Krassimir Markov, Vitalii Velychko, Oleksy Voloshin
(editors)

Natural and Artificial Intelligence

ITHEA
SOFIA
2010

Krassimir Markov, Vitalii Velychko, Oleksy Voloshin (ed.)

Natural and Artificial Intelligence

ITHEA®

Sofia, Bulgaria, 2010

ISBN 978-954-16-0043-9

First edition

Recommended for publication by The Scientific Concil of the Institute of Information Theories and Applications FOI ITHEA

This book is engraved in prof. Zinoviy Lvovich Rabinovich memory. He was a great Ukrainian scientist, co-founder of ITHEA International Scientific Society (ITHEA ISS). To do homage to the remarkable world-known scientific leader and teacher this book is published in Russian language and is concerned to some of the main areas of interest of Prof. Rabinovich.

The book is opened by the last paper of Prof. Rabinovich specially written for ITHEA ISS. Further the book maintains articles on actual problems of natural and artificial intelligence, information interaction and corresponded intelligent technologies, expert systems, robotics, classification, business intelligence; etc. In more details, the papers are concerned in:conceptual problems of the natural and artificial intelligent systems: structures and functions of the human memory, ontological models of knowledge representation, knowledge extraction from the natural language texts;network technologies;evolution and perspectives of development of the mechatronics and robotics;visual communication by gestures and movements, psihology of vision and information technologies of computer vision, image processing;object classification using qualitative characteristics;methods for comparing of alternatives and their ranging in the procedures of expert knowledge processing;ecology of programming – a new trend in the software engineering;decision support systems for economics and banking;systems for automated support of disaster risk management;and etc.

It is represented that book articles will be interesting for experts in the field of information technologies as well as for practical users.

General Sponsor: Consortium FOI Bulgaria (www.foibg.com).

Printed in Bulgaria

ISBN 978-954-16-0043-9

C\o Jusautor, Sofia, 2010

ОБ АВТОМАТИЗАЦИИ ПРОЦЕССА ИЗВЛЕЧЕНИЯ ПЕРВИЧНЫХ ЗНАНИЙ ИЗ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТОВ

Александр Палагин, Сергей Крывый, Дмитрий Бибиков

Аннотация. *Предлагается формализация процесса извлечения знаний из естественно-языковых текстов, на основании которой строится автоматизированная система анализа определений, временных и событийных отношений, имеющихся в текстах.*

Ключевые слова*: обработка естественно языковых текстов, онтологии, формализация*

ACM Classification Keywords*: H4m. Miscellaneous*

Введение

Бурное развитие науки и техники на протяжении последних десятилетий 20-го столетия и начала 21 столетия привела к огромному накоплению количества научно-технической информации. Информационный бум и лавинообразный рост информации приводит к сильному «зашумлению» несущественной информацией интересующей нас предметной области. В связи с этим одному человеку, даже очень высокой квалификации, не под силу освоить, понять и воспользоваться этой информацией с целью проведения научных исследований. Единственным видимым выходом из сложившейся ситуации является автоматизация процесса поиска и обработки необходимой информации. Для решения этой проблемы создаются технологии, ориентированные на смысловые структуры. Ведущей парадигмой структурирования информации на сегодняшний день являются онтологии или иерархические концептуальные структуры, представляющие собой модель предметной области, состоящей из иерархии понятий (концептов) предметной области, связей между ними и законов, действующих в рамках этой модели. Успешное решение проблемы автоматизации процесса построения такого типа структур зависит от успешного решения следующих проблем:

- проблема анализа естественно-языковой текстовой информации с целью извлечения знаний [1,5,6,7];
- проблема построения автоматизированной системы поиска и извлечения знаний, ее архитектуры и инструментария пользователя [1,2,6,7];
- проблема интеграции знаний из нескольких предметных областей с целью обеспечения эффективности проведения исследований междисциплинарного характера [1].

