# Разработка комплекса автоматизированного построения терминосистемы

В. А. Суворова, В. В. Бахтин, Е. В. Исаева
(Пермский государственный национальный
исследовательский университет)

В статье описан разработанный научным коллективом программный комплекс, предназначенный для построения и анализа терминосистемы. Проект представляет собой экспертную систему, содержащую словарь-тезаурус (TSReader), систему идентификации и классификации терминов (TSBuilder), терминологическую базу данных. Автоматизированное создание терминосистемы достигается благодаря использованию эвристического алгоритма идентификации терминов и видоизмененного персептрона Розенблатта для классификации терминов.

**Ключевые слова:** интеллектуальный анализ текста, нейросетевое моделирование, моделирование терминосистемы, автоматизированная идентификация терминов, автоматизированная категоризация терминов.

## Введение

Процесс систематизации знаний имеет огромное значение в освоении любой научной области, добиться увеличения эффективности данного процесса можно, в частности, через классификацию и систематизацию терминологии [1]. Именно поэтому имеется потребность в создании терминосистем различных научных областей, а вопрос об автоматизации построения и пополнения терминосистемы на сегодняшний день является особенно актуальным.

В статье представлен опыт упорядочения терминологии на примере конкретной области знаний — компьютерной вирусологии, посредством разработки программного комплекса, включающего:

- терминологическую базу данных;
- словарь-тезаурус;
- систему идентификации и классификации терминов.

### Терминологическая база данных

Основополагающее значение в разработке комплекса автоматизированного построения терминосистемы имеет база терминов компьютерной вирусологии, содержание которой отражено в словаре-тезаурусе, а также используется для автоматизированной идентификации и классификации терминов. Разработанная база данных включила в себя следующие основные разделы: термин на английском языке, перевод на русский язык, дефиницию на английском языке, контекст на английском языке, ссылку на источник контекста (автор статьи/монографии, название публикации, издание, дата публикации, страница употребления термина, URL электронной публикации), ссылку на источник дефиниции [2].

Информационное наполнение описываемой базы осуществлялось студентами ПГНИУ компьютерных специальностей путем выделения терминов в процессе прочтения англоязычных книг по компьютерной безопасности.

Взаимодействие с базой данных осуществлялось при помощи интерфейса ИС «Семограф» (http://semograph.com/), на сервере которой хранятся данные.

Далее необходимо было разработать классификацию терминов, соответствующую экспертному знанию о компьютерной вирусологии. Выделение категорий производилось по результатам изучения теоретического материала по компьютерной вирусологии [3, 4, 5] и экспертного анализа содержимого БД. Таким образом, были выделены и теоретически обоснованы 15 семантических полей, применяемые для классификации. Для возможности выполнения автоматизированной категоризации терминов требуется обучающее множество, поэтому затем была выполнена ручная классификация идентифицированных терминов. В результате была создана база данных, включившая более 1000 категирозированных терминов.

### Словарь-тезаурус TSReader

В рамках проекта был создан электронный отраслевой словарь (названый TSReader), который отражает содержание разработанной базы дан-

ных и автоматически обновляется при пополнении базы новыми терминами.

Разработанный словарь-тезаурус по компьютерной вирусологии предоставляет пользователям следующие возможности:

- просмотр предметно ориентированных дефиниций слов;
- просмотр контекстного употребления;
- просмотр аналога термина на русском языке;
- определение онтологических связей, посредством полевого анализа терминологии.

Графический интерфейс словаря-тезауруса является интуитивно понятным, главная страница словаря представлена на рис. 1.

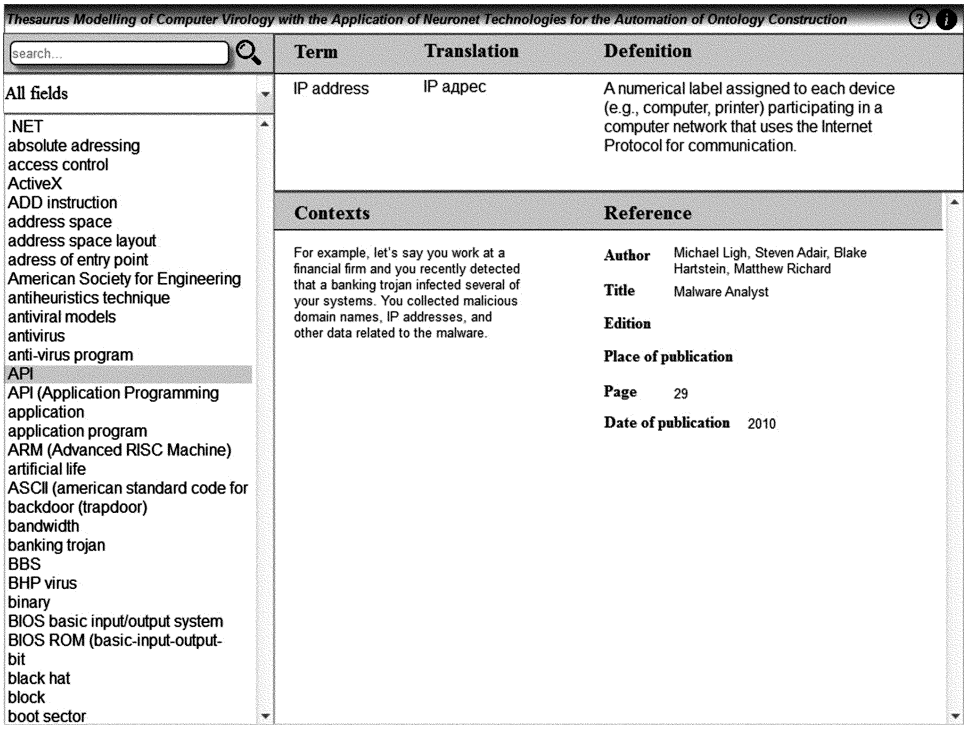


Рис. 1. Основная страница словаря-тезауруса.

Англо-русский словарь-тезаурус по компьютерной вирусологии представляет интерес, прежде всего, для студентов компьютерных специальностей, преподавателей компьютерных дисциплин и иностранных язы-

ков для специальных целей, а также и для широкого круга пользователей, сталкивающихся с вопросами компьютерной безопасности в повседневной жизни [6].

### Инструмент интеллектуального анализа TSBuilder

Программный продукт TSBuilder (свидетельство о государственной регистрации программы для ЭВМ № 2016612898) является инструментом интеллектуального анализа текстов с использованием коллекции категоризированных элементов (терминов) из описанной ранее базы данных. Позволяет осуществлять автоматизированную идентификацию терминов и их последующую классификацию.

Модуль интеллектуальной идентификации терминов основан на разработанном эвристическом алгоритме. Алгоритм автоматической идентификации терминов получает на вход текст для анализа, а также список терминов, импортированных из БД (ИС «Семограф»).

Основная идея алгоритма заключается в следующем: необходимо для каждого слова из текста определить его основу (основа слова находится путём отбрасывания наиболее часто встречающихся приставок, суффиксов, окончаний), а затем попытаться найти в базе данных термин с той же основой. Если в базе найден термин с той же основой, то слово из текста также будет считаться простым термином, состоящим из одного слова. Затем просматриваются пары (тройки) слов, уже являющиеся односложными терминами, и в случае, если они стоят друг за другом в анализируемом тексте, данная пара (тройка) слов считается составным термином.

Очевидным недостатком данного алгоритма является возможность поиска только тех терминов, которые являются однокоренными словами с терминами в базе данных.

Далее запускается процесс категоризации найденных терминов, представляющий собой элемент контролируемого машинного обучения с заданными классами. В основе алгоритма классификации лежит технология нейросетевого моделирования [7].

Термин, состоящий из одного слова, относится к семантическому полю, к которому принадлежит его основа в базе данных. При категоризации составных терминов сначала определяются семантические поля каждого слова, а в качестве входных данных нейросети выступают весо-

вые коэффициенты соответствующих полей. На основании этих данных нейронная сеть определяет поле и передает на выход его идентификатор.

Для категоризации двухсловных терминов используется сеть из одного слоя (рис. 2 а)). Нейрон переходит в возбужденное состояние при одновременном выполнении следующих условий:

$$\begin{cases} \omega_1 \geqslant \omega_2, \\ \frac{1}{e^{-(\omega_1+\omega_2)}+1} > \frac{1}{2}, \end{cases}$$

где $\omega_1, \omega_2$ — весовые коэффициенты полей для первого и второго слова составного термина.

В этом случае двухсловный термин будет отнесен к семантическому полю первого слова, иначе — второго.

