множество R всех предикатных операций. Алгеброй предикатных операций над R называется любая алгебра, заданная на носителе R.

Пусть $F(X_1, X_2, \ldots, X_k) = Y$ – предикатная операция, отображающая множество P^k в множество P. Здесь $X_1, X_2, \ldots, X_k$ – предикатные переменные, выступающие в роли аргументов операции F; Y – предикатная переменная, являющаяся значением операции F. Отрицанием $\neg F = \overline{F}$ предикатной операции F называется такая предикатная операция, значения которой определяются по правилу

$$(\neg F)(X_1, X_2, \ldots, X_k) = \neg F(X_1, X_2, \ldots, X_k), \qquad (6)$$

для любых $X_1, X_2, \ldots, X_k \in P$. Пусть F и G – предикатные операции, отображающие P^k в P. Дизъюнкцией $F \vee G$ предикатных операций F и G называется предикатная операция, значения которой определяются по правилу

$$(F \vee G)(X_1, X_2, \ldots, X_k) = F(X_1, X_2, \ldots, X_k) \vee G(X_1, X_2, \ldots, X_k), \qquad (7)$$

для любых $X_1, X_2, \ldots, X_k \in M$. Конъюнкцией $F \wedge G$ предикатных операций F и G называется предикатная операция, значения которой определяются по правилу

$$(F \wedge G)(X_1, X_2, \ldots, X_k) = F(X_1, X_2, \ldots, X_k) \wedge G(X_1, X_2, \ldots, X_k), \qquad (8)$$

для любых $X_1, X_2, \ldots, X_k \in M$. В последних трех равенствах слева от знака равенства фигурируют операции $\neg$, $\vee$ и $\wedge$ над предикатными операциями; справа знаки $\neg$, $\vee$ и $\wedge$ обозначают операции над предикатами. Булевой алгеброй предикатных операций называется любая алгебра предикатных операций с базисом операций, состоящим из отрицания, конъюнкции и дизъюнкции.

Основные результаты и выводы. В отличие от других, ранее применяемых методов синтаксического анализа, предложенный логико-семантический метод опирается на структуру предложений и семантику текста в целом. Он позволит качественно аннотировать (реферировать) полнотекстовые документы. Качество аннотирования (реферирования) обеспечивается за счет синтаксического анализа, реализованного с помощью алгебры конечных предикатов и предикатных операций.

Список литературы: ***1.*** *Удо Хан, Индерджиет Мани*. Системы автоматического реферирования. "Открытые системы", 2000, № 12. ***2.*** *Hahn U., Mani I.* The challenges of automatic summarization. IEEE Computer, 33(11):29-35, 2000. ***3.*** *Башмаков А. И., Башмаков И. А.* Интеллектуальные информационные технологии: Учеб. Пособие. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2005. – 304 с. ***4.*** *Барсегян А. А., Куприянов М. С., Степаненко В. В., Холод И. И.* Технология анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. – СПб.: БХВ-Петербург, 2007. – 384 с.

5. *Браславский П., Колычев И.* Автоматическое реферирование веб-документов с учетом запроса. Грант ООО "Яндекс" № 102707, company/yandex. ru/grant/2005/ll_Braslavski_102707.pdf. ***6.*** *Марчук Ю. Н.* Компьютерная лингвистика: учебное пособие /Ю.Н. Марчук. – М.:АСТ: Восток – Запад, 2007. – 317 с. ***7.*** *Дударь З. В., Рассадникова А. В., Шабанов-Кушнаренко Ю. П.* Тексты естественного языка как формулы лингвистической алгебры //АСУ и приборы автоматики. – 1998. – № 107. – С. 135-144. ***8.*** *Баталин А. В.* Формальное описание структуры естественного языка как алгебры предикатных операций и его применение в системах искусственного интеллекта. – Дис. … канд. Техн. Наук. – Харьков: ХНУРЭ, 2004. – 168 с. ***9.*** *Хайрова Н. Ф., Замаруева І. В.* Машинний переклад: Навч. посіб. – Харків: Око, 1998. – 82 с. ***10.*** *Бондаренко М. Ф., Шабанов-Кушнаренко Ю. П.* Теория интеллекта: Учебник. – Харьков: ООО «Компания СМИТ», 2006. – 267-281. ***11.*** *Шабанов – Кушнаренко Ю. П.* Теория интеллекта: Проблемы и перспективы – Х.: Вища шк., 1987. ***12.*** *Шабанов-Кушнаренко Ю. П., Шаронова Н. В.* Компараторная идентификация лингвистических объектов – К., ИСИО, 1993.

