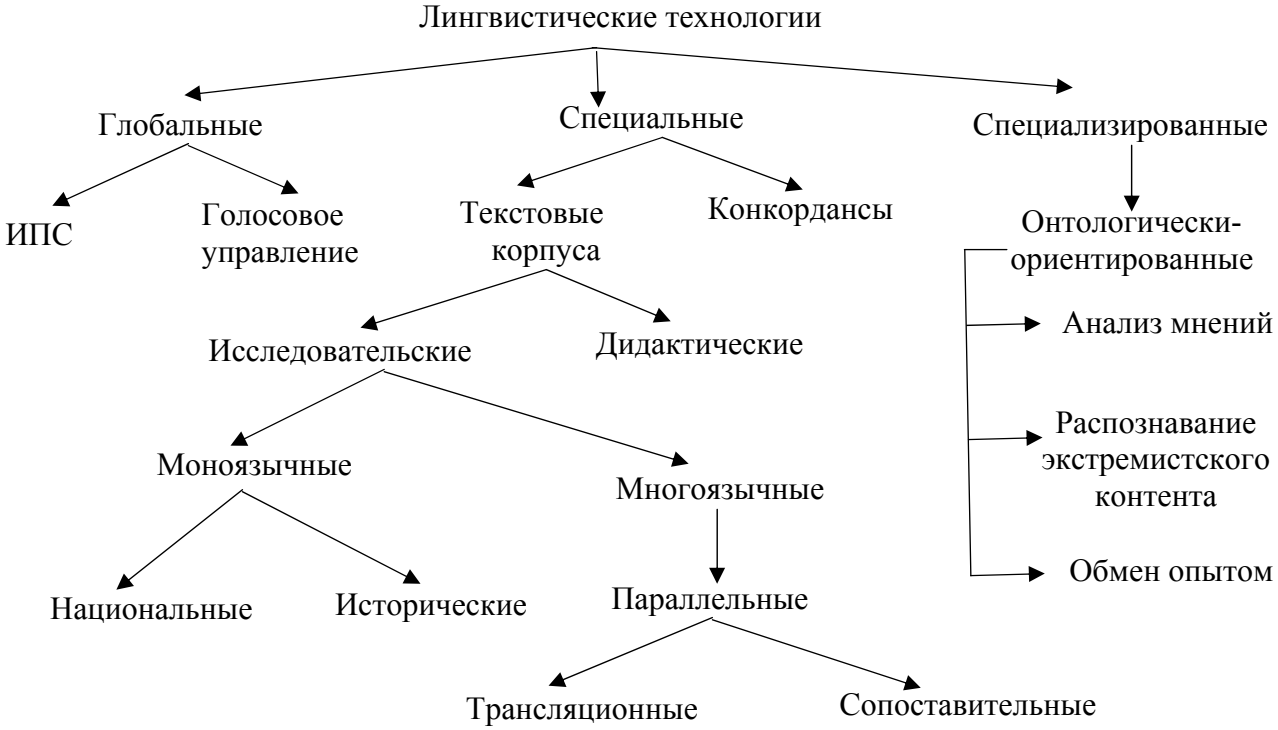

Рис. 1. Виды лингвистических технологий, выделяемые по категориям пользователей

В медицине системы такого типа выдают врачу диагнозы, которые ставились на основе определённых симптомов, применявшиеся методы лечения, назначавшиеся медикаментозные средства. Анализ больших массивов данных, зафиксированных в амбулаторных картах, может выявить ранее неизвестные зависимости, например, между течением болезни и индивидуальными особенностями пациентов [12]. При разработке систем такого типа применяются методы машинного обучения и различные вероятностно-статистические метрики. На рис. 1 представлена классификация лингвистических технологий по прагматическому критерию.

ТЕХНОЛОГИЧЕСКИЕ КРИТЕРИИ КЛАССИФИКАЦИИ

Лингвистические технологии могут идентифицироваться в зависимости от типа входного документа (т. е. текстового файла, загружаемого пользователем в систему или приложение); типа выходных данных, выдаваемых пользователю; типа алгоритмов, применяемых для обработки входных текстов.

В зависимости типа носителя текста общепринятым является деление технологий, реализуемых в лингвистическом программном обеспечении, на технологии обработки текстов и речи. В первом случае на входе программы или системы – письменный текст в виде символов некоторого естественного языка, во втором – устная речь, представленная высказываниями в звуковой форме.

