Определение семантического содержания предметной области на основе формирования тезауруса

В.В. Баранюк, А.В. Богорадникова, О.С. Смирнова

Аннотация **– В статье представлены результаты исследования по отбору терминов для включения в тезаурус в рамках работ по созданию основы для построения базы знаний в области бионики. В исследовании извлеченный из текстового корпуса тезаурус использовался для построения семантической сети. С целью формирования легко воспринимаемого определения предметной области по средствам соотнесения слова с другими понятиями и их группами размерность семантической сети в последующем была снижена за счет исключения элементов, обладающих низкой релевантностью.**

Ключевые слова **– тезаурус, система информационной поддержки, база знаний, бионика, бионические технологии.**

При исследовании вопросов создания интеллектуальной системы информационной поддержки процессов создания и развития перспективных бионических технологий основной предметной областью является область бионики, науки о применении в технических системах принципов организации, свойств, функций и структур живой природы.

Бионика рассматривает биологию и технику совсем с новой стороны, объясняя, какие общие черты и какие различия существуют в природе и в технике. В настоящее время различают:

– *биологический* аспект бионики, изучающий процессы, происходящие в биологических системах;

– *теоретический (модельный)* аспект бионики, который строит математические модели этих процессов;

– *технический (инженерный)* аспект бионики, применяющий модели теоретического аспекта для решения инженерных задач.

Бионика изучает биологические системы с целью применения полученных знаний для решения различных инженерных задач. Для этого сначала строятся бионические модели, на которых можно исследовать реализуемость технической системы или устройства, за короткое время обработать различные параметры и устранить конструктивные недостатки.

Накопленный в бионике опыт моделирования чрезвычайно сложных систем имеет большое значение. Для создания таких моделей требуется большое количество сведений о живых организмах, различных характеристик об их состояниях, движении, жизнедеятельности, органах чувств и др. Накопление таких сведений целесообразно осуществлять в специальных базах знаний. Составной частью интеллектуальной системы информационной поддержки процессов создания и развития перспективных бионических технологий является база знаний в области бионики. В ряде публикаций [1, …, 9] рассматривались различные вопросы, связанные с её построением.

В настоящей статье представлены результаты исследования по отбору терминов для включения в тезаурус в рамках работ по созданию основы для построения базы знаний в области бионики.

В рамках исследований, проводимых при создании интеллектуальной системы, рассматривался процесс формирования базы знаний предметной области с использованием тезауруса, извлекаемого из текстового корпуса. Фрагмент текстового корпуса включал пять следующих книг:

– Асташенков П.Т. «Что такое бионика»;

– Варшавский В.И., Поспелов Д.А. «Оркестр играет без дирижера»;

– Литинецкий И.Б. «Беседы о бионике»;

– Решодько Л.В. «Бионика. Биологические аспекты»;

– Агнес Гийо, Жан-Аркади Мейе. «Бионика. Когда наука имитирует природу».

Вначале был проведён последовательный анализ каждого раздела и формирование на основании каждого из них лингвистического корпуса из терминов данной предметной области. Получен агрегированный результат, отражающий тезаурус предметной области, небольшой фрагмент которого представлен в таблице 1. Он включает список ключевых понятий с выделением термина, его определения, частоты встречаемости и ссылки на термины, имеющиеся в его определении (выделены курсивом).

Статья получена 10.08.2016 г.

Исследование выполнено федеральным государственным бюджетным образовательным учреждением высшего образования «Московский технологический университет» (МИРЭА) за счет гранта за счет гранта Российского научного фонда (проект №14-11-00854).

К.т.н., с.н.с., В.В. Баранюк, МИРЭА (e-mail: valentina_bar@mail.ru).

А.В. Богорадникова, МИРЭА (e-mail: bogoradnikova@mirea.ru).

О.С. Смирнова, МИРЭА (e-mail: mail.olga.smirnova@yandex.ru).