Третья из вышеприведенных проблем, в частности, тесно соприкасается с проблемой эффективного использования систем автоматического поиска доказательств теорем в формальных логических теориях и с подобными ей проблемами. Основу систем автоматического поиска доказательств составляют пруверы – программы, выполняющие доказательство. Успешное применение прувера будет только тогда, когда в его распоряжении имеется вся необходимая информация для успешного проведения доказательства. Например, если пруверу нужно доказать теорему Лагранжа о делимости порядка конечной группы на индекс ее подгруппы, то пруверу недостаточно иметь только аксиоматику теории групп. Ему также понадобиться аксиоматика теории делимости, а может и аксиоматика Пеано вместе с некоторыми дополнительными фактами из других областей. Решение этой проблемы облегчается, если имеется интегрированная система, включающая в себя необходимые сведенья из других областей знаний. Тогда прувер сам находит нужную ему информацию и использует ее для успешного проведения доказательства.

В данной работе предлагается некоторая формализация процесса анализа естественно-языковых текстов (ЕЯТ) и представления результатов этого анализа.

Формальная постановка проблемы

Процесс автоматизации какой-либо деятельности, как правило, требует формализованной постановки задачи, которая дает возможность выполнения анализа данной задачи с целью выработки метода ее решения. Когда речь идет об автоматизации процесса извлечения знаний из ЕЯТ и построения соответствующей онтологии, то необходимо определить понятия «знание» и «извлечение знания». С целью формализации вышеуказанных понятий, введем следующие определения, пользуясь нотацией констрейнтного программирования [3].

Пусть дано некоторое множество D, на котором определена конечная совокупность $R = \{R_1, \ldots, R_k\}$ отношений $R_i \subseteq D^n, i = 1, 2, \ldots, k$, конечной арности. Языком ограничений L на D называется непустое множество $L \subseteq R$. Проблема выполнимости ограничений из L формулируется следующим образом.

Для произвольного множества D и языка ограничений L на D проблемой выполнимости огра-ничений $CSP(L)$ является решение такой комбинаторной задачи:

дана тройка $P = (V, D, C)$, где

- $V = \{v_1, \ldots, v_m\}$ - конечное множество переменных;

- $C = \{c_1, \ldots, c_q\}$ - конечное множество ограничений, где ограничение c_i из *C* является парой (s_i, R_i), где $s_i = (v_{i1}, ..., v_{ij})$ - кортеж, состоящий из переменных, $R_i \in L - n_j$ -арное отношение на D;

найтифункцию $\varphi : V \to D$ такую, что $\forall (s_i, R_i) \in C$ кортеж $(\varphi(v_{i1}), \ldots, \varphi(v_{ij})) \in R_i$ либо убедится в том, что её не существует, $i = 1, 2, \ldots, k$. Множество D в этом случае называется областью проблемы, а функция φ называется интерпретацией $CSP(L)$.

В случае анализа ЕЯТ с целью извлечения знаний множество D, как область проблемы, интерпретируется как множество объектов, извлечённых из входного текста *T*, которое факторизовано по некоторому отношению эквивалентности *R* (это отношение будем называть отношением синонимии) и в котором «закодированы» отношения $R_i, i = 1, 2, \ldots, k$. Множество переменных $V = \{v_1, v_2, \ldots, v_m\}$ принимает значения в этом факторизованном множестве объектов, фигурирующих в тексте *T* (это могут быть лексико-грамматические разряды, конкретные объекты (люди, даты, предметы и т.п.)).

Проблемой извлечения знаний из ЕЯТ называется проблема поиска интерпретации $\varphi : V \to D$ с явным построением отношений R_i совокупности $L \subseteq R$. При этом, отношения $R_i \in L, i = 1, 2, ..., k,$ извлеченные из текста *T,* будем называть ***знаниями.***

Приведенное определение достаточно общее и его необходимо конкретизировать. Конкретизация интерпретации и отношений в таком случае определяется целями, которые преследуются при анализе данного текста *T*. В качестве примеров конкретизации можно привести следующие.

A) *Лексико-грамматический анализ* приводит к конкретизации интерпретации $\varphi : V \to T$ и отношений $R_i \in L$. Интерпретация φ в данном случае представляется в виде суперпозиции двух функций φ_1 и φ_2,

т.е. $\varphi(V) = \varphi_2(\varphi_1(V)) = \varphi_1 * \varphi_2(V)$, где $*$ означает суперпозицию функций. Функции φ_1 и φ_2 реализуют процесс синтаксического и семантического анализа предложений текста *T*, а отношения R_1 и R_2 - это синтаксические ограничения (синтаксические правила языка, в котором представлен текст *T*) и семантические ограничения.

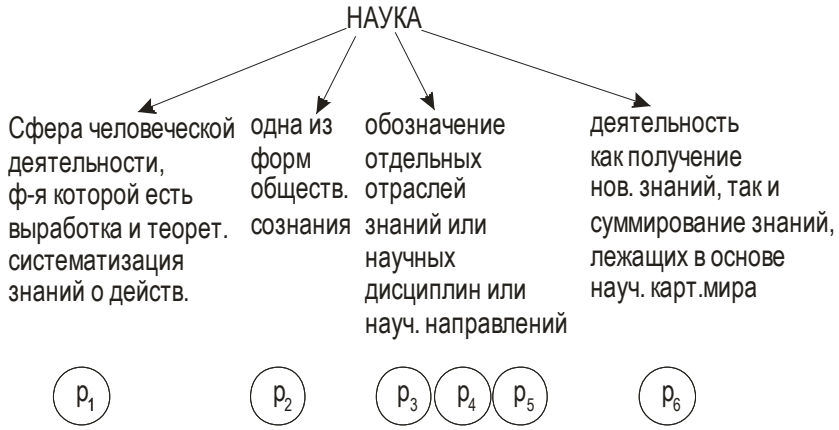
Функцию φ_1 тоже можно рассматривать как суперпозицию отображений φ_{11} и φ_{12}, которые реализуют соответственно морфологический и синтаксический анализ предложений ЕЯТ *T* и которые вместе с отображением φ_2 составляют классическую систему лексико-грамматического анализа [4].

Б) *Силлогистика Аристотеля* является другим примером уточнения интерпретации φ и отношений $R_i \in L$. В этом случае интерпретация φ носит теоретико-множественный характер, а отношения $R_i \in L$ - это отношение включения для множеств и его свойства. Более полное описание этого уточнения можно найти в работах [5,6].

В) *Текст библиографического характера* является примером хорошо структурированного текста. Это значит, что проблема извлечения знаний из такого текста решается относительно просто. Детальное описание этого процесса приведено в работе [5].

Обработка определений в ЕЯТ и их формальное представление

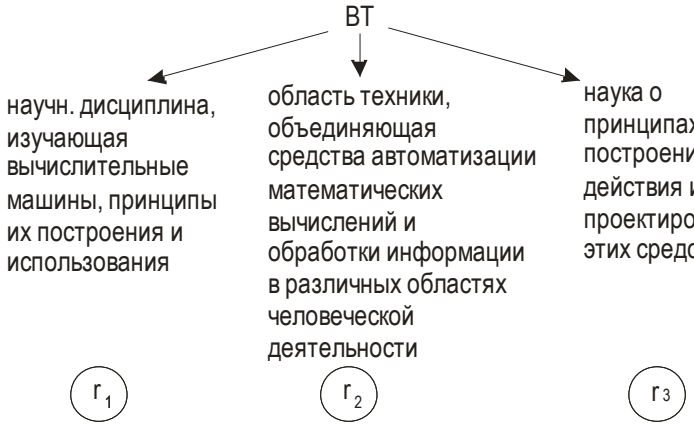
При анализе ЕЯТ первоочередным является обнаружение двух фундаментальных отношений, которые присутствуют практически в любом ЕЯТ. Это отношения эквивалентности и частичного порядка. Первое из этих отношений определяет классы синонимичных объектов, а второе отношение – иерархию подчиненности классов эквивалентности. Оба эти отношения составляют основу построения онтологий, а знания, полученные на этом этапе – будем называть первичными. В отношение частичного порядка может вкладываться различный семантический смысл: это может быть *отношение таксономии* («принадлежать» множеству, классу, группе и т. п.), *отношение партономии* («состоит из»), *отношение генеалогии* («отец-сын»), *причинно-следственное отношение* («если - то»), *атрибутное отношение* и т. д. [8]. В качестве примера рассмотрим множество определений. В энциклопедическом словаре находим такое определение понятия “НАУКА”.



$K(a) = \{p_1, p_2, p_3, p_4, p_5, p_6\}$ - класс эквивалентности.

Последующие понятия также взяты из энциклопедического словаря, где они определяются.

$$K(b) = \{q_1, q_2, q_3, q_4, q_5\}$$



$$K(c) = \{r_1, r_2, r_3\}$$

Приведенные примеры показывают, что построение классов эквивалентности не составляет особых трудностей. В результате построения классов появляются объекты:

$K(a) = \{p_1, p_2, p_3, p_4, p_5, p_6\}$, $K(b) = \{q_1, q_2, q_3, q_4, q_5\}$, $K(c) = \{r_1, r_2, r_3\}$.

