Использование семиотического подхода для описания интеллектуальных информационных систем

В.Б.Барахнин, А.М.Федотов

Аннотация

В статье на базе семиотического подхода уточняется смысл, вкладываемый в термины «информация», «знание», «тезаурус», «онтология», применительно к той области информатики, которая изучает процессы взаимных преобразований данных, информации и знаний. Это позволяет дать краткое описание функционирования интеллектуальных информационных систем в терминах теоретической информатики.

Ключевые слова: информационный поиск, аналогия, сходство, кластеризация.

В настоящее время одним из основных инструментов информационного обеспечения научной деятельности являются *интеллектуальные информационные системы* (ИнтС), функционирующие по схеме [1]:

ИнтС = РИС + ИПС + ИнИн, (1)

где РИС – рассуждающая информационная система (формализующая правила логического вывода), ИПС традиционная информационно-поисковая система, а ИнИн – интеллектуальный интерфейс (диалог, графика и т.д.). Более развитые ИнтС должны обладать и механизмом пополнения базы данных, функционируя по схеме

ИнтС = РИС + ИПС + ИнИн + АП, (2)

где АП – автоматическое извлечение данных из текстов и соответствующее пополнение базы данных посредством этих фактов.

В процессе функционирования интеллектуальных информационных систем компьютер в диалоговом режиме усиливает комбинаторное мышление и логические возможности человека, благодаря чему интеллектуальная система обладает по сравнению с обычной информационно-поисковой системой новыми возможностями, позволяя удовлетворить квалифицированного пользователя в соответствии со схемой «документ – факт – рассуждение» [1, с. 343].

Однако при переходе к более общей постановке вопроса, когда интеллектуальная информационная система рассматривается как некое технологическое устройство компьютерной переработки информации, неизбежно возникнут определенные сложности терминологического характера. Действительно, под технологией в соответствии с [7] следует понимать определенную последовательность методов обработки, изготовления, изменения состояний и свойств сырья или материалов в процессе производства продукции. Иными словами, любая технология по своей сути – инструмент, применяемый для превращения потребляемых факторов в продукцию, или, вообще говоря, для достижения планируемых результатов (см. [5]). И, наконец, совсем кратко: «технология – способ преобразования данного в необходимое» [17].

Что же выступает исходным материалом для технологии переработки информации? Ответ, на первый взгляд, очевиден: сама информация. Однако и на вопрос о конечном продукте напрашивается тот же ответ! Разумеется, человек, владеющий теоретическими основами информатики, даст ответ, что исходным материалом служат данные, а конечным продуктом – знания (или, по крайней мере, информация). Однако при этом следует учитывать, что существует множество подходов к понятию «информация» с философских, социологических, биологических, физико-математических или кибернетических позиций ([1, с. 393]), включая так называемую «техническую» теорию информации, которая является, по сути, теорией передачи и хранения данных. Поэтому можно обнаружить десятки порой противоречащих друг другу определений того, что является информацией или знанием. Даже специалисты по информатике, работающие в разных ее областях, например документальной информации и экспертных систем, вкладывают в термин «знания» несколько разный смысл (ср., в частности, [1] и [2]). Заметим, что в трактовке термина «данные» (понимаемые как факты и идеи, представленные в формализованном виде [16]) столь значительных расхождений обычно не наблюдается.

Еще одна терминологическая проблема, возникающая при описании интеллектуальных информационных систем, предназначенных для работы с документами той или иной предметной области, связана с наименованием структуры, содержащей базовые понятия этой предметной области и связи между ними. В настоящее время в соответствующей роли выступают как термин «тезаурус», так и термин «онтология», причем нередко они используются как синонимы.

Цель данной работы заключается в том, чтобы уточнить смысл, вкладываемый в термины «информация», «знание», «тезаурус», «онтология», применительно к той области информатики, которая изучает процессы взаимных преобразований данных, информации и знаний, установив при этом (на базе семиотического подхода) основания выбора определений, принятых именно в этой области. Это позволяет дать краткое описание функционирования интеллектуальных информационных систем в терминах теоретической информатики.

