# ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

**4**(188)
**2012**

**ТЕОРЕТИЧЕСКИЙ И ПРИКЛАДНОЙ НАУЧНО-ТЕХНИЧЕСКИЙ ЖУРНАЛ**

**Издается с ноября 1995 г.**

**УЧРЕДИТЕЛЬ**
**Издательство "Новые технологии"**


## СОДЕРЖАНИЕ




Информация о журнале доступна по сети Internet по адресу **http://novtex.ru/IT.**
Журнал включен в систему Российского индекса научного цитирования.
Журнал входит в Перечень научных журналов, в которых по рекомендации ВАК РФ должны быть опубликованы научные результаты диссертаций на соискание ученой степени доктора и кандидата наук.

УДК 004.912

**М. В. Бочков**, д-р техн. наук, проф.,
НОУ ДПО "Центр предпринимательских рисков", г. Санкт-Петербург,
**А. Ю. Бородащенко**,
канд. техн. наук, преподаватель,
Академия ФСО России, г. Орел,
e-mail: bay55@mail.ru

# Перспективы развития методов семантической фильтрации текстовых документов

*Рассмотрены существующие и перспективные подходы к поиску близких по содержанию документов по отношению к текстам, наиболее релевантным пользовательским потребностям.*

***Ключевые слова:*** *текст, обработка текста, семантическое сходство, семантическое расстояние, мера близости, алгоритм фильтрации текстов*

## Введение

Современное развитие информационных технологий и массовое использование глобальных информационных ресурсов приводит к ежегодному лавинообразному росту объемов хранимой и передаваемой информации. При этом разнородность информации, ее объемы и уровень достоверности существенно усложняют своевременное и актуальное представление о ситуации в конкретной предметной области.

Быстро развивающееся высокотехнологичное производство требует постоянного мониторинга ситуации, поиска и отслеживания новейших разработок, статистической и аналитической обработки публикаций и научных фондов.

Анализ указанной информации предполагает поиск источников данных, наиболее полно и объективно отражающих реальные рыночные процессы. Основными видами такой информации являются статистические, коммерческие, биржевые, финансовые, профессиональные и научно-технические данные. Для перечисленных категорий текстовая информация является преобладающим видом, требующим применения соответствующих технологий обработки.

Значительные объемы информационных потоков делают невозможным непосредственное ознакомление человека с каждым текстом и тем более глубокое осмысление его содержания. Отбор релевантной информации сопряжен со значительными затратами временных и трудовых ресурсов. Эти обстоятельства затрудняют принятие обоснованных и своевременных решений, в основу которых должно быть положено изучение всего массива информации, отражающей ситуацию в аспекте поставленной руководителем задачи. В связи с этим разработка и внедрение на предприятии информационно-аналитических систем и технологий, ориентированных на автоматизированную обработку текстовой информации на основе методов интеллектуального анализа данных (ИАД), являются актуальной задачей.

## Обзор технологий обнаружения знаний в текстах (Text Mining)

Для решения указанных выше информационно-аналитических задач в настоящее время активно используются такие технологии и системы, как "компьютерная конкурентная разведка" [1], идея которой заключается в автоматизации и ускорении процессов извлечения необходимой для конкурентной борьбы информации из открытых источников и ее аналитической обработки, широко применяются новые направления науки и технологий, получившие названия: "управление знаниями" (*knowledge management*), "обнаружение знаний в базах данных" (*knowledge discovery in databases* или *Data*, *Text* и *Web Mining*) [2].

Информационно-поисковые системы (ИПС) и информационно-аналитические системы (ИАС) позволяют собирать и накапливать всю доступную информацию как из внутренних, так и из внешних источников. Технологии Data Mining, Text Mining и Web Mining выявляют неочевидные закономерности в данных или текстах — так называемые скрытые знания. Эти технологии также помогают обнаружить в "информационном сырье" ранее неизвестные данные, полезные знания. Системы этого класса способны осуществлять анализ больших массивов документов и формировать предметные указатели понятий и тем, освещенных в этих документах.

Структура и характеристика методов технологии обнаружения знаний в текстах представлена на рис. 1.

В соответствии с уже сложившейся на протяжении нескольких десятков лет методологией к основным элементам Text Mining относятся аннотирование (*summarization*), выделение объектов, понятий (*feature extraction*), кластеризация (*clustering*), классификация (*classification*), ответ на запросы (*question answering*), тематическое индексирование (*thematic indexing*) и поиск по ключевым словам (*keyword*

*searching*). Также в некоторых случаях набор дополняют средства поддержки и создания таксономии (*oftaxonomies*) и тезаурусов (*thesauri*) [3].

Для решения задач класса *Text Mining* в настоящее время в основном применяют лингвистические и математические методы.

В настоящее время известно, как минимум, три десятка программных продуктов, специализирующихся в этой области [4—6]. Как правило, это масштабируемые системы, в которых реализованы различные математические и лингвистические алгоритмы анализа текстовой информации. Они имеют развитые графические интерфейсы, богатые возможности визуализации и манипулирования с данными, предоставляют доступ к различным источникам данных, функционируют в архитектуре клиент — сервер.

Подробный обзор и сравнительный анализ ряда информационно-поисковых и информационно-аналитических систем обработки текстовой информации представлен в работе [7]. Возможности большинства из приведенных выше систем во многом пересекаются. Поэтому их можно достаточно полно охарактеризовать на примере нескольких общедоступных систем (см. ниже таблицу).

Таким образом, в большинстве ИАС в достаточно полной мере реализованы основные функции технологии Text Mining и практически не реализована функция семантической фильтрации информации. Единственной коммерческой системой, поддерживающей функцию семантической фильтрации, является программный продукт SearchInform (Server и Desktop) компании СофтИнформ. Однако и в этой системе поиск реализован простым способом на основе запросов по ключевым словам без учета связей между ними. Достоинством системы SearchInform является возможность сортировки результатов поиска по степени релевантности документов поисковому запросу.


**Рис. 1. Структура и характеристика методов Text Mining**

### Семантическая фильтрация на основе многоаспектного рассмотрения текста

Анализ формальных методов, применимых для решения задачи поиска близких по содержанию документов, представленных на рис. 1, и возможностей существующих систем (см. таблицу) показал ряд их существенных недостатков.

1. Семантическая фильтрация не реализована в большинстве систем либо реализована простым способом на основе запросов по ключевым словам, что не удовлетворяет потребностям пользователя.

2. Рассмотрение текста осуществляется в изоляции от его информационного окружения, в качестве которого выступают другие текстовые документы информационного массива, формируемого традиционными средствами поиска.

3. Использование ключевых слов или отдельных словосочетаний для оценки семантической близости текстов вместо рассмотрения отдельного текста как взаимосвязанной последовательности всех его слов, порожденной источником с определенными статистическими свойствами, существенно зависящими от тематической направленности.

Первые два недостатка существующих ИАС обусловливают высокий уровень неопределенности относительно статистических образов анализируемых текстов и приводят к недостаточной чувствительности используемых критериев семантического сходства при сравнении текстов, относящихся к одной тематической рубрике.

В связи с этим принципиально новым подходом в работе является многоаспектное рассмