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:

Ведущая организация:

Защита диссертации состоится _____ в ___ часов на заседании диссертационного Совета Д002.073.01 при Учреждении Российской академии наук Институт проблем информатики РАН
РАН по адресу: 119333, Москва, ул. Вавилова, 44, корп. 2.

С диссертацией можно ознакомиться в библиотеке Института проблем информатики Российской академии наук.

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

Автореферат разослан « ___» _____ 2009 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 деятельности субъектов сферы 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 мониторинга и определяемых с их помощью значений индикаторов. На сегодняшний день в какой-то мере уже накоплен отечественный и зарубежный опыт проведения мониторинга, анализа и оценки результативности в сфере науки. Изучение этого опыта позволяет утверждать, что повышение роли систем мониторинга и определяемых с их помощью значений индикаторов придает весьма актуальный характер задаче построения словаря терминов для описания и решения широкого спектра задач мониторинга, анализа и оценки. Здесь особую значимость приобретает разработка наиболее эффективного информационного ресурса для проведения мониторинга – семантического словаря. Таким образом, проектирование и разработка семантического словаря системы информационного мониторинга РАН являются актуальными вопросами.

Цель работы состоит в создании и исследовании технологии разработки семантического словаря для системы информационного мониторинга, в том числе для разрешения конфликтных ситуаций между экспертами области мониторинга, анализа и оценки при отсутствии согласованных систем соответствующих терминов (индикаторов, параметров, критериев и экспертных оценок), их дефиниций и отношений, а также другой релевантной информации, позволяющей получить четкое представление о понятиях в системе информационного мониторинга РАН на протяжении фиксированных отрезков времени.

Разработанный семантический словарь ИТСМ РАН должен удовлетворять следующим требованиям:

- Независимость от других подсистем ИСТМ РАН, оперирующих индикаторами и соответствующими ресурсами;
- Фиксация основных параметров, необходимых для полноценной дефиниции и вычисления выбранных пользователем терминов (в частности, индикаторов);
- Категоризация заранее выявленных и добавляемых в процессе работы системы терминов и понятий мониторинга;
- Обеспечение связей и отношений как между категориями и подкатегориями терминов, так и между терминами и необходимыми для прояснения их смысла ресурсами (например, алгоритмическими ресурсами – для демонстрации вариантов вычисления индикаторов; информационными ресурсами – для указания необходимых статей, схем, диаграмм, документов, где фигурирует выбранный

пользователем индикатор; нормативных ресурсов – для указания источника, легализующего тот или иной индикатор и, возможно, содержащего пояснительную информацию по толкованию и использованию данного термина);

- Прояснение смысла терминов мониторинга (индикаторов) и обеспечение экспертов и пользователей ИТСМ их согласованными дефинициями и доступной релевантной информацией.

Достижение поставленной цели даст возможность использовать разработанный семантический словарь в системе информационного мониторинга РАН для получения согласованных дефиниций индикаторов и других терминов мониторинга, отслеживания их эволюции во времени (что связано со спецификой формирования терминов и понятий в области информационного мониторинга, дефиниции которых могут меняться достаточно часто, но возникает необходимость сохранения предыдущих версий и вариантов) и интеграции новых индикаторов и их категорий в словарь ИТСМ.

Методы исследования. Методологической основой для выполнения настоящей работы послужили современные исследования в области представления знаний, методы категоризации данных, методы организации реляционных баз данных, методы обработки запросов инструментария Mimosa V5, технологии проектирования диалоговых программ, достижения в области использования новых информационных технологий обработки текстовых документов.

Новизна работы. Выполненная диссертационная работа является одной из первых попыток разработки лингвистического обеспечения для системы информационного мониторинга, в частности, ИТСМ РАН. При ее реализации автором достигнуты новые результаты, основные из которых заключаются в следующем:

- исследована предметная область информационного мониторинга, анализа и оценки (далее – информационного мониторинга), определен круг ее специфических задач и очерчена проблематика;
- исследованы и апробированы возможности применения лингвистических методов представления знаний для решения задач области;
- исследована область патентоведения и изучена задача мониторинга научной деятельности в патентах;
- проведены теоретические исследования, результаты которых позволили применить метод категоризации к массиву утвержденных Министерством образования и науки индикаторов;
- предложен подход к построению гибкой и легко модифицируемой классификационной схемы, куда были встроены обработанные методом категоризации индикаторы и на базе которой был впервые спроектирован и разработан семантический словарь системы информационного мониторинга;

- впервые предложена и реализована возможность установления взаимосвязей между индикаторами (и другими терминами мониторинга) и алгоритмическими, информационными и нормативными ресурсами для прояснения их смысла и выработки согласованных терминов мониторинга;
- впервые в качестве статей семантического словаря предложены и реализованы параметризуемые статьи с интегрированными параметризуемыми дефинициями;
- реализована параметризуемая статья семантического словаря для индикатора «индексы самоцитирования в патентах»;
- предложен новый комбинированный метод построения запросов в сформированных параметризуемых статьях семантического словаря ИТСМ;
- впервые осуществлена программная реализация диалогового лингвистического обеспечения с использованием комплекса запросов системы Mimosa V5.

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. В целом диссертационная работа имеет экспериментальный характер, хотя предложенная теория и алгоритмы целиком реализованы в эксплуатируемой системе ИТСМ РАН. Эта система представляет собой совокупность программных модулей, предназначенных для обеспечения работы ряда подсистем: «Диаграммы», «Статистика» и «Семантический словарь». Все подсистемы активно задействованы при проведении процедур мониторинга, анализа и оценки научной деятельности РАН. Использование спроектированного и разработанного в результате выполнения работы семантического словаря ИТСМ РАН облегчает проведение вышеперечисленных процедур экспертами и позволяет прояснять смысл индикаторов (реализовано на примере индикатора «индексы самоцитирования в патентах»).

Значительная часть результатов, полученных в ходе выполнения данной работы, вошла в нижеперечисленные проекты Российского фонда фундаментальных исследований и Российского Гуманитарного Научного Фонда.

На защиту выносятся:

1. Подход на основе детального анализа результатов исследований в области компьютерной лингвистики, предполагающий применение семантического словаря как инструмента разрешения проблем информационного мониторинга.
2. Метод категоризации, примененный к массиву утвержденных Министерством образования и науки индикаторов.

3. Построение и реализация классификационной схемы, на которой основана архитектура разработанного семантического словаря ИТСМ РАН.

4. Механизм установления взаимосвязей между индикаторами и алгоритмическими, информационными и нормативными ресурсами ИТСМ РАН для прояснения их смысла и выработки согласованных терминов мониторинга и отслеживание их эволюции во времени.