1. Данные, информация и знания как разные аспекты одного сообщения

Еще в конце 1960-х годов в информатике был принят подход (см. [10]), восходящий к определению Л. Бриллюэна [23], согласно которому информация «…есть сырой материал и состоит из простого собрания данных, тогда как знание предполагает некоторое размышление и рассуждение, организующее данные путем их сравнения и классификации». Однако уже к середине 1970-х годов созрело понимание (см., например, работу П. Чена [24] и приведенную в ней библиографию), что содержательная ценность данных, не сопровождаемых семантикой в виде модели предметной области, которую эти данные описывают, весьма не высока, т.е. такие данные фактически не являются информацией в традиционно-узком (не шенноновском) значении этого слова.

Решение этой проблемы возникло на пути применения в информатике методов семиотики – общей теории знаковых систем. Об использовании семиотических методов речь идет уже в [10], но лишь применительно к построению формализованных информационных

языков. Однако связь семиотики и информатики гораздо глубже. В классических работах [3, 23] уже присутствовало осознание того обстоятельства, что ценность информации (в отличие от ее количества) зависит от субъекта, ее воспринимающего, но поскольку семиотические аспекты этого вопроса явно ускользали от внимания авторов, он не мог найти решения в указанных работах и оставался на уровне постановки. Так, в работе [3] приводится следующий пример (который цитируется и в [10]): «…Сообщение о том, родила ли жена Джона Смита мальчика или девочку, содержит столько же двоичных единиц информации, сколько и сообщение о том, кого родила ваша жена. Вместе с тем последнее сообщение представляет для вас неизмеримо большую ценность, чем первое». Вполне очевидно, однако, что два указанных сообщения имеют куда большее сходство, чем простое равенство количества битов.

Реальное осознание сложностей в преодолении разрыва между шенноновским понятием информации и концепцией семантической информации как средства социальной коммуникации возникло в середине 1960-х годов: см., например, работы Ю.А. Шрейдера [18-20] и У. Шрамма [27], причем Ю.А. Шрейдер показал, что о количестве семантической информации в данном сообщении есть смысл говорить лишь применительно к конкретному приемнику сообщения.

Попытка «телеологического» (с точки зрения объективных внечеловеческих целей и целесообразности) описания особенностей восприятия сообщения субъектом была предпринята Р. Акоффом и Ф. Эмери [22]. Они предложили классифицировать сообщения по видам изменений в получателе, которые делятся на несколько типов (при этом сообщение может принадлежать сразу к нескольким типам):

1. *информация* (изменения вероятности выбора);
2. *инструкция* (изменения в эффективности выбора);
3. *мотивация* (изменения в удельных ценностях).

Были предложены формулы для количественной оценки каждой из названных величин, однако практическая ценность этих формул оказалась не слишком высокой ввиду чрезвычайной сложности реальной оценки изменения состояния субъекта – носителя интеллекта. Основная заслуга Р. Акоффа и Ф. Эмери состоит в том, что они, по-видимому, одними из первых специалистов в области информатики обратили на многоуровневость восприятия сообщения получателем и сделали попытку описать эти уровни.

Наконец, в начале 1980-х годов немецким исследователем В. Гитом была предложена пятиуровневая модель [25], наиболее полно отражающая различные аспекты термина «информация». Структура модели представлена на рис. 1.