Технологии обработки речи включают два основных направления: распознавание речи и синтез речи. Распознавание речи широко используется в автоответчиках, системах голосового управления, в электронных устройствах в режиме "*hands-free*", а также для идентификации характеристик личности, таких как возраст, пол, эмоции [13] и даже степень алкогольного или наркотического опьянения [14]. Синтез речи предусматривает конвертацию текста в речь и используется в приложениях, предназначенных для пользователей с нарушениями зрения и расстройствами, не позволяющими читать письменный текст. Озвучка книг применяется и обычными пользователями с целью релаксации или изучения иностранных языков. В современных вопросно-ответных системах применяются как распознавание, так и синтез речи – это голосовые помощники, разработанные основными ИТ-корпорациями. *Google Assistant*, *Cortana*, *Alexa*, *Siri* – наиболее востребованные виртуальные голосовые помощники, с помощью которых миллионы пользователей осуществляют управление электронными устройствами, отдавая команды системам умного дома, выполняя поиск в Интернете [15]. *Google Assistant* преобразует устную команду или вопрос пользователя в письменный запрос, который выполняется поисковой системой *Google*; результаты поиска озвучиваются с помощью системы синтеза речи.

Входные тексты можно классифицировать не только по типу носителей информации, но и по другим признакам. Одна из проблем обработки документов на естественном языке состоит в том, что они содержат неструктурированные данные. Соответственно возникает необходимость их предварительной обработки [16], которая может существенно осложниться, если тексты не редактируются в процессе их подготовки и публикации. Тексты блогов, чатов, форумов, размещаемые в Интернете никак не вычитываются и не

редактируются, в отличие, например, от научных статей и монографий. Качество орфографии и грамматики таких текстов напрямую зависит от уровня образования и языковой подготовки их авторов. В то же время тексты этого типа представляют собой ценный источник информации для современных лингвистических технологий, таких как интеллектуальный анализ мнений пользователей, обнаружение экстремистского и террористического контента.

При обработке и анализе исходных текстов появляются две проблемы. Во-первых, эти тексты относятся к диалогической речи, как форме коммуникации, и, чтобы понять смысл сообщения автора, нужно проанализировать связи этого сообщения с сообщениями других участников диалога, что требует разработки сложных алгоритмов анализа связной диалогической речи. Во-вторых, эти тексты могут содержать орфографические ошибки, лексические и грамматические варианты, типичные для такого типа речи. Соответственно, должны создаваться специальные словари и дополнительные правила предварительной обработки текстовых документов [17]. Обработка чатов представляет особые трудности, так как эти тексты продуцируются спонтанно, в то время как тексты блогов и форумов могут готовиться предварительно. Таким образом лингвистические технологии могут быть разделены на предназначенные для обработки редактируемых и нередактируемых документов, при этом последние делятся на технологии обработки подготовленных и спонтанных текстов. На рис. 2 представлена классификация лингвистических технологий по типу входных текстов.

Выходные данные также могут служить критерием дифференциации лингвистических технологий. Применение технологий поиска и реферирования позволяет получать вторичные документы, отражающие содержание первичных текстов. В первом случае вторичный документ представляет собой ссылку на веб-ресурс с его описанием, во втором – он представляет содержание первичного документа в сокращённом виде.

Системы машинного перевода генерируют эквивалентный текст на другом языке/языках, который должен адекватно передавать полное содержание оригинала. Технологическое развитие существенно повлияло на машинный перевод, в котором в настоящее время применяются разнообразные технологии. В устном (синхронном и последовательном) переводе всё ещё доминируют ручные методы, в то время как в письменном применяются компьютерно-опосредованные, полуавтоматические и автоматизированные технологии. Компьютерно-опосредованный перевод основывается на технологиях переводческой памяти, которые позволяют добавлять использованные раннее варианты перевода, существенно экономя время человека – переводчика. Полуавтоматические технологии предполагают ручное пред-редактирование текста оригинала или пост-редактирование выходного текста. Особенностью современных методов автоматического перевода является использование технологий дополненной реальности, когда перевод выполняется при наведении фото/видео камеры на текст [18].

В результате автоматической классификации документов определяется имя класса, к которому относится входной текст [19]. Это может быть жанр (жанровая классификация), тематическая категория (категоризация), имя автора текста, если оно неизвестно – авторская атрибуция, распознавание спама или плагиата. Ещё это может быть некоторая модель, отражающая имплицитное содержание текста, например, модель его тематической структуры. Сюда же можно отнести технологии анализа мнений пользователей о коммерческих продуктах, технологии распознавания текстов экстремистского или террористического содержания, технологии обмена опытом в конкретной предметной области. Применение этих технологий не связано с решением классификационных задач и предполагает достаточно сложный анализ связного текста, соотносящегося с определённой предметной областью (дискурса), см. рис.1.