Проблема появляется при вычисление второго отношения, которое определяет отношение подчиненности (иерархии) между имеющимися классами эквивалентности. Однако при таком абстрактном

представлении классов $K_i(x)$ это отношение определить нельзя. Для этого необходимо знать структурные характеристики элементов из классов $K_i(x)$. Поэтому естественным образом появляется необходимость в структуре элементов из классов эквивалентности. Например, если вернуться к вышерассмотренным примерам, то каждый элемент из класса $K(a)$ принимает вид:

$p_1 = (p_{11}, p_{12}, p_{13})$, $p_2 = (p_{21})$, $p_3 = (p_{31}, p_{32}, p_{33})$, $p_4 = (p_{41}, p_{42})$, $p_5 = (p_{51}, p_{52})$,
$p_6 = (p_{61}, p_{62})$, где

p_{11} = «сфера человеческой деятельности», p_{12} = «выработка знаний о объективной действительности»,

p_{13} = «система знаний о объективной действительности», p_{21} = «форма общественного сознания»,

p_{31} = «отрасль знаний», p_{32} = «научная дисциплина» , p_{33} = «научное направление»

p_{41} = «деятельность по получению новых знаний» p_{42} = «суммирование знаний о НКМ».

Аналогично структурируются и остальные элементы в классах эквивалентности.

$r_1 = (r_{11}, r_{12}, r_{13}, \ldots)$, $r_2 = (r_{21}, r_{22}, r_{23}, \ldots)$, $r_3 = (r_{31}, r_{32}, r_{33}, \ldots)$,

$q_1 = (q_{11}, q_{12}, q_{13}, \ldots)$, $q_2 = (q_{21}, q_{22}, q_{23}, \ldots)$, $q_3 = (q_{31}, q_{32}, q_{33}, \ldots)$, $q_4 = (q_{41}, q_{42}, q_{43}, \ldots)$

$q_5 = (q_{51}, q_{52}, q_{53}, \ldots)$

Из приведенной структуризации вытекает следующая формализация. Если класс эквивалентности относится к объекту a, то его формальное определение выглядит как дизъюнкция элементов, составляющих этот класс. Каждый элемент, входящий в тот или иной класс эквивалентности описывается соответствующим предикатом, т.е. если $K(a) = \{p_1, p_2, p_3, p_4, p_5, p_6\}$, то $p(a) \Leftrightarrow p_1(a) \vee \ldots \vee p_6(a)$, где p_i-предикаты, характеризующие элементы из класса $K(a)$, а их дизъюнкция характеризует весь класс понятия a.

Далее, если $q_i \in K(a)$ и $q_i = (q_{i1}, q_{i2}, \ldots, q_{ik})$, то элемент p_i (или объект p_i), характеризую-щийся атрибутами p_{ij}, представляется в виде конъюнкции вида

$$p_i(a) \Leftrightarrow p_{i1}(a) \wedge \ldots \wedge p_{ik}(a),$$

где $p_{ij}(a)$ - предикат, характеризующий отдельный атрибут понятия a, $i = 1, \ldots, l; j = 1, \ldots, k.$

Следовательно, каждый класс $K(a)$ описывается дизъюнктивной формой вида

$$p(a) \Leftrightarrow (p_{11}(a) \wedge \ldots \wedge p_{1m_1}(a)) \vee \ldots \vee (p_{l1}(a) \wedge \ldots \wedge p_{lm_l}(a)).$$

Введенная формализация определяет отношение частичного порядка, которое вводится следующим образом:

$$K(a) \leq K(b) \Leftrightarrow (\exists p_i(a))(\exists q_j(b))(q_j(b) \leq p_i(a)),$$

где $q_j(b) \leq p_i(a)$ означает, что $q_j(b)$ входит в виде конъюнктивного члена в $p_i(a)$.

Введенное таким образом отношения частичного порядка естественным образом требует предикатно-реляционного представления объектов из классов эквивалентности и самих этих классов [7]. Для

иллюстрации сказанного, вернемся к вышеприведенному примеру. Выясним, каким образом определяется тот факт, что классу «НАУКА» подчиняется класс «ИНФОРМАТИКА».