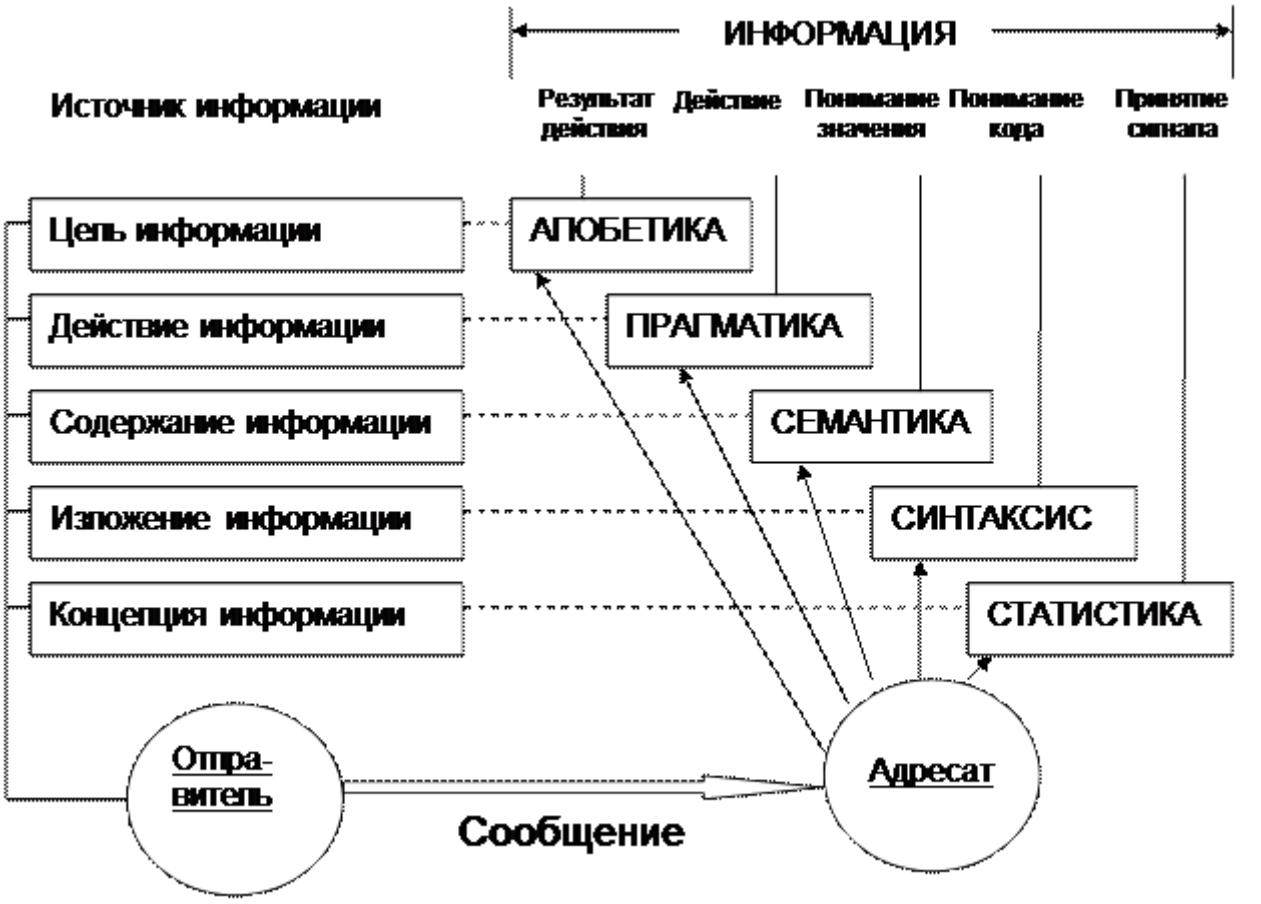

Рис. 1. Пятиуровневая модель информации.

Анализируя эту модель, нетрудно видеть, что ее нижний уровень соответствует шенноновскому значению термина «информация», три последующих – семиотической триаде (синтактика – семантика – прагматика), а верхний уровень носит, скорее, философский характер. При этом наличие в некотором сообщении информации высокого уровня влечет за собой наличие информации всех низших высоких уровней, но, разумеется, не наоборот (еще раз напомним: объем информации зависит, в том числе, от характеристик адресата, причем это касается всех уровней информации).

Следует отметить, что модель В. Гитта в ее полном объеме не получила широкого распространения (во многом потому, что он пытался с ее помощью, делая акцент на пятый уровень, доказать невозможность самопроизвольного возникновения такой сложной информации, как генетический код, что явно противоречит общепринятым в современной науке представлениям).

Идеи, весьма близкие к тем, которые воплощены в модели В. Гитта, однако в несколько менее стройной форме, были высказаны в монографии Ю.А. Шрейдера и А.А. Шарова [21], изданной в 1982 году.

Таким образом, с начала 1980-х годов семиотическая триада заняла прочное место в кибернетике, о чем свидетельствуют соответствующие статьи в «Словаре по кибернетике» [16], хотя в указанное время семиотическая терминология применялась, скорее, при описании языка (понимаемого как частный случай знаковой системы) в целом, нежели при анализе отдельных сообщений. К настоящему моменту описание непосредственно информации с помощью семиотической терминологии получило широкое распространение в отечественной литературе, как научной, так и учебной.