5. Программная реализация диалогового лингвистического обеспечения с использованием комплекса запросов системы Mimosa V5.

6. Проектирование, разработка и интеграция семантического словаря в качестве функционального модуля в ИТСМ РАН.

7. Проектирование параметризуемых статей с интегрированными параметризуемыми дефинициями для использования в качестве статей семантического словаря.

8. Разработка и программная реализация параметризуемой статьи семантического словаря для индикатора «индексы самоцитирования в патентах».

9. Комбинированный метод построения запросов в сформированных параметризуемых статьях семантического словаря ИТСМ.

Апробация работы и публикации. Проведенные исследования и полученные результаты работы вошли в проекты РФФИ и РГНФ:

1. РФФИ, грант №09-07-00156;
2. РФФИ, грант№06-07-07001ано;
3. РГНФ, грант №05-03-03230а;
4. РГНФ, грант №06-02-04043а;
5. РГНФ, грант № 05-03-12328в.

Материалы диссертации докладывались на следующих международных конференциях:

1. Международная конференция по компьютерной лингвистике «Диалог-2006»;
2. Международная конференция по компьютерной лингвистике «Диалог-2007»;
3. Международная конференция по компьютерной лингвистике «Диалог-2008»;
4. Международная конференция «MEGALING-2006» «Горизонты прикладной лингвистики и лингвистических технологий»;
5. Международная конференция «MEGALING-2007» «Горизонты прикладной лингвистики и лингвистических технологий»;
6. Atlanta Conference on Science, Technology and Innovation Policy (ATLC-2007);
7. Atlanta Conference on Science and Innovation Policy (ATLC-2009);
8. 10th International Conference on Science and Technology Indicators (Vienna-2008);
9. The 2009 World Congress in Computer Science, Computer Engineering, and Applied Computing (WORLDCOMP’09);

10. Information and Brokerage Conference on Information and Communication Technologies in the EU's 7th Framework Programme (Moscow-2008);
11. ICT Proposers' Day (Budapest-2009).

Основные результаты диссертации опубликованы в 17 публикациях, в том числе три в рекомендованных ВАК журналах, и в двух научно-исследовательских отчетах ИПИ РАН – № гос. регистрации, № гос. регистрации за 2008, 2009 гг.

Структура диссертации. Диссертация состоит из введения, четырех глав, заключения, списка литературы и приложений (60 наименований). Работа изложена на 109 страницах, включающих 43 рисунка и 1 таблицу.

СОДЕРЖАНИЕ РАБОТЫ

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

В первой главе проводится анализ и обзор видов лингвистического обеспечения, в частности, словарей и тезаурусов, поскольку последние 15-20 лет именно они наиболее часто используются в качестве средств представления информационного контента в различных областях для решения ряда задач. Среди рассмотренных словарей приведены традиционные и электронные словари, идеографические словари и тезаурусы. Они проанализированы с точки зрения задач, функций и назначения в сравнении с аналогичными аспектами семантического словаря ИТСМ РАН с целью позиционирования разработанного словаря среди множества существующих словарей и тезаурусов.

Кроме того, поскольку описываемый в данной работе семантический словарь ИТСМ РАН содержит в себе некоторые черты формальных и неформальных онтологий, то возникла необходимость сопоставительного анализа различных видов словарей и онтологий с разработанным словарем (рис. 1). В частности, в параграфе 1.1 подробно описан ресурс EuroWordNet – автоматизированные словари (подтип лексических онтологий), построенные по модели WordNet. Словари такого типа объединяют в себе результаты современных разработок в области компьютерной лингвистики и широкое применение для решения различных задач, в том числе в качестве справочной системы и инструмента для проведения лингвистических исследований. При разработке семантического словаря ИТСМ автор стремился максимально учесть функциональность и преимущества таких словарей.

Далее на основании проведенного анализа в работе приводится описание специфики предлагаемого семантического словаря. Он содержит иерархию понятий (показатели, индикаторы, параметры, экспертные оценки и критерии) системы информационного мониторинга в сфере науки, определяя связи между терминами при помощи иерархических отношений и ассоциаций. Кроме того, словарь содержит формально определенное

отношение класс-подкласс (показатели-индикаторы, показатели-экспертные оценки, индикаторы-индикаторы результатов программ фундаментальных научных исследований, и т.д.).

В следующем параграфе главы 1 рассматривается лингвистическое обеспечение систем информационного мониторинга на примере семантического словаря ИТСМ РАН и приводится обоснование его функций, мотивированное особенностями проведения процедур мониторинга, анализа и оценки. В частности, различия в понимании экспертами смысла индикаторов являются серьезным препятствием в реализации всех трех процедур: информационный мониторинг, анализ, получение экспертных оценок результатов и результативности. Это вызвало необходимость решения задачи согласования понимания индикаторов. Кроме того, в силу особенностей формирования терминов мониторинга возникает задача частной референции, когда, например, одно название индикатора может обозначать целый класс индикаторов (индексы цитирования, которые ранжируются по авторам, по изданиям, по годам и т.п.).

Сложность решения вышеописанных задач информационного мониторинга обусловила выбор лингвистического обеспечения в виде семантического словаря.

Рис. 1. Спектр онтологий и семантический словарь ИТСМ.

Для расширения 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 этого индикатора в системе оценки;
- 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, содержится список всех вариантов и некоторая классификация вариантов алгоритма, которые могут использоваться для вычисления значений этого индикатора. Упомянутая функция дополняет классификационную функцию словаря по отношению ко всему набору индикаторов и других показателей системы мониторинга.

Расширение функциональности семантического словаря с целью разрешения разногласий между экспертами (пользователями системы мониторинга), предполагает, что словарь должен быть средством унификации терминов для пользователей этой системы.

Помимо включения в словарные статьи внешних ссылок смысл индикаторов раскрывается при помощи иллюстраций. Это позволяет также продемонстрировать зависимость их семантики от видов референции и конкретных референтов. Использование иллюстраций в словарных статьях (графики, диаграммы, блок-схемы и пр.) способствует облегчению процесса согласования понимания индикаторов пользователями системы мониторинга.

В результате обзора словарей и онтологий на основе проанализированных традиционных словарей и электронных ресурсов показано, что конструктивная новизна разработанного семантического словаря состоит в том, что он позволяет не только решать вышеперечисленные проблемы компьютерной лингвистики, свойственные данной области, но и наглядно демонстрирует один из методов проектирования и реализации вспомогательного ресурса для новой предметной области. Словарь содержит ссылки на внешние ресурсы, иллюстрации и нормативные документы, содержащие источники терминов области. Инструмент с таким сочетанием функций для области информационного мониторинга разработан впервые.