Таким образом, можно выделить технологии, которые полностью воспроизводят содержание оригинального текстового документа и технологии, предоставляющие пользователю некоторую метаинформацию о нём. Мета-информационные лингвистические технологии могут быть разделены на те, которые воспроизводят содержание оригинала и те, что выдают информацию, имплицитно содержащуюся в тексте. Первые можно назвать презентативными, вторые – интеллектуальными. В зарубежных источниках для обозначения технологий, позволяющих раскрыть имплицитную текстовую информацию, используется термин *text mining* [20] (см. рис.1).

На рис. 3 представлена классификация лингвистических технологий по типу выходных данных.

Алгоритмы, применяемые при разработке и создании лингвистического программного обеспечения, также могут выступать в качестве критериев классификации. По этим критериям можно идентифицировать статистические технологии и технологии, основанные на правилах. И те, и другие используются для распознавания единиц текста и/или их характеристик. Наиболее типичный пример – стохастические теггеры и теггеры на правилах (регуляторные теггеры). Последние функционируют на основе морфологических, лексических, синтаксических и контекстуальных правил. Стохастические теггеры вычисляют для каждого слова вероятность его использования с тем или иным тегом и выбирают тег с наибольшим вероятностным коэффициентом [21]. Применение как вероятностно-статистических, так и чисто статистических алгоритмов предполагает приписывание весовых коэффициентов входным лингвистическим единицам, что предопределяет широкое использование различных методов взвешивания терминов для выявления наиболее статистически значимых из них, определяющих специфику определенного текста, что особенно важно для решения классификационных задач. В настоящее время наиболее распространены такие вероятностные технологии, как марковские и байесовские модели, хи-квадрат, отношение шансов, опорные машинные вектора. В некоторых предметных областях вероятностные технологии стали использоваться по умолчанию. Например, в системах фильтрации спама используется наивный байесов-

ский классификатор [22]. К наиболее распространённым чисто статистическим методам относится TF*IDF, широко применяемый с целью классификации единиц текста [23].

К вероятностно-статистическим методам обычно прибегают в том случае, если обрабатываются единицы текста синтаксического или дискурсивного уровня, в то время как чисто статистические методы используются для обработки лексических единиц по технологии списка слов (*bag of words*). Большая сложность вероятностно-статистических методов обусловливает широкое применение нейросетей, которые требуют машинного обучения на некоторых эталонных текстах. В настоящее время интенсивно развиваются и широко применяются технологии глубокого обучения, предполагающие многоуровневый анализ структуры текста [24]. На первом уровне обычно создаётся объектная модель, включающая списки лингвистических объектов и/или параметров входного текста. На втором уровне устанавливаются некоторые особенности распределения объектов/параметров на основе вероятностно-статистического анализа.