Класс «НАУКА», обозначенный как $K(a)$, описывается формулой

$p(a) \Leftrightarrow p_1(a) \vee p_2(a) \vee p_3(a) \vee \ldots \vee p_6(a)$, где

$p_1(a) \Leftrightarrow$ СФЕРА-ЧЕЛОВЕЧ-ДЕЯТ(*a*), $p_2(a) \Leftrightarrow$ ОТРАСЛЬ ЗНАНИЙ(*a*),

$p_3(a) \Leftrightarrow$ НАУЧНАЯ-ДИСЦИП(*a*), $p_4(a) \Leftrightarrow$ НАУЧН-НАПРАВЛ(*a*),

$p_5(a) \Leftrightarrow$ ДЕЯТ-ПОЛУЧ-НОВЫХ-ЗН(*a*), $p_6(a) \Leftrightarrow$ ДЕЯТ-СУММИР-ЗН-НКТ(*a*),

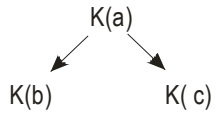
Класс «ИНФОРМАТИКА», обозначенный $K(b)$, описывается формулой

$$q(b) \Leftrightarrow q_1(b) \vee q_2(b) \vee q_3(b) \vee \ldots \vee q_l(b),$$

где $q_1(b) \Leftrightarrow p_3(a) \wedge$ ОБЩ-СВОЙ-НАУЧ-ИНФ(b) $\wedge$ ЗАКОН-ПРОЦ-НАУЧ-ДЕЯТ(b) $\wedge$ ....

Используя определенное выше отношение частичного порядка, находим, что $p_3(a)$ входит в определение класса $K(a)$, причем $q_1(b) \vee p_3(b) = p_3(b)$ (в силу закона поглощения), а это значит, что $K(b) \leq K(a)$.

Аналогично определяется подчинение класса $K(c)$ классу $K(a)$, в результате чего получаем граф



Полученную таким образом иерархию можно изменять или модифицировать путем диалога с пользователем с целью достижения более правильного представления.

Подводя итог всему сказанному, вышеописанный процесс обработки текстов определений формализуется следующим образом.

Пусть *T* – множество текстов определений. По этому множеству *T* строится множество классов, которое определяется отношением эквивалентности R и составляет фактор множество D. На полученном таким образом множестве D определяется специальные отношения $R_2, \ldots, R_k$, которые описывают характеристические свойства элементов из D, т.е. элементов из классов эквивалентности. Эти отношения представляются в виде предикатов, которые определяют отношение частичного порядка R_1. Это отношение является вторым отношением, по которому строится онтология. Более точно, онто-логия строится по транзитивному замыканию R_1^* отношения R_1, которое согласовано с отношением R.

Определение 1.Отношение R и R_1^* будемназывать согласованными, если $\forall a, b \in D$ имеет место включение $(a,b) \in R * R_1^*$, где $R * R_1^*$ - суперпозиция отношений R и R_1.

Из этого определения естественным образом следует первичная онтология: $O = (D = T / R, \Re = \{R_1, R_2, \ldots, R_k\}, \varphi, A)$, где $\varphi : D \to T$ - интерпретация, A – множество аксиом, которое определяется предикатами, описывающими характеристические свойства элементов из D, $R_2, \ldots R_k$ -соответствующие им отношения, а R_1 - отношение частичного порядка.

Обработка ЕЯТ относительно временных и объектных отношений

Рассмотрим еще один пример конкретизации отношений R и R_1, которые будем называть временными и объектными.

Отношение эквивалентности *R* определяется конкретным объектом (отсюда и название этого отношения), фигурирующим в тексте *T* (например, личность, театр, институт, вуз, кафедра и т.п.), а все объекты с ним связанные в тексте *T,* составляют класс эквивалентности по этому отношению. Отношение R_1 определяет связь объектов из классов эквивалентности отношения *R* либо в хронологическом порядке (временные зависимости), либо генеалогическом, либо каком-нибудь ином подобном порядке.

Отношение *R* и R_1 являются источником построения новых отношений. Например если R_1 описывает временные зависимости, то можно определить отношение, которое связывает объекты, относящиеся к данному конкретному моменту времени (год, месяц, день и т.п.). Поясним это примером.

