Математичні моделі та методи

УДК 519.7

М.К. Ефимов[1], В.А. Лещинский[1], Л.Г. Петрова[2], С.Ю. Шабанов-Кушнаренко[1]

[1]*Харьковский национальный университет радиоэлектроники, Харьков*
[2]*Сумской областной институт последипломного педагогического образования, Сумы*

ПРИМЕНЕНИЕ АППАРАТА ОБОБЩЕННЫХ ПРОСТРАНСТВ ДЛЯ МОДЕЛИРОВАНИЯ МОРФОЛОГИИ ЕСТЕСТВЕННОГО ЯЗЫКА

Исследована проблемная область, включающая в себя задачи алгебры конечных предикатов и описание естественного языка с помощью этого математического аппарата. Проанализирована текущая ситуация, сложившаяся в современной науке по отношению к решению данного круга задач. Применен математический аппарат алгебры конечных предикатов для реализации обобщенных пространств.

Ключевые слова: *теория интеллекта, алгебра конечных предикатов и предикатных операций, линейный логический оператор.*

Введение

Появление развитых диалоговых систем, языков общения и специальных средств редактирования требуют нового подхода к проблеме общения. Для человека идеально было бы общение на обычном естественном языке. Кроме того, естественный язык – это единственная известная на сегодня моделирующая система, средствами которой можно описать окружающий мир.

Подходящий математический аппарат для описания механизмов, процессов и правил естественного языка был разработан в 80-х годах 20 столетия. Таким аппаратом является алгебра конечных предикатов, на языке которой могут быть записаны в виде уравнений любые конечные отношения, а так же алгебра предикатных операций, на языке которой можно описывать действия над отношениями. Уже накопился обширный ряд моделей, которые описывают различные части русского языка с помощью данного формального аппарата.

Не все процессы обработки информации могут быть автоматизированы до тех пор, пока не будут найдены способы формального представления смысла. С начала 80-х г.г. область ИИ привлекла внимание многочисленных исследователей и причина этого в открывшейся возможности реализации ИИ как новой технологии обработки информации. Проблема управления информацией лежит не только в управлении фактами, чем занимается обычная СУБД, она состоит также в управлении смыслом (значением) или семантикой данных [1– 3]. В связи с этим в настоящее время популярны системы, обладающие интеллектуальными функциями, и привлекшие внимание как воплощение новой технологии в области обработки информации.

В последние десятилетия было разработано множество систем, обладающих специфическими чертами обработки знаний, и получивших название экспертных систем (ЭС). Под экспертной системой понимается система, объединяющая возможности компьютера со знаниями и опытом эксперта в такой форме, что система может предложить разумный совет или осуществить разумное решение поставленной задачи. Такое формальное определение ЭС одобрено комитетом группы специалистов по ЭС Британского компьютерного общества. Основным отличием экспертной системы от других диалоговых человеко-машинных систем является подсистема объяснения – ЭС должна иметь способность к объяснению своих решений и тех рассуждений, на основе которых эти решения были приняты. Часто от экспертной системы требуют, чтобы она могла работать с неточной и неполной информацией.

1. Алгебра конечных предикатов

Разработка алгебры конечных предикатов (АКП) открывает возможность перехода от алгоритмического описания информационных процессов к описанию их в виде уравнений, а уравнения задают отношения между переменными. Кроме того, уравнения позволяют узнать реакцию системы даже при неполном определении входных сигналов, тогда как не полностью разработанный алгоритм является недееспособным. Единственным недостатком такого описания с помощью уравнений АКП является то, что число переменных в уравнениях большое (а в уравнениях интеллектуальных процессов это число, без сомнения, огромное), а это значительно усложняет процедуру обработки таких уравнений. Для избавления от данного недостатка используются методы декомпозиции.

АКП определена на множестве всех n-местных k-значных предикатов, т.е. функций вида $y = f(x_1, x_2, \ldots, x_n)$, где $x_1, x_2, \ldots, x_n$ заданы на множестве $A = \{a_1, a_2, \ldots, a_\kappa\}$, $y \in \{0, 1\}$. Роль базисных элементов играют предикаты узнавания букв вида

$$x_i^{aj} = \begin{cases} 1, & \text{åñëè } x_i = a_j; \\ 0, & \text{åñëè } x_i \neq a_j; \quad (1 \le i \le n; 1 \le j \le k). \end{cases}'$$

для которых справедливы законы истинности, отрицания и ложности.

