На правах рукописи

Соколина Ксения Александровна

АВТОМАТИЗИРОВАННОЕ УПРАВЛЕНИЕ АГЕНТНЫМ ПОИСКОМ ТЕМАТИЧЕСКОЙ ИНФОРМАЦИИ В ИНТЕРНЕТЕ (НА ПРИМЕРЕ НАУЧНОГО НАПРАВЛЕНИЯ «ФИЗИКА ПЛАЗМЫ»)

Специальность 05.13.01 – системный анализ,
управление и обработка информации
(в информационных системах)

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Автор:

Москва – 2018

Работа выполнена в Национальном исследовательском ядерном университете «МИФИ».

Научный руководитель:	**Оныкий Борис Николаевич**, доктор технических наук, профессор, НИЯУ МИФИ, заведующий кафедрой «Анализ конкурентных систем»
Официальные оппоненты:	**Синицын Владимир Игоревич**, доктор физико-математических наук, Федеральный исследовательский центр «Информатика и управление» Российской академии наук, заведующий отделом «Информационные технологии управления»
	Григорьева Мария Александровна, кандидат технических наук, Научно-исследовательский центр «Курчатовский институт», Курчатовский комплекс НБИКС-технологий, Лаборатория технологий больших данных для проектов в области мега-сайенс, старший научный сотрудник
Ведущая организация:	Объединенный институт ядерных исследований, г. Дубна

Защита диссертации состоится «18» апреля 2018 г. в 15 час. 00 мин. на заседании диссертационного совета Д 212.130.03 на базе Национального исследовательского ядерного университета «МИФИ»: 115409, г. Москва, Каширское шоссе, д. 31. Тел.: +7 (499) 324-84-98.

С диссертацией можно ознакомиться в библиотеке Национального исследовательского ядерного университета «МИФИ» и на сайте http://ods.mephi.ru.

Отзывы и замечания по автореферату в двух экземплярах, заверенные печатью, просьба высылать по вышеуказанному адресу на имя ученого секретаря диссертационного совета.

Автореферат разослан «13» марта 2018 г.

Ученый секретарь диссертационного совета Д 212.130.03, д.т.н. Леонова Н.М.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Агентные технологии являются очередным шагом в создании систем интеллектуальной обработки данных. Научные работы по этому направлению начались в конце прошлого века. В 1996 г. было создано агентство FIPA (Международное сообщество разработчиков интеллектуальных агентов), которое взяло на себя функции стандартизации и координации работ в области агентных технологий. В России агентные технологии являются частью широкого фронта работ по системам «Big Data». В этом направлении известны работы таких отечественных ученых, как Кореньков В.В., Крюков Ю.А., Воеводин В.В., Будзко В.И., Сенаторов М.Ю., Солдатов А.А.

По определению FIPA «Агент – это сущность, которая находится в некоторой среде, от которой она получает данные, которые отражают события, происходящие в среде, интерпретирует их и исполняет команды, которые воздействуют на среду». В настоящей работе агент – это поисковая программа, которая автоматически инициируется по определенному сценарию, сканирует заданные тематические группы источников информации и доставляет эту информацию в полнотекстовую базу данных Мультиагентной информационно-аналитической системы (МИАС).

Актуальность темы. Быстрый рост количества научно-технической информации (НТИ) в Интернете сопровождается ее распределением по многочисленным сайтам университетов, научных центров и их тематических подразделений. Увеличивается количество сайтов различных научных групп и отдельных специалистов. В этих условиях сбор и обработка НТИ интерактивными методами, с помощью глобальных поисковых систем становятся практически невозможными из-за большой размерности поисковых задач. Одним из наиболее перспективных путей решения этой проблемы является использование агентных технологий, реализующих эти операции круглосуточно, без участия пользователя. Поэтому тема данной диссертации, в которой решаются задачи управления агентной технологией сбора и обработки научно-технической информации в Интернете, является актуальной.

Работы по исследованию и разработке МИАС выполнялись в рамках Федеральной целевой программы «Научные и научно-педагогические кадры инновационной России» по проекту

«Мультиагентные информационно-аналитические системы по естественнонаучным и технологическим направлениям» №16.740.11.0129 от 02 сентября 2010 года.

Объектом исследования в данной работе является Мультиагентная информационно-аналитическая система по естественнонаучным и технологическим направлениям.

Предметом исследования является технология агентного поиска и обработки тематической научно-технической информации в сети Интернет.

Цель диссертационной работы состоит в создании и экспериментальном исследовании «ядра» Мультиагентной информационно-аналитической системы на примере тематического направления «Физика плазмы». Ядром МИАС будем называть необходимое количество программно-технических средств и баз данных, обеспечивающих все функции системы по ограниченному числу тематических направлений и используемых языков. Ядро системы должно оставаться неизменным при тематическом и лингвистическом масштабировании (развитии) системы.

Достижение поставленной цели предполагает решение следующих **основных задач:**

– разработка концептуальной модели мультиагентной системы, позволяющей осуществлять регулярное автоматизированное информирование пользователя по его предметной специализации;

– формирование тематических баз данных для управления агентным поиском - маршрутной базы данных, содержащей адреса для обращения агентов, тематического тезауруса для фильтрации и рубрикации поступающей информации;

– разработка и исследование методов и средств актуализации маршрутной базы данных на основе обработки регулярно поступающих агентных коллекций документов;

– разработка и исследование методов и средств актуализации многоязычных тематических тезаурусов, включая подключение новых иностранных языков (алфавитных и иероглифических);

– исследование отношений между терминами тематических тезаурусов по введенному в работе индексу общности;

– разработка и реализация методов регулярного выпуска типовых информационно-аналитических отчетов для пользователей:

дайджестов (тематических новостных подборок), семантических сетей с различными типами отношений между объектами, досье объекты профессионального интереса пользователя.

Научная новизна полученных результатов:

– Поставлены и решены задачи управления агентным поиском НТИ в Интернете в мультиагентных информационно-аналитических системах. Существенной новизной предложенных решений является расширение функций информационно-аналитического обслуживания пользователей – наряду с запросно-ответным режимом в них реализованы функции обеспечения пользователей информационно-аналитическими отчетами.