Вторая глава посвящена описанию особенностей основных составляющих семантического словаря системы информационного мониторинга – индикаторов информационного мониторинга в сфере науки, а также целям создания семантического словаря индикаторов анализа и оценивания, его функциям и особенностям построения классификационной схемы индикаторов, лежащей в основе архитектуры словаря (рис. 2).

В первом параграфе главы рассматривается специфика формирования основных составляющих словаря – 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 результативности.

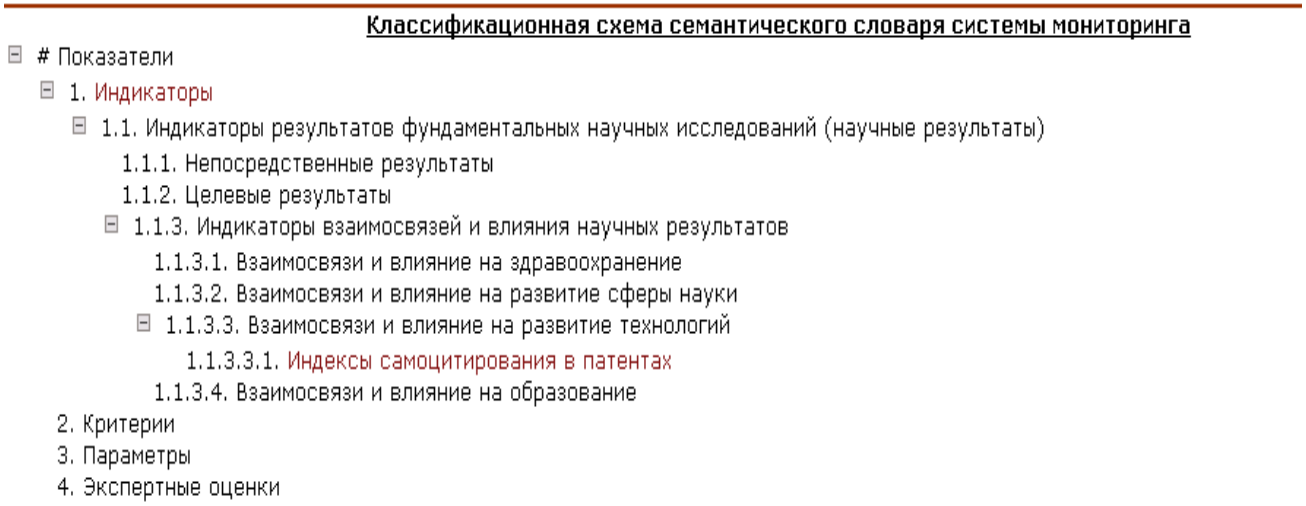

Рис. 2. Классификационная схема семантического словаря ИТСМ.

Далее описываются особенности процедуры мониторинга и соответствующие аспекты использования индикаторов. В частности, при проведении анализа данных и получении экспертной оценки нередко возникают проблемы, связанные с различными подходами существующих категорий пользователей к построению и эксплуатации систем индикаторов. Различия в подходах нередко являются следствием неоднозначного понимания пользователями смысла 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 именно в этой системе мониторинга.

Иногда содержательную информацию об отдельных индикаторах можно найти в нормативных документах или научных публикациях, но это встречается крайне редко.

В следующем параграфе главы описаны цели создания семантического словаря, которые связаны со следующими организационными стадиями выполнения программ НИР: планирование, выполнение и описание (демонстрация) полученных результатов. На каждой из описанных стадий традиционно проводится три процедуры:

1. Сбор данных о 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.

Чем сложнее научные программы и проекты, тем большее значение приобретает экспертиза и качественные (экспертные) оценки результативности научной деятельности. Результаты экспертизы и качественные оценки часто должны быть сформулированы экспертами 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 конвенциональными. Различия в понимании экспертами смысла различных показателей являются серьезным препятствием в реализации всех трех процедур (мониторинг, анализ, получение экспертных оценок). Поэтому основной целью создания семантического словаря ИТСМ РАН является получение адекватных экспертных оценок.

Далее описываются проблемы, возникающие на этапе получения экспертной оценки. Сложность проблемы согласования понимания проиллюстрирована в рамках классической модели К. Поппера.

Наложение модели К. Поппера на четыре стадии общения наглядно иллюстрирует тот этап общения экспертов, где возникают противоречия между ними при интерпретации индикаторов результативности (рис.3).

Таким образом, Рис. 3 иллюстрирует именно тот случай, которого желательно избежать. Поскольку логическая структура семантического словаря основана на результатах категоризации индикаторов, он обеспечивает согласование понимания экспертами смысла индикаторов системы мониторинга.

В третьем параграфе главы подробно описаны характерные функции семантического словаря и построение его классификационной схемы. Специфическими функциями семантического словаря являются:

- классификационная функция на уровне вариантов алгоритма;
- функция унификации терминологии области информационного мониторинга;
- репрезентативная функция прояснения смысла индикаторов.

