СТРУКТУРИРОВАНИЕ ОНТОЛОГИИ АССОЦИАЦИЙ ДЛЯ КОНСПЕКТИРОВАНИЯ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТОВ

Виктор Гладун, Виталий Величко, Леонид Святогор

***Аннотация**: Рассмотрен подход к конспектированию ЕЯ текстов с использованием трехуровневой онтологии ассоциаций. Предложенная структура онтологии позволяет улучшить связность конспекта.*

***Ключевые слова**: тематический анализ текста, конспектирование текста, онтология ассоциаций.*

***ACM Classification Keywords**: I.2.7 Natural Language Processing - Text analysis*

***Conference**: The paper is selected from XIVth International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008*

Введение

Процедуры репрезентации и получения новых знаний занимают в общей проблеме искусственного интеллекта ведущее место. В поисках универсальных средств конструктивного описания окружающего материального мира, адекватного человеческому мышлению и познанию, всё чаще обращаются к онтологиям. Онтология представляет собой совокупность концептов, отношений и функций интерпретации. К достоинствам онтологий можно отнести: а) глубокое взаимодействие объектов и явлений с контекстной средой; б) экономное хранение информации, требующее запоминания концептов и отношений, а не сцен; в) универсальный характер онтологии, допускающий использование её структуры в качестве инструмента для решения задач семантического анализа естественно-языковых (ЕЯ) текстов. Одной из таких задач является *конспектирование текста* как способ сжатого и релевантного представления содержания дискурса. В качестве примера программы выполняющей конспектирование русскоязычного текста можно привести программу КОНСПЕКТ [1], в которой для указания семантических отношений между словами используются ассоциативные связи, а множество ассоциативных признаков отражает некоторые категории внешнего мира (например, ассоциации по роду деятельности, по времени и другие). Однако слабым местом системы ассоциативных связей является упрощённая, одноуровневая структура, которая не учитывает глубинную иерархию понятий, описывающих реальный мир. Поле ассоциаций представляет собой просто *список концептов*, размещённых в алфавитном порядке.

Отсутствие структуры над полем ассоциаций приводит к потере глубоких ассоциаций и, кроме того, не позволяет в рамках выбранной схемы оперировать элементами смыслового сопровождения при раскрытии содержания текста через конспект.

Задача заключается в том, чтобы превратить систему ассоциативных связей в иерархическую *структуру репрезентации знаний*, адекватную иерархии понятийного аппарата человека, и использовать её затем для семантического анализа ЕЯ текстов.

Обоснование подхода и решаемые задачи

В настоящее время предложено много вариантов описания мира, которые опираются на тезаурусы и онтологии [2]. Для репрезентации знаний об Универсуме они широко используют философские категории и предназначены либо для систематизации лексического богатства естественного языка (словари Роже, Дорнзайфа, Идеографический словарь русского языка О.С.Баранова [3]), либо для структуризации знаний человека о мире (онтологии «Mikrokosmos», SUMO, Дж. Совы [4]). Существуют и проблемно-ориентированные онтологии, которые приспособлены, например, для построения языково-онтологических информационных систем [5, 6]. Для построения многоуровневых онтологий используются методы и

математические модели, содержащие модели онтологии, знаний и действительности предметной области (Про) [7]. Предлагаемая здесь структурированная онтология ассоциаций (ОнтА) служит для решения более частной задачи – *тематического анализа текстов* [8].

При построении ОнтА авторы исходили из следующих предпосылок: а) научной методологией репрезентации окружающего мира должен быть *системный анализ;* б) терминология (категории и концепты онтологии) должны в основном базироваться на понятиях, которые установились в *естественных науках,* с привлечением, в необходимых случаях, философских категорий; в) для конструктивного практического использования структура онтологии должна быть иерархической и содержать верхний, средний и нижний уровень иерархии понятий.

Системный анализ. Понятийно-содержательный подход представления знаний, присущий системному анализу, в отличие от формально-математического, важен именно с позиций выявления семантических отношений. При анализе *системно-информационной картины мира* [9] рассматривают следующие основные *типы ресурсов* в природе и обществе: *Вещество* – субстанция, отображающая состояние материи; *Энергия* – характеристика движения материи; *Информация* – мера порядка и самоорганизации материи; *Человек* – уникальный ресурс общества, субъект осознания материи, мера интеллекта; *Организация* – форма упорядоченности ресурсов и существования системы; *Пространство* – мера протяжённости (распространения и распределения) материи; *Время* – мера существования состояния материи (вещества).