Пусть $D = \{a, b, c, d\}$ - объекты, фигурирующие в данном тексте, который обрабатывается. Тогда $D / R = \{K(a), K(b), K(c), K(d)\}$ состоит из классов, элементами которых являются моменты времени t_{xi}, связанные с объектом $x \in D$, $i = 1, \ldots, j_x$. Допустим, что нас интересуют объекты, которые фигурируют в моменты времени t. Тогда момент времени t определяет отношение R_t, состоящее из объектов, связанных этим моментом времени. Арность этого отношения определяется мощностью множества D. Если $D = \{K(a), \ldots, K(b)\}$, то $R_t = \{(a', \ldots, b') \mid a' \in K(a) \wedge \ldots \wedge b' \in K(b)\}$.

Далее, на отношениях R_t определяется отношения линейного порядка: $R_t \le R_{t'} \Leftrightarrow t \le t'$. Пользуясь этим отношением порядка, можно ввести следующее понятие.

Определение 2.*Элементарным временным сценарием* для временного интервала $[t_1, t_k]$ называется цепь $R_{t_1} \le R_{t_2} \le \ldots \le R_{t_k}$, где R_{t_i} - отношения, определенные выше.

Это определение можно модифицировать в зависимости от семантики отношений R и R_1. Действительно, цепь отношений $R_{t_1} \le R_{t_2} \le \ldots \le R_{t_k}$ описывает хронологию некоторых событий во времени, связанных с конкретными людьми, объектами и т.п. Такого типа цепь в действительности может служить основой некоторого реального сценария.

Заключение

Описанные в данной работе способы автоматизации обработки ЕЯТ составляет основу как теоретического, так практического анализа извлечения знаний из ЕЯТ. Используя эту основу и прежде всего ее реализацию, предполагается наращивание ее мощности за счет построения новых мета отношений над построенными отношениями, являющимися отдельными частями знаний, имеющихся в исследуемом тексте.

Литература

1. Палагин А.В., Кривий С.Л., Петренко Н.Г., Знание ориентированные информационные системы с обработкой естественно-языкових объектов: основы методологии и архитектурно-структурная организация. – ж. УСиМ. – 2009. - №3. – С.42-55

2. Палагин А.В., Петренко Н.Г. Системно-онтологический анализ предметной области. – ж.УСиМ – 2009.№4.–С.3-14.

3.Cohen D. Jeavons P. The Complexity of Constraint Languages. In “Handbook of Constraint Programming . - Edited by F. Rossi, P. van Beek and T. Walsh. -2006. – P. 245 - 280.

4. Апресян Ю.Д. Лингвистический процессор для сложных информационных систем. -М.: Наука.-1992.- 324с.

5. Палагін О.В., Кривий С.Л., Петренко М.Г., Бібіков Д.С. Алгебро-логічний підхід до аналізу та обробки текстової інформації.-ж. «Проблемы программирования». - 2010. -№ 2.- (в печати).

6. Кулик Б.А. Логика естественных рассуждений.- С.-Петербург: Невский диалект.- 2001.- 127 с.

7. Рубашкин В.Ш. Представление и анализ смысла в информационных системах. – М.: Наука.- 1989.-188 с.

8. Gavrilova T., Laird D. Practical Design of Business Enterprise Ontologies. In Industrial Applications of Semantic Web. Eds. Bramer m., Terzyn V. - 2005. – Springer. – P. 61-81.

Информацияобавторах

Александр ВасильевичПалагин*– академик НАН Украины, заместитель директора Института кибернетики им. В.М. Глушкова НАН Украины, Украина, Киев; проспект акад. Глушкова, 40 e-mail: palagin_a@ukr.net*

Область научных интересов: Интеллектуальные информационные системы

Сергей ЛукьяновичКрывый*– профессор Киевского национального университета имени Тараса Шевченко; Украина, Киев; ул. Владимирская, 40 e-mail: krivoi@i.com.ua*

Область научных интересов: Дискретная математика, обработка знаний, анализ, верификация, оптимизация и проектирование программного обеспечения

Дмитрий СергеевичБибиков*–аспирант Института кибернетики им. В.М. Глушкова НАН Украины; e-mail: bb_coff@mail.ru*

Область научных интересов: Искусственный интеллект, автоматизация поиска доказательств в формальных логических языках