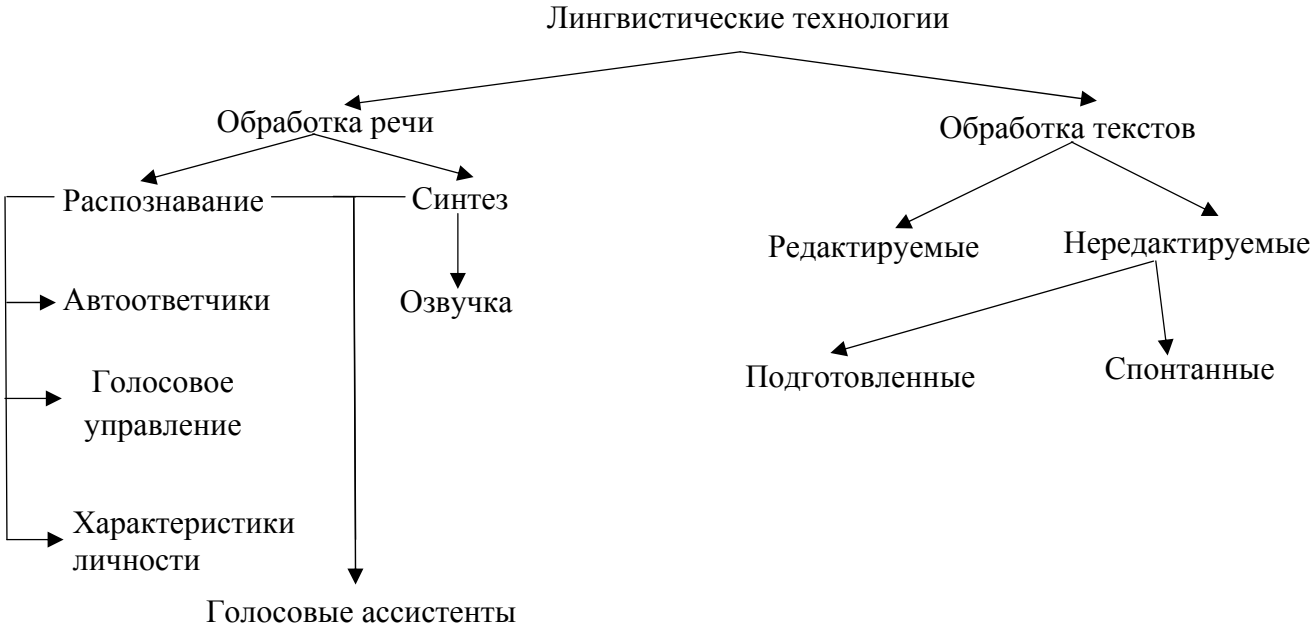

Рис. 2. Виды лингвистических технологий, выделяемые по типу входных данных

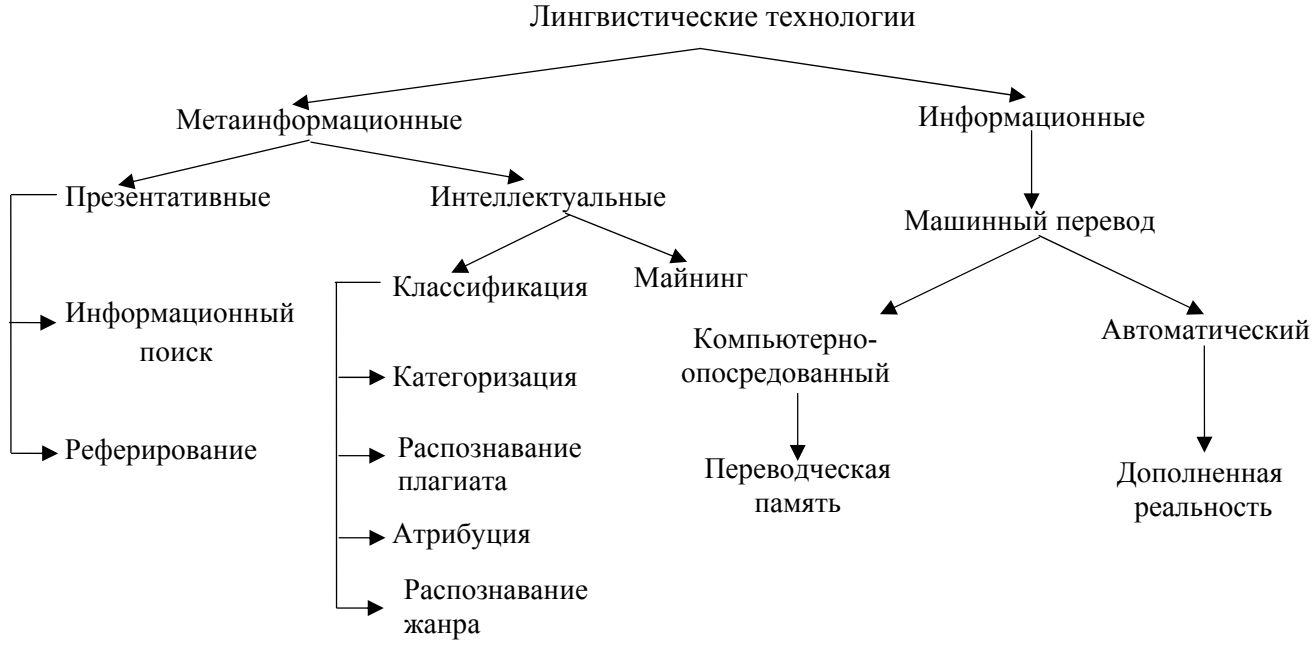

Рис. 3. Виды лингвистических технологий, выделяемые по типу выходных данных

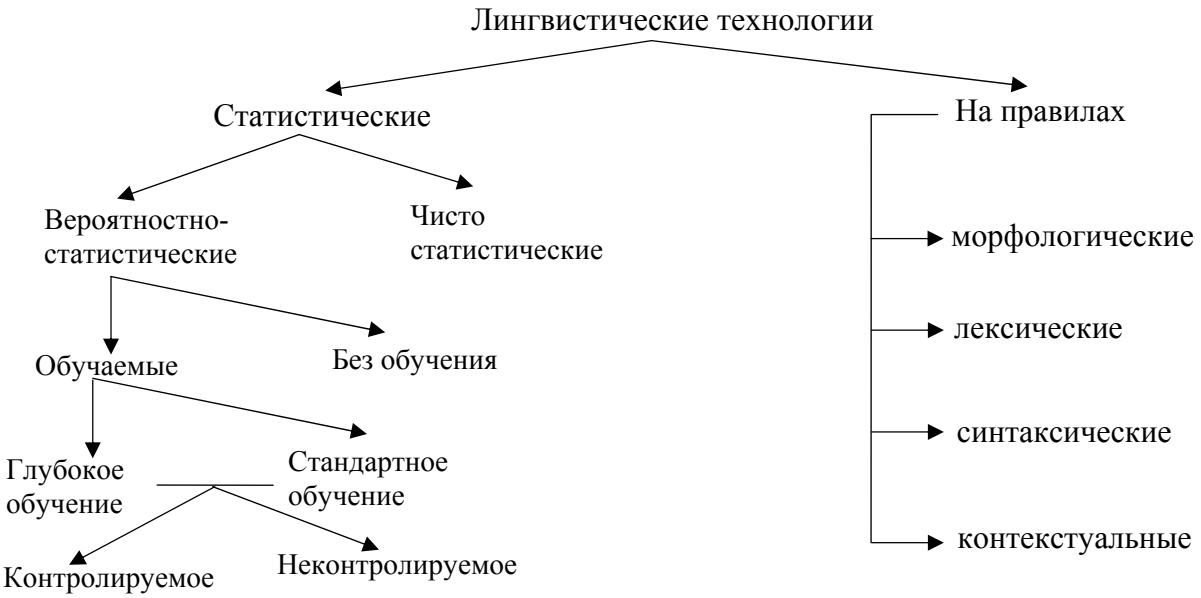

Рис. 4. Виды лингвистических технологий, выделяемые по типам алгоритмов

На третьем уровне генерируются данные (например, тематическая структура текста). Нейросетевое обучение, как правило, проводится на втором уровне анализа. Оно может быть контролируемым (*supervised*), либо неконтролируемым (*unsupervised*). В первом случае создаётся обучающая выборка и разрабатывается алгоритм обучения. Обучающие выборки обычно предполагают ручную разметку текстов экспертами и требуют больших затрат как средств, так и времени. В этой связи всё более популярными становятся методы неконтролируемого обучения.

На рис. 4 представлены лингвистические технологии, выделяемые по особенностям применяемых алгоритмов.

ЗАКЛЮЧЕНИЕ

В настоящей статье предложена оригинальная классификации современных лингвистических технологий на основе двух выделенных нами основных групп критериев: технологического и семиотического. При этом мы исходили из понимания технологий как методов и средств обработки и трансформации исходного объекта по определённым правилам и стандартам с целью получения продукта, необходимого пользователю. В случае информационных технологий таким исходным объектом является некоторая информация, а продуктом, получаемым на выходе, – новая информация. Специфика лингвистических технологий состоит в том, что входная информация представлена текстом на естественном языке. Исходя из такой интерпретации, можно выделить три критерия классификации, представленные в статье: по типу входных данных, по типу выходных данных, по методам, способам, алгоритмам обработки входных