Эти ресурсы, на наш взгляд, обладают необходимой *содержательной строгостью*, поскольку являются объектами исследований в физических и социальных науках. Они могут быть использованы при синтезе онтологии ассоциаций в качестве категорий верхнего уровня. Тем самым намечено существенное отличие ОнтА от известных моделей описания мира, которые были упомянуты выше.

Принципы построения онтологии. Задачи исследования и цель

С учётом рекомендаций, которые необходимо выполнять при построении онтологии [6, 10, 11], в данной работе предлагается трёхуровневая онтология для решения задачи конспектирования ЕЯ текстов. В основу разработки положены следующие принципы.

1. *Принцип полноты.* Категории верхнего уровня должны исчерпывающим образом представлять Материю; за пределами этих категорий не должно существовать никаких проявлений сущего.

2. *Принцип естественнонаучности и проблемной ориентации*. Все категории и концепты онтологии должны быть выражены понятиями, которые установились в естественных и математических науках при изучении материального мира и являются общепринятыми. При этом часть онтологии должна быть представлена концептами, которые широко используются в междисциплинарных текстах (с нейтральной, общедоступной лексикой), а вторая часть онтологии структурируется под конкретную область знаний (ПрО). Первая часть имеет постоянный статус, а проблемно-ориентированная онтология формируется специалистом и носит переменный характер.

3. *Принцип взаимосвязанности уровней.* Категории онтологии верхнего уровня раскрываются наборами концептов среднего уровня. В свою очередь, концепты нижнего уровня должны служить определителями для терминов словаря ПрО. Связь между средним и нижним уровнями организуется с помощью именованных отношений вида: «быть частью», «принадлежать множеству», «совпадать с», «находиться в семантическом отношении с».

4. *Принцип ассоциативности*. Концепты онтологии нижнего уровня должны служить полем для индексирования терминов ПрО. При этом используются семантические отношения вида: «находиться в ассоциативной связи с».

5. *Принцип отражения антагонизмов*. Концепты, которые отражают свойства или понятия, имеющие свою противоположность или дополнительность по равному основанию, входят в онтологию парами или тройками полярных обозначений.

На основе сформулированных предпосылок становится ясной следующая перспектива действий. Необходимо выбрать категории, концепты и связи между ними для верхнего и среднего уровней ОнтА. Необходимо создать поле концептов нижнего уровня – внести в него термины ПрО. Согласовать поле концептов нижнего уровня с концептами среднего уровня. На заключительном этапе следует связать нижний уровень онтологии с базой данных естественного языка, для чего необходимо найти актуальные ассоциативные связи между словарём основ русского языка и терминами, выбираемыми из поля концептов нижнего уровня. Процесс установления ассоциативных связей называется индексацией словаря.

В результате выполнения этих действий онтологическая структура становится конструктивной для процедур семантического анализа и раскрытия темы. Она замкнёт слова русского языка (слова, взятые из текста) через их индексы (связи с концептами нижнего уровня) на *траектории* внутри сетевых структур нижнего, среднего и верхнего уровней. Это позволит, кроме составления самого конспекта, сопроводить его комментариями, которые будут активизированы на траекториях сетевых структур и тем самым улучшить семантическую компоненту конспекта.

Цель структурирования Онтологии ассоциаций заключается в том, чтобы создать трёхуровневую иерархическую онтологическую систему, которая в сжатом виде отражает актуальные знания о структуре внешнего мира, ориентирована на обработку корпуса текстов как общего (междисциплинарного), так и проблемно-ориентированного характера, и позволяет более глубоко раскрыть тему при конспектировании текста.

Выбор категорий верхнего уровня онтологии ассоциаций

Существует много подходов к дихотомии Мира. Нам представляется наиболее конструктивной идея, выдвинутая академиком Вернадским, который построил материалистическое мировоззрение как единство *Косного вещества, Биосферы и Ноосферы.* В качестве методологии онтологического синтеза, как указывалось выше, принят системно-аналитический подход.

Узловой точкой общей картины мира является философская категория *Материя*. Она может быть исчерпывающим образом представлена тремя категориями: *Вещество* (косное), *Энергия, Жизнь* (субстанция живого). Каждая из трех категорий представлена рядом подкатегорий, как показано на рис. 1. Используемые выше категории верхнего уровня образуют *кластеры понятий* для развития и усложнения онтологии. Приведенную онтологическую структуру необходимо снабдить некоторыми пояснениями.