Поступила в редколлегию 20.01.09

УДК 519.7

С. В. ГОНЧАРОВ, аспирант НТУ «ХПИ»,
САЙЕД МОХАММАД ТАУХИД СИДДИКИ, аспирант НТУ «ХПИ»

ИСПОЛЬЗОВАНИЕ ПРЕДИКАТНЫХ КАТЕГОРИЙ ДЛЯ ПРЕДСТАВЛЕНИЯ ИНФОРМАЦИИ В ДОКУМЕНТАХ

В статті пропонується модель представлення знань, яка базується на використанні предикатних категорій та алгебри скінчених предикатів. Зроблені висновки відносно того, що теорія категорій та її предикатна інтерпретація дозволяють описувати процеси формування множин знань у базах даних логічного типу.

В статье предлагается модель представления знаний, основанная на использовании предикатных категорий и алгебры конечных предикатов. Сделаны выводы о том, что теория категорий и ее предикатная интерпретация позволяют описывать процессы формирования множества знаний в базах знаний логического типа.

In the article the model of representation of the knowledge, based on use of predicate categories and algebras of final predicates is offered. Conclusions that the theory of categories and its predicate interpretation allow to describe processes of formation of set of knowledge in bases of knowledge of logic type are made.

Введение. Существует несколько основных моделей представления текстовой информации (знаний): логические, сетевые продукционные и фреймовые. Каждая из них имеет свои средства и свой язык представления информации в виде текстов естественного языка. В основе всех этих моделей лежит использование правил логического вывода для получения новых знаний на основе имеющихся. Главной задачей при этом является разработка такого способа представления знаний, который бы охватывал различные

модели представления знаний и отличался от них единым формальным языком представления и обработки информации. В качестве такого единого языка представляется логичным использовать язык алгебры предикатов и предикатных операций [1, 2].

Еще более абстрактным и мощным инструментом, который можно использовать для нужд информатизации, в том числе для машинного представления и обработки знаний, является теория категорий. Теория категорий была разработана на основе исследований в области гомологии и гомологической алгебры. Авторами теории категорий Эленбергом и Маклейном были введены понятия безобъектной категории и категории с объектами. Для расширения возможностей классической теории категорий в области информатизации разработаны понятия предикатной категории, модифицированной категории и квазикатегории [3].

Цель работы. Создание модели представления знаний, основанной на использовании предикатных категорий и алгебры конечных предикатов.

Основной материал. В толковом математическом словаре [4] дается определение категории как совокупности однотипных математических объектов (множеств, пространств, групп и т. д.) и их отображений друг на друга (морфизмов). Класс объектов категории K обозначается $Ob\ K$, а класс морфизмов – $Mor\ K$. Безобъектная классическая категория является одним из видов алгебр и задается множеством M, элементы которого называются *морфизмами* и единственной частичной бинарной операцией fg умножения морфизмов, отображающей декартово произведение $M \times M$ в M. Морфизмы рассматриваются как некоторые бесструктурные элементы множества M. Существуют также тождественные или единичные морфизмы $e \in M$ такие, что существует произведение $ee = e$ и для любых $f, g \in M$, для которых существует произведение fe и eg выполняются равенства $fe = f$ и $eg = g$. Так как операция умножения является в общем случае частичной, существует единственный левый единичный морфизм e_f и единственный правый единичный морфизм $e_{f'}$ для любого морфизма $f \in M$ такие, что $fe_f = e_f f = f$. Произведение fg морфизмов f и g существует только в том случае, если правый единичный морфизм морфизма f совпадает с левым морфизмом морфизма g таким образом, что $fe = f$ и $eg = g$.