данных. В семиотическую группу входят семантический, синтаксический и прагматический критерии идентификации лингвистических технологий. По семантическому критерию выделяются уровни единиц языка и соотносимые с ними программы. По синтаксическому критерию идентифицируется последовательность выполнения программ. По прагматическому критерию выделяются технологии, предназначенные для разных групп пользователей.

Лингвистические технологии могут быть реализовываны в различном программном обеспечении: программах, приложениях, системах. Лингвистические приложения представлены рассмотренными в статье конкордансами, предоставляющими пользователям статистические данные о распределении лингвистических единиц текста. Лингвистические программы представляют собой реализацию определённого алгоритма и используются в качестве модуля систем. Обычно они не реализуются в качестве отдельных приложений, хотя иногда это делается в учебных целях. Лингвистические системы включают ряд разноуровневых программных модулей и выдают информацию, необходимую определённым категориям пользователей. Сложность этих систем определяется количеством программных модулей разного уровня, а также используемыми алгоритмами. В наиболее сложных системах используются все программы, представленные в табл. 1, а также технологии машинного обучения, лежащие в основе систем искусственного интеллекта. Заметим, что если исходить из теста Тьюринга, то основным критерием искусственного интеллекта будет способность адекватного общения на естественном языке, что предполагает использование именно лингвистических технологий. В этом смысле вызывает сомнение распространившаяся в последнее время практика рассматривать в качестве

искусственного интеллекта любые системы, использующие технологии машинного обучения, в том числе и те, которые не имеют никакого лингвистического содержания.

