На правах рукописи

КОЖУНОВА ОЛЬГА СЕРГЕЕВНА

ТЕХНОЛОГИЯ РАЗРАБОТКИ СЕМАНТИЧЕСКОГО СЛОВАРЯ СИСТЕМЫ ИНФОРМАЦИОННОГО МОНИТОРИНГА

Специальность 05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Москва, 2009

Работа выполнена в Учреждении Российской академии наук
Институт проблем информатики РАН

Научный руководитель –	кандидат технических наук Зацман Игорь Моисеевич
Официальные оппоненты –	доктор технических наук, профессор Хорошевский Владимир Федорович
	кандидат технических наук, доцент Тарасов Валерий Борисович
Ведущая организация –	Всероссийский научно-исследовательский институт проблем вычислительной техники и информатизации (ВНИИПВТИ)

Защита диссертации состоится декабря 2009 года в часов мин. на заседании диссертационного совета Д002.073.01 при Учреждении Российской академии наук Институт проблем информатики РАН по адресу: 119333, Москва, ул. Вавилова, 44, корп. 2.

С диссертацией можно ознакомиться в библиотеке Учреждения Российской академии наук Институт проблем информатики РАН.

Отзывы в одном экземпляре, с заверенной подписью, просим направлять по адресу: 119333, г. Москва, ул. Вавилова, 44, корп. 2, в диссертационный совет.

Автореферат разослан «____» _____ 200__ г.

Ученый секретарь
диссертационного совета Д002.073.01
доктор технических наук, профессор С.Н. Гринченко

Общая характеристика работы

Актуальность темы. В настоящее время существенно изменилась значимость данных информационного мониторинга научных исследований и программной деятельности в сфере науки. Ранее данные мониторинга и определенные на их основе значения индикаторов практически не влияли на бюджетный процесс. Однако уже через несколько лет планируется значительную часть научного бюджета распределять с учетом значений индикаторов результативности научных исследований. Это коренным образом меняет роль систем информационного мониторинга, анализа и оценивания программной деятельности в сфере науки (далее - систем информационного мониторинга) и определяемых с их помощью значений индикаторов. На сегодняшний день уже накоплен отечественный и зарубежный опыт проведения мониторинга, анализа, индикаторного и экспертного оценивания результативности в сфере науки. Изучение этого опыта позволяет утверждать, что повышение роли систем мониторинга придает весьма актуальный характер задаче построения словаря показателей мониторинга как для описания, так и для решения широкого спектра задач индикаторного и экспертного оценивания результативности в сфере науки. Здесь особую значимость приобретает создание технологии разработки средств лингвистического обеспечения системы информационного мониторинга, основанного на использовании семантического словаря показателей мониторинга.

Целью диссертационного исследования является создание и исследование технологии разработки семантического словаря показателей для систем информационного мониторинга.

Для достижения поставленной цели диссертационного исследования были решены следующие задачи:

- разработка структуры семантического словаря показателей для систем информационного мониторинга;
- создание технологии разработки семантического словаря показателей и ее интеграция в системы информационного мониторинга;
- разработка метода построения комплексных запросов на поиск в БД и вычисление значений индикаторов;
- представление комплексных запросов на поиск в БД, вычисление индикаторов в виде статей семантического словаря и программная реализация запросов.

Методы исследования. Теоретические и практические исследования базируются на методах системного анализа, искусственного интеллекта, в том числе, методах классификации показателей мониторинга и методах построения и обработки запросов на поиск слабоструктурированных полнотекстовых документов.

Новизна работы. Выполненная диссертационная работа является одной из первых попыток создания технологии разработки лингвистического ресурса для системы информационного мониторинга. При ее реализации автором

достигнуты новые результаты, основные из которых заключаются в следующем:

- осуществлено исследование возможностей ДСМ-метода[1] для модификации семантических словарей на разработанном автором макете системы пополнения семантического словаря, в основе которой лежит механизм пополнения и порождения гипотез разного уровня на основе готового списка понятий и примеров понятий;
- проведен когнитивно-лингвистический анализ экспериментального массива текстов, содержащих термины области мониторинга, анализа и оценки научной деятельности, и согласование извлеченных из них понятий с классификационной схемой показателей мониторинга;
- разработана структура семантического словаря системы информационного мониторинга на основе гибкой и легко модифицируемой классификационной схемы;
- впервые предложена и реализована возможность установления взаимосвязей между словарными статьями индикаторов и алгоритмическими, информационными и нормативными ресурсами для прояснения их смысла и выработки согласованных терминов мониторинга;
- впервые в качестве статей семантического словаря предложено использовать параметризуемую статью: текстовые дефиниции на естественном языке с интегрированными параметрами на поиск в базах данных и вычисление значений индикаторов;
- предложен новый комплексный метод построения запросов на поиск в базах данных и вычисление значений индикаторов в виде параметризуемых статей семантического словаря Информационно-технологической системы мониторинга РАН (ИТСМ РАН).

Разработанный в ходе выполнения данной работы программный модуль ИТСМ РАН «Семантический словарь», функционирующий совместно с основными модулями этой системы, но независящий от других структур классификации индикаторов мониторинга, является уникальным как по самой разработке, так и по своему назначению.

Практическая значимость работы заключается:

- в разработке структуры семантического словаря показателей мониторинга;
- в создании технологии разработки семантического словаря показателей, обеспечивающей построение комплексных запросов на поиск в БД и вычисления значений индикаторов информационного мониторинга;
- в разработке и программной реализации функционального модуля «Семантический словарь», интегрированного в экспериментальный макет ИТСМ РАН;

---

[1] *Финн В.К.* О базах знаний интеллектуальных систем типа ДСМ // II Всесоюзная конференция «Искусственный интеллект-90», Минск, 1990 – с. 180-182.

- в программной реализации параметризуемой статьи семантического словаря для индикатора «индексы самоцитирования в патентах»;
- в использовании результатов, полученных в ходе выполнения диссертационной работы, в следующих проектах Российского фонда фундаментальных исследований и Российского гуманитарного научного фонда: РФФИ, грант № 09-07-00156; РФФИ, грант № 06-07-07001ано; РГНФ, грант № 05-03-03230а; РГНФ, грант № 06-02-04043а; РГНФ, грант № 05-03-12328в.