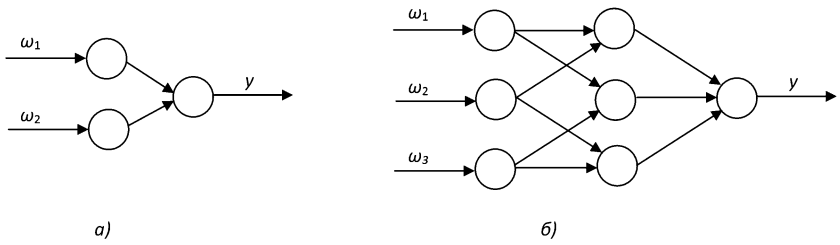


Рис. 2. Схема нейронной сети для а) двухсловного термина, б) трехсловного термина.

Для категоризации трехсловных терминов используется нейронная сеть из двух слоев (рис. 2 б)), на вход которой также подаются весовые коэффициенты полей каждого из трех слов $\omega_1, \omega_2, \omega_3$. Первый слой определяет семантические поля для пар слов (1,2),(1,3),(2,3) аналогично тому, как работает однослойная сеть. После этого подключается второй слой нейросети, функцией возбуждения которого будет функция выбора максимума. На вход подаются поля, определенные для каждой из трех пар, определяется вес каждого из полей, поле с максимальным весом передается как выходные данные, это поле становится полем всего термина. При обучении нейронной сети применяется метод поощрения/наказания: результат системы сравнивается с результатом, полученный от эксперта, при совпадении результатов изменений не производится, иначе — верное поле получает увеличение веса, а неверное — уменьшение. Изначальный весовой коэффициент каждого поля равен 0.

Составные термины из описанной ранее базы данных были использованы для обучения нейронной сети, при этом среднеквадратичная ошибка обучения составила 0,167. Апробация программного комплекса TSBuilder была проведена на текстовых данных, принадлежащих предметной области «Компьютерная вирусология». В результате тестирования нейронной сети среднеквадратичная ошибка составила приблизительно 0,255.

### Заключение

Программные продукты, разработанные научным коллективом, могут найти широкое применение в различных областях, могут использоваться для создания и упорядочения терминосистем различных научных областей.

Данный проект имеет ярко выраженный прикладной характер, но также несет теоретическую значимость как для исследователей в области лингвистики (изучение терминологии, фреймовое и онтологическое моделирование), так и для специалистов в области информационных технологий (разработка экспертной системы), прикладной математики (алгоритмы интеллектуального анализа данных), компьютерной безопасности (создание базы знаний по компьютерной вирусологии).

Работа выполнена в рамках проекта «Тезаурусное моделирование предметной области компьютерной вирусологии с применением нейросетевых технологий для автоматизации разработки онтологий» при поддержке гранта Российского фонда фундаментальных исследований 2014–2015 гг. (№ 14–06–31143).

## Список литературы

[1] Литовченко В. И. Классификация и систематизация терминов // Вестник Сибирского государственного аэрокосмического университета им. академика А. Ф. Решетнева. Сер.: Педагогика, филология, право, экология. — Красноярск: Изд-во СибГАУ, 2006. — С. 156–159.

[2] Исаева Е. В., Суворова В. А., Бахтин В. В. Автоматизированная разработка отраслевого словаря по компьютерной вирусологии // Евразийский вестник гуманитарных исследований. — Пермь: Пермский институт экономики и финансов, 2016. — С. 61–65.

[3] Козлов Д. А., Парандовский А. А., Парандовский А. К. Энциклопедия компьютерных вирусов. — М.: СОЛОН – Р, 2010.

[4] Вирусная энциклопедия лаборатории Касперского. [Эл. ресурс]. — URL: https://securelist.com/encyclopedia/ (дата обращения: 16.11.2016).

[5] Вирусная библиотека компании Dr.Web. [Эл. ресурс]. — URL: http://vms.drweb.ru/search/ (дата обращения: 16.11.2016).

[6] Исаева Е. В., Суворова В. А., Бахтин В. В. Машинное обучение с заданными классами в когнитивном терминоведении: автоматизация разработки отраслевого словаря // Научно-техническая информация. Сер. 2: Информационные процессы и системы. — 2016. — № 5. — С. 35–42.

[7] Ясницкий Л. Н. Введение в искусственный интеллект: учеб. пособие для ВУЗов. — М.: Изд. центр «Академия», 2005.