С помощью применения АКП для изучения ЕЯ уже получены некоторые теоретические результаты. Основываясь на гипотезе, гласящей, что естественный язык есть некоторая алгебра, была разработана и описана так называемая лингвистическая алгебра. Эта алгебра рассматривалась как алгебра предикатов. В содержательном плане предложения естественного языка, как и математические утверждения, представляют собой отношения.

Ключевой проблемой при создании автоматизированных информационных систем является проблема понимания смысла сообщения. Легко предположить, что проанализировав предложение в соответствии с его синтаксисом, можно установить и его смысл. К сожалению, на деле все обстоит не так просто. Анализ предложения, которому нас учат в школе, зависит от понимания семантики (т.е. смысла) предложения в гораздо большей степени, чем это предполагалось до того, как были предприняты попытки автоматизировать этот анализ. Между смыслами и текстами имеется соответствие: каждому смыслу отвечает более или менее определенная совокупность текстов, а каждому тексту – более или менее определенное множество смыслов. Правила, определяющие, какие тексты соответствуют каким смыслам, и образуют то, что принято называть языком. На самом деле смысл предложения складывается из смысла входящих в него слов. По мнению многих ведущих специалистов в области языкознания и прикладной лингвистики, смысл слова не является суммой смыслов составляющих его морфем, это даже не функция, это отношение более общего вида. В связи с этим становится актуальной задача исследования и математического моделирования межморфемных отношений, т.е. тех связей, которые существуют в слове между префиксами и корнями, корнями и суффиксами, основами и окончаниями (флексиями).

В связи с вышеизложенным объектом настоящего рассмотрения является морфология, в центре внимания которой стоит слово с его грамматическими изменениями, так как способность человека владеть языком зависит от его способности обрабатывать отдельное слово, а уже после словосочетания, предложения и тексты. Словоизменение – один из самых важных и сложных разделов языкознания. Словарный состав обогащается все время преимущественно путем деривации, т.е. образования новых слов на основе и с использованием уже имеющегося в языке материала.

2. Обобщение понятия пространства

Понятие пространства определяется как логически мыслимая форма (или структура), служащая средой, в которой осуществляются другие формы и те или иные конструкции.

Структурой называется родовое понятие, объединяющее понятия, общей чертой которых является то, что они применимы к множествам, природа которых не определена. Чтобы определить структуру, задают отношения, в которых находятся множества (типовая характеристика структуры), а затем постулируют, что данные отношения удовлетворяют условиям – аксиомам структуры.

Решеткой (структурой) называется любое множество A, на котором заданы две двухместные операции:$\vee$ – дизъюнкция и $\wedge$ – конъюнкция, удовлетворяющие законам идемпотентности, коммутативности, ассоциативности и элиминации. Если операции $\vee$, $\wedge$ удовлетворяют также и законам дистрибутивности, то решетка называется дистрибутивной.

Если в множестве A содержатся также элементы 0 и 1 на нем введена и операция отрицания , удовлетворяющая законам противоречия и исключенного третьего, то получаем дистрибутивную решетку с дополнениями или булеву решетку. Дополнением элемента $a \in A$ называется элемент $b \in A$ $b=a$.

Логическая математика попыталась дать определение общего понятия пространства в менее расплывчатой форме. Она выделила в известных примерах пространств такое их качество, как размерность и пытается сформировать предельно общее понятие, к которому еще применимо свойство размерности. Классическая же (числовая) математика пошла по другому пути: она ухватилась за такое важное для нее понятие как непрерывность известных пространств и выделила в них понятие топологии. При этом обобщении уже теряется понятие размерности. Но для логической математики понятие непрерывности не столь важно. Она обобщает понятие пространства так, что в нем уже теряется свойство непрерывности, но сохраняется свойство размерности. Из конкретных примеров пространств, введенных в математике, выделяем два обобщения. Одно из них можно было бы назвать размерным пространством, а другое – топологическим. Но то и другое представляются не очень удачным. За первым сохраняем название топология, а за вторым – просто пространство (принимаемое теперь в обобщенном смысле), рис. 1.

Логическая математика предлагает это понятие как уточненный заменитель не очень четкого приведенного выше понятия пространства. В нем момент непрерывности не упоминается. Рассмотрим понятие топологии.