– Впервые для агентного поиска созданы управляющие базы данных: «Мировые научно-исследовательские и технологические организации по физике плазмы» и «Тезаурус по физике плазмы в международном стандарте TMX 1.4b». Разработаны алгоритмы человеко-машинного управления актуализацией баз данных.

– Приоритет и авторские права автора диссертации на управляющие базы данных зарегистрированы в Федеральной службе по интеллектуальной собственности (Роспатент) и Бюро регистрации авторских прав при Библиотеке Конгресса США (US Copyright Service).

– Разработанная методика построения многоязычных тезаурусов позволила, в частности, создать трехъязычный (англо-русско-китайский) тезаурус и продолжить его расширение на другие языки.

– Введена новая характеристика терминов тезауруса – индекс общности – и предложен метод ее вычисления. Показано, что научные тематические тезаурусы имеют сравнительно небольшое (порядка двух-трех десятков) количество понятий с высоким индексом общности. Показано, что этот результат имеет практически полезное следствие при тематическом и лингвистическом масштабировании агентной системы.

– Совокупность решенных задач позволила создать ядро мультиагентной информационно-аналитической системы по естественнонаучным и технологическим направлениям.

Практическое значение полученных результатов:

– Разработана и реализована методика построения систем управления агентным поиском тематической научно-технической

информации в Интернете, инвариантная по отношению к различным тематическим направлениям и национальным языкам.

– Результаты диссертационной работы использованы в научной и учебной деятельности кафедр «Физика плазмы» и «Анализ конкурентных систем» НИЯУ МИФИ и в производственной деятельности компании «Аналитические бизнес решения».

– Системное решение поставленных в диссертации задач определило возможность для масштабирования системы по другим тематическим направлениям («грид-системы», «лазерные промышленные технологии», «фотоника» и т.д.), а также по используемым национальным языкам.

Методологической основой работы является системный анализ и системное проектирование. Использованы методы теории вероятностей и статистики, методы регрессионного анализа, экспериментальные методы исследования, а также решение тестовых задач для оценки полноты и качества выполнения информационно-аналитических функций системы. Также использована методология, заложенная в международные стандарты FIPA и стандарт TMX 1.4b. В диссертации представлены примеры автоматизированного формирования дайджестов, различных вариантов построения семантических сетей и досье на объекты профессионального интереса пользователя.

Основные положения диссертации, выносимые на защиту:

– методика формирования и поддержания в актуальном состоянии базы данных «Мировые научно-исследовательские и технологические организации по физике плазмы»;

– методика формирования многоязычных тематических тезаурусов, в частности, Тезауруса по физике плазмы на русском, английском и китайском языках;

– технология агентного поиска новостной тематической информации с использованием управляющих баз данных («Мировые научно-исследовательские и технологические организации по физике плазмы» и «Тезаурус по физике плазмы в международном стандарте TMX 1.4b») и метод динамического управления их актуализацией;

– экспериментальная реализация агентного поиска и технологии автоматизированного формирования типовых информационно-аналитических отчетов в МИАС.

Апробация работы. Основные результаты диссертации докладывались на следующих всероссийских и международных научных конференциях:

– School on Nuclear Electronics & Computing Based on XXV International Symposium on Nuclear Electronics & Computing (Будва, Черногория, 2015 г.);

– International School JINR-CERN-MEPHI on Information Technologies «GRID and Advanced Information Systems» (Дубна, 2015 г.);

– Международная молодёжная конференция «Современные проблемы прикладной математики и информатики», MPAMCS 2014 (Дубна, 2014 г.);

– Всероссийская научная Интернет-конференция с международным участием «Современные системы искусственного интеллекта и их приложения в науке» (Казань, 2013 г.);

– Научные сессии НИЯУ МИФИ (Москва, 2012-2015 гг.).

Публикация результатов. Основные результаты диссертации опубликованы в 18 печатных работах, из них 6 статей в периодических научных изданиях, рекомендованных ВАК России (в том числе 1 работа в журнале, входящем в реферативную базу данных SCOPUS), 9 работ в статьях и материалах конференций и 3 свидетельства о регистрации баз данных.

Достоверность результатов, представленных в диссертации, подтверждается результатами экспериментальных исследований, использованием приведенных данных в опытной эксплуатации МИАС. Все данные прошли апробацию в научных изданиях, на международных конференциях. На базы данных, управляющие агентным поиском, получены свидетельства о регистрации авторских прав.

Личный вклад автора. Научные результаты, вынесенные на защиту, принадлежат лично автору. При создании и вводе в эксплуатацию МИАС автор самостоятельно разработал все компоненты системы, составившие ее ядро. Соавторами в публикациях являются коллеги из команды по разработке МИАС.

Объем и структура работы. Диссертация состоит из введения, четырех разделов, заключения, списка литературы и 6 приложений. Общий объем работы – 173 страницы машинописного текста, из них

129 страниц основного текста. Работа иллюстрирована 12 таблицами и 54 рисунками. Список литературы содержит 58 источников, в том числе, 20 на иностранных языках.

Диссертация представляется к защите по специальности 05.13.01 – системный анализ, управление и обработка информации (в информационных системах), т.к. ее содержание соответствует областям исследований, указанных в паспорте специальности (пункты 2, 3, 6, 7, 9, 12).

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** рассматривается актуальность задачи управления агентной технологией сбора и обработки научно-технической информации в Интернете, определяются цели и задачи диссертации, научная новизна полученных результатов. Обосновывается необходимость построения динамической системы управления агентным поиском, в частности, разработки методов и средств актуализации тематической маршрутной базы данных для управления агентным поиском и многоязычного тематического тезауруса для фильтрации агентных сообщений.

В **первом разделе** проведен обзор работ, посвященных автоматизированным системам агентного поиска и обработки научно-технической информации в Интернете. Выявлено два принципиально разных подхода к построению таких систем.

Первый подход состоит в использовании глобальных поисковых систем, рассчитанных на произвольный запрос пользователя, и последующем ранжировании документов в выдаче. Основными проблемами данного подхода являются: неизбежность больших выдач на запрос, неуправляемость со стороны пользователя принципами ранжирования документов, ограниченный охват источников информации (только источники, проиндексированные поисковыми системами), режим обслуживания пользователя только в формате «запрос-ответ».