Как всякая таксономия, предложенная в данной работе классификационная схема является абстрактной и не учитывает некоторые гибридные и промежуточные формы. В частности, в одной системе могут сочетаться как регуляторные, так и вероятностные алгоритмы, сочетание которых может повышать эффективность обработки текста, уменьшая количество ошибок. Такая гибридная технология реализована в частеречном теггере, используемом в Британском национальном корпусе. Наряду с методами контролируемого и неконтролируемого обучения может применятся и частично-контролируемое обучение (*semi-supervised learning*). Технология TF*IDF, которая относится к чисто статистическим методам, может быть модифицирована для применения в качестве вероятностно-статистического метода [25]. В некоторых случаях трудно соотнести применяемую технологию с одним из видов технологий. Взвешивание терминов может применяться на всех уровнях системы языка, поскольку может относиться ко всем лингвистическим единицам, указанным в таб. 1.

Отметим, что настоящая статья не претендует на полное и всестороннее представление всех существующих лингвистических технологий. Это достаточно сложно, учитывая динамику их развития и проникновение в другие предметные области, которое идёт по двум направлениям: 1) применение лингвистических технологий в нелингвистических дисциплинах таких как, например, клиническая лингвистика, занимающаяся исследованием нарушений речи и разработкой методов их коррекции [26]; 2) обработка мультимодальных сообщений, включающих не только устный и/или письменный текст, но и экстралингвистические объекты (жесты, изображения). В некоторых источниках утверждается, что внедрение мультимодальных технологий, обусловленное распространением мобильных сенсорных устройств, привело к сдвигу парадигмы развития информационных технологий [27]. Это мнение имеет под собой основу, так как современный этап развития мультимодальных технологий характеризуется повсеместным использованием мультимедийных средств и их интеграцией с текстовой информацией. Заметим, однако, что уже сейчас можно прогнозировать вытеснение сенсорных устройств устройствами виртуальной реальности на основе голографических технологий [28], которые позволяют генерировать трёхмерные изображения, масштабируемые пользователем по своему усмотрению. Это позволит устранить зависимость эргономических характеристик аппаратных средств от диагонали дисплея и приведёт к их миниатюризации.

Независимо от степени развития мультимедиа, текстовые данные в ближайшей перспективе будут составлять основу целого ряда информационных технологий. Надеемся, что предложенные нами критерии идентификации и классификации лингвистических технологий внесут вклад в дальнейшее развитие предметной области, связанной с автоматической обработкой текстовых документов. В первую очередь, имеется в виду создание общепринятой классификационной схемы, которая может послужить основой для разработки онтологии языкознания, отражающей структуру этой научной дисциплины. Такая онтология, в свою очередь, может стать основной частью лингвистической информационной системы, предназначенной для удовлетворения запросов различных категорий пользователей. Очевидно, что эти направления развития требуют серьёзных усилий со стороны научного сообщества.

СПИСОК ЛИТЕРАТУРЫ

1. Вернадский В.И. О науке. Т.1. Научное знание. Научное творчество. Научная мысль / сост.: Г.П. Аксёнов. – Дубна: Феникс, 1997. – 576 с.
2. Яцко В.А. Методы и алгоритмы автоматического анализа текста // Научно-техническая информация. Сер. 2. – 2011. – № 9. – С. 12-19.
3. Яцко В.А. Предметная область компьютерной лингвистики // Вестник Иркутского государственного лингвистического университета. – 2014. – № 2. – С. 24-35.
4. Modi H.K., Parikh M.C. A Review on optical character recognition techniques // International journal of computer applications. – 2017. – Vol. 160, № 6. – P. 20-24. DOI:10.5120/ijca2017913061
5. Hecking T., Leydesdorff L. Topic modelling of empirical text corpora: Validity, reliability, and reproducibility in comparison to semantic maps. – 2018. – URL: https://arxiv.org/ftp/arxiv/papers/1806/1806.01045.pdf .
6. Anandarajan M., Hill Ch., Nolan T. Text Preprocessing // Practical text analytics. Advances in Analytics and Data Science. – Vol 2. – Cham, Switzerland, 2018. – P. 45-59. DOI https://doi.org/10.1007/978-3-319-95663-3_4
7. Lyding V., Schöne K. Design and development of the MERLIN learner corpus platform // Proceedings of the Tenth international conference on language resources and evaluation. – Portoro, Slovenia, 2016. – P. 2471-2477. – URL: https://www.aclweb.org/anthology/L16-1392.pdf
8. Connector usage in the English essay writing of native and non‐native EFL speakers of English // World Englishes. – 1996. – Vol. 15, Iss. 1. – P. 17-27. DOI https://doi.org/10.1111/j.1467-971X.1996.tb00089.x
9. Kim Y., Ross S. Examining variations of prominent features in genre classification // Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008). – Waikoloa, HI, 2008. – P. 132-132.
10. Mahlberg M., Stockwell P., de Joode, J., et al. CLiC Dickens: novel uses of concordances for the integration of corpus stylistics and cognitive poetics // Corpora. – Vol. 11, Iss. 3. – P. 433- 463. – URL: https://www.euppublishing.com/doi/full/10.3366/cor.2016. 0102
11. Alkahtani M., Choudhary A., De A., Harding J.A. A decision support system based on ontology and data mining to improve design using warranty data // Computers & industrial engineering. – 2019. – Vol. 128. – P. 1027-1039. DOI https://doi.org/10.1016/j.cie.2018.04.033