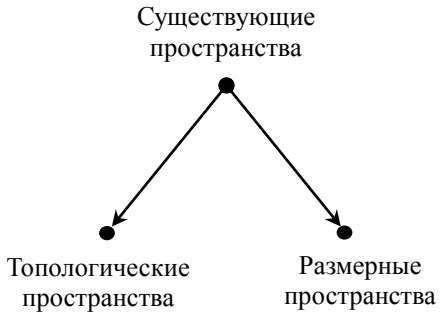

Рис. 1. Классификация пространств

Пусть A – некоторое множество, а $\tau=\{A_i, i\in I\}$ – семейство его подмножеств (I – семейство индексов – имен подмножеств оно может иметь произвольную мощность). Говорят, что семейство τ определяет в множестве A топологию или топологическую структуру, если оно обладает следующими свойствами:

- все множество A и пустое множество $\varnothing$ принадлежит семейству τ;

- пересечение конечного числа множеств из τ принадлежит семейству τ;

- объединение конечного числа множеств из τ принадлежит семейству τ.

Эти три условия называются аксиомами топологической структуры.

Говорят, что задана топология в множестве A. И в топологии появляются те же, что и раньше, две операции – аналоги дизъюнкции и конъюнкции и элементы – аналоги нуля и единицы.

Множество A, рассматриваемое вместе с заданной на нем топологией τ, называется топологическим пространством.

Обычно в топологических пространствах, используемых на практике, всегда присутствует какая-то размерность, т.е. они и в самом деле суть пространства. Лучше было бы сказать не «топологическое пространство», а «топологическая структура». При этом элементы множества A называются точками, а подмножества A_i из семейства τ – открытыми множествами топологического пространства A. Само множество A называется носителем топологии τ.

3. Интерпретация пространства как декартовой системы координат

С помощью понятия пространства мы выходим на новые методы анализа и синтеза произвольных отношений. А такого рода задачи в информатике – главное. В частном случае пространство превращается в декартову систему координат, рис. 2.

B – носитель пространства – множество всех точек плоскости; y – вектор пространства – точка плоскости. x_1, x_2 – координаты вектора (координаты точки плоскости). $(x_1, x_2)=y$ – координатное представление точки. Строго говоря, $y\neq(x_1, x_2)$. Линии могут появляться, исчезать, образовывать области.

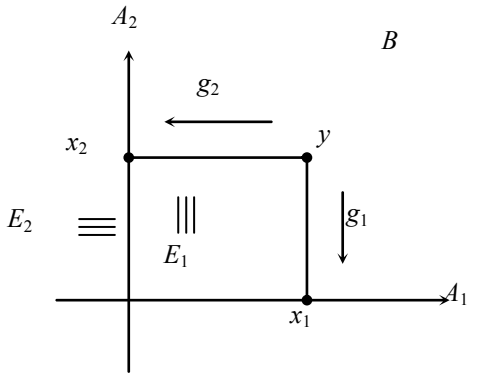

Рис. 2. Декартова система координат

В каждой точке имеется целое множество векторов (быть может, даже пустое). $y=S(x_1, x_2)$ – отображение пространства в его носитель. У декартова пространства S – функция (но у более общего пространства могла бы быть не функция, а произвольное отображение). Координатная система пространства $A_1\times A_2$ – множество всех пар вида (x_1, x_2). У декартова пространства координатная система пространства с точностью до изоморфизма совпадает с самим пространством, но в общем случае это не так. Предикат пространства $S(x_1, x_2, y)$ связывает x_1, x_2 с y. Его отождествляем с самим пространством. Он задан на $A_1\times A_2\times B$. $g_1(y)=x_1$; $g_2(y)=x_2$ – проекторы пространства. Для декартова пространства – это функции (S, g_1, g_2). В общем случае – произвольные отображения. Проекционные предикаты $G_1(y, x_1)$, $G_2(y, x_2)$ – соответствуют проекторам пространства. $E_1(y_1, y_2)$, $E_2(y_1, y_2)$ – сопровождающие квазитолерантности. Для декартова пространства сопровождающие предикаты превращаются в сопровождающие эквивалентности. E_1 – разбиение – система вертикальных линий; E_2 – система горизонтальных линий. Множество точек, имеющих одну и ту же координату x_1 (или x_2). В общем случае – это система пересекающихся областей: для каждого значения x_1 и x_2 – своя область. В декартовой системе на пересечении линий – вектор единственный.