На защиту выносятся следующие результаты:

1. когнитивная технология разработки семантического словаря системы информационного мониторинга;
2. подход к модификации семантических словарей на основе ДСМ-метода применительно к разработанному автором макету системы пополнения семантического словаря, в основе которой лежит механизм порождения гипотез разного уровня на основе списка понятий и примеров понятий;
3. метод когнитивно-лингвистического анализа экспериментального массива текстов, содержащих термины области мониторинга, анализа и оценки;
4. механизм извлечения понятий из текстов и их согласовывания в соответствии с классификационной схемой показателей мониторинга;
5. метод построения комплексных запросов на поиск в БД и вычисления значений индикаторов на основе статей семантического словаря системы информационного мониторинга;
6. структура словарных статей семантического словаря с интегрированными параметрами поиска в БД и вычисления значений индикаторов (параметризуемых словарных статей);
7. программная реализация функционального модуля семантического словаря и технология его интеграции в систему информационного мониторинга;

Апробация работы и публикации. Материалы диссертации докладывались на следующих международных конференциях и семинарах:
Международная конференция по компьютерной лингвистике «Диалог-2006; Международная конференция по компьютерной лингвистике «Диалог-2007»; Международная конференция по компьютерной лингвистике «Диалог-2008»; Международная конференция «MEGALING-2006» «Горизонты прикладной лингвистики и лингвистических технологий»; Международная конференция «MEGALING-2007» «Горизонты прикладной лингвистики и лингвистических технологий»; Atlanta Conference on Science, Technology and Innovation Policy (ATLC-2007); Atlanta Conference on Science and Innovation Policy (ATLC-2009); 10th International Conference on Science and Technology Indicators; The 2009 World Congress in Computer Science, Computer Engineering, and Applied Computing (WORLDCOMP’09); Information and Brokerage Conference on Information and Communication Technologies in the EU's 7th Framework Programme (Moscow-2008); ICT Proposers' Day (Budapest-2009).

Основные результаты диссертации опубликованы в 18 публикациях, в том числе в трех публикациях в рекомендованных ВАК журналах, и в двух научно-исследовательских отчетах плановых НИР ИПИ РАН.

Структура диссертации. Диссертация состоит из введения, четырех глав, заключения, списка литературы (80 наименований) и 4 приложений. Работа изложена на 146 страницах, включающих 43 рисунка и 1 таблицу.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность исследования, сформулированы цель и задачи исследования, научная новизна исследования и приведены основные результаты работы.

В первой главе проводится анализ и обзор видов лингвистического обеспечения, в частности, словарей и тезаурусов, поскольку именно они наиболее часто используются в качестве средств лингвистического обеспечения информационных систем. Среди словарей были рассмотрены традиционные и электронные словари, идеографические словари и тезаурусы. Они проанализированы с точки зрения задач, функций и назначения в сравнении с аналогичными аспектами технологии разработки семантического словаря ИТСМ РАН с целью его позиционирования среди рассмотренных словарей и тезаурусов.

Кроме того, поскольку описываемый в данной работе семантический словарь ИТСМ РАН содержит в себе некоторые черты формальных и неформальных (лингвистических) онтологий, то был проведен соответствующий сопоставительный анализ и позиционирование словаря в рамках системы классификации онтологий, предложенной McGuiness[2]. В частности, в параграфе 1.1 подробно описан ресурс EuroWordNet, построенный по модели WordNet. Словари такого типа объединяют в себе результаты современных разработок в области компьютерной лингвистики и широко применяются для решения различных задач, в том числе в качестве справочной системы и инструмента для проведения лингвистических исследований.

Далее на основании проведенного анализа в работе приводится описание специфики предлагаемого семантического словаря, существенной для создания технологии его разработки. При этом, использовались такие базовые понятия как показатели, индикаторы, параметры, экспертные оценки и критерии, используемые в процессе информационного мониторинга в сфере науки, в том числе связи между ними, заданные при помощи иерархических отношений и ассоциаций[3]. Отметим, что словарь содержит формально определенное отношение класс-подкласс (показатели-индикаторы, показатели - индикаторы результатов программ фундаментальных научных исследований, и т.д.).

---

[2] *McGuinness, D. L*. (2003) Ontologies come of age. In: Fensel, D.; Hendler, J.; Lieberman, H.; Wahlster, W. (Eds). *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*. MIT Press, Cambridge, MA, USA, pp. 171–194.

[3] Зацман И.М Терминологический анализ нормативно-правового обеспечения создания систем мониторинга в сфере науки // Экономическая наука современной России. № 4, 2005. - С. 114-129.

В следующем параграфе главы 1 рассматривается назначение средств лингвистического обеспечения систем информационного мониторинга и приводится описание функций семантического словаря. В частности, различия в понимании экспертами смысла индикаторов являются серьезным препятствием в реализации всех трех основных процедур, необходимых для оценивания программной деятельности в сфере науки: информационный мониторинг, анализ, получение количественных и экспертных оценок ее результатов, эффективности и результативности. Это вызвало необходимость решения задачи согласования понимания индикаторов разными экспертами. Отмечено, что в силу особенностей формирования терминов мониторинга возникает также задача частной референции, когда одно название индикатора может обозначать целый класс индикаторов (например, индексы цитирования, смысл которых зависит от учета самоцитирования, а также цитирования соавторами и т.п.).

Для обеспечения необходимой функциональности словаря необходима экспликация видов референции для индикаторов. Для этого в системе мониторинга в момент времени использования каждого названия индикатора предлагается различать три основных вида референции:

- название индикатора относится ко всем вариантам алгоритма, которые могут использоваться для вычисления его значений;
- название индикатора относится только к части (подклассу) вариантов алгоритма;
- название индикатора относится только к одному варианту.

Для экспликации используемого вида референции в словарной статье семантического словаря, посвященной некоторому индикатору, предлагается включить список всех вариантов алгоритма, которые могут использоваться для вычисления его значений. Упомянутая функция дополняет классификационную функцию словаря по отношению ко всему набору индикаторов и других показателей системы мониторинга.

В результате обзора традиционных словарей и электронных лингвистических ресурсов показано, что новизна предлагаемого семантического словаря состоит в том, что он содержит ссылки на алгоритмические и информационные ресурсы системы информационного мониторинга, а также нормативные документы как источники терминов рассматриваемой предметной области. Инструмент с таким сочетанием функций для области информационного мониторинга разработан впервые.

В последнем параграфе главы 1 рассматриваются аспекты обработки исходных ресурсов словаря системы информационного мониторинга. По результатам отбора текстовых массивов, содержащих термины информационного мониторинга, был произведен когнитивно-лингвистический анализ текста. Формализация процесса извлечения знаний об индикаторах из текстовых массивов (в частности, нормативных документов) была осуществлена автором (в режиме экспертного анализа) на языке логики предикатов первого порядка для лексического и семантического уровней

анализа текстов и извлечения терминов. Основные результаты для лексического уровня приведены ниже.