Таблица 1 – Фрагмент тезауруса

Термин	Определение	Частота
Система	Группа взаимодействующих *объектов*, выполняющих общую *функциональную задачу*. В ее основе лежит некоторый *механизм связи*.	854
Бионика	*Наука*, занимающаяся изучением принципов построения и *функционирования биологических систем* и их *элементов* и применением полученных знаний для коренного усовершенствования существующих и создания принципиально новых *машин*, приборов, аппаратов, строительных *конструкций и технологических процессов*.	629
Модель	*Система, исследование* которой служит средством для получения *информации* о другой *системе*. Представление некоторого реального *процесса, устройства* ил и концепции.	535
Автомат	Техническое *устройство*, которое может выполнять *действия* заложенной в нём программы без непосредственного участия человека.	533
Сигнал	Носитель *информации*, используемый для передачи сообщений в *системе связи*.	517
Эксперимент	*Метод исследования* некоторого явления в управляемых наблюдателем *условиях*. Отличается от наблюдения активным *взаимодействием* с изучаемым *объектом.*	455
Информация	Любые сведения и данные, отражающие *свойства объектов* в природных (биол., физ. И др.), социальных и техн. *системах* и передаваемые *звуковым*, графическим (в т. Ч. Письменным) или иным способом без применения или с применением техн. средств.	409
Задача	Проблемная ситуация с явно заданной целью, которую необходимо достичь; в более узком смысле задачей также называют саму эту цель, данную в рамках проблемной ситуации, то есть то, что требуется сделать.	388
Метод	Способ теоретического *исследования* или практического осуществления чего-либо.	366

С использованием полученных результатов была сформирована семантическая сеть, фрагмент которой представлен на рисунке 1.

Далее ставилась задача снижения размерности семантической сети за счёт исключения элементов, обладающих низкой релевантностью.

Для этого было необходимо:

– определить коэффициент связности для каждого термина;

– исключить термины, обладающие низкой связностью.

Для снижения размерности графа использовался принцип Парето, названный в честь итальянского учёного Вильфредо Парето. Этот принцип в наиболее общей формулировке звучит следующим образом: *«20 % усилий дают 80 % результата, а остальные 80 % усилий — лишь 20 % результата»*. Принцип нашёл широкое применение в разных областях деятельности и помог снизить размерности при описании и изучении сложных систем.

Рассматриваемый фрагмент тезауруса данной предметной области составляет сто семьдесят два термина. Для того, чтобы выделить ключевые слова, составляющие его основу, использован принцип Парето, который интерпретирован следующим образом: двадцать процентов наиболее встречаемых терминов определяют остальные восемьдесят процентов. Эти двадцать процентов выделены на графике, представленном на рисунке 2.

Двадцать процентов тезауруса составляют 35 ключевых слов, с высокими частотами встречаемости. Данная выборка позволила построить более простой и легко воспринимаемый семантический граф, представленный на рисунке 3.

Однако, анализ этого графа показал, что данный результат не в полной мере отражает желаемое представление о предметной области и требует дальнейшей доработки применительно к рассматриваемой задаче построения базы знаний интеллектуальной системы информационной поддержки процессов создания и развития перспективных бионических технологий.

При этом графическое отображение полученного тезауруса в области бионики и бионических технологий и его программная реализация позволят наглядно отобразить место термина в системе понятий рассматриваемой предметной области, ссылки и связи данного термина, а также привязку термина к источникам (информационным ресурсам).