Произведение морфизмов ассоциативно $(fg)h = f(gh)$ для любых $f, g, h \in M$, для которых существуют произведения $(fg)h$ и $f(gh)$. Множество M морфизмов с единичными морфизмами и с действующей на нем операцией умножения, обладающей выше приведенными свойствами, называется *безобъектной категорией* K. Обозначают $M = MorK$, $f \in M$,

$f \in MorK$. $MorK$ – это множество всех морфизмов категории K. Если $f \in MorK$, то говорят, что морфизм f является K – морфизмом.

В объектной категории дополнительно к морфизмам вводится понятие объектов. Множество объектов категории K обозначается $Ob\ K$. Объекты обозначаются буквами $A, B, C, \ldots$. Если $A \in Ob\ K$, то A является объектом категории K или K – объектом. Говорят, что f есть морфизм из объекта A в объект B и пишут $f : A \to B$ или $A \xrightarrow{f} B$. Объект A называется началом морфизма f, а объект B – его концом. Каждой паре объектов A, $B \in Ob\ K$ ставится в соответствие некоторое, возможно даже пустое, множество $H_K(A, B)$ морфизмов категории K. Иначе его обозначают $Hom_K(A, B)$, $MorK(A, B)$ или проще $H(A, B)$, $Hom(A, B)$, $Mor(A, B)$. Для каждого морфизма $f \in MorK$ существует единственная пара объектов A и B такая, что $A, B \in ObK$ и $f \in H_K(A, B)$. Это утверждение говорит о том, что, если морфизмы интерпретировать как некоторые функции, то каждая функция должна иметь область определения A и область значений B. Иначе функция будет задана не полностью.

Над множеством $MorK$ определена, вообще говоря, частичная двухместная операция умножения морфизмов. Произведение fg морфизмов $f : A \to B$ и $g : C \to D$ определено только в том случае, если $B = C$, то есть конец морфизма f совпадает с началом морфизма g. В этом случае произведение fg является морфизмом из объекта A в объект D. В данном случае для объектов $A, B, C \in K$ определено отображение $H_K(A, B) \times H_K(B, C) \to H_K(A, C)$. Знак $\times$ в данном случае обозначает декартово произведение множеств морфизмов. Морфизмы f, g категории K вида $f : A \to B$ и $g : B \to C$ называются последовательными, а морфизмы вида $f : A \to B$ и $g : A \to B$ – параллельными. Умножение морфизмов ассоциативно

$$(fg)h = f(gh), \tag{1}$$

когда $f : A \to B$, $g : B \to C$, $h : C \to D$.

Равенство (1) выражает категорный закон ассоциативности. Закон ассоциативности можно наглядно отобразить в виде категорной диаграммы (рис. 1).

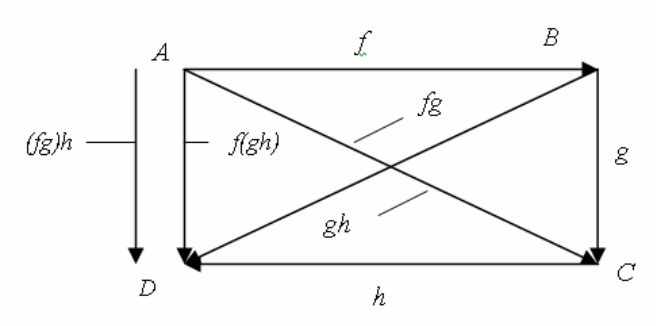

Рис. 1. Категорная диаграмма закона ассоциативности

Любая категорная диаграмма состоит из объектов категории и стрелок (морфизмов) и представляет собой ориентированный раскрашенный граф. Категорные диаграммы делятся на замкнутые и разомкнутые. Замкнутые категорные диаграммы называются коммутативными, так как результат действия морфизмов при их последовательном выполнении зависит только от начального и конечного положения объектов категории на диаграмме. Категорные диаграммы делятся на общие и частные. Общие коммутативны для всех объектов и морфизмов категории. Общими коммутативными диаграммами выражаются свойства какой-либо конкретной категории. Частные категорные диаграммы относятся к конкретным объектам и морфизмам. Они могут быть как замкнутыми, так и разомкнутыми. Для каждого объекта $B \in Ob\ \ K$ существует морфизм $e_B : B \to B$, называемый единичным или тождественным морфизмом объекта B, такой, что

$$fe_B = f \text{ и } e_B g = g, \tag{2}$$

для всех морфизмов $f : A \to B$, и $g : B \to C$. Тождества (2) называются категорными законами тождества. Они выражаются следующей коммутативной диаграммой тождества (рис. 2).