$M=\{t_1, t_2, t_3,\ldots t_n\}$ – исходный корпус текстов, содержащих термины мониторинга, в частности, имена индикаторов (t_j – *некоторый нормативный документ*);

$I=\{i_1, i_2, i_3,\ldots i_y\}$ – система терминов мониторинга, построенная на основе корпуса *M,* где *y* – число терминов системы;

$Def=\{def_1, def_2,\ldots def_v\}$ – множество определений системы терминов мониторинга, заданных в корпусе М, где *v* – количество определений;

$Def_j = \{def_{j_1}, def_{j_2},\ldots def_{j_z}\}$ – множество определений смысла индикаторов, содержащихся в корпусе М (*z* – количество определений индикаторов, $0 \le z \le y$) определено для некоторого $t_j \in M$, $Def \supseteq Def_j$;

$I_{ind} = \{i_{ind_1}, i_{ind_2}, i_{ind_3}, \ldots. i_{ind_u}\}$ – подсистема терминов мониторинга – множество индикаторов, $I \supseteq I_{ind}$, *u – число индикаторов*;

$I_j = \{i_{j_1}, i_{j_2}, i_{j_3}, \ldots. i_{j_d}\}$ – множество индикаторов документа t_j, $I \supseteq I_j$, *d – число индикаторов (дефиниции этих индикаторов приведены в M)*;

Базовая аксиома 1 (о существовании текста для произвольного термина из системы терминов мониторинга) для $I=\{i_1, i_2, i_3,\ldots i_y\}$, *y-число терминов системы мониторинга,* $1<m<y$

$\forall i_m \in I\ (\exists t_m \in M(t_m \supseteq i_m))$

Базовая аксиома 2 (о включении индикаторов документа t_j**) для** $I_{ind} = \{i_{ind_1}, i_{ind_2}, i_{ind_3}, \ldots. i_{ind_u}\}$, *u – число индикаторов в системе терминов мониторинга,* $I_{ind} \neq \varnothing$

$\forall i_{ind_x} \in I_{ind}\ \&\ \forall t_j \in M\ (I_j \subseteq I_{ind})$

Базовая аксиома 3 (о существовании определений не для всех индикаторов документа t_j**) для** $I_j = \{i_{j_1}, i_{j_2}, i_{j_3}, \ldots. i_{j_d}\}$, *d – число индикаторов,* $1 \le x \le d$, $Def_j = \{def_{j_1}, def_{j_2},\ldots def_{j_z}\}$

$\exists i_{j_x} \in I_j\ (!\exists\ def_{j_x} \in Def_j)$

Базовая аксиома 4 (о существовании терминов, не являющихся именами индикаторов)

$\exists i_j \in I\ \&\ \notin I_j$

Лемма о лексической полноте системы терминов (при использовании базовых аксиом 1, 3 и 4)

$M=\{t_1, t_2, t_3,\ldots t_n\}$ *– корпус текстов, содержащий термины мониторинга*

$I=\{i_1, i_2, i_3,\ldots i_y\}$, *y – число терминов мониторинга*

$I_j = \{i_{j_1}, i_{j_2}, i_{j_3}, \ldots. i_{j_d}\}$, *d – число индикаторов,* $I_j \neq \varnothing$

$1 \le x \le n,\ 1 \le d \le y$

$[\exists i_x \subseteq t_j\ (!\exists def_x \in Def)\ \&\ (\exists i_{j_x} \in I_j)] \rightarrow (i_{j_x} \equiv i_x)$

Построенная формализация процесса извлечения терминов мониторинга и их определений из массива текстов (нормативные документы, научные статьи и т.д.) позволила выделить необходимые индикаторы, связи между ними и их определения. Это облегчило дальнейшую интеграцию индикаторов и других показателей мониторинга в классификационную схему.

Вторая глава посвящена описанию особенностей одной из основных категорий показателей информационного мониторинга – индикаторов результатов, эффективности и результативности в сфере науки, а также целям создания семантического словаря, его функциям и особенностям классификационной схемы индикаторов и других категорий показателей, лежащей в основе структуры словаря (рис. 1).

В первом параграфе главы рассматривается специфика формирования одной из основных категорий показателей информационного мониторинга – индикаторов. Прежде всего, оценивается их количество. Приблизительную оценку количества индикаторов результативности для сферы науки в целом, включая индикаторы результативности программ и проектов научных исследований, финансируемых на конкурсной основе, можно определить на основе данных сборника «Наука России в цифрах» за 2005 год, содержащего 168 *статистических индикаторов*, к которым необходимо добавить несколько десятков *информационных индикаторов*. При этом каждый индикатор может иметь несколько вариантов алгоритма его вычисления. Тот же порядок оценки количества индикаторов результативности можно получить на основе данных сборника «Индикаторы науки» за 2006 год, содержащего 209 статистических индикаторов. Аналогичный сборник Евросоюза (2003 год) содержит 374 индикатора результатов, эффективности и результативности.

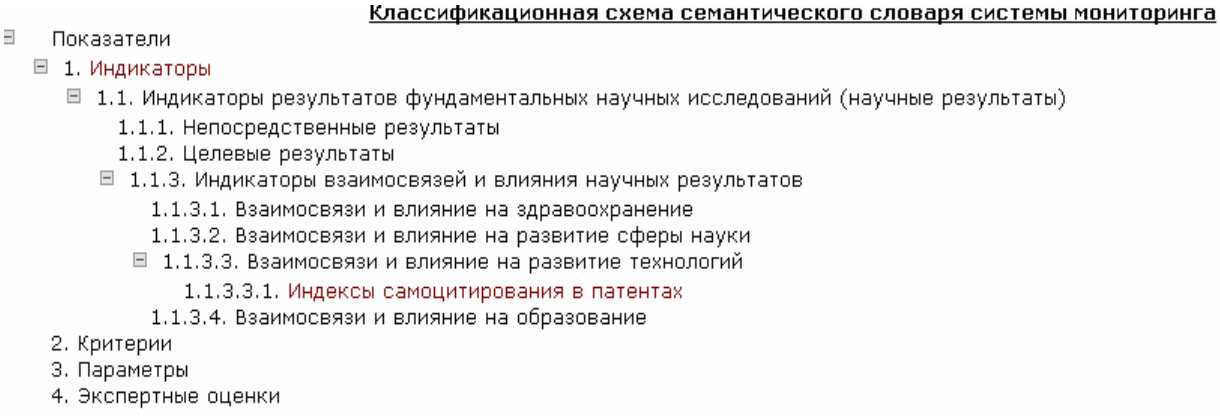

Рис. 1. Классификационная схема, использованная для построения структуры семантического словаря.

Далее в диссертации описываются особенности процедур мониторинга и соответствующие им аспекты использования индикаторов и других категорий показателей. В частности, при проведении анализа данных и получении экспертной оценки нередко возникают проблемы, связанные с различными подходами существующих групп пользователей к использованию наборов (систем) индикаторов и других категорий показателей. Различия в подходах к

использованию нередко являются следствием различий в понимании разными пользователями смысла одних и тех же индикаторов.

Кроме того, было выявлено, что для индикаторов в большинстве случаев характерна новизна и слабые ассоциативные связи имен индикаторов и обозначаемых ими понятий. То есть, по аналогии с топонимами (географическими названиями), из названия которых не выводится географическое место нахождения, которое они обозначают, смысл индикатора зачастую трудно понять, зная только его название. Например, иногда пользователи систем информационного мониторинга считают тождественными «индикатор результативности» и «индикатор эффективности». Чтобы понять смысл этих индикаторов при работе с системой, необходимо ознакомиться со способами (алгоритмами) их вычисления, реализованными в системе мониторинга.

В следующем параграфе главы описаны цели создания семантического словаря, соотнесенные со следующими организационными стадиями выполнения программ научных исследований: 1) планирование, 2) выполнение, 3) описание (демонстрация) полученных результатов и их публикация. На каждой из описанных стадий традиционно проводится три процедуры:

1. Сбор данных о планируемых (ожидаемых) или полученных результатах научной деятельности (мониторинг);
2. Обработка собранных данных, включая определение значений индикаторов и других категорий показателей, в том числе характеристик ресурсов, использованных для получения этих результатов (анализ);
3. Экспертная оценка результативности научной деятельности с использованием полученных значений индикаторов, характеристик и параметров (оценивание).

Согласно аналитическому обзору "Evaluating for Science: Processes & Protocols", подготовленному Европейской федерацией национальных академий наук ALLEA[4], чем сложнее научные программы и проекты, тем большее значение приобретает экспертиза и качественные (экспертные) оценки результатов, эффективности и результативности научной деятельности. Результаты экспертизы и другие качественные оценки иногда должны быть получены экспертами на основе количественных индикаторов в условиях явного или неявного использования ими слабых ассоциативных связей имен индикаторов и обозначаемых ими понятий. При этом имеющиеся дефиниции индикаторов и других категорий показателей часто не являются общепринятыми. Различия в понимании экспертами смысла различных индикаторов и других категорий показателей являются существенным препятствием в реализации всех трех процедур (мониторинг, анализ, получение оценок). Поэтому одной из основных целей создания семантического словаря является предоставление экспертам средств лингвистического обеспечения процесса.

---

[4] Hackmann, H., Drenth, P.J.D., Schroots, J.J.F. Evaluating for Science: Processes & Protocols. – Amsterdam: ALLEA, 2004.

Далее в диссертации описываются проблемы, возникающие на этапе получения экспертных оценок.

В третьем параграфе главы описаны функции семантического словаря и построение его структуры. Основными функциями семантического словаря являются:

- классификационная функция индикаторов и других категорий показателей с учетом вариантов алгоритмов их вычислений;
- унификация терминологии области информационного мониторинга;
- репрезентативная функция прояснения смысла индикаторов.

Эти функции во многом предопределили структуру словаря в целом и его словарных статей. При этом его структура была построена на основе классификационной схемы.

В *параграфе 4* проводится анализ некоторых алгоритмов ДСМ-метода и обсуждается возможность их применения в технологии разработки семантического словаря ИТСМ РАН. Приводится описание отдельных алгоритмов и процедур, а также макета интеллектуальной системы пополнения семантического словаря, разработанной автором для апробации возможности применения ДСМ-метода в задачах построения лингвистических ресурсов. В параграфе отдельно рассматривается алгоритм Норриса, который реализует процедуру индуктивного обобщения, и его вариант, адаптированный к задаче пополнения семантического словаря в разработанном макете.

Краткое описание алгоритма Норриса следующее. Имеется набор множеств (объектов). Введем на этом наборе множеств какой-нибудь линейный порядок и фиксируем его. На n-ом шаге алгоритм достраивает диаграмму Хассе для n-го множества, используя построенную диаграмму для первых $n-1$ множеств.

Обозначим через $n(m)$ номер множества m (сами множества и их пересечения будут обозначаться маленькими буквами, а подмножества номеров множеств – большими буквами).

Пусть $L(k)$ – множество понятий, полученных при обработке первых k множеств. Очевидно, что $L(0)$ пусто.

Рассмотрим n-й шаг алгоритма.

Пусть x – очередное множество, номер которого равен n, т. е. $n(x) = n$.

Действие 1. Для множества x просматриваем каждое понятие из $L(n-1)$. Пусть (Y, y) – очередное понятие (y – это пересечение множеств, номера которых составляют Y).

1.1. Если y является подмножеством множества x ($y \subset x$), т. е. $y \cap x = y$, то при добавлении множества x к множествам с номерами из Y их общее пересечение будет y. Поэтому пару (Y, y) заменим на пару $(Y \cup \{n\}, y)$.

1.2. Если же $y \not\subset x$, то найдем пересечение этих множеств: $z = y \cap x$.

Так как $z \subset y$, то z является подмножеством любого множества с номером из Y. И если номера всех ранее рассмотренных множеств, подмножеством которых является z, входят в Y, то добавляется новое понятие $(Y \cup \{n\}, z)$, т. е. для пересечения z на этом шаге набор $Y \cup \{n\}$ – максимальный.

Если это условие нарушено, то набор $Y \cup \{n\}$ не является максимальным для z, возвращаемся на начало Действия 1.

Это условие также можно записать в виде:

$\{v \mid n(v) < n(x), n(v) \notin Y, z \subset v\} = \varnothing$.

Если данное условие для множества z выполнено, то говорят, что вхождение пересечения z в множество построенных пересечений является *относительно каноническим*.

Действие 2. Если множество x не является подмножеством никакого из предыдущих множеств, т. е. верно условие:

$\{v \mid n(v) < n(x), x \subset v\} = \varnothing$, то добавляется новое понятие: $(\{n\}, x)$.

В этом случае говорят, что вхождение множества x в множество построенных пересечений является *каноническим*.

Для случая пополнения семантического словаря алгоритм Норриса был адаптирован с учетом потребности предметной области, для которой был разработан макет интеллектуальной системы пополнения семантического словаря.

Особенности алгоритма в работе продемонстрированы на следующем частном примере. В начале работы системы пользователь в качестве пополняемого понятия выбирает «природные катастрофы». Для этого он выделяет некоторое число примеров этого понятия: тайфун, оползень и смерч. Система считывает выбранные примеры понятия и ищет в корпусе текстов предложения, содержащие их. Если такие предложения не найдены, то система сообщает пользователю об ошибке. В случае если эти предложения обнаружены, система передает их для обработки морфологическому нормализатору: «На Африку налетел сильный тайфун, вызвавший многочисленные жертвы среди населения», «В результате оползня разрушено целое селение», «Смерч вызвал многочисленные жертвы среди рогатого скота». Нормализованные предложения: «На Африка налететь сильный $, вызвать многочисленный жертва среди население», «В результат $ разрушить целое селение», «$ вызвать многочисленный жертва среди рогатый скот». Нормализованные предложения (+-примеры, т.е. предложения, содержащие примеры пополняемых понятий) записываются в базу фактов. Затем они обрабатываются процедурой индуктивного обобщения – алгоритмом Норриса:

База фактов заполняется файлами с +-примерами:

plus_1.txt: На Африка налететь сильный $, вызвать многочисленный жертва среди население

plus_2.txt: В результат $ разрушить целое селение

plus_3.txt: $ вызвать многочисленный жертва среди рогатый скот

База знаний содержит +-гипотезы, сгенерированные в результате пересечения +-примеров в соответствии с алгоритмом Норриса и встроенной в него процедурой сходства.

1.txt: На Африка налететь сильный $, вызвать многочисленный жертва среди население – 1 пример, взятый целиком (не является подмножеством других множеств и не входит ни в один пример целиком) – гипотеза 1

2.txt: В результат $ разрушить целое селение – 2 пример целиком; – гипотеза 2

3.txt: $ вызвать многочисленный жертва среди рогатый скот – 3 пример целиком; – гипотеза 3

4.txt: 0 (то есть файл пуст)– пустое пересечение 1 и 2 примеров; – гипотеза 4

5.txt: $ вызвать многочисленный жертва – пересечение 1 и 3 примеров; – гипотеза 5

Результатом работы алгоритма Норриса являются сгенерированные гипотезы, которые передаются для проверки процедуре абдукции (у каждой гипотезы должен быть пример, из которого она образована, иначе абдукция неверна и дальнейшая работа с этими гипотезами невозможна). В случае успешного завершения абдукции гипотезы из базы знаний передаются процедуре аналогии, которая и осуществляет их наложение на текст, т.е. на предложения, которые могут содержать потенциальные примеры для пополняемого понятия. Например, «Потоп вызвал многочисленные жертвы». Это предложение совпадает с гипотезой № 5. Значит, «потоп», совпадающий с $ в предложении гипотезы, и есть искомый пример понятия «природные катастрофы» и будет успешно добавлен в базу понятий.

В конце параграфа приведены выводы о том, что несовпадение задач разрабатываемых систем выявило неприменимость формализмов ДСМ-метода при формировании технологий создания и интеграции семантического словаря в ИТСМ РАН. Это мотивировало дальнейший поиск в направлении выбора оптимальных алгоритмов и формализмов технологии разработки семантического словаря ИТСМ РАН. Формализмы ДСМ-метода могут быть использованы при дальнейших модификациях семантического словаря ИТСМ РАН для его пополнения новыми терминами мониторинга. В заключение параграфа приводятся результаты сопоставления задач системы пополнения семантического словаря и ИТСМ РАН.

В последнем параграфе главы 2 описываются особенности классификационного метода применительно к формированию структуры словаря в системе информационного мониторинга. Подробно представлен процесс уточнения смысла индикаторов (метод классификации), включающий две стадии. На первой стадии осуществляется позиционирование каждого предлагаемого к использованию индикатора в рамках структуры семантического словаря ИТСМ РАН, полученной в результате использования классификационной схемы показателей. На второй стадии эксперты могут уточнять смысл индикаторов посредством анализа содержания словарных статей и их связей с нормативными, информационными и алгоритмическими компонентами системы мониторинга.

Использование метода классификации для распределения перечисленных индикаторов по структуре семантического словаря ИТСМ РАН позволили не только предложить процедуру согласования смысла индикаторов, но также продемонстрировать возможности структуры семантического словаря ИТСМ РАН на одном из списков индикаторов, утвержденном и используемом в сфере науки.

Третья глава посвящена технологическим принципам разработки семантического словаря. *В первом параграфе главы* рассматриваются задачи определения значений индикаторов в системах мониторинга. Представлены

особенности формирования и использования индикаторов в целом и применительно к различным видам программ. В рамках системы базовых терминов системы мониторинга[5] индикаторы определены как показатели количественной оценки, вычисляемые на основе информационных ресурсов системы мониторинга

В качестве вывода к первому параграфу приводится заключение о том, что возникла потребность в формировании инструмента, который бы учитывал специфику формирования, вычисления и использования наборов индикаторов и других категорий показателей в системе мониторинга. В качестве такого инструмента предлагается семантический словарь систем информационного мониторинга.

В следующем параграфе описывается ИТСМ РАН, в рамках создания которой и была опробована технология разработки семантического словаря. В соответствии с Программой фундаментальных научных исследований РАН на 2008-2012гг. эта система должна обеспечивать руководителей и сотрудников Президиума РАН информацией о результатах формирования, реализации и мониторинга этой Программы.

В третьем параграфе главы представлена предлагаемая структура семантического словаря ИТСМ РАН. В настоящее время из предлагаемой структуры в рамках ИТСМ РАН в семантическом словаре реализована следующая часть: Показатели → Индикаторы → Индикаторы результатов Программы фундаментальных научных исследований → Непосредственные результаты, Целевые результаты, Индикаторы взаимосвязей и влияния научных результатов → Взаимосвязи и влияние на развитие технологий → Индикаторы самоцитирования в описаниях изобретений.

Реализация этой части предлагаемой структуры выполнена впервые в рамках диссертационного исследования, описанного в данной работе.

В параграфе 4 описывается процесс создания словарных статей в семантическом словаре. Словарная статья имеет несколько параметров поиска и обработки информационных полей при исполнении запроса на вычисление значений индикаторов. Изменение значений параметров позволяет получать различные варианты вычисления индикатора.

Сочетание параметров (значения которых задаются с помощью полей на электронной форме вычисляемого индикатора) предложено впервые на основе проанализированных и обработанных автором информационных ресурсов Роспатента. Каждый параметр, участвующий в вычислении индикатора, связан с различными аспектами его вычисления (рис. 2).

Заключительный параграф главы посвящен информационным ресурсам, используемым для вычисления значений индикаторов.

Далее описывается связь процедуры вычисления индикаторов в семантическом словаре ИТСМ РАН с информационными ресурсами. Поскольку основной целью работы является технология разработки такого

[5] Зацман И.М., Веревкин Г.Ф., Шубников С.К. Моделирование систем мониторинга. - М.: ИПИ РАН, 2008. – 115 с.

словаря систем мониторинга, то автором была продемонстрирована ее реализуемость на примере вычисления группы индикаторов «Индексы самоцитирования в описаниях изобретений».

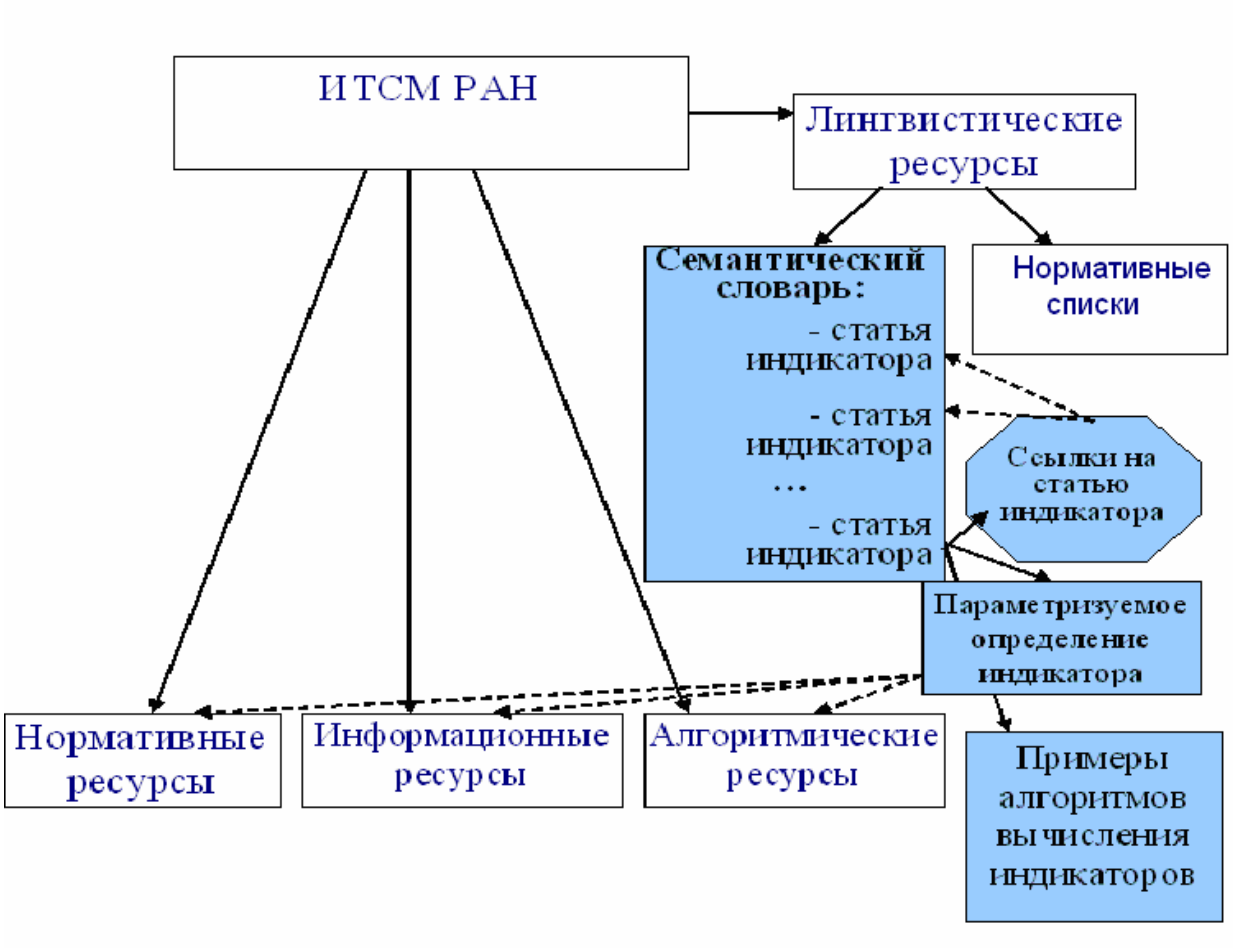

Рис. 2. Связь статей семантического словаря с ресурсами ИТСМ РАН.

В четвертой главе представлены результаты вычислительных экспериментов, проведенных в рамках предложенной технологии разработки семантического словаря ИТСМ РАН. Словарь был протестирован на группе индикаторов – «Индексы самоцитирования в описаниях изобретений». Индикаторы этой группы вычисляются по следующему правилу: определяются абсолютная и относительная частотности встречаемости фамилий авторов изобретений в библиографических ссылках на их собственные статьи в полнотекстовых описаниях этих изобретений.

Индексы самоцитирования могут быть вычислены для одного автора или группы авторов. В качестве параметров поиска могут быть также заданы временные промежутки и тематика изобретений. Эта группа индикаторов и варианты алгоритма вычисления их значений применительно к патентной сфере описываются впервые.

Индексы самоцитирования имеют большое значение для оценки прикладного использования результатов научной деятельности, так как они позволяют выявить те научные результаты, которые тем или иным образом связаны с новыми техническими решениями, в прикладном использовании

которых участвуют авторы этих научных результатов. Таким образом, индексы самоцитирования могут быть использованы для исследования инновационной активности научных сотрудников в тех или иных технологических областях.

В ИТСМ РАН для вычисления различных вариантов индексов самоцитирования автором были подготовлены следующие тестовые информационные ресурсы:

1. Массив полных описаний изобретений сотрудников РАН, отобранных по тематическим классам технологических областей.

2. Xml-файл, сформированный программой обработки изобретений Patanalysis[6], которая была модифицирована автором данной работы специально для подготовки тестовых информационных ресурсов. В этом файле содержатся поля из html-описаний изобретений, которые необходимы для вычисления группы индикаторов «Индексы самоцитирования в описаниях изобретений». Программа Patanalysis модифицированной версии в качестве входных параметров использует массив кратких описаний патентов, сформированных в результате выполнения запроса к БД Роспатента.

Запрос к БД Роспатента был необходим для формирования массива описаний изобретений, патентообладателями которых являются учреждения РАН. Полученные по запросу описания изобретений позволяют вычислить значения группы индикаторов «Индексы самоцитирования в описаниях изобретений». При этом использовались следующие параметры: фамилии, имена, отчества авторов патентов, их гражданство, номера патентов, даты подачи заявки на патент, учреждения-патентообладатели и т.д.

Таким образом, подготовка тестовых информационных ресурсов включала следующие два этапа. На первом этапе был сформулирован и отлажен запрос, который позволил получить полнотекстовые описания изобретений, где в качестве патентообладателей указаны учреждения РАН. Необходимость в таком отборе была мотивирована тем, что целью этой группы индикаторов является исследования инновационной активности научных сотрудников в тех или иных технологических областях.

На втором этапе был проведен анализ полнотекстовых описаний изобретений. В результате этого анализа было выявлено, что наибольшее количество изобретений, в которых содержатся ссылки на публикации их авторов, относится к областям химии, биологии, физики и радиоэлектроники.

Помимо вышеперечисленных ресурсов в рамках работы был также реализован механизм извлечения тегов из xml-файла для заполнения полей параметров вычисления индикаторов в семантическом словаре ИТСМ РАН.

В следующем параграфе приводятся особенности построения словарных статей индикаторов. Семантический словарь ИТСМ РАН, основанный на классификационной схеме показателей мониторинга, позволяет просматривать все уровни иерархии этой структуры словаря. Кроме того, ввиду особенностей

---

[6] Программа структуризации ссылок цитирования в описываемых изобретений РФ: Свидетельство об официальной регистрации программы для ЭВМ N 2007613549 / С.К. Шубников. Зарегистрировано в Реестре программ для ЭВМ 22.08.2007.

структуры словаря в ней существует возможность расширения как существующих категорий индикаторов и других показателей, так и добавления новых. Xsd-представление с рекурсивной ссылкой позволяет формировать уровни иерархии внутри структуры словаря с необходимой разработчику степенью подробности.

Это делает возможным описание новых индикаторов внутри самой структуры словаря посредством их последовательной интеграции на нужные уровни иерархии (рис. 3). Такой подход расширяет функциональные возможности семантического словаря ИТСМ РАН и позволяет не только описывать в нем новые термины и понятия, но и устанавливать между ними необходимые связи и наглядно демонстрировать их в общей структуре словаря.

В параграфе 3 приведены результаты вычисления значений индикаторов. Вычисление значений индикаторов происходит во вкладке «Параметризуемая статья семантического словаря» ИТСМ РАН для группы индикаторов «Индексы самоцитирования в описаниях изобретений».

На этом примере продемонстрирована работа алгоритма вычисления значений индикаторов. В полнотекстовых описаниях изобретений анализируются следующие поля: фамилии, имена и отчества авторов изобретений, дата публикации изобретений и тематические классы изобретений (на основе Международной патентной классификации - МПК). Эти поля в дальнейшем отбираются и обрабатываются в соответствии со значениями заданных параметров этой группы индикаторов. Иначе говоря, происходит поиск и подсчет встречаемости фамилий авторов изобретений с целью определения значений индексов самоцитирования за некоторый временной промежуток по заданной тематике.