Рисунок 1 – Понятия и связи между ними, представленные в виде графа

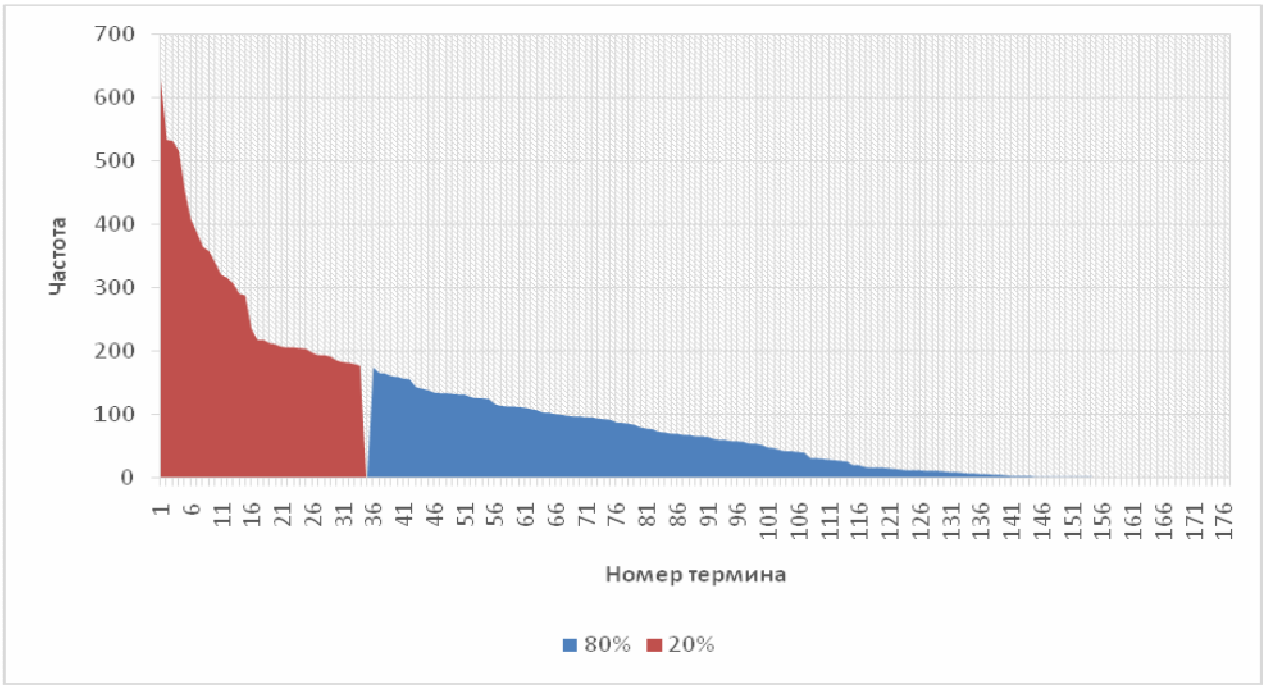

Рисунок 2 – Графическая интерпретация закона Парето применительно к фрагменту тезауруса

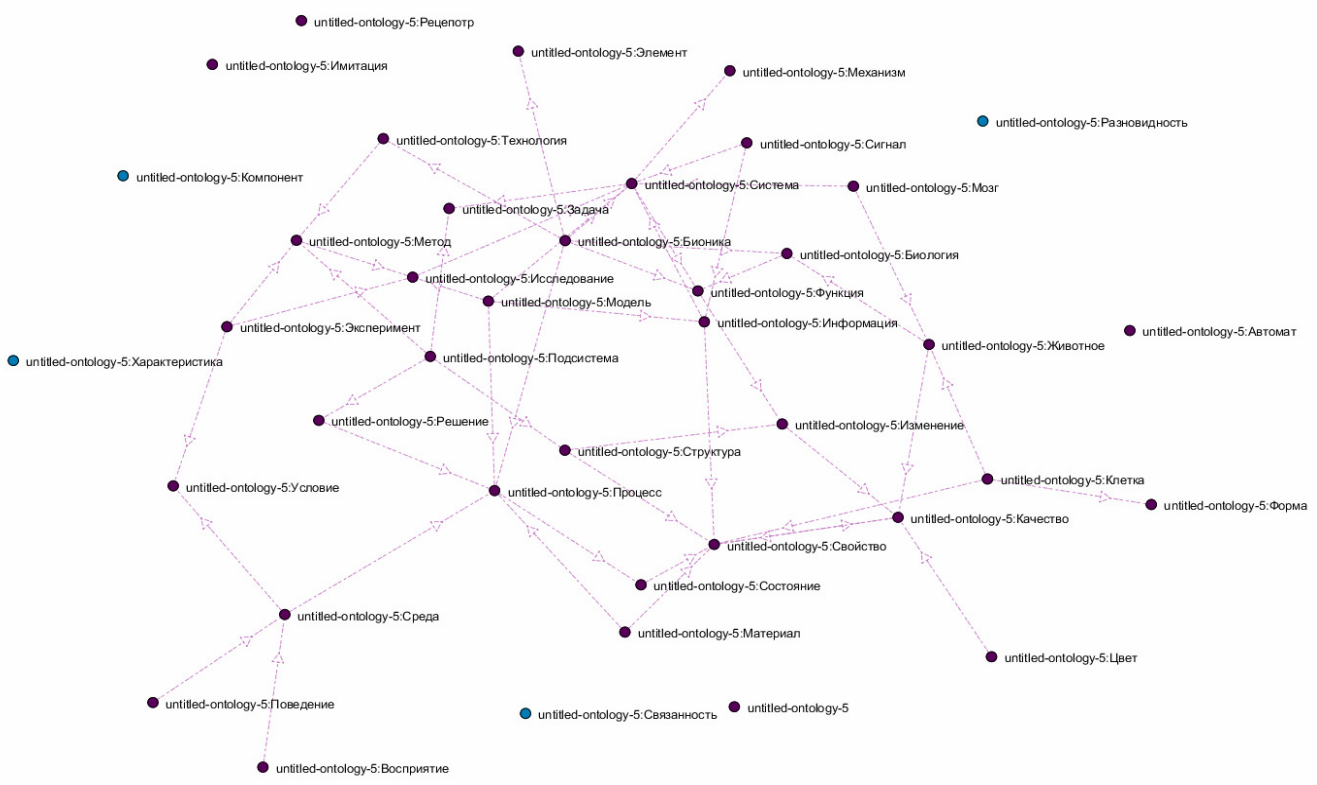

Рисунок 3 – Граф, состоящий из 20% наиболее встречаемых терминов

БИБЛИОГРАФИЯ

[1] A.S. Sigov, V.V. Nechaev, V.V. Baranyuk, M.I. Koshkarev, A.A. Melikhov, O.S. Smirnova, A.V. Bogoradnikova Architecture of domain-specific data warehouse for bionic information resources // Ecology, Environment and Conservation Paper. Vol. 21 Nov. 2015 Suppl. Issue, pp 181 – 186.

[2] Мелихов А.А., Нечаев В.В. Пополнение базы знаний интеллектуальной системы информационной поддержки развития перспективных бионических технологий: формирование перечня источников. Научный и общественно-информационный журнал «Информационные и телекоммуникационные технологии» №28, 2015. с. 16 – 20.

[3] Баранюк В.В., Смирнова О.С. Роевой интеллект как одна из частей онтологической модели бионических технологий. International Journal of Open Information Technologies. Том 3, № 12 (2015), с. 13 – 17.

[4] БаранюкВ.В., Смирнова О.С. Детализация онтологической модели по роевым алгоритмам, основанным на поведении насекомых и животных. International Journal of Open Information Technologies. Том 3, № 12 (2015), с. 18 – 27.

[5] Смирнова О.С., Богорадникова А.В., Блинов М.Ю. Описание роевых алгоритмов, инспирированных неживой природой и бактериями, для использования в онтологической модели. International Journal of Open Information Technologies. Том 3, № 12 (2015), с. 28 – 37.

[6] Смирнова О.С., Елисеева Е.И., Ершова О.А., Сесин И.Ю. Подходы к классификации информационных ресурсов в области бионических технологий. Национальная ассоциация ученых (НАУ). Ежемесячный научный журнал№ 4 (9) / 2015. Часть 3. Труды IX Международной научно-практической конференции «Отечественная наука в эпоху изменений: постулаты прошлого и теории нового времени». г. Екатеринбург, 15-17 мая 2015 г, с.18-22.

[7] Нечаев В.В., Баранюк В.В., Смирнова О.С., Кошкарев М.И., Володина А.М., Богорадникова А.В., Маркелов К.С. Учебное пособие «Информационные ресурсы и технологии» по курсам «Базы данных», «Хранилища данных и OLAP-технологии», «Системный анализ», «Информационные технологии» для студентов, обучающихся по направлению 09.03.04 «Программная инженерия». Изд.: ФГБОУ ВО «МИРЭА». 2015. – 92 с.

[8] Баранюк В.В., Смирнова О.С., Богорадникова А.В. Интеллектуальная система информационной поддержки развития перспективных бионических технологий: основные направления работ по созданию. International Journal of Open Information Technologies. Том 2, №12, 2014, с.17 – 19.

[9] Сигов А.С., Нечаев В.В., Кошкарев М.И. Архитектура предметно-ориентированной базы знаний интеллектуальной системы. International Journal of Open Information Technologies. Том 2, №12, 2014, с.1 – 6.

[10] Войшвилло Е.К. Понятие как форма мышления: логико-гносеологический Анализ. – М.: Изд-во МГУ, 1989. – 239 с.

[11] Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – Москва: МГУ, 2011. – 512 с.

[12] Онтологии и тезаурусы: модели, инструменты, приложения: учебное пособие /Б.В. Добров, В.В. Иванов, Н.В. Лукашевич, В.В. Соловьёв. – М.: Интернет-Университет Информационных Технологий; БИНОМ. Лаборатория знаний. 2009. – 173 с.: ил. – (Серия «Основы информационных Технологий»).

Defining the scope semantics by forming its thesaurus

V.V. Baranjuk, A.V. Bogoradnikova, O.S. Smirnova

Abstract – In this paper, we present results of the study aimed for bionics knowledge base research and development. The former article regards the problem of classification and tagging for certain chunks of text, related to bionics. The proposed method involves application of a scope-specific thesaurus which is extracted from the from the initial text corpora. This thesaurus is then used for matching with the data extracted from the text chunks and populating a set of tags, which can be then reduces by eliminating units with low relevance.

Keywords – thesaurus, information support system, knowledge base, bionics, bionic technologies.