4. Морфологическая интерпретация пространства

Введем носитель пространства B – это множество всех словоформ имен существительных.

Введем оси пространства:

A^1 - множество всех имен существительных, представленных своими словарными формами (конь, сани); A^2 - падеж с шестью значениями (именительный, родительный, дательный, винительный, творительный, предложный); A^3 - число со значениями “единственный”, “множественный”.

Определим предикат S на $A_1\times A_2\times A_3\times B$: $S(x^1, x^2, x^3, y)=\{1$, если все данные согласуются; 0, если не согласуются$\}$. Предикат S называется морфологическим предикатом.

Словоформа выделяется из текста, при помощи анализа пробелов и знаков препинания.

S (стул, творительный, множественное, стульями)=1.

S (стол, тв., мн., стульями) =0.

S (стул, тв., мн., стулья)=0

Носитель русского языка всегда может, в меру своего знания языка, реализовать этот предикат.

Морфологический предикат связывает слово, грамматические признаки и словоформу.

Анализ словоформы – это отыскание словоформы по слову и грамматическим признакам.

$S(x^1, x^2, x^3)=y$.

Нормализация словоформы – переход от словоформы к слову (словарной форме) $g^1(y)=x^1$.

Например:

Пень $\rightarrow$ пеня, пень

кого, чего?

кто, что?

Парадигмой слова (стол, стола, столом, столами, ...) называется множество всех решений уравнения $g^1(y)=x^1$ при фиксированном x^1 относительно переменной y.

Анализ словоформы: отыскание падежа и числа словоформы $x^2=g^2(y)$; $x^3=g^3(y)$. То же для глаголов и других частей речи. Все неоднозначно и частично.

Например: Боже, отче - какой падеж? Не существует среди шести – звательный. (Кто - что - б. м. именит.). Отображение $x^2=g^2(y)$ - частичное. Здания - (именит., множ.) или (родит., единств.).

$x^2=g^2(y)$, $x^3=g^3(y)$ - многозначное отображение.

(Спутник, винит., ед.)={спутника (одуш.), спутник (неодуш.)

Синтез - это отыскание словоформы по x^1, x^2, x^3. Например: «сахар» - «чай без сахара», «дай сахару».

$S(x^1, x^2, x^3)=y$ - неоднозначное отображение.

(Лев, род., ед.)=льва (животное), лева (достоинство монеты (болгария)).

(Сани, ед., или - нет.)

Семантика морфа:

Строитель - одуш. лицо, деятель.

Слова: дом - неодуш., строение.

Выводы

Можно усомниться, будет ли естественным называть описанную выше конструкцию пространством. Ведь обычно в математике векторы пространства можно складывать, а здесь о сложении элементов множества В нет и речи. Но в науке незримо присутствует еще и другое - “фольклорное” или интуитивное понятие пространства, под которым понимается любое многомерное образование, т.е. просто координатная система. Когда исследователь встречается с какими-то объектами, представленными, к примеру, на плоскости, то он часто говорит, что они расположены в двумерном пространстве, даже в том случае, если объекты эти не подвергаются действию операции сложения. Представляется, что в его сознании понятие многомерности неразрывно связано с интуитивным понятием пространства. Поэтому любое многомерное образование естественно именовать пространством. Понятие же отношения явно основано на идее многомерности (о чем свидетельствует наличие у него арности), а значит, мы не совершим ничего предосудительного, если отождествим его с понятием пространства.

Возникает вопрос: нужны ли кому-нибудь такие обобщенные пространства, ощущается ли в них практическая потребность? Нам представляется, что такие пространства будут полезны. Понятия отношения и обобщенного пространства равносильны, вместе с тем на понятии отношения основана вся логическая математика. Логическая же математика выполняет роль основного языка, с помощью которого формально описываются информационные объекты и процессы. Взгляд на отношение как на обобщенное пространство удобен из-за своей естественности и привычности как для математиков, так и для специалистов по информационным процессам. И те и другие владеют классическим понятием пространства, которое глубоко изучено, обросло удобной и разветвленной терминологией. Опираясь на аналогию между традиционными и обобщенными пространствами, можно будет быстро развивать учение об отношениях как теорию обобщенных пространств.