Отметим, что семиотический подход фактически использован при определении базисных понятий и в цитированной выше монографии [1], изданной ВИНИТИ. *Данные* понимаются в ней (в соответствии с традиционным подходом) как факты и идеи, представленные в символьной форме, позволяющей проводить их передачу, обработку и интерпретацию, а *информация* – как смысл, приписываемый данным на основании известных правил представления фактов и идей. Структурированная (связанная причинно-следственными и иными отношениями) информация, образующая систему, составляет *знания*.
Исходя из этого понимания терминов «данные», «информация», «знания», которого мы будем придерживаться в дальнейшем, можно сказать что *данные соответствуют синтаксическому уровню сообщения, информация (в узком смысле!) – семантическому, а знания – прагматическому*.

Среди новейших исследований семиотических оснований информатики можно выделить работы И.М. Зацмана [8, 9]. По мнению автора, сообщение может относиться к одной из трех основных сред, соответствующих семиотической триаде:

1. *цифровая среда* (двоично кодируемые объекты которой соотносятся со знаками, их формами и значениями);
2. *среда социальных коммуникаций* (объекты которой сенсорно воспринимаемы человеком);
3. *ментальная среда знаний человека*.

Названные среды достаточно четко разграничены (т.е. конкретные сообщение или сведения принадлежат только к одной определенной среде). С учетом этого автор формулирует определения базисных терминов информатики, а также типологизирует основные классы элементарных технологий обработки информации, переводящие сообщение (сведения) из одной среды в другую.

Изложенная концепция позволила И.М. Зацману сделать интересные выводы гносеологического и общеметодологического характера, хотя в ее рамках довольно сложно обосновать методы и алгоритмы, предназначенные для «массовой» автоматической переработки информации.

2. О соотношении понятий «тезаурус» и «онтология»

Еще одной важной задачей является установление определенности в понимании и разграничении использования терминов «тезаурус» и «онтология». Изначально в информатике использовался лишь термин «тезаурус», пришедший из лингвистики, где он с начала XIX века обозначал особый тип словаря, в котором понятию ставилось в соответствие слово (в т.ч. *различные* слова), его обозначающие. Со середины 1950-х годов термин «тезаурус» прочно вошел в профессиональную лексику специалистов в области информатики, причем определения тезауруса несколько варьировались в зависимости от класса задач, для решения которых предназначался тезаурус. В частности, применительно к задачам информационного поиска под тезаурусом обычно понимается так называемый *нормативный* тезаурус[10, с. 432] – словарь-справочник, содержащий все лексические единицы информационно-поискового языка – дескрипторы (вместе с ключевыми словами, которые в пределах данной информационно-поисковой системы считаются синонимами этих дескрипторов), причем дескрипторы в словаре должны быть систематизированы по смыслу, а смысловые связи между ними эксплицитно выражены.

Однако в 1990-х годах в информатике, наряду с термином «тезаурус», стал употребляться близкий по смыслу термин «онтология». Первоначально этот термин, заимствованный из философии, появился в среде специалистов по искусственному интеллекту, при этом наиболее широко известно следующее определение Т. Грубера [26]: «онтология – это явная спецификация концептуализации» (т.е. абстрактного представления предметной области). Довольно быстро термин «онтология» перешел и в другие области информатики, включая разработку информационных систем, и стал обозначать [14] «способ, который используется для описания некоторой области знаний…, в частности базовых понятий этой области, их свойств и связей между ними». В настоящее время, как отмечено в [4], под онтологией нередко стали понимать широкий спектр структур, представляющих знания о той или иной предметной области с разной степенью формализации [28]:

1. словарь с определениями;
2. простая таксономия;
3. тезаурус (таксономия с терминами);
4. модель с произвольным набором отношений;
5. таксономия и произвольный набор отношений;
6. полностью аксиоматизированная теория.

В работах многих авторов термин «онтология» начал употребляться вместо термина «тезаурус» (что, в общем, неудивительно, ибо определения онтологии в той или иной степени сходны с определением тезауруса, а первоначальное значение термина «онтология» – «учение о бытии», звучит куда более многообещающе, чем заурядный «тезаурус» – «запас»).

Возникла ситуация, когда разными терминами стали называть один и тот же объект. Попытка разрешения коллизии сделана в работах А.С. Нариньяни [11, 12], причем в основе проделанного анализа лежит семиотическая методология. Подчеркивая, что «еще недавно сегодняшняя онтология именовалась тезаурусом», автор предлагает следующее разграничение этих понятий: «тезаурус скорее более закреплен за лексикой в проекции на семантику, а онтология в ее новом, информационном употреблении – это семантика и прагматика, возможно до известной степени в проекции на язык», причем система сущностей, описываемая онтологией, должна быть связана универсальными зависимостями типа «общее – частное», «часть – целое», «причина – следствие» и т.п. [11].