Второй подход заключается в сканировании тематической группы сайтов для регулярного информирования пользователей-специалистов о новых публикациях в их предметной области. Основная идея состоит в том, что поиск информации осуществляется агентами

автоматически по заранее установленным маршрутам и без прямого участия пользователя. Наиболее существенным недостатком данного подхода является статичность контента маршрутной базы данных агентного поиска и, как следствие, снижение ее актуальности. В данной диссертации предлагается устранить этот недостаток путем перехода к динамическим методам актуализации баз данных, управляющих агентным поиском – тематических маршрутных баз и многоязычных тезаурусов для фильтрации полученных агентных коллекций.

Идея динамического управления агентным поиском положена в основу концептуальной модели Мультиагентной информационно-аналитической системы по естественнонаучным и технологическим направлениям, представленной на рис. 1.

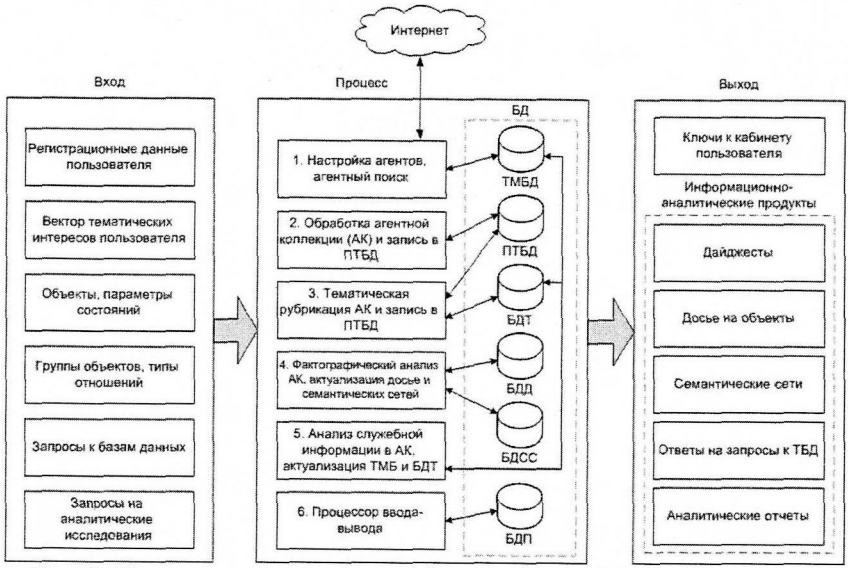

Рис. 1. Концептуальная модель Мультиагентной информационно-аналитической системы (МИАС) по естественнонаучным и технологическим направлениям

На рис. 1: ТМБД – тематическая маршрутная база данных, ПТБД – полнотекстовая база данных, БДТ – база данных тезаурусов, БДД – база данных досье, БДСС – база данных семантических сетей, БДП – база данных пользователей.

Главное отличие МИАС от предшествующих разработок состоит в том, что это не только запросно-ответная система, а, прежде всего, система регулярного информационно-аналитического обеспечения профессиональных пользователей тематической информацией по направлениям их деятельности.

Модель пользователя МИАС описывается индексом его тематических интересов:

$$\rho_{U_i}^{(K,P)}\left(v_1 I_1, v_2 I_2, \dots, v_j I_j, \dots, v_N I_N\right), \quad (1)$$

где $K, P-$ индекс коллективного или персонального пользователя;

$U_i -$ регистрационный индекс i-го пользователя, $i = 1, 2, \dots, M, M -$ общее число зарегистрированных пользователей в системе;

$I_j -$ индекс тематической рубрики, $j = 1, 2, \dots, N$, N – размерность тематического индекса пользователя;

$v_j -$ относительная значимость для пользователя j-ой тематической рубрики;

$\sum_{j=1}^{N} v_j = 1$, оценки даются самим пользователем при регистрации.

Качество регулярного тематического информационного обслуживания предложено оценивать с помощью коэффициента полноты тематического обслуживания пользователя:

$$K_{U_i} = \sum_{j=1}^{N_\Phi} v_j, \quad 0 \le K_{U_i} \le 1, \quad (2)$$

где $N_\Phi -$ фактическое количество обслуживаемых тематических рубрик.

Производным параметром является коэффициент средней полноты тематического агентного обслуживания пользователей в системе:

$$K_S = \frac{\sum_{i=1}^{M} K_{U_i}}{M}, \quad (3)$$

где M – общее число пользователей, зарегистрированных в системе.

Значения этого коэффициента, отнесенные к отрезкам времени, т.е. $K_{U_i}(t_i)$ позволяют исследовать динамику полноты обслуживания и сравнивать различные системы по параметру качества обслуживания пользователей.

Второй раздел посвящен решению вопросов управления агентным поиском и, прежде всего, вопросам формирования тематической группы источников.

Тематическая маршрутная база данных (ТМБД), определяющая для агентов пространство поиска, формируется итерационным методом. Стартовое состояние определяется экспертным путем. Впоследствии сетевые адреса новых источников тематической информации выявляются при обработке поступающих в систему агентных коллекций. Архитектурная модель агентной системы с контуром динамического управления ТМБД представлена на рис. 2.