12. Metsker O., Bolgova E., Yakovlev A., Funkner A.A., Kovalchuk S. Pattern-based mining in electronic health Records for complex clinical process analysis // Procedia computer Science. – 2017. – Vol. 119. – P. 197-206. DOI: https://doi.org/10.1016/j.procs.2017.11.177. – URL: https://reader.elsevier.com/reader/sd/pii/S1877050917323876?token=1417A1C4412BD36F1A607C6C6145A3AB8C6A9E7DF028679B22EF2CC906F431B204C0CF16E88529855E40D789BD1CFEDC
13. Shaqra F.A., Duwairi R., Al-Ayyoub M. Recognizing emotion from speech based on age and gender using hierarchical models // Procedia computer science. – 2019. – Vol. 151. – P. 37-44. DOI https://doi.org/10.1016/j.procs.2019.04.009. – URL: https://reader.elsevier.com/reader/sd/pii/S1877050919304703?token=23FB457FC4DC75AB59E7AB29813FEBCD5BEDDB7F135C42644CB0B198E27D0B5A314C7433CC770265BF396E666FDAB34A
14. Wang W., Wu H., Li M. Deep neural networks with batch speaker normalization for intoxicated speech detection. – 2019. – URL: https://www.researchgate.net/profile/Ming_Li372/publication/338448146_Deep_Neural_Networks_with_Batch_Speaker_Normalization_for_Intoxicated_Speech_Detection/links/5e156cd44585159aa4bcf504/.pdf
15. Tulshan A.S., Tulshan S.N. Survey on virtual assistant: Google Assistant, Siri, Cortana, Alexa // Advances in Signal Processing and Intelligent Recognition Systems. SIRS 2018. Communications in Computer and Information Science. Vol. 968 / eds. S. Thampi, O. Marques, S. Krishnan, K.C. Li, D. Ciuonzo, M. Kolekar. – Singapore: Springer, 2019. – P. 190-201. DOI https://doi.org/10.1007/978-981-13-5758-9_17
16. Das T.K., Kumar P.M. BIG data analytics: A framework for unstructured data Analysis // International journal of engineering and technology. – 2013. – Vol 5, № 1. – P. 153-156. –URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.411.6697 &rep=rep1&type=pdf
17. Farzindar A., Inkpen D. Natural language processing for social media. Second edition. – Morgan & Claypool publishers, 2011. – 195 p. DOI: 10.2200/S00809ED2V01Y201710HLT038. – URL: https://www.morganclaypoolpublishers.com/catalog_Orig/samples/9781681736136_ sample. pdf .
18. Wołk K., Agnieszka Wołk A., Marasek K. Implementing statistical machine translation into mobile augmented reality systems // Multimedia and Network Information Systems. Advances in Intelligent Systems and Computing. Vol 506 / eds. A. Zgrzywa, K. Choroś, A. Siemiński. – Cham: Springer, 2016. – P. 61-73. DOI: https://doi.org/10.1007/978-3-319-43982-2_6
19. Mita K., Dalal Mukesh A. Zaveri. Automatic Text Classification: A Technical Review // International Journal of Computer Applications. – 2011. – Vol. 28, № 2. – P. 37-40. – URL: https://www.researchgate.net/profile/Mukesh_Zaveri/publication/ 266296879_Automatic_Text_Classification_A_Technical_Review/links/ 54e74a0a0cf2b199060ae1c5.pdf
20. Talib R., Hanif M.K., Shaeela Ayesha Sh., Fatima F. Text mining: Techniques, applications and issues // International journal of advanced computer science and applications. –2016. – Vol. 7, № 11. – P. 414-418. – URL: https://pdfs.semanticscholar.org/f3ed/ac7ee49a6a6d8eb3deb073589aeaf36ab45d.pdf
21. Nambiar S.K., Leons A., Jose S. Natural language processing based part of speech tagger using hidden Markov model // Third international conference on I-SMAC. – Palladam, India, 2019. – P. 782-785.
22. Scneider K. A comparison of event models for naive Bayes anti-spam e-mail filtering // Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics, 1. – Stroudsburg, PA, USA, 2003. – P. 307-314. – URL: http://www.aclweb.org/anthology/E03-1059
23. Gautam J., Kumar E. An integrated and improved approach to terms weighting in text classification // International journal of computer science issues. – 2013. – Vol. 10, Iss. 1. – P 310-314. – URL: https://www.researchgate.net/publication/299168493_An_Integrated_and_Improved_Approach_to_Terms_Weighting_in_Text_Classification
24. Young T., Hazarika D., Poria S., Cambria E. Recent trends in deep learning based natural language processing // IEEE computational intelligence magazine. – 2018. – Vol. 13, № 3. – P. 55-75. – URL: https://arxiv.org/pdf/1708.02709.pdf
25. Roelleke T., Wang J. TF-IDF uncovered: a study of theories and probabilities // Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval. – New York, 2008. – P. 435–442. DOI: https://doi.org/10.1145/1390334.1390409. – URL: https://dl.acm.org/doi/pdf/10.1145/1390334.1390409?download=true
26. Priyadarshi B., Mahesh B.V.M. Neurodevelopmental disorders from a clinical linguistics perspective // Emerging trends in the diagnosis and intervention of neurodevelopmental disorders. – Hershey, Pennsylvania, USA, 2018. – P. 85-101. DOI: 10.4018/978-1-5225-7004-2.ch005
27. Oviatt Sh., Cohen Ph.R. The Paradigm shift to multimodality in contemporary computer interfaces. – Williston: Morgan&Claypool, 2015. – 243 p. DOI: https://doi.org/10.2200/S00636ED1V01Y201503HCI030
28. Wright S. Move over, voice: Holograms are the next user interface. – 2017. – URL: https://venturebeat.com/2017/12/24/move-over-voice-holograms-are-the-next-user-interface.