Развивая и уточняя эту концепцию, А.С. Нариньяни показал в [12], что соотношение между понятиями

Тезаурус – Онтология – Модель предметной области

симметрично известному треугольнику Фреге

Слово – Понятие – Сущность,

то есть онтология рассматривается как общая часть модели предметной области и тезауруса, связывающая знания о мире со знаниями о языке, причем полноценный тезаурус невозможен без онтологии, поскольку она, пусть даже в простейшей форме, является скелетом всякой системы данных и/или знаний.

Заметим, что ранее подобные функции приписывались непосредственно тезаурусу. Так, еще в середине 1970-х годов Н.С. Панова и Ю.А. Шрейдер писали [15], что тезаурус можно интерпретировать как запас семантической информации, содержащейся в документах на данную тему, то есть как описание структуры знаний. Аналогичный подход используется и некоторыми современными исследователями. В частности, в диссертации С.В. Жмайло [6] показано, что информационно-поисковый тезаурус можно с достаточным основанием считать моделью предметной области знаний, или базой знаний, при этом подчеркивается, что в тезаурусе обязательно должны быть явно указаны парадигматические отношения между лексическими единицами: синонимы, квазисинонимы, «род – вид», «часть – целое» и др.

Обобщая сказанное, можно сделать следующий практический вывод: *тезаурус становится онтологией тогда, когда связи между дескрипторами не просто эксплицированы* (как это предусмотрено в классическом определении [10*]), но и классифицированы*. Следует отметить, что многие тезаурусы, например по науковедению и лексикографии, созданные С.Е. Никитиной в [13], или по безопасности инженерных систем, созданный С.В. Жмайло [6], ввиду своей структурной сложности могут быть охарактеризованы как онтологии.

Выводы

С учетом сделанных терминологических уточнений формулы (1) и (2), согласно которым отличительной особенностью интеллектуальных информационных систем является их функционирование в соответствии со схемой «документ – факт – рассуждение», могут быть интерпретированы следующим образом*: интеллектуальные информационные системы позволяют не только извлекать из данных информацию, но и получать новые знания*. Иными словами, функционирование интеллектуальной информационной системы основано на двух противоположных процессах: *при пополнении интеллектуальной системы новыми сведениями происходит преобразование семантической информации в данные, однако непосредственно потребности пользователя удовлетворяет обратный процесс – извлечение из данных нужной пользователю информации и знаний*. При этом в качестве базы знаний выступает *онтология* предметной области, то *есть информационно-поисковый тезаурус*, в котором обязательно должны быть явно указаны парадигматические отношения между лексическими единицами.

Список литературы

[1] *Арский Ю.М., Гиляревский Р.С., Туров И.С., Черный А.И.* Инфосфера: Информационные структуры, системы и процессы в науке и обществе. – М.: ВИНИТИ, 1996.

[2] *Гаврилова Т.А., Хорошевский В.Ф*. Базы знаний интеллектуальных систем. – СПб.: Питер, 2000.

[3] *Голдман С*. Теория информации / Пер.с англ. – М.: Иностранная литература, 1957.

[4] *Добров Б.В., Лукашевич Н.В., Синицын М.Н., Шапкин В.Н.* Разработка лингвистической онтологии по естественным наукам для решения задач информационного поиска // Труды Седьмой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2005). – Ярославль, 2005. – С. 70-79.

[5] *Желены М*. Управление высокими технологиями // В кн.: Информационные технологии в бизнесе. Энциклопедия / Пер.с англ. – СПб.: Питер, 2002. – С. 81-89.

[6] *Жмайло С.В*. Исследование и разработка теории и методики построения тезаурусов для информационного поиска в полнотекстовых базах данных. Автореф. … кандидата техн. наук: 05.13.17. – Москва, 2005.

[7] *Жукова Е.А., Мелик-Гайказян И.В.* Философские проблемы технологий и феномен Hi-Tech // В кн.: Философия математики и технических наук. – М.: Академический Проект, 2006. – С. 557-586.

[8] *Зацман И.М*. Концептуальный поиск и качество информации. – М.: Наука, 2003.