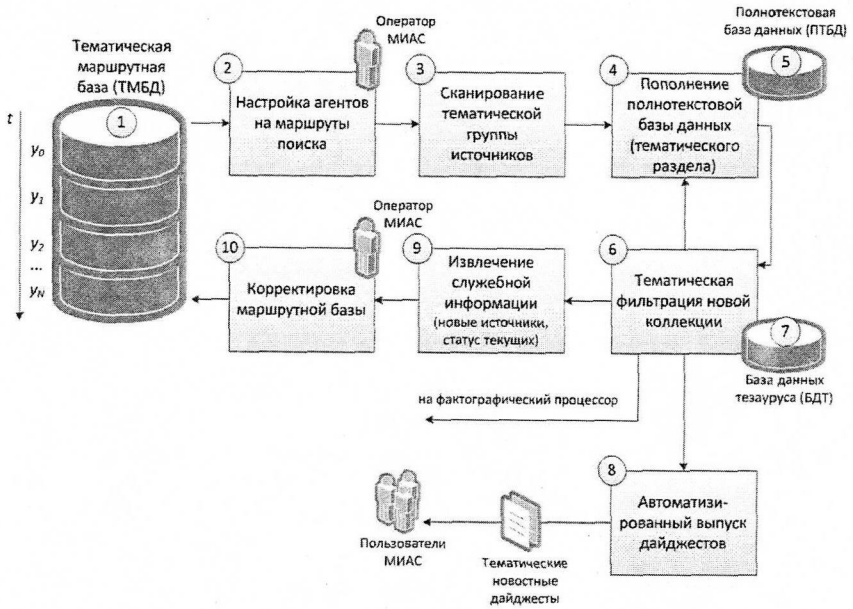

Рис. 2. Архитектурная модель МИАС

В системе 10 функциональных модулей. Обработка данных осуществляется в двух контурах: целевом, имеющем выход на пользователя (модули 1–8) и служебном, предназначенном для контроля за источниками и актуализации ТМБД (модули 6, 7, 9, 10, 1). Работа служебного контура находится под контролем оператора, который принимает решения об актуализации ТМБД – включение

новых источников или архивация бездействующих. В данном разделе приведены алгоритмы действий оператора для обоих случаев.

Для тематического направления «Физика плазмы» автором сформирована и поддерживается ТМБД «Мировые научно-исследовательские и технологические организации по физике плазмы» (зарегистрирована в патентных ведомствах РФ и США). Она представляет собой набор структурированных сведений об организациях, занимающихся исследованиями в области физики плазмы, включая веб-адреса для настройки поисковых агентов и получения актуальной новостной информации об исследованиях. Структура ТМБД разработана как типовая и используется в МИАС для различных тематических направлений («грид-системы», «лазерные промышленные технологии», «фотоника» и т.д.).

При настройке агентов по адресам ТМБД появляются следующие трудности: каждый источник имеет свой интерфейс и условия доступа к информации. Все эти особые условия должны быть отражены в поисковых предписаниях агентов. Для преодоления этой трудности в работе предложено использовать три типа шаблонов для настройки агентов и три оценки успешности обращения к определенному сайту. На рис. 3 представлен фрагмент ТМБД по физике плазмы, где доступность источников отражена по принципу светофора.

№	Наименование на русском языке	Подразделение	Ссылка на новостную ленту	Доступность источника
1	Австралийский национальный университет	Research School of Physics and Engineering - Plasma Research Laboratory	http://physics.anu.edu.au/prl/news.php	зеленый
2	Сиднейский университет	School of Physics	http://sydney.edu.au/news/physics/1736.html	красный
3	Университет Флиндерс	School of Chemical & Physical Sciences	http://blogs.flinders.edu.au/flinders-news/feed/	
4	Инсбрукский университет имени Леопольда и Франца	Institut für Ionenphysik und Angewandte Physik	http://www.uibk.ac.at/ionen-angewandte-physik/aktuelles.html	
5	Венский технический университет	Institute of Applied Physics	http://www.iap.tuwien.ac.at/www/news/index	желтый
6	Университет Буэнос-Айреса	Instituto de Física del Plasma	No newsline	
7	Институт прикладных проблем физики НАН Армении	Plasma Physics and Acoustics Laboratory	No newsline	
8	Институт физики им. Б.И.Степанова НАН Беларуси	Plasma physics laboratories	http://ifanbel.bas-net.by/russian/index.html	
9	Центр ядерных исследований Бельгии	Nuclear fusion research	http://www.sckcen.be/en/NewsPageRSSFeed	

Рис. 3. Фрагмент ТМБД с цветовой маркировкой доступности источников

Динамика состояний ТМБД по физике плазмы представлена на рис. 4. В настоящее время в базе содержится 575 источников информации по физике плазмы.

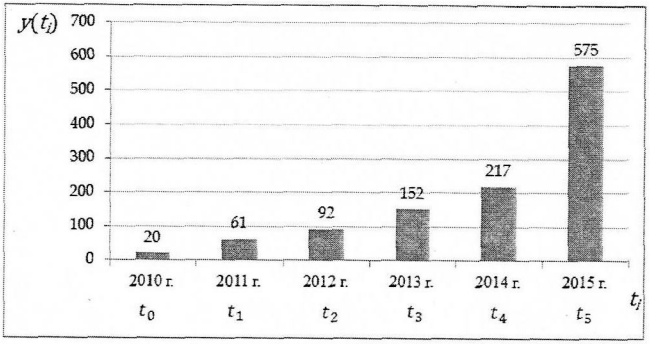

Рис. 4. Динамика роста количества информационных источников в ТМБД по физике плазмы

Методами регрессионного и дисперсионного анализа автором установлено, что процесс пополнения ТМБД новыми источниками, описывается случайным временным рядом, для которого вычислен линейный тренд и коридор стандартных отклонений:

$$\tilde{y}(t) = (20 + 49t) \pm 16. \qquad (4)$$

График функции представлен на рис. 5.

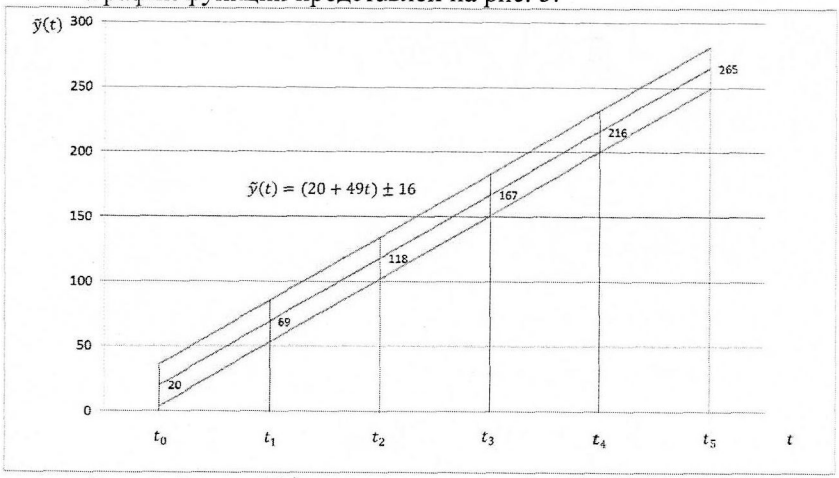

Рис. 5. Тренд $\tilde{y}(t)$ в коридоре стандартных отклонений

Случайные отклонения от тренда связаны со случайностью потока ссылок на новые источники в агентных коллекциях.