Материал поступил в редакцию 06.05.20.

Сведения об авторе

ЯЦКО Вячеслав Александрович – доктор филологических наук, профессор Хакасского государственного университета им. Н.Ф. Катанова, г. Абакан
e-mail: viatcheslav-yatsko@rambler.ru

УВАЖАЕМЫЕ КОЛЛЕГИ!

ВИНИТИ РАН предлагает Вашему вниманию Реферативный Журнал в электронной форме

РЖ в электронной форме (ЭлРЖ) выпускается по всем разделам естественных, технических и точных наук.

Каждый номер ЭлРЖ является полным аналогом печатного номера РЖ по составу описаний документов, их оформлению и расположению. Он сопровождается оглавлением, указателями.

ЭлРЖ представляет собой информационную систему, снабженную поисковым аппаратом и позволяющую пользователю на персональном компьютере:

- читать номер РЖ, последовательно листая рефераты;
- просматривать рефераты отдельных разделов по оглавлению;
- обращаться к рефератам по указателям авторов, источников, ключевых слов;
- проводить поиск документов по словам и словосочетаниям;
- выводить текст описаний документов во внешний файл.

ЭлРЖ в версии Windows Вы можете получить за текущий год с любого номера, а также за предыдущие годы.

Подробную информацию Вы можете получить:

Адрес: 125190, Россия, Москва, ул. Усиевича, 20, ВИНИТИ РАН

Телефон 499-155-42-85 499-151-78-61

E-mail: Contact@viniti.ru, Feo@viniti.ru

ВНИМАНИЮ ЧИТАТЕЛЕЙ!

ИЗДАНИЕ УДК

УНИВЕРСАЛЬНАЯ ДЕСЯТИЧНАЯ КЛАССИФИКАЦИЯ

АЛФАВИТНО-ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

в 2-х томах

Алфавитно-предметный указатель (АПУ) к 4-му полному изданию УДК на русском языке:

Том I содержит АПУ от буквы А до Н;

Том II содержит АПУ от буквы М до Я и указатель латинских наименований к классам УДК 56 Палеонтология, 57 Биологические науки, 58 Ботаника, 49 Зоология, 61 Медицинские науки.

АПУ содержит около 100 000 понятий, представленных в полных таблицах УДК.

При его составлении были учтены изменения, опубликованные в Выпусках № 1 – 6 «Изменения и дополнения к УДК»

Для подписки необходимо направить заявку для оформления счета по адресу:

125190, Россия, Москва, ул. Усиевича, 20, ВИНИТИ РАН

Телефоны: 499 155-42-85, 499 151-78-61

E-mail: feo@viniti.ru

http://www.udcc.ru