Первое. Общепринятыми в теории познания являются такие важные философские понятия, как *Материя, Бытие, Сознание, Субстанция, Субстрат, Мера, Пространство, Время, Состояние, Свойство, Количество, Качество* и другие. В ОнтА почти все эти понятия, или эквивалентные им, перенесены на средний уровень, благодаря чему они освобождаются от чисто философского смысла и «работают» как термины естественных наук. Например, пространство и время присутствуют как конкретные признаки *локализации* объектов и явлений в четырёхмерном координатном пространстве. Термины количество и качество определяются при помощи *меры* и так далее.

Второе. На прагматическом уровне можно показать, что предлагаемая онтология обладает свойством *полноты.* Будем исходить из того, что всё, что мы знаем о свойствах *Материи*, заключено в следующих четырёх постулатах. *Материя*: а) *существует* как объективная реальность, б) *проявляет себя* в движении и развитии, в) *распределена* в пространственно-временном континууме и г) *отображается* разумом. В таком случае формами проявления *Материи* служат *Вещество, Жизнь и Энергия*. В свою очередь, *пространство* и *время* служат формами распределения *Материи*. Все пять форм проявления и распределения материи интегрируются в едином поглощающем понятии – *Бытие Материи*. Следовательно, онтология, замыкаясь на понятие *Бытия Материи*, исчерпывающим образом отображает все известные (или сущие) свойства данной субстанции, то есть – является *полной системой* верхнего уровня репрезентации знаний.

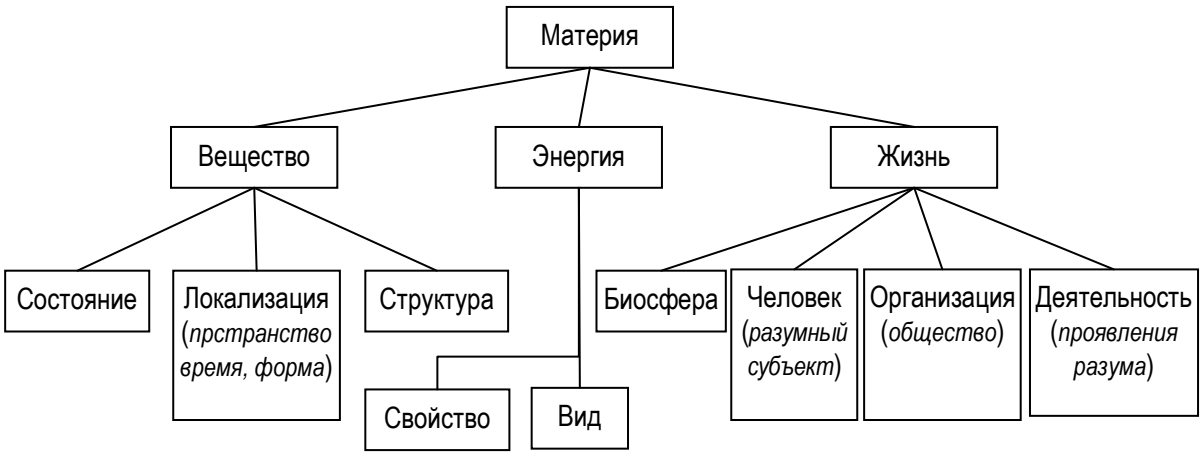

Рис. 1. Структура онтологии ассоциаций верхнего уровня.

Третье. В онтологии ассоциаций важную роль играют такие понятия, как *мера* и *имя*: они пронизывают все уровни онтологии и характеризуют большинство концептов. Здесь понятие меры используется не в математическом и не в философском смысле (как связь между количеством и качеством), а как результат измерения; имя обозначает множество или кластер. Иногда можно проследить тесную связь между *мерой* и *именем*: если число служит точечной оценкой количества и часто несёт избыточную информацию, то имя задаёт сразу диапазон измерений, то есть – кластер. Например, на шкале температуры различают состояния: *жара, тепло, нормально, прохлада, холод, мороз;* все они обладают ясной и непосредственной семантикой. В онтологии ассоциаций перечислять все состояния системы (кроме случаев, когда они существенны для представления ПрО) нет необходимости; они будут обобщены понятиями: *мера тепла* или *имя состояния* или *свойство системы*. Хотя ОнтА оперирует с однословными концептами, в некоторых случаях применяются сложные концепты (например, *рождение-гибель*), когда пара или тройка связанных (по равному основанию и противоположных по смыслу) имён подчёркивает область определения сложного концепта. В общем случае имя служит идентификатором состояния системы, идентификатором множества или абстрактным понятием.