Для оценки результативности обращения агентов по маршрутам ТМБД введен коэффициент успешности агентного поиска S_a, представляющий собой отношение результативных маршрутов к общему числу маршрутов в базе. Для ТМБД по физике плазмы:

$$S_a^{\text{ФП}} \cong 0{,}7 \div 0{,}9. \qquad (5)$$

В процессе пополнения ТМБД абсолютная величина неуспешных агентных поисков также растет, однако, значительно медленнее, чем количество успешных поисков. Это объясняется действиями оператора в контуре управления ТМБД. Исключаются источники, не имеющие отношения к тематическому направлению и приводящие к неудачам агентного поиска. Это подавляет эффекты устаревания контента ТМБД. Таким образом, контур динамического управления ТМБД, описанный в архитектурной модели МИАС, должен работать постоянно. В этом случае, показатели агентного поиска будут иметь высокие значения на всех этапах жизненного цикла ТМБД.

В **третьем разделе** решаются задачи формирования тематических тезаурусов, предназначенных для фильтрации и рубрикации поступающих в систему коллекций агентных сообщений. Тезаурусы представляют собой специальные тематические словари, которые позволяют выявлять смысл понятий не только с помощью их толкований, но и посредством их соотнесения с другими понятиями и их определениями. Кроме того, они могут включать в себя терминологию на нескольких различных языках.

В диссертации предложено создавать и вести тезаурусы в виде баз данных типа «память переводов». Международным стандартом в этой области является формат Translation Memory Exchange (TMX), который используется в большинстве программных продуктов для автоматизированной обработки текстов на разных языках. С использованием этого формата, по сравнению, например, с таблицами Microsoft Excel, расширяются возможности хранения и группировки терминов и их определений, а также управления ревизиями данных при коллективной работе экспертов.

В соответствии с предложенным форматом, автором создана и зарегистрирована в ФИПС БД «Тезаурус по физике плазмы в международном стандарте TMX 1.4b». Термины тезауруса структурированы в виде набора языковых пар, которые представляют

собой эквивалентные по смыслу сегменты текста на исходном языке и сегменты на целевом языке (языках), рис. 6.

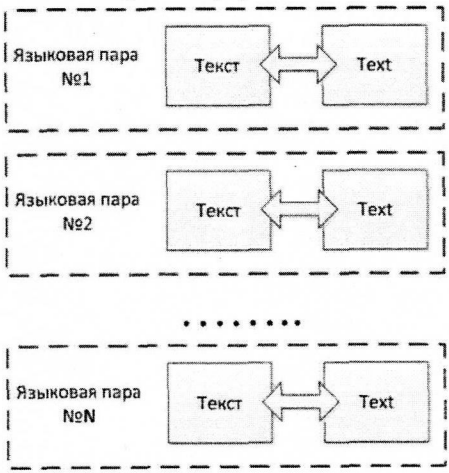

Рис. 6. Схема организации памяти переводов в МИАС

В настоящее время в тезаурусе по физике плазмы представлен 271 термин на русском, английском и китайском языках. Как и ТМБД, он является динамическим объектом и пополняется аналогичным описанному в разделе 2 способом.

Анализ результатов экспериментального сканирования тематической группы источников показал, что получаемый объем информации существенно зависит от языков в используемых тезаурусах. На примере источников из китайского сегмента Интернета показано, что при поиске информации только на английском языке теряется до 90% тематической информации, т.к. публикационная активность значительно выше именно на национальных языках. Отсюда следует вывод, что максимальная полнота агентного поиска может быть достигнута только при использовании многоязычных тезаурусов.

Исследование структуры тематических тезаурусов привело к необходимости введения новой характеристики терминов – «индекс общности». Индексом общности термина тезауруса с номером n из общего числа терминов N предложено называть число D_n^N, которое показывает, сколько раз термин с номером n используется для определения других терминов тезауруса.

На рис. 7 представлено распределение значений индекса общности в тезаурусе по физике плазмы. На диаграмме видно, что количество терминов с высоким индексом общности невелико – 6% (16 терминов).

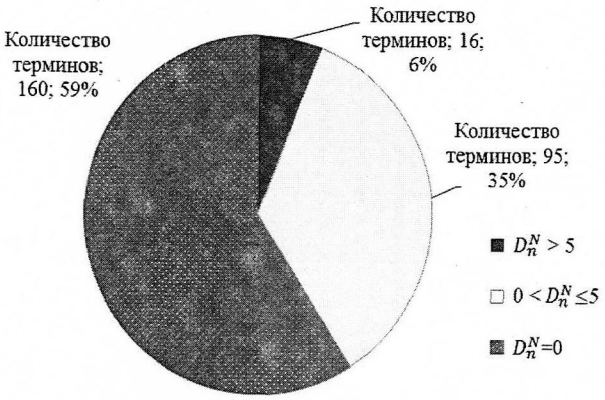

Рис. 7. Диаграмма распределения терминов по индексу общности

Однако именно эти термины с наибольшей частотой встречаются в тематических текстах, следовательно, именно они должны в первую очередь использоваться в качестве ключевых слов при тематической фильтрации агентных коллекций. На рис. 8 представлено экспериментальное подтверждение данного утверждения.

Агентные коллекции документов составлены из трех типов источников по 1000 документов: научные журналы по физике плазмы, новостные ленты специализированных организаций, новостные ленты политематических организаций. Каждая группа столбцов представляет собой результат тематической фильтрации в зависимости от набора ключевых слов с различным индексом общности.