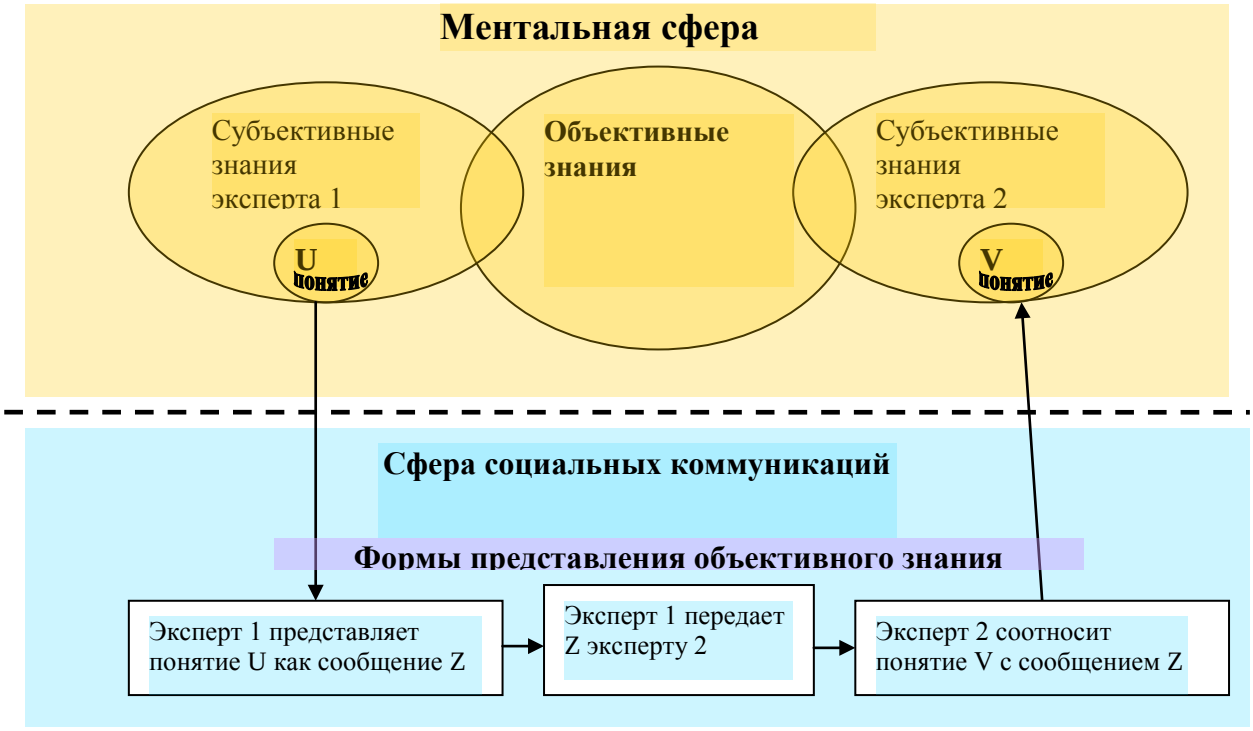

Рис. 3. Коммуникативная неудача при попытке передать смысл индикатора от одного эксперта другому.

Эти функции предопределили структуру словаря, которая была построена на основе классификационной схемы, поскольку методика согласования смысла индикаторов основана на категоризации знаний о мониторинге, анализе и оценке в сфере науки, в том числе, знаний о результатах категоризации и методе классификации. Разные категории показателей могут быть связаны между собой родовидовыми и функциональными тезаурусными отношениями.

При построении классификационной схемы была учтена и нашла свое отражение такая особенность структуры показателей сферы науки и системы информационного мониторинга как учет институциональности структуры РАН. Экспериментальный вариант системы мониторинга должен обеспечить анализ и оценку результатов и результативности деятельности РАН на четырех институциональных уровнях:

- макроуровень – Российская академия наук в целом;
- мезоуровень – институты Российской академии наук;
- микроуровень – научные коллективы институтов;
- наноуровень – ученые и инженеры.

Далее в параграфе 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 встраивание каждого предлагаемого к использованию индикатора в

классификационную схему, полученную в результате категоризации показателей. На второй стадии происходит уточнение смысла индикаторов посредством использования словарных статей, связанных с нормативными, информационными и алгоритмическими компонентами системы мониторинга.

Разработанная итеративная процедура согласования смысла индикаторов одновременно использует следующие компоненты системы мониторинга:

- нормативный компонент системы;
- семантический словарь с названиями и определениями видов индикаторов, характеристик, критериев, параметров и экспертных оценок;
- информационный компонент системы;
- библиотеку алгоритмов и программ (алгоритмический компонент системы).

Использование метода классификации для распределения перечисленных индикаторов по 6 категориям (рубрикам) классификационной схемы позволили не только решить проблему согласования смыслов индикаторов для индикаторов результатов научной деятельности РАН, но также продемонстрировать некоторые особенности распределения разработанных Приказом 68 индикаторов.

Третья глава посвящена технологическим принципам разработки семантического словаря. *В первом параграфе главы* рассматриваются задачи определения индикаторов в системах мониторинга. Представлены особенности формирования и использования индикаторов в целом и применительно к различным программам (социальным, научным и т.п.). В рамках системы базовых терминов системы мониторинга индикаторы определены как показатели количественной оценки любого из следующих трех видов, вычисляемые на основе информационных ресурсов системы мониторинга:

- результативности бюджетных программ, включая, как частный случай, результативность бюджетных расходов;
- ресурсной эффективности бюджетных программ, включая, как частный случай, эффективность бюджетных расходов;
- степени достижения целей и запланированных результатов деятельности.

В качестве вывода к первому параграфу приводится заключение о том, что изначальное отсутствие дефиниции понятия «индикатор» в нормативных документах, мотивирующих проведение мониторинга или создание систем мониторинга определенной предметной области или программы, также существенно затрудняет процесс мониторинга и оценки. Кроме того, отсутствие определений и примеров значений и смысла индикаторов, а также их согласованных, четких дефиниций, привязанных к определенным ресурсам, лишает лиц, участвующих в проведении мониторинга, возможности получить адекватные оценки той или иной деятельности и успешно завершить процесс мониторинга. Поэтому возникла потребность в

формировании инструмента, который бы учитывал специфику формирования и использования индикаторов в системе мониторинга, позволив свести к минимуму те погрешности, которые содержатся в нормативных документах, сопровождающих эту систему, а также наглядно и четко иллюстрировать смысл индикаторов и все существующие варианты их вычисления (в зависимости от изменения параметров и от вариантов рассматриваемого индикатора, которые варьируются год от года). Таким инструментом стал семантический словарь ИТСМ.

В следующем параграфе описывается система информационного мониторинга РАН, в которую был интегрирован семантический словарь. В соответствии с программой фундаментальных научных исследований РАН, утвержденной Президиумом РАН, эта система должна обеспечивать руководителей и сотрудников Президиума РАН информацией о результатах формирования, реализации и мониторинга Программы ФНИ РАН.

Далее в параграфе описываются цели ИТСМ РАН.

В заключение данного параграфа приводится пример работы одного из функциональных модулей ИТСМ РАН – подсистемы «Диаграммы».

В третьем параграфе главы представлена архитектура разработанного семантического словаря ИТСМ РАН, основанная на построенной автором классификационной схеме. В настоящее время из построенной схемы в рамках системы мониторинга ИСТМ 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 патентах. Реализация этой части схемы выполнена впервые в рамках исследований, описанных в данной работе. Ранее индикатор самоцитирования в патентах для системы мониторинга не вычислялся.