Уникальным свойством человеческого языка является передача смысла «по умолчанию». Эту функцию выполняет *мера*, когда она сопровождается контекстом. Например, слово «нормально» обозначает не только ситуацию, но и подразумевает *отклонения от нормы* в обе стороны, то есть – несёт значительную семантическую нагрузку. Авторы исследования разделяют гипотезу, что познание человеком бесконечно-разнообразной внешней среды возможно благодаря его умению классифицировать. Отсюда следует необходимость и универсальность семантической категории *имя*, которое может существовать только в общей языковой среде.

С учётом приведенных выше универсальных категорий, понятий и комментариев строится продолжение онтологической схемы.

Построение онтологии среднего уровня

Онтология среднего уровня (ОСУ) должна связывать категории верхнего уровня сложности с концептами, которые описывают конкретные свойства ПрО на нижнем уровне. Промежуточный уровень иерархии необходим для более глубокого и разветвлённого раскрытия общих связей и закономерностей, накопленных при изучении в разных дисциплинах. По-сути, он представляет собой *слой междисциплинарного человеческого знания* и обобщает коллективный опыт. Более того, проекция нижнего уровня онтологии на средний позволяет раскрыть содержательную компоненту ЕЯ текста и

одновременно усилить её объяснительной компонентой онтологии ассоциаций. От удачного построения этой (средней) части онтологической структуры зависят в целом интерпретационные возможности системы семантического анализа.

ОСУ представляет собой конструкт, заполняемый один раз концептами общего назначения. Это не исключает его доработку специалистами разных областей знания. Инженер по знаниям имеет право согласовывать общий уровень онтологии ассоциаций с теми профессиональными знаниями ПрО, которые он будет детально формулировать на нижнем уровне иерархии. Фактически ОСУ выступает в качестве постоянной составляющей онтологии ассоциаций. В отличие от неё, онтология нижнего уровня является переменной составляющей.

Описание структуры ОСУ. Онтология среднего уровня представляет собой совокупность сетевых структур: именем каждой структуры служит категория верхнего уровня, узлами являются концепты среднего уровня, а внутренние связи раскрывают (характеризуют) основные свойства категории.

Заполнение узлов ОСУ произведено такими понятиями, которые являются общеупотребительными для обозначения элементов знания и имеют определённый смысл для специалистов разных, в том числе гуманитарных, областей. Однако задачей данного слоя не является полный охват этих элементов, наоборот: углубление в предметную область достигается средствами нижнего уровня. Ориентиром для выбора концептов ОСУ является семантический анализ дискурсов общетематической направленности.

Каждая структура ОСУ содержит как безусловные, так и сомнительные связи, которые могут быть скорректированы экспертом. Это не является недостатком онтологии ассоциаций, а придаёт ей динамический характер. Кроме того, эксперт по своему усмотрению может свободно выбирать и смешивать концепты и отношения разных типов (род-вид, часть-целое и т.д.); при этом он преследует цель выявления значимых семантических ассоциаций для выразительной репрезентации знаний о данной категории. Многие отношения в категориях среднего уровня представлены отглагольными существительными и могут повторяться в раскрытии структуры различных категорий. «Сила связи» между родовидовыми понятиями, понятиями типа часть-целое выше, чем между понятиями, связанными отглагольными существительными. Введение связей, представленных отглагольными существительными, позволяет расширить цепочки ассоциаций и повысить связность текста, отобранного в конспект.

Здесь, в целях экономии места, полные списки концептов ОСУ (от десяти до тридцати слов на категорию) вместе с их связями не приводятся, а концепты в скобках даны выборочно. В качестве примера представлен фрагмент графа связи понятий категории *Биосфера* (см. рис. 2).

Состояние = **{**устойчивое – изменчивое; покой – движение; конечное – промежуточное; твёрдое – жидкое – газообразное;…**}**. *Структура* = **{**система; однородность – неоднородность; форма; содержание;…**}**. *Локализация* = **{**пространство; время; распределение; начало – конец;…**}**. *Свойство энергии* = **{**движение – покой; процесс – акт; действие – противодействие; превращение – сохранение;…**}**. *Вид энергии* = **{**тепловая; кинетическая – потенциальная; созидание – разрушение;…**}**. *Биосфера* = **{**существо; организм; растение; животное; популяция; среда; экология; существование; выживание; эволюция; развитие-вырождение; размножение; потомство; рождение – гибель; опасность – безопасность; борьба;…**}**. *Человек* = **{**организм; эмоции; разум; характер; воля; занятие;…**}**. *Организация* = **{**социум; управление; семья; закон; свобода – необходимость; …**}**. *Деятельность* = **{**теория – практика; работа – занятие; познание; информация; тактика – стратегия; перемещение; …**}**.