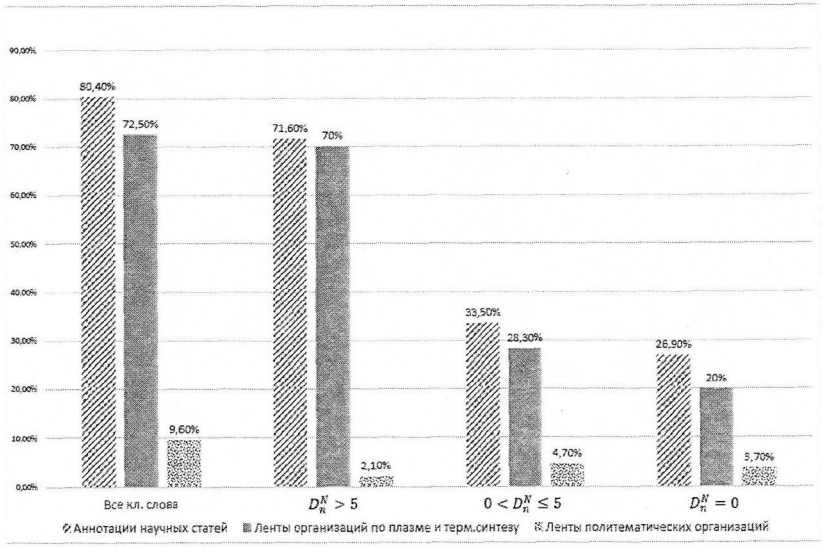

Рис. 8. Степень фильтрации агентных коллекций из разных типов источников в зависимости от используемых ключевых слов

Исходя из этого, термины с высоким индексом общности рекомендуется использовать в первую очередь при подключении к тезаурусу новых языков, т.к. это существенно сокращает трудовые затраты на выполнение перевода.

В данном разделе работы также приведены результаты экспериментов по оценке полноты (R) и точности (P) агентного информационного поиска с использованием тематического тезауруса.

Результаты эксперимента, включающего поиск по специализированным и по политематическим источникам на английском языке, представлены в табл. 1. Для фильтрации агентных коллекций использовались все термины тезауруса по физике плазмы. Высокие значения показателей для всех типов агентных коллекций объясняются тематической ориентацией и регулярной актуализацией ТМБД и тезауруса.

Таблица 1. Характеристики полноты и точности поиска для английского языка

Параметр	Спец. источники	Политем. источники	Все
R – полнота	0,88	0,90	0,89
P – точность	0,98	0,69	0,92

Показатели полноты и точности поиска для русского, немецкого, французского, итальянского, испанского и китайского языков представлены в табл. 2. В ходе данного эксперимента использовались только термины с высоким индексом общности.

Таблица 2. Экспериментальные значения полноты и точности поиска на нескольких языках

Язык выборки	Полнота (*R*)	Точность (*P*)
Русский	0,89	0,89
Немецкий	0,94	0,94
Французский	0,97	0,91
Итальянский	0,93	0,93
Испанский	0,85	0,92
Китайский	0,91	0,88

Таким образом, подтвердилась гипотеза о достаточности и эффективности использования терминов с высоким индексом общности для тематического поиска и фильтрации агентных коллекций на различных языках.

Четвертый раздел посвящен информационно-аналитическим продуктам и технологическим услугам МИАС, которые включают в себя:

– регулярный выпуск дайджестов (тематических новостных подборок),

– формирование фактографических досье на объекты интереса пользователей,

– формирование и ведение семантических сетей.

Базы данных и прикладные функциональные системы в составе ядра МИАС, обеспечивающие выполнение вышеописанных функций, представлены на рис. 9.

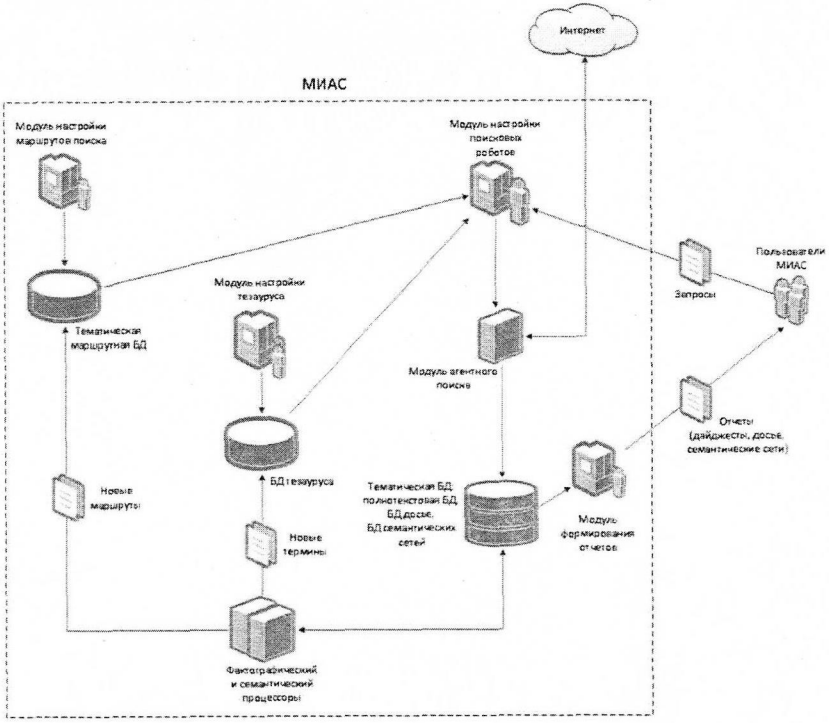

Рис. 9. Функциональная схема ядра МИАС

Здесь описаны типовые информационно-аналитические продукты, генерируемые МИАС, приведены содержательные примеры для области «Физика плазмы» и показана их практическая значимость для решения различных информационно-аналитических задач.

Дайджесты, как тематические подборки публикаций за некоторой период времени, решают задачу регулярного новостного информирования пользователей-специалистов. Высокий уровень автоматизации при агентной обработке информации из большого числа источников позволяет решать проблему «больших данных», а использование пользовательских словарей ключевых слов и

нескольких уровней фильтрации – проблему качества обслуживания на различных уровнях (организация – подразделение – специалист).

Частота формирования дайджестов определяется потребностями конкретного пользователя, а также плотностью входного потока тематической информации. Опытная эксплуатация системы показывает, что для регулярного информационного обеспечения кафедры или научного подразделения университета достаточен ежемесячный выпуск тематического дайджеста. При этом у пользователя остается возможность самостоятельно в интерактивном режиме просмотреть все агентные коллекции, полученные за месяц и ранее.