В параграфе 4 описывается процесс создания словарных статей в семантическом словаре. Словарная статья имеет динамический характер, поскольку содержит несколько редактируемых параметров обработки информационных полей при формировании запроса для вычисления индикаторов. Их изменение позволяет получать различные варианты вычисления одного индикатора. Все варианты сохраняются в ИТСМ и могут быть использованы в различных задачах системы. Сочетание параметров (полей на электронной форме вычисляемого индикатора) также впервые отобрано на основе проанализированных и обработанных автором информационных ресурсов – массива текстов патентов (данные РосПатента) – посредством программы аккумуляции и отбора патентов Mimosa v5 и специально разработанных автором модулей анализа. Каждый параметр, участвующий в вычислении индикатора, связан с различными аспектами его рассмотрения, что обусловлено сложной структурой показателей и институциональными уровнями РАН (Рис.4).

Заключительный параграф главы посвящен информационным ресурсам для вычисления индикаторов. Индикаторы представляют собой специфичные для области мониторинга агрегированные понятия, смысл которых затруднительно прояснить при помощи обычных дефиниций. В работе был проиллюстрирован их смысл при помощи нормативных документов, в которых они описываются, диаграмм, на которых они вычисляются с указанием параметров, рисунков, где изображены их связи с другими показателями или случаи использования, вариантов вычисления алгоритмов, и др. ресурсами. Такой подход к описанию смысла индикаторов был разработан в данной работе впервые. Разработка подхода была мотивирована исследованием ряда нормативных документов и документов отчетного характера, которые в той или иной степени содержали упоминания, ссылки и указания на индикаторы.

Далее описывается связь процедуры вычисления индикаторов в семантическом словаре ИТСМ с информационными ресурсами. Поскольку основной целью работы является технология разработки такого словаря для систем мониторинга, автором спроектирован семантический словарь и реализовано вычисление группы индикаторов «Индексы самоцитирования в патентах».

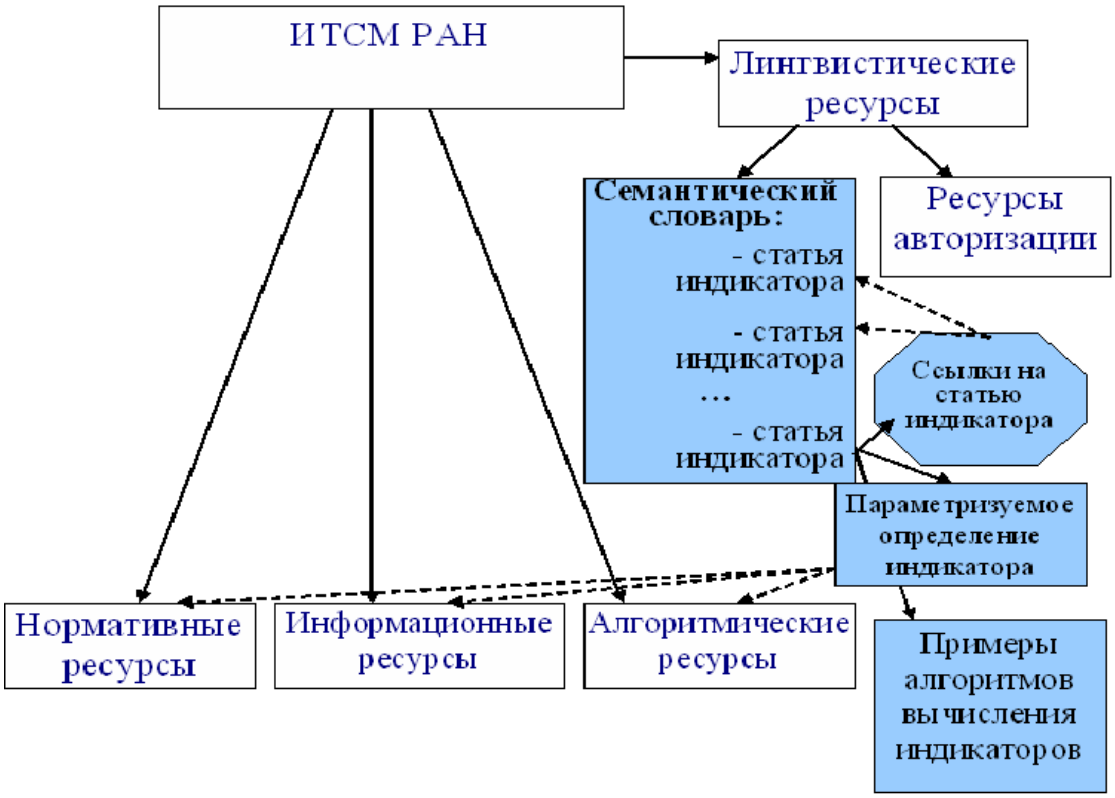

Рис. 4. Связь ИТСМ с ее ресурсами.

В четвертой главе представлены результаты вычислительных экспериментов, проведенных в рамках предложенной технологии разработки

семантического словаря ИТСМ в системах мониторинга. Словарь был протестирован на группе индикаторов – «индексах самоцитирования в патентах». Индикаторы этой группы вычисляются по принципу подсчета частоты встречаемости фамилий авторов патентов в полнотекстовых описаниях их изобретений. Кроме того, индексы могут быть сгруппированы по фамилии автора (или по всем доступным в базе данных патентных описаний фамилиям авторов), по временным промежуткам и по тематике публикуемых изобретений. В данной работе эти индикаторы и алгоритм их вычисления применительно к патентной сфере описываются впервые.

Индексы самоцитирования имеют большое значение для оценки результатов научной деятельности РАН, так как они позволяют выявить ту степень новизны и актуальность патентных изобретений, которые характеризуют наличие результатов в патентной сфере в ракурсе изобретений учреждений РАН. Возможно также использование индексов для исследования специфики отдельных предметных областей.

В семантическом словаре ИТСМ для вычисления различных вариантов этих индикаторов автором были подготовлены следующие информационные ресурсы:

1. Массив описаний патентов, отобранных по принципу максимального количества цитирований по представленным тематическим классам предметных областей наук;

2. Xml-файл, сформированный программой обработки патентов Patanalysis, которая была модифицирована автором данной работы специально для подготовки информационных ресурсов. В файле содержатся поля из html-описаний патентов, которые необходимы для вычисления группы индикаторов «Индексы самоцитирования в патентах». Программа Patanalysis модифицированной версии в качестве входных параметров использует массив кратких описаний патентов, сформированных в результате выполнения запроса в программе Mimosa v5.

Кроме того, была произведена настройка инструментария Mimosa V5 для визуализации, отбора и редактирования кратких патентных описаний по тематикам, по годам выдачи патентов, по стране патентования, и другим параметрам. Запрос с помощью системы Mimosa V5 был необходим для формирования массива описаний изобретений, патентообладателями которых являются учреждения РАН, и для предварительного отбора и просмотра патентов с максимальным количеством цитирований в их текстах. Полученные посредством Mimosa v5 описания патентов позволили выбрать параметры для вычисления группы индикаторов «индексы самоцитирования в патентах» в семантическом словаре ИТСМ. К таким параметрам относятся фамилии, имена, отчества авторов патентов, их национальная принадлежность, номер патентов, дата подачи заявки на патент, учреждение-патентообладатель и т.д.