Главное назначение ОСУ состоит в том, чтобы раскрывать категории верхнего уровня и одновременно интерпретировать в сжатом виде концепты нижнего уровня онтологии ассоциаций.

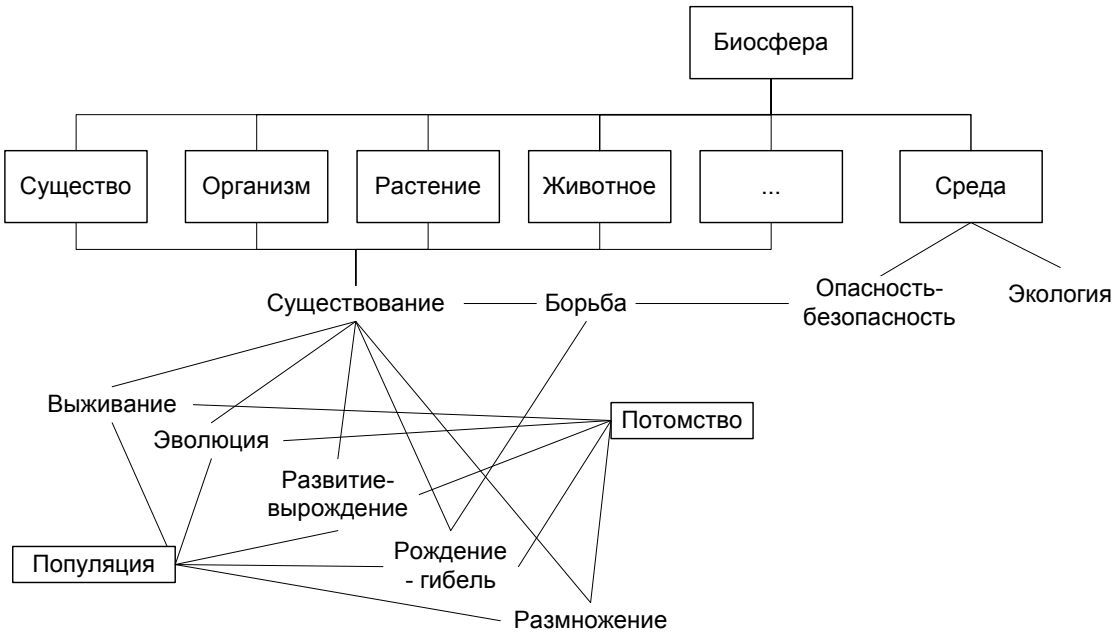

Рис. 2. Структура онтологии ассоциаций среднего уровня для категории *Биосфера*.

Онтология нижнего уровня.

Онтология нижнего уровня (ОНУ) предметной области представляет собой таблицу, в которой слова из словаря основ русского языка (из базы данных) напрямую связаны с ограниченным множеством концептов предметной области, причём концепты ПрО, в свою очередь, имеют восходящие связи к онтологии среднего уровня. Благодаря двусторонним связям в ОНУ база данных через таблицу включается в полную онтолого-ассоциативную структуру. Слова русского языка, которые помещены в словарь основ и употребляются в текстах, оказываются с помощью семантических отношений сопряженными с категориями и концептами всех уровней. Созданный конструкт позволяет решать задачи семантического анализа ЕЯ текста.

Проблема выбора поля концептов нижнего уровня здесь подробно не рассматривается, поскольку выбор зависит от специалиста, который решает определённую задачу. Специалист (эксперт) в своей области выбирает термины и определения из доступных ему источников: учебников, толковых словарей, монографий и т.д. и формирует поле ПрО. Это поле может быть структурировано. Важно, чтобы термины ПрО были некоторым образом связаны с ОСУ, то есть, чтобы ПрО не оказалась изолированной от верхних структур ОнтА.

Например, в определении семантической сети, взятом из толкового словаря по вычислительным системам [12], эксперт выбрал термины: *представление знаний, помеченный граф, вершина графа, понятие, концепт.* Именно эти термины будут соединены с концептом *семантическая сеть* ОНУ и использованы для отбора предложений из дискурса в конспект по теме «семантическая сеть». Концепт *семантическая сеть*, в свою очередь, может быть связан с концептами среднего уровня онтологии, например – *информация* или *наука* из категории *деятельность* ОВУ. Структура онтологии нижнего уровня создаётся в зависимости от требуемой детализации в выделении тематической направленности текста.