[9] *Зацман И.М*. Семиотические основания и элементарные технологии информатики // Информационные технологии. – 2005. – № 7. – С. 18-31.

[10] *Михайлов А.И., Черный А.И., Гиляревский Р.С*. Основы информатики. – М.: Наука, 1968.

[11] *Нариньяни А.С*. Кентавр по имени ТЕОН: Тезаурус + Онтология // Труды международного семинара Диалог'2001 по компьютерной лингвистике и ее приложениям. – Аксаково, 2001. – Т. I. – С. 184-188.

[12*] Нариньяни А.С*. ТЕОН-2: от Тезауруса к Онтологии и обратно // Труды международного семинара Диалог'2002 по компьютерной лингвистике и ее приложениям. – Протвино, 2002. – Т. I. – С. 307-313.

[13] *Никитина С.Е.* Семантический анализ языка науки. – М.: Наука, 1987.

[14] *Овдей О.М., Проскудина Г.Ю.* Обзор инструментов инженерии онтологий // Труды Шестой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2004). – Пущино, 2004. – С. 59-68.

[15] *Панова Н.С., Шрейдер Ю.А*. Принцип двойственности в теории классификации // Научно-техническая информация. Сер. 2. – 1975. – № 10. – С. 3-10.

[16] *Словарь* по кибернетике. 2-е изд., переработанное и дополненное. – Киев: Главная редакция Украинской Советской Энциклопедии им. М.П. Бажана, 1989.

[17] *Технология* // В: Тезаурус по образованию и педагогике. – Институт информатизации образования в составе Московского государственного гуманитарного университета им. М.А. Шолохова. http://www.mgopu.ru/ininfo/r1_thesaurus.htm#technology.

[18] *Шрейдер Ю.А.* О количественных характеристиках семантической информации // Научно-техническая информация. Сер. 2. – 1963. –

№ 10. – С. 35-39.

[19] *Шрейдер Ю.А.* О семантических аспектах теории информации // В сб.: Информация и кибернетика. – М.: Советское радио, 1967. – С. 15-47.

[20] *Шрейдер Ю.А.* Об одной модели семантической информации // В сб.: Проблемы кибернетики. – Вып. 13. – М.: Наука, 1965. – С. 233-240.

[21] *Шрейдер Ю.А., Шаров А.А.* Системы и модели. – М.: Радио и связь, 1982.

[22] *Ackoff R., Emery F.* On Purposeful Systems. – Ch.-N.Y.: Aldine-Atherton, 1972. /Рус. пер. *Акофф Р., Эмери Ф.* О целеустремленных системах – М.: Советское радио, 1974.

[23] *BrillouinL.* Science and information theory. – N.Y.: Academic Press, 1956. / Рус.пер. *Бриллюэн Л.* Наука и теория информации. – М.: Физматгиз, 1960.

[24] *Chen P.P.* The entity-relational model. Toward a unified view of data // ACM TODS. – 1976. – № 1. – P. 9-36. / Рус. пер. *Чен П. П.-Ш.* Модель «сущность-связь» – шаг к единому представлению данных // СУБД. – 1995. – № 3. – С. 137-158.

[25] *Gitt W.* Ordnung und Information in Technik und Natur // In: Gitt W. (Hrsg.): Am Anfang war die Information. Graefeling: Resch KG, 1982. – S. 171-211.

[26] *Gruber T.* A translation Approach to Portable Ontology Specifications // Knowledge Acquisition Journal. – 1993. – V. 5. – № 2. – P. 199-220.

[27] *Schramm W.* Information Theory and Mass Communication // In: Communication and Culture. N.Y.: Holt, Rinehart & Winston, 1966. – P. 521-534.

[28] *Welty C., McGuinness D., Uschold M., Gruninger M., Lehmann F.* Ontologies: Expert Systems all over again // AAAI-1999 Invited Panel Presentation. – 1999.

Работа выполнена при частичной финансовой поддержке РФФИ (грант № 08-07-00229, 09-07-00277, 10-07-00302), Президентской программы «Ведущие научные школы РФ» (грант № НШ-6068.2010.9), ФЦП «Научные и научно-педагогические кадры инновационной России» на 2009-2013 гг. (госконтракт ГК № П484 от 04.08.2009 г.) и интеграционных проектов СО РАН.

Об авторах

В.Б.Барахнин - Институт вычислительных технологий СО РАН

А.М.Федотов - Новосибирский государственный университет