Далее в работе описываются два этапа подготовки информационных ресурсов. На первом этапе средствами инструмента Mimosa v5 был сформулирован и отлажен запрос, который позволил выдавать патентные

описания, где в качестве патентообладателей указаны учреждения РАН. Необходимость в таком отборе была мотивирована тем, что разработанный семантический словарь является частью ИТСМ РАН, то есть системы, осуществляющий мониторинг деятельности российской науки.

Целью второго этапа отбора патентов являлся анализ патентных документов. В результате было выявлено, что наибольшее количество патентов, в которых содержатся ссылки на статьи и другие публикации, относится к областям химии, биологии, физики и радиоэлектроники. Поскольку основной задачей было вычислить различные индикаторы группы «индексы самоцитирования в патентах», первоначально тестовый массив патентов был сформирован именно из документов наиболее частотных классов в отношении ссылок на внешние источники. Кроме того, на этом этапе была осуществлена модификация программы Patanalysis, осуществляющая структуризацию ссылок цитирования в описываемых патентах РФ. Информация о структурированных ссылках в патентах была необходима в силу специфики вычисляемой в словаре группы индикаторов («индексов самоцитирования в патентах»).

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.

В следующем параграфе приводятся особенности построения индикаторов. Семантический словарь ИТСМ, основанный на классификационной схеме показателей мониторинга (содержащей индикаторы), позволяет просматривать все уровни иерархии этой схемы и вычислять отдельные доступные индикаторы (в частности, «индексы самоцитирования в патентах»). Кроме того, ввиду особенностей проектирования и реализации схемы в ней существует возможность расширения как существующих категорий, так и добавления новых показателей. Xsd-представление с рекурсивной ссылкой позволяет формировать уровни иерархии внутри схемы с необходимой разработчику степенью подробности.

Это делает возможным построение новых индикаторов внутри самой схемы посредством их последовательной интеграции на нужные уровни иерархии и дальнейшего определения в семантическом словаре (рис.5). Такой подход расширяет функциональные возможности ИТСМ и позволяет не только описывать новые термины и понятия в семантическом словаре, но и устанавливать между ними необходимые связи и наглядно демонстрировать их на общей схеме классификации.

В параграфе 3 приведены результаты процедуры вычисления значений индикаторов. Вычисление значений индикаторов происходит во вкладке «Параметризуемая статья семантического словаря» ИТСМ для группы индикаторов «индексы самоцитирования в патентах».

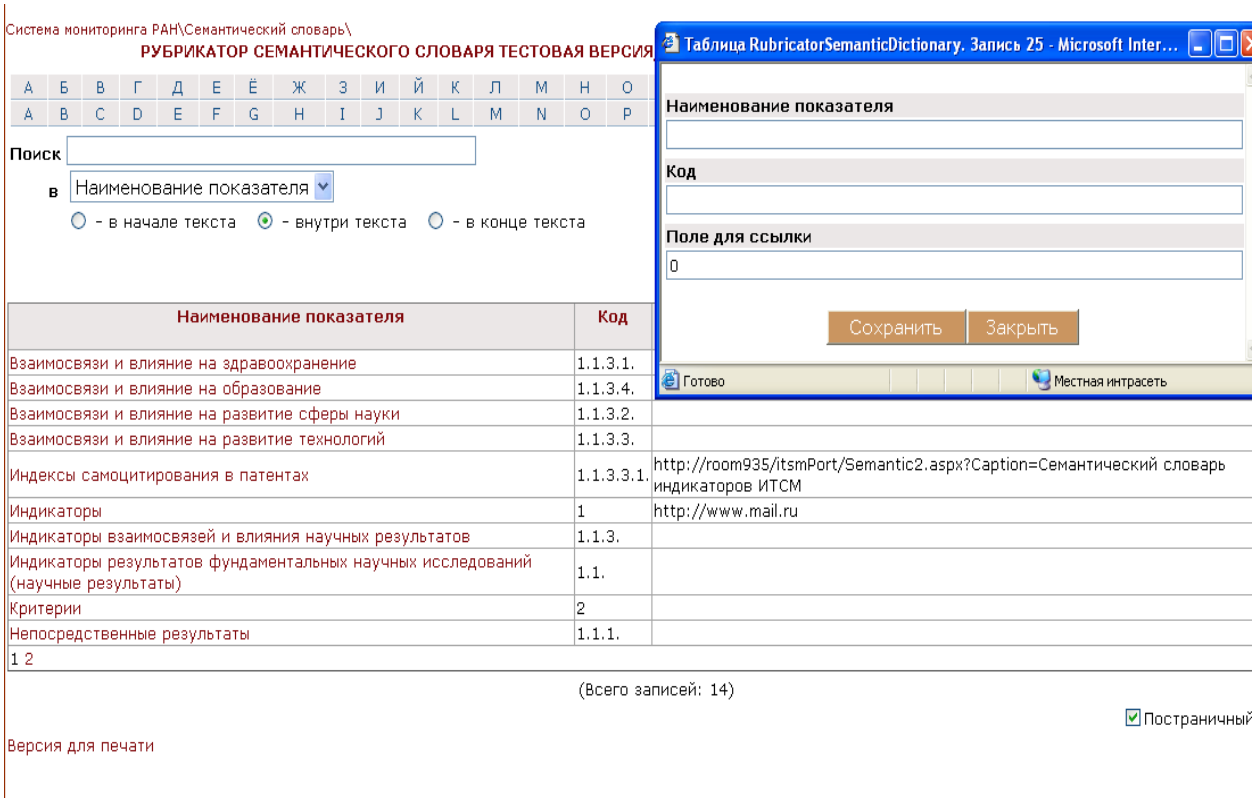

Рис.5. Добавление новых категорий индикаторов.