Технология конспектирования текста состоит в следующем. На вход системы семантического анализа поступает очередное значимое слово, которое выбрано из дискурса. Оно активизирует нужные связи онтологии ассоциаций и на каждом уровне иерархии возбуждает определённые фрагменты сети. Траектория возбуждения запоминается и используется затем либо для более глубокой интерпретации текста, либо как инструмент для нового раскрытия темы – с учётом уже найденных концептов.

Заключение

Разработанная трёхуровневая иерархическая структура онтологии ассоциаций является сетью, определяемой совокупностью связанных между собой категорий, понятий, концептов и основ русского языка. Её создание подчинено задаче тематического анализа текстов как общей природы, так и проблемно-ориентированных. Онтология верхнего уровня отражает системную картину мира и служит базой для развития онтологии. Онтология среднего уровня обслуживает, в основном, континуум и структуру междисциплинарных знаний. Проблемная область конструируется специалистом на нижнем уровне онтологии, где систематизируются терминологические и специальные знания, взятые из профессиональных источников. Все три уровня иерархии знаний замыкаются на базу данных естественного языка. Процедура концептуального тематического анализа текста состоит в том, что для очередного значимого слова, которое выделено в тексте, на всех уровнях иерархии активизируются соответствующие ему концепты, принадлежащие определённой семантической траектории, после чего данная траектория используется для более глубокого раскрытия темы при конспектировании текста.

Литература

1. Гладун В.П., Величко В.Ю., Святогор Л.А. Тематический анализ естественно языковых текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006» Бекасово, 2006 г. – М.: Изд-во РГГУ.–2006. – с.115-118.
2. Нариньяни А.С. Кентавр по имени ТЕОН: Тезаурус + Онтология// Труды международного семинара Диалог'2001 по компьютерной лингвистике и ее приложениям. -Т.1. –Аксаково, 2001. – с.184-188.
3. Баранов О.С. Идеографический словарь русского языка. М.–2002, 1200 с.
4. John F. Sowa, Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA, ©2000. – 594p.
5. Палагін О.В., Петренко М.Г. Модель категоріального рівня мовно-онтологічної картини світу //Математичні машини і системи. –2006. – №3. – с.91-104.
6. Палагін О.В., Петренко М.Г. Розбудова абстрактної моделі мовно-онтологічної інформаційної системи //Математичні машини і системи. –2007. – №1. – с.42-50.
7. Артемьева И.Л. Многоуровневые модели предметных областей и методы их разработки // Десятая нац. конф. по искусственному интеллекту с междунар. участием, Обнинск, 25-28 сентября 2006: сб. тр. в 3-х томах. Москва: Физматлит. –2006. Т.1. – с.44-51.
8. Штерн І.Б. Вибрані топіки та лексикон сучасної лінгвістики. Енцикл. словник. – К.: “АртЕк”. –1998. – 336 с.
9. Казиев В.М. Введение в анализ, синтез и моделирование систем. – ИНТУИТ.ру, БИНОМ. Лаборатория знаний. – 2006. – 248с.
10. Соловьева Е.А. Естественная классификация: системологические основания. Под ред. М.Ф. Бондаренко. – Харьков. ХТУРЭ. –1999. – 222с.
11. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. – СПб.: Питер. –2001. – 384с.
12. Толковый словарь по вычислительным системам. Под ред. В.Иллингуорта и др.: Пер. с англ. А.К. Белоцкого и др. Под ред. Е.К.Масловского и др.: –М.: Машиностроение. –1989. – 568с.

Информация об авторах

Гладун Виктор Поликарпович *- Ин-т кибернетики им. В.М. Глушкова НАН Украины, Киев-187 ГСП, 03680, просп. акад. Глушкова, 40,e-mail: glad@aduis.kiev.ua*

Величко Виталий Юрьевич *- Ин-т кибернетики им. В.М. Глушкова НАН Украины, Киев-187 ГСП, 03680, просп. акад. Глушкова, 40,e-mail: vitaly@aduis.kiev.ua*

Святогор Леонид Александрович *- Ин-т кибернетики им. В.М. Глушкова НАН Украины, Киев-187 ГСП, 03680, просп. акад. Глушкова, 40,e-mail: glad@aduis.kiev.ua*