Чтобы продемонстрировать естественность использования обобщенных пространств при исследовании информационных объектов, рассмотрим их применение для изучения механизма словоизменения в русском языке. Русские тексты составляются из отдельных словоформ. Так, предыдущее предложение состоит из словоформ «русские», «тексты» и т.д. Конкретный вид словоформы определяется заданным словом (точнее - его словарной формой) и значениями грамматических признаков. К грамматическим признакам относятся падеж, число, род, лицо, время и т. п. Словоформа «русском», входящая в словосочетание «в русском языке», определяется словом «русский», предложным падежом, мужским родом и единственным числом; словоформа механизма характеризуется словом механизм, родительным падежом и единственным числом. Словоформу естественно рассматривать как вектор, а набор компонентов, состоящий из слова и грамматических признаков, - как ее координатное представление. К примеру, в качестве координатного представления словоформы y=механизма принимаем набор определяющих ее компонентов (x_1, x_2, x_3)=(механизм, родительный, единственное). В роли переменной x_1 используется словарная форма слова, в роли x_2 - падеж, x_3 - число.

Во многих случаях словоформа однозначно определяется своим координатным представлением, а координатное представление однозначно определяется своей словоформой. Так, набор компонентов

(механизм, родительный, единственное) определяет только одну словоформу механизма. И обратно, словоформа механизма однозначно разлагается в набор компонентов (механизм, родительный, единственное). Но так бывает далеко не всегда. К примеру, набор компонентов (сахар, родительный, единственное) порождает две словоформы: сахара (чай без сахара) и сахару (дай кусочек сахару). Словоформе собаки соответствуют два координатных представления: (собака, родительный, единственное) и (собака, именительный, множественное). Набору компонентов (терпение, именительный, множественное) не соответствует ни одной словоформы. Эти примеры наглядно показывают, что механизм русского словоизменения (и не только русского) естественным образом охватывается понятием обобщенного пространства; вместе с тем, в классическое понятие декартова пространства этот механизм не укладывается. Мы полагаем, что существует значительное число информационных объектов, требующих для своего математического описания привлечения понятия обобщенного пространства.

Список литературы

1. Бондаренко М.Ф. Теория интеллекта: учебник / М.Ф. Бондаренко, Ю.П. Шабанов-Кушнаренко. – Х.: Изд-во СМИТ, 2007 – 576 с.

2. Марчук Ю.Н. Лингвистическая прагматика и общение с ЭВМ / Отв. ред. Ю.Н. Марчук. – М., 1989. – 142 с.

3. Бондаренко М.Ф. Автоматическая обработка информации на естественном языке / М.Ф. Бондаренко, А.Ф. Осыка. – К.: УМК ВО, 1994. – 144 с.

Поступила в редколлегию 24.10.2012

Рецензент: д-р техн. наук, проф. С.Ф. Чалый, Харьковский национальный университет радиоэлектроники.

ЗАСТОСУВАННЯ АПАРАТУ УЗАГАЛЬНЕНИХ ПРОСТОРІВ ДЛЯ МОДЕЛЮВАННЯ МОРФОЛОГІЇ ПРИРОДНОГО МОВИ

М.К. Ефимов, В.А. Лещинский, Л.Г. Петрова, С.Ю. Шабанов-Кушнаренко

Досліджена проблемна область, що включає в себе завдання алгебри скінченних предикатів та опис природної мови за допомогою цього математичного апарату. Проаналізована поточна ситуація, що склалася в сучасній науці у відношенні до вирішення даного кола завдань. Застосовано математичний апарат алгебри скінченних предикатів для реалізації узагальнених просторів.

Ключові слова: *теорія інтелекту, алгебра скінченних предикатів і предикатних операцій, лінійний логічний оператор.*

THE APPARATUS OF GENERALIZED SPACES USE FOR NATURAL LANGUAGE MORPHOLOGY MODELING

M.K. Efimov, V.A. Leschynskiy, L.G. Petrova, S.Yu. Shabanov-Kushnarenko

The investigated problem area that includes tasks algebra of finite predicates and description of natural language by means of this mathematical apparatus. Analyzed the current situation in modern science in relation to the resolution of the range of tasks. Applied Mathematics algebra of finite predicates to implement generalized spaces.

Keywords: *theory of intelligence, algebras of finite predicates and predicate operations, linear logical operator.*