На примере продемонстрирована работа алгоритма вычисления значений индикаторов. В просматриваемом текстовом описании патента отбираются значимые данные: фамилии, имена и отчества авторов патента, дата публикации патента и тематические классы патента (классы Международной патентной классификации (МПК)). Эти данные в дальнейшем выступают в качестве параметров вычисления этой группы индикаторов: ими пополняется xml-файл с полями, сформированными в соответствии с xsd-схемой при помощи разработанного автором механизма извлечения тегов. Xsd-схема была построена автором в соответствии со структурой патентных описаний и адаптирована для задач вычисления индикаторов в семантическом словаре ИТСМ. Кроме того, параллельно с отбором и сохранением параметров, необходимых для вычисления индексов самоцитирования в патентах, происходит поиск и подсчет встречаемости фамилий авторов просматриваемых патентов в текстах их изобретений с целью подсчета индексов их самоцитирований за выделенный временной промежуток по определенной тематике. По всему массиву имеющихся патентов отбираются те патенты, в которых в качестве авторов патентов фигурируют выделенные пользователем в параметризуемой статье словаря авторы, отвечающие фиксированному промежутку времени и соответствующие обозначенной им же тематике МПК. Для каждого патента, отвечающего такому условию, подсчитывается частота встречаемости автора в тексте описания патента. Индексы, посчитанные для нескольких патентов, суммируются для фиксированного набора параметров, то есть, для каждого автора, каждого временного периода и отдельно взятой тематики формируется отдельный,

суммарный индекс по релевантным запросу патентам. Поэтому все возможные варианты подсчета индексов самоцитирования патентов образуют группу индикаторов с соответствующим названием, а не один отдельный, фиксированный индикатор с единственным алгоритмом вычисления.

Полученная группа индикаторов, вычисленная вне зависимости от существующих в ИТСМ структур, инвариантна как по времени, так и относительно других данных ИТСМ, что позволяет вычислять индикатор в целом, а не по выделенным аспектам, которые могут меняться. Такая структура дает возможность вычислять значения индикаторов в независимом режиме и получать представления о смысле индикатора и его связях с внешними ресурсами.

Параграф 4 посвящен словарным статьям семантического словаря. Словарные статьи семантического словаря в системе информационного мониторинга структурированы в соответствии с ресурсами, с которыми связан словарь, с учетом потребностей пользователей системы. Каждая статья содержит параметризуемую дефиницию (то есть определение значения индикатора, зависящее от нескольких модифицируемых параметров) и поэтому носит название параметризуемой. Для отдельно вычисляемого индикатора параметризуемая статья выглядит в соответствии с необходимыми для него параметрами. Каждая статья связана с отдельной группой индикаторов в классификационной схеме, в которой представлены разные уровни классификации индикаторов с возможностями выбора вариантов их вычисления и ссылкой на внешние информационные, алгоритмические и нормативные компоненты системы.

В основе формы параметризуемой статьи лежит запрос, результатами выполнения которого являются не фиксированные информационные поля, как это происходит при выполнении типичных запросов, а результат обработки информационных полей, поскольку при его построении используются не только поля, но и параметры их обработки. Например, для предоставления пользователю выбора фамилии, имени и отчества автора патента при задании параметров запроса для вычисления индикатора дальнейший выбор одного из авторов в этом списке, по сути, является параметром обработки информационных полей PatentNumber (номер патента), Surname (фамилия автора патента), Name (имя автора патента) и Patronimic (отчество автора патента). То есть параметром обработки этих полей является комбинированное поле FIO (фамилия, имя, отчество), которое заполняется, если найден номер патента, соответствующий выделенной пользователем фамилии автора. Таким образом, формируемый запрос является комбинированным запросом на поиск с последующей обработкой при заданных параметрах ее проведения.

На данный момент в семантическом словаре с целью демонстрации результатов работы реализована параметризуемая статья для группы индикаторов «индекс самоцитирования авторов патентов». В состав параметров этой статьи включены фамилия, имя, отчество авторов патентов,

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 для фиксированных параметров обработки информационных полей схемы данных.

В последнем параграфе главы описывается программная реализация макета семантического словаря. Семантический словарь системы мониторинга был реализован средствами проектирования и разработки ИТСМ. Одним из основных принципов построения системы мониторинга или ИТСМ является разделение всех функциональных подсистем на две категории (базовые и прикладные). Сначала проектируются базовые функциональные подсистемы, а затем создаются прикладные функциональные подсистемы методом порождения, в том числе, порождение на основе графического интерфейса.

ИТСМ объединяет в себе множество ресурсов, в том числе и семантический словарь (рис. 6, 7). Поскольку единицами словаря являются индикаторы, адекватное вычисление и согласование которых, как было выяснено автором в ходе теоретических исследований в рамках данной работы, напрямую влияет на успешность процесса мониторинга, то наличие в составе ИТСМ внешних по отношению к словарю ресурсов позволило продемонстрировать один из способов решения этой проблемы. Иллюстрация смысла индикаторов именно посредством внешних ресурсов: алгоритмической, нормативной и информационной компонент – является одним из принципиально новых решений, впервые описанных в данной работе.

Роль семантического словаря в ИТСМ не ограничивается функцией экспликации несогласованности пониманий терминов мониторинга экспертами и ее разрешением. Семантический словарь позволяет элиминировать неточности ключевых понятий, задействованных в работе системы, указывать ссылки на наиболее полные и релевантные современному состоянию предметной области ресурсы, а также анализировать, отслеживать и корректировать как сами понятия и алгоритмы их определяющие, так и связи между существующими и новыми показателями ИТСМ.