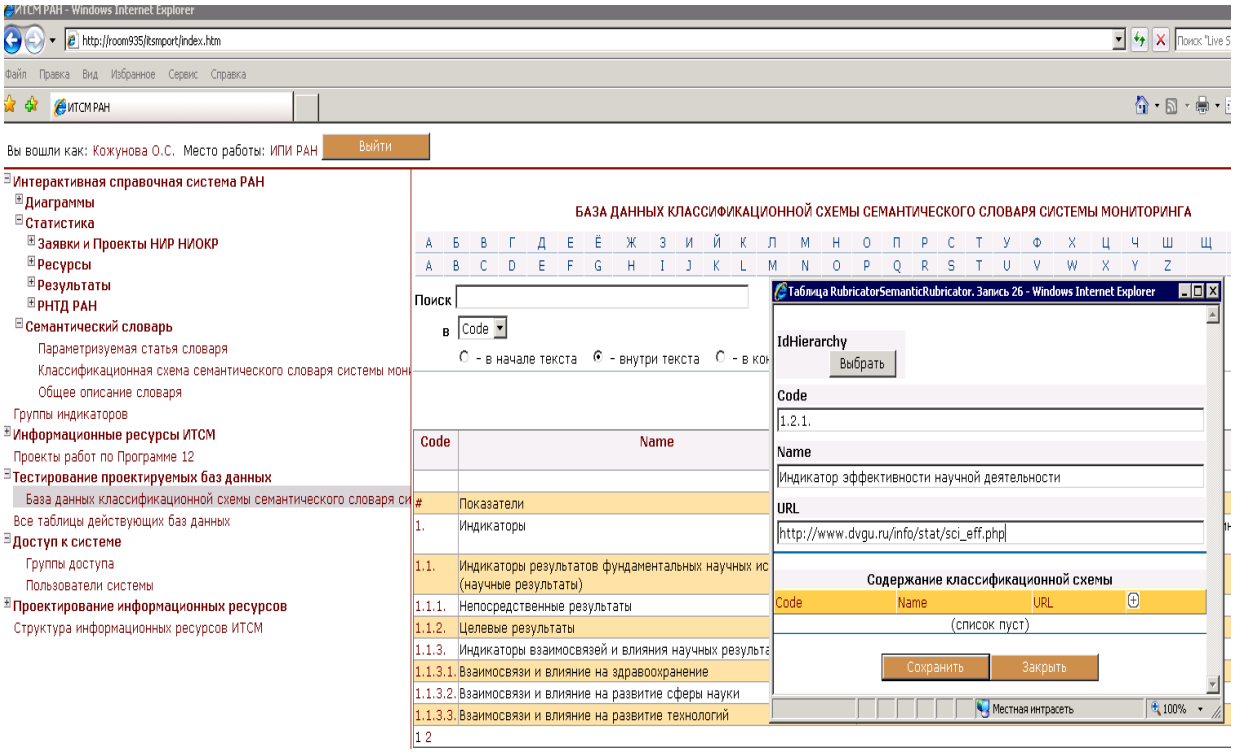

Рис. 3. Добавление новых категорий индикаторов.

По всему массиву имеющихся изобретений отбираются те, в которых в качестве авторов патентов фигурируют заданные пользователем в параметризуемой статье словаря авторы, отвечающие фиксированному промежутку времени и соответствующие обозначенной им же тематике, заданной с помощью рубрик МПК. Для каждого патента, отвечающего такому условию, подсчитывается абсолютная и относительная частотности встречаемости фамилий авторов изобретений в библиографических ссылках на их собственные статьи в полнотекстовых описаниях этих изобретений. Индексы, посчитанные для нескольких патентов одного автора, суммируются, то есть, для каждого автора, каждого временного периода и отдельно взятой тематики формируется отдельный, суммарный индекс по релевантным запросу изобретениям. Поэтому все возможные варианты подсчета индексов самоцитирования образуют группу индикаторов с соответствующим названием, а не один отдельный, фиксированный индикатор с единственным алгоритмом вычисления.

Параграф 4 посвящен описанию параметризуемых словарных статей семантического словаря (рис. 4). Словарные статьи семантического словаря в системе информационного мониторинга структурированы в соответствии с ресурсами, с которыми связан словарь. Каждая статья содержит параметризуемую дефиницию (то есть определение значения индикатора, зависящее от нескольких параметров) и поэтому носит название параметризуемой. Каждая статья связана с отдельной группой индикаторов. В структуре словаря представлены разные уровни классификации индикаторов.

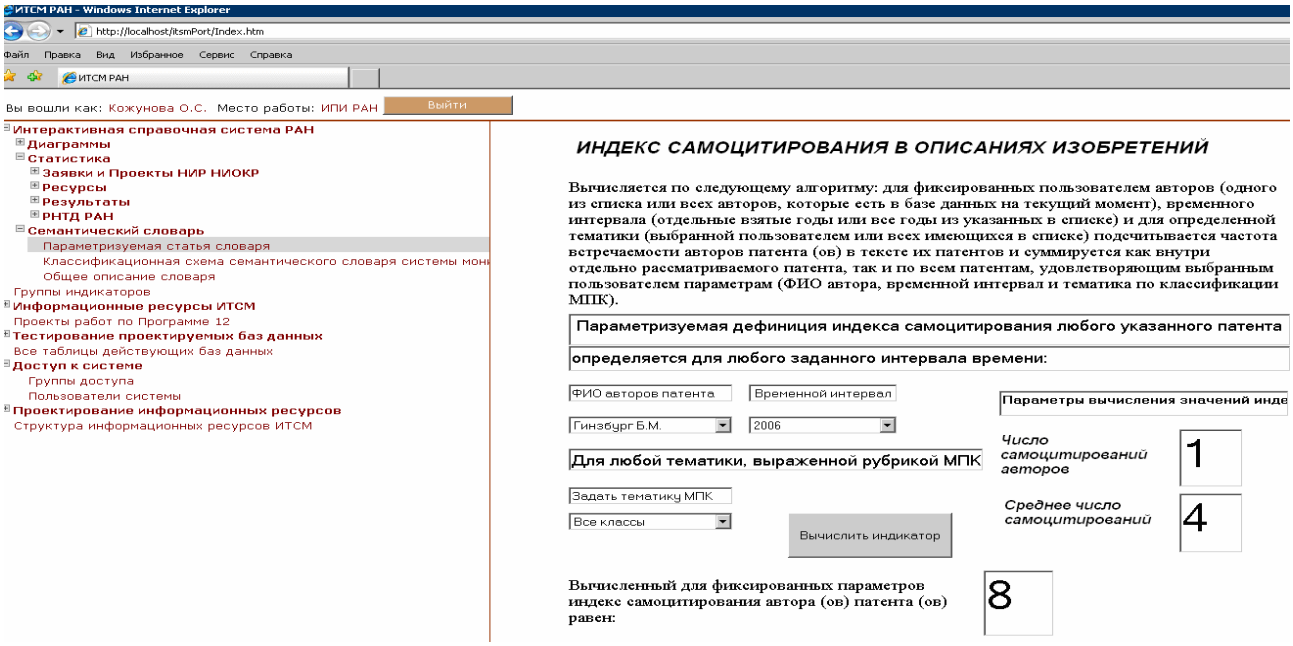

Рис. 4. Параметризуемая словарная статья семантического словаря

В основе параметризуемой статьи лежит комплексный запрос, результатами выполнения которого являются не совокупность информационных полей, как это происходит при выполнении поисковых запросов, а результат поиска и последующей обработки информационных полей. Поэтому комплексный

запрос включает и параметры поиска, и параметры вычисления значений индикаторов (на рис.4 параметры вычисления значений индикаторов – число самоцитирований авторов изобретений и среднее число самоцитирований, то есть соотношение индекса самоцитирования автора изобретений и его патентной активности).