В качестве примера рассмотрен дайджест по физике плазмы, сформированный за три месяца только по китайским источникам. Автоматический подсчет частоты упоминания организаций в текстах сообщений позволяет оценивать их активность на заданных направлениях исследований. Аналогичным образом можно прослеживать активность отдельных персон, коллективов или целых стран. Располагая структурированной новостной информацией, можно решать и другие аналитические задачи, поставленные пользователем.

Для информирования пользователей о состояниях объектов профессионального интереса разработан формат «досье». Такими объектами могут быть объявлены любые сущности, например, организации (вуз, НИИ, промышленная компания, исследовательская лаборатория), некоторое событие или процесс (проведение долгосрочного эксперимента, строительство крупного промышленного объекта) и т.д. В досье содержится краткая фактографическая информация, выделенная оператором-аналитиком из текстов агентных сообщений. Формирование досье является полезным в процессе ведения тематически устойчивых баз данных. В качестве примеров приведены фрагменты различных форматов досье на объекты: Кафедра физики плазмы НИЯУ МИФИ, Исследовательский центр Кадараш, Международный проект ИТЭР.

Рассмотрены примеры визуализации информации в виде семантических сетей – графов, представляющих объекты и отношения между ними. Преимуществами такого способа представления данных является краткость, наглядность, возможность поиска неочевидных взаимосвязей. В диссертации представлены фрагменты следующих семантических сетей, выполненных в МИАС: визуализация отношений

между объектами ТМБД «Мировые научно-исследовательские и технологические организации по физике плазмы», визуализация фактографических данных из новостной статьи, сеть «Кооперация в проекте ИТЭР».

В **заключении** сформулированы основные результаты диссертации.

В **приложениях** приведены следующие документы:

– копии свидетельств о регистрации управляющих баз данных в Федеральной службе по интеллектуальной собственности (Роспатент) и Бюро регистрации авторских прав при Библиотеке Конгресса США;

– копии актов об использовании результатов диссертации в учебной и научной деятельности кафедр «Физика плазмы» и «Анализ конкурентных систем» НИЯУ МИФИ, а также в производственной деятельности компании «Аналитические бизнес решения»;

– фрагмент тематического дайджеста по физике плазмы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ РАБОТЫ

В диссертационной работе решена актуальная научно-техническая задача исследования и разработки методов и средств автоматизированного управления регулярным агентным поиском и обработкой тематической научной и технологической информации в Интернете.

На примере тематического направления «Физика плазмы» разработаны и экспериментально исследованы человеко-машинные процедуры создания и поддержания в актуальном состоянии баз данных для управления многоязычным агентным поиском тематической информации и автоматизированного выпуска прикладных информационных продуктов для пользователей. В результате специалистам по физике плазмы стала доступна тематическая информация из 575 мировых источников. При этом достигнуты высокие показатели полноты и точности агентного поиска: по результатам экспериментов $R = 0{,}90$; $P = 0{,}92$.

Основные научно-технические результаты, полученные в диссертационной работе, заключаются в следующем:

1. Разработана концептуальная модель мультиагентной информационно-аналитической системы, отличающаяся тем, что в ней,

наряду с режимом интерактивного поиска информации по запросу пользователя, реализованы:

– автоматизированный режим агентного поиска тематической информации по заданному сценарию,

– фильтрация агентных коллекций и формирование тематических баз данных,

– режим регулярного обеспечения пользователей новостной агентной информацией в форме дайджестов, объектных досье и семантических сетей.

2. На примере тематического направления «Физика плазмы» автором созданы и зарегистрированы базы данных для управления агентными технологиями поиска и обработки информации:

– Тематическая маршрутная база данных «Мировые научно-исследовательские и технологические организации по физике плазмы», свидетельства о государственной регистрации №2014620346 от 26.02.2014 (РФ) и №TXu001904126 от 13.02.2014 (США)

– «Тезаурус по физике плазмы в международном стандарте TMX 1.4b», свидетельство о государственной регистрации №2015620043 от 12.01.2015 (РФ).

3. Разработаны и исследованы человеко-машинные процедуры динамической актуализации тематических маршрутных баз для тематического агентного поиска. Показана достаточность использования трех типов шаблонов для создания поисковых предписаний агентам.

4. Поставлена и решена задача автоматизированного построения и периодической актуализации многоязычных тематических тезаурусов для алфавитных и иероглифических языков. Предложенное решение соответствует международному стандарту TMX 1.4b. На примере китайского сегмента Интернета проведены эксперименты по агентному поиску с использованием трехъязычного (англо-русско-китайского) тезауруса. Получена оценка возможных потерь тематической информации при использовании в тезаурусе только английского и русского языков (до 90%).

5. Предложена специальная характеристика терминов в тематических тезаурусах – «индекс общности», позволяющая выделять наиболее «ценные» термины для использования их в качестве ключевых слов в задачах тематической фильтрации агентных

коллекций. Показана полезность этой характеристики при лингвистическом масштабировании агентной системы.

6. Результаты диссертационной работы явились научной основой для создания ядра «Мультиагентной информационно-аналитической системы (МИАС) по естественнонаучным и технологическим направлениям», разработанной в НИЯУ МИФИ по Федеральной целевой программе «Научные и научно-педагогические кадры инновационной России» (2009-2013 гг.). Автор был ответственным исполнителем по данной НИР.

7. Результаты диссертационной работы непосредственно использованы в учебной и научной деятельности кафедр «Физика плазмы» и «Анализ конкурентных систем» НИЯУ МИФИ, а также в производственной деятельности компании «Аналитические бизнес решения», о чем имеются соответствующие акты.

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации, представленные в базах данных Scopus и Web of Science

1. Artamonov A., Leonov D., Nikolaev V., Onykiy B., Pronicheva L., Sokolina K. Ushmarov I. Visualization of semantic relations in multi-agent systems // Scientific Visualization. – 2014. – 6 (3). – pp. 68-76 (Scopus).