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

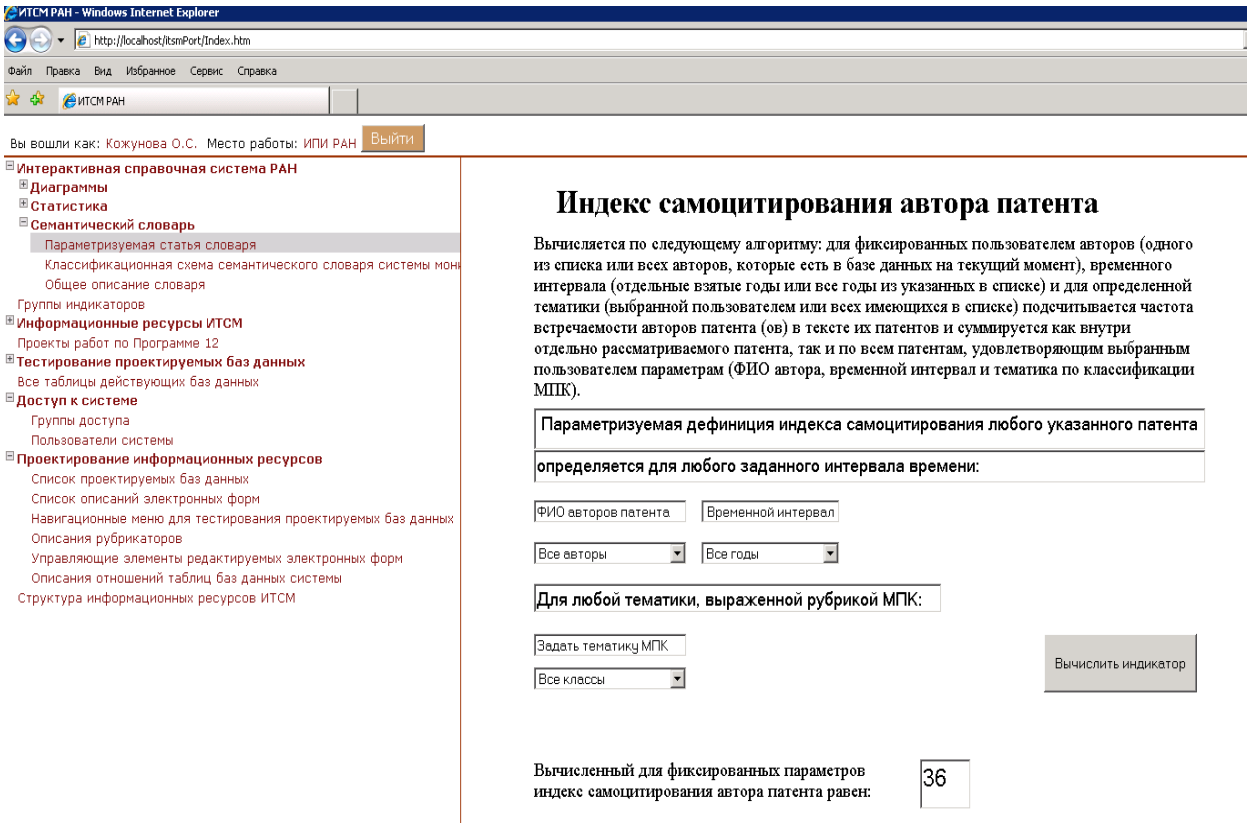

Рис. 6. ИТСМ, Семантический словарь. Вкладка «Параметризуемая статья словаря».

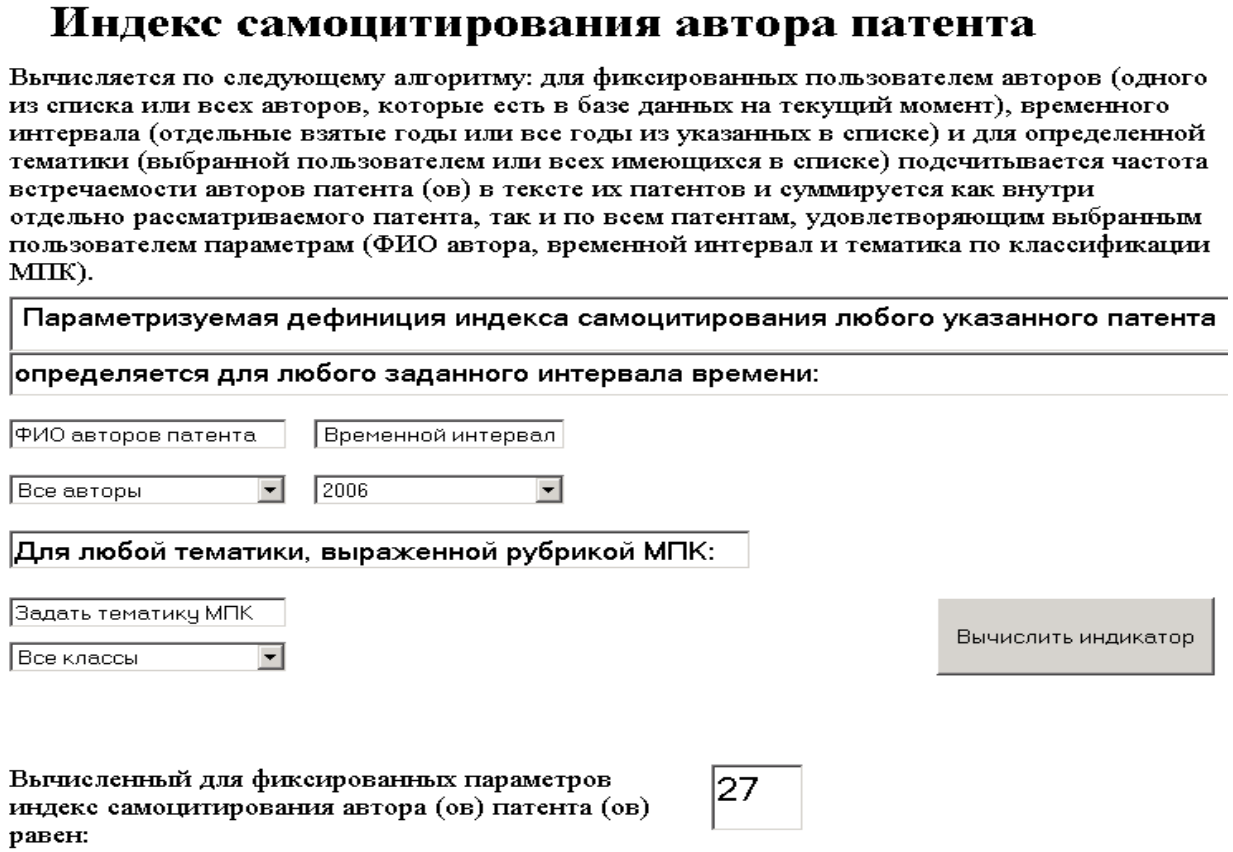

Рис. 7. Параметризуемая статья для вычисления индикаторов «индексы самоцитирования в патентах».

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 логических оснований информационной и компьютерной наук, исследование основных стадий становления каждой из наук и методов, разработанных главными представителями научных направлений)
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: проектирование макета семантического словаря ИТСМ РАН, 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 в системе 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 классификационной схемы и архитектуры 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 и факторы конвергенции информационной и компьютерной наук // **Информатика и ее применения,** том 2, вып. 1. – М.: Торус Пресс, 2008. – с.77-98. (Личный вклад диссертанта: анализ логических оснований информационной и компьютерной наук, исследование основных стадий становления каждой из наук и методов, разработанных главными представителями научных направлений)

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 // Atlanta Conference on Science, Technology, and Innovation Policy 20097. Atlanta, USA, Georgia Institute of Technology, 2009. (Личный вклад диссертанта: описание модифицированной классификационной схемы, семантического словаря и части структуры ИТСМ)

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 и его функциональности в ИТСМ РАН и обзор аналогичных инструментальных средств)