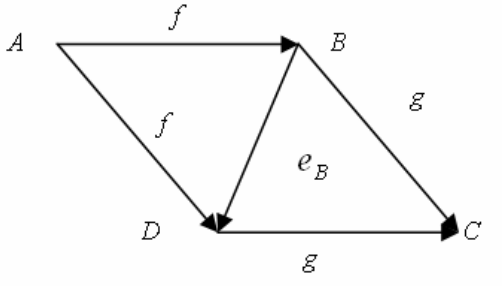

Рис. 2. Коммутативная диаграмма тождества

Классическая категория допускает различные интерпретации, в том числе проективную и предикатную интерпретации [3]. Предикатная интерпретация является наиболее важной для нужд информатизации. Предикатная категория $Pred$ задается на некотором универсуме U. В роли объектов $A,B,C,\ldots$ категории $Pred$ используются произвольные подмножества универсума U. В роли множества объектов $Ob\ \ Pred$ выбираем систему всех подмножеств универсума U. В роли морфизма $f : A \to B$ категории $Pred$ используется произвольный линейный логический оператор $F_f(P) = Q$, преобразующий предикат P в предикат Q. Линейный логический оператор записывается в виде:

$$\exists x \in A\left(K_f(x,y)P(x)\right) = Q(y). \tag{3}$$

Предикат $P(x)$ задан на множестве A, а предикат $Q(y)$ задан на множестве B. Предикат $P(x)$, заданный на множестве A рассматривается как экземпляр объекта A, предикат $Q(y)$ на B – как экземпляр объекта B. Морфизм $f : A \to B$ преобразует экземпляры объекта A в экземпляры объекта B. Предикат $K_f(x,y)$ является ядром линейного логического оператора F_f. Он полностью характеризует вид преобразования (3). Предикат $K_f(x,y)$ задан на декартовом произведении $A \times B$ множеств A и B. Морфизм f вида (3) полностью задан предикатом $K_f(x,y)$. В роли множества $Mor(A,B)$ берется система всевозможных операций вида (3). В категории $Pred$ каждому морфизму $f \in Pred$ взаимно однозначно соответствует ядро $K_f(x,y)$ преобразования (3). Каждый морфизм $f : A \to B$ категории $Pred$ можно задать, указав предикат $K_f(x,y)$, заданный на $A \times B$. Множество $Mor\ \ Pred$ получается объединением множеств $Mor\,Pred(A,B)$, где (A,B) всевозможные пары множеств $A,B \subseteq U$ или в виде совокупности преобразований (3) со всевозможными ядрами $K_f(x,y)$, заданных на всевозможных декартовых произведениях $A \times B$ множеств $A,B \subseteq U$.

Примером ядра морфизма, заданного на декартовом произведении $A \times B$ множеств $A = \{a,o,y\}$ и $B = \{д,ж,m,x\}$ может служить предикат

$$K_f(x,y) = x^a(y^д \vee y^ж \vee y^x) \vee x^o(y^m \vee y^x) \vee x^y(y^ж \vee y^x). \tag{4}$$

Двудольный граф предиката $K_f(x,y)$ представлен на рис. 3.

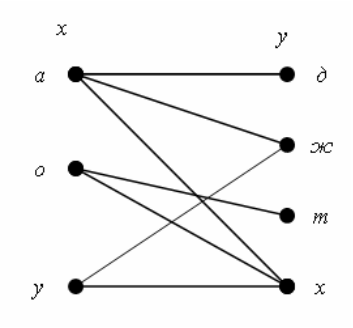

Рис. 3. Двудольный граф предиката $K_f(x,y)$

Линейный логический оператор с этим ядром запишется так:

$$Q(y) = \exists x \in \{a,o,y\}(x^a(y^д \vee y^ж \vee y^x) \vee x^o(y^m \vee y^x) \vee \vee x^y(y^ж \vee y^x))P(x) \quad (5)$$