Публикации в журналах из перечня ВАК Российской Федерации

2. Оныкий Б.Н., Соколина К.А. Концептуальные вопросы проектирования мультиагентных информационно-аналитических систем для поиска и обработки научно-технической информации // Системы высокой доступности. – 2017. – Т. 13. № 1. – С. 40-51.

3. Артамонов А.А., Галин И.Ю., Леонов Д.В., Михина Е.К., Оныкий Б.Н., Соколина К.А. Поисковые агентные технологии с многоязычным тезаурусом // Вестник Национального исследовательского ядерного университета «МИФИ». – 2015. – Т.4, №4. – С. 369-376.

4. Артамонов А.А., Галин И.Ю., Ионкина К.В., Курнаев В.А., Соколина К.А., Черкасский А.И. Тематические тезаурусы в агентных технологиях поиска научно-технической информации в интернете (на

примере тезауруса по теме «Физика плазмы») // Математическое моделирование. – 2015. – Т.27, №7. – С. 4-9.

5. Николаев В.С., Оныкий Б.Н., Соколина К.А., Ушмаров И.А. Агентное сканирование мировых интернет-ресурсов по естественнонаучным и технологическим направлениям // Системы высокой доступности. – 2014. – №2. – С. 50-53.

6. Будзко В.И., Леонов Д.В., Николаев В.С., Оныкий Б.Н., Соколина К.А. Развитие информационно-аналитической поддержки научно-технической деятельности в Национальном исследовательском ядерном университете «МИФИ» // Системы Высокой доступности. – 2011. – Т. 7, № 4. – С. 4-17.

Тезисы научных докладов

7. Артамонов А.А., Леонов Д.В., Соколина К.А., Черкасский А.И. Формирование тематических кластеров для агентного поиска научно-технической информации в интернет (на примере тематического направления «Физика плазмы») // Современные проблемы прикладной математики и информатики (MPAMCS'2014): Тезисы докладов международной конференции (Дубна, 25-29 августа 2014 г.) – 2014. – С. 41.

8. Артамонов А.А., Галин И.Ю., Николаев В.С., Соколина К.А., Черкасский А.И. Тематические тезаурусы в агентных технологиях поиска научно технической информации в Интернет (на примере тезауруса по «Физике плазмы») // Современные проблемы прикладной математики и информатики (MPAMCS'2014): Тезисы докладов международной конференции (Дубна, 25-29 августа 2014 г.) – 2014. – С. 160.

9. Михина Е.К., Соколина К.А. Тематический тезаурус по физике плазмы для управления поиском научно-технической информации на китайском языке // Научная сессия НИЯУ МИФИ – 2015. – 2015. – Т.3. – С. 201.

10. Баламутенко А.Б., Николаев В.С., Соколина К.А, Суслина И.В. Регистрация авторских прав на базы данных в Российской Федерации и Соединенных Штатах Америки // Научная сессия НИЯУ МИФИ – 2014. – 2014. – Т.3. – С. 175.

11. Соколина К.А. Развитие информационно-аналитической поддержки научно-технической деятельности в НИЯУ МИФИ // Научная сессия НИЯУ МИФИ – 2013. – 2013. – Т.3. – С. 50.

12. Ананьева А.Г., Артамонов А.А., Соколина К.А., Черкасский А.И. Экспериментальные исследования эффективности тематического агентного поиска // Современные системы искусственного интеллекта и их приложения в науке. Всероссийская научная Интернет-конференция с международным участием: материалы конф. (Казань, 25 июня 2013 г.) / Сервис виртуальных конференций Pax Grid; сост. Синяев Д. Н. – Казань: ИП Синяев Д. Н. – 2013. – С. 8-10.

13. Соколина К.А., Курнаев В.А., Артамонов А.А., Черкасский А.И. Интеллектуальная агентная система по физике плазмы // Современные системы искусственного интеллекта и их приложения в науке. Всероссийская научная Интернет-конференция с международным участием: материалы конф. (Казань, 25 июня 2013 г.) / Сервис виртуальных конференций Pax Grid; сост. Синяев Д. Н. – Казань: ИП Синяев Д. Н. – 2013. – С. 68-72.

14. Соколина К.А., Оныкий Б.Н., Ионкина К.В., Торопкина А.М. Применение мультиагентных информационно-аналитических систем в учебно-научной деятельности // III Всероссийская научно-практическая конференция «Информационные технологии в образовании XXI века». Сборник научных трудов. – Москва: НИЯУ МИФИ. – 2013. – С. 132-135.

15. Соколина К.А., Леонов Д.В., Николаев В.С., Оныкий Б. Н.. Мультиагентные информационно-аналитические системы (МИАС) как инструмент решения современных информационно-аналитических задач // Научная сессия НИЯУ МИФИ – 2012. – 2012. – Т.3. – С. 48.

Свидетельства о государственной регистрации баз данных

16. Артамонов А.А., Соколина К.А., Баламутенко А.В., Николаев В.С., Леонов Д.В., Суслина И.В., Ананьева А.Г., Проничева Л.В., Ушмаров И.А., Черкасский А.И. Мировые научно-исследовательские и технологические организации по физике плазмы. Свидетельство о регистрации базы данных №2014620346, Февраль 26, 2014.

17. Артамонов А.А., Курнаев В.А., Оныкий Б.Н., Галин И.Ю., Соколина К.А., Курнаев А.А., Николаев В.С., Баламутенко А.В., Леонов Д.В., Проничева Л.В., Фомина Ю.Е. Тезаурус по физике плазмы в международном стандарте TMX 1.4b Specification. Свидетельство о регистрации базы данных №2015620043, Январь 12, 2015.

18. Artamonov A., Balamutenko A., Nikolayev V., Sokolina K., Leonov D., Suslina I., Ananieva A., Pronicheva L., Ushmarov I., Cherkasskij A. World Plasma Physics Research Organisations Database. TXu 1-904-126, February 13, 2014.

Подписано в печать 27.02.2018г.

Усл.п.л. – 1.5
Заказ № 41630
Тираж: 100 экз.

Копицентр «ЧЕРТЕЖ.ру»
ИНН 7701723201
107023, Москва, ул.Б.Семеновская 11, стр.12
(495) 542-7389
www.chertez.ru