В семантическом словаре с целью демонстрации реализуемости результатов работы спроектирована параметризуемая статья для группы индикаторов «Индекс самоцитирования в описаниях изобретений». В состав параметров этой статьи включены фамилия, имя, отчество авторов изобретений, временной промежуток (отбор патентов по дате публикации на сайте Роспатента) и отбор патентов по тематике МПК. В заключение параграфа приводится пример словарной статьи семантического словаря.

В последнем параграфе главы описывается программная реализация действующего макета семантического словаря, который был реализован на основе средств проектирования ИТСМ РАН. Одним из основных принципов ее проектирования является разделение всех функциональных подсистем на две категории (базовые и прикладные). Сначала проектируются базовые функциональные подсистемы, а затем создаются прикладные функциональные подсистемы, их модули и задачи. Действующий макет семантического словаря реализован в виде функционального модуля интерактивно-справочной подсистемы ИТСМ РАН.

В **заключении** приводятся основные выводы, полученные в работе. В **приложения** вынесены поясняющие и вспомогательные материалы.

Публикации по теме работы

1. Финн В.К., Виноградов Д.В, Кожунова О.С. Интеллектуальная система пополнения семантических словарей // **Программные продукты и системы,** № 2, 2006. – с.27-30. (Личный вклад диссертанта: проектирование и разработка макета семантического словаря)

2. Кожунова О.С. Моделирование пополнения семантического словаря // Системы и средства информатики. Вып. 16.- М.: Наука, 2006.– С. 339-354.

3. Кожунова О.С. Применение правдоподобных рассуждений ДСМ – метода для пополнения семантического словаря // Труды международной конференции Диалог-2006 "Компьютерная лингвистика и интеллектуальные технологии". - М.: Изд-во РГГУ, 2006. – С.243-247.

4. Кожунова О.С. Опыт применения правдоподобных выводов ДСМ-метода для пополнения семантического словаря // Материалы международной конференции «MegaLing’2006», Партенит, 2006. – с.209-210.

5. Зацман И.М., Кожунова О.С. Предпосылки конвергенции информационной и компьютерной наук // Системы и средства информатики. Тематический выпуск «Научно-методологи ческие вопросы информатики».- М.: Наука, 2006.– С. 112-139. (Личный вклад диссертанта: анализ оснований информационной и компьютерной наук)

6. Зацман И.М., Кожунова О.С. Семантический словарь системы информационного мониторинга в сфере науки: задачи и функции. // Системы и средства информатики. Вып. 17.- М.: Наука, 2007.- С. 124-141. (Личный вклад диссертанта: принципы разработки семантического словаря систем информационного мониторинга, анализ задач и функций существующих идеографических словарей, сопоставительное исследование задач и функций семантического словаря и других словарей)

7. Кожунова О.С., Зацман И.М. Прагматические аспекты создания семантического словаря терминов информационного мониторинга // Труды международной конференции Диалог-2007 "Компьютерная лингвистика и интеллектуальные технологии".- М.: Изд-во РГГУ, 2007. - С. 278-285. (Личный вклад диссертанта: описание специфики создания семантического словаря систем информационного мониторинга и анализ основных аспектов его построения)

8. Кожунова О.С. Опыт применения правдоподобных выводов ДСМ-метода для пополнения семантического словаря // Сборник научных трудов «MegaLing'2006». – Киев: Довира, 2007. – с.149-161.

9. Кожунова О.С. Семантический словарь терминов системы оценки результативности в сфере науки // Материалы международной конференции «MegaLing'2007», Партенит, 2007. – с.170-171.

10. Zatsman Igor, Kozhunova Olga. Evaluation system for the Russian Academy of sciences: clarification tools // Atlanta Conference on Science, Technology, and Innovation Policy 2007. Atlanta, USA, Georgia Institute of Technology, 2007. (Личный вклад диссертанта: описание макета семантического словаря, его структуры и основных категорий индикаторов)

11. Кожунова О.С. Eurowordnet: задачи, структура и отношения // **Информатика и ее применения,** том 2, выпуск 4. – М.: Торус Пресс, 2008. – С. 85-92.

12. Зацман И.М., Кожунова О.С. Предпосылки конвергенции информационной и компьютерной наук // **Информатика и ее применения,** том 2, вып. 1. – М.: Торус Пресс, 2008. – с.77-98. (Личный вклад диссертанта: анализ оснований информационной и компьютерной наук)

13. Кожунова О.С. Семантический словарь системы информационного мониторинга в сфере науки и ресурс Eurowordnet: структура, задачи и функции // Системы и средства информатики. Вып. 18.- М.: Наука, 2008.- С. 156-171.

14. Кожунова О.С. Классификационная схема семантического словаря системы мониторинга: опыт применения в процессе оценки результативности научной деятельности // Труды международной конференции Диалог-2008 "Компьютерная лингвистика и интеллектуальные технологии". - М.: Изд-во РГГУ, 2008. – с.210 – 216.

15. Zatsman, I., Kozhunova O. Evaluating for institutional academic activities: classification scheme for R&D indicators // Proceedings of the 10th International Conference on Science and Technology Indicators (17th - 20th September 2008, University of Vienna, Austria). - Vienna: Austrian Research

Center GmbH, 2008. - Pp. 428-431. (Личный вклад диссертанта: описание макета семантического словаря и модифицированного варианта его классификационной схемы)

16. Zatsman Igor, Kozhunova Olga. Evaluation System for the Russian Academy of Sciences: Objectives-Resources-Results Approach and R&D Indicators // Proceedings of the Atlanta Conference on Science, Technology, and Innovation Policy 20097. Atlanta, USA, Georgia Institute of Technology, 2009 (in e-print at http://ieeexplore.ieee.org/Xplore/guesthome.jsp). (Личный вклад диссертанта: описание модифицированной классификационной схемы семантического словаря)

17. Igor M. Zatsman, Olga S. Kozhunova. Emerging personal concepts and tracing their evolution by computer: semiotic foundations // Proceedings of WorldComp'09. 2009. (Личный вклад диссертанта: описание макета семантического словаря, его функциональности в системах информационного мониторинга и обзор аналогичных инструментальных средств)

18. Zatsman Igor, Kozhunova Olga. Evaluation System for the Russian Academy of Sciences: Objectives-Resources-Results Approach and R&D Indicators // Abstracts of the Atlanta Conference on Science, Technology, and Innovation Policy 20097. Atlanta, USA, Georgia Institute of Technology, 2009/ (http://conferences.library.gatech.edu/acsip/index.php/acsip09/atlc09/paper/view/104/147). (Личный вклад диссертанта: описание модифицированной классификационной схемы, семантического словаря и части структуры ИТСМ)