Если вместо предиката $P(x)$ подставить в формулу (4) его конкретное значение, например $P(x) = x^a \vee x^y$, то в результате получим следующее значение предиката $Q(y)$:

$$Q(y) = \exists x \in \{a,o,y\}(x^a(y^д \vee y^ж \vee y^x) \vee x^o(y^m \vee y^x) \vee \vee x^y(y^ж \vee y^x))(x^a \vee x^y) = y^д \vee y^ж \vee y^x \quad (6)$$

Этот результат можно получить и графически, если элементам множества $P = \{a,y\}$ с помощью ребер двудольного графа предиката $K_f(x,y)$ поставить в соответствие связанные с ними элементы множества $Q = \{д,ж,x\}$. Таким образом, морфизм (5) преобразует множество $P = \{a,y\}$ в множество $Q = \{д,ж,x\}$.

Данный пример иллюстрирует возможность использования морфизмов предикатной категории для хранения знаний о том, какие двухбуквенные слова русского языка могут быть образованы на множествах гласных и согласных букв $A = \{a,o,y\}$ и $B = \{д,ж,m,x\}$, а также для выполнения запросов типа «Какие двухбуквенные строки образуют слова, если на первом месте стоят буквы *а* и *у*?». В данном случае может быть образованно пять таких слов: *«ад», «аж», «ах», «уж», «ух»*. Ядро линейного логического оператора можно рассматривать как знания или правила получения знаний, а сам линейный логический оператор как механизм выполнения запроса для получения новых знаний.

Если ядро линейного логического оператора реализовать аппаратно в виде переключательной цепи, то получим один из блоков лингвистического процессора, который может хранить знания о правилах образования двухбуквенных слов русского языка и может выполнять запросы выше приведенного типа.

Подобные линейные логические операторы могут быть использованы для хранения знаний о правилах образования морфемных цепочек, словосочетаний, предложений, а также для получения семантических значений словоформ по известным семантическим значениям морфем, семантических значений словосочетаний по известным семантическим значениям отдельно взятых слов и т. д.

Таким образом, семантическое значение любого фрагмента текста данного уровня может быть получено по семантическим значениям фрагментов текста предыдущего уровня с помощью соответствующих линейных логических операторов.

Основные результаты и выводы. На основании всего выше изложенного можно сделать следующие выводы. Теория категорий и ее предикатная интерпретация позволяют описывать процессы формирования множества знаний в базах знаний логического типа. Логические принципы формирования баз знаний присущи не только логическим моделям представления знаний. Аналогичные механизмы логического вывода работают при формировании множеств знаний в сетевых, продукционных и фреймовых моделях представления знаний.

Используя предикатные категории для описания формирования баз знаний, множество правил вывода можно хранить в виде ядер линейных операторов, а сам механизм формирования знаний в виде линейных операторов, представленных с помощью формул алгебры предикатов. Схемная реализация линейных операторов позволит создать процессор обработки и формирования знаний, включающий базу знаний и блок логического вывода на знаниях.

Теория категорий дает возможность ясно и наглядно описывать процессы формирования и обработки знаний в виде категорных диаграмм. Теория категорий может стать реальной основой создания систем интеллектуальной обработки текстовой информации.

Список литературы: ***1.*** *Бондаренко М. Ф., Шабанов-Кушнаренко Ю. П.* Теория интеллекта: Учебник. – Харьков: ООО «Компания СМИТ», 2006. – 576 с. ***2.*** *Шабанов-Кушнаренко Ю. П., Шаронова Н. В.* Компараторная идентификация лингвистических объектов: Монография. – Киев: Изд-во Ин-та системных иссл. образования Украины, 1993.– 116 с. ***3.*** О модифицированных категориях / *М. Ф. Бондаренко, З. В. Дударь, А. А. Іванилов, В. В. Маникин, Ю. П. Шабанов-Кушнаренко* // Радиоэлектроника и информатика. 2005. № 1, с 87-99. ***4.*** *Першиков В. И.* Толковый словарь по информатике: Свыше 10 000 терминов/ *В. И. Першиков, В. М. Савинков*; рец. Л. Д. Райков. – М.: Финансы и статистика, 1995.-543 с.

Поступила в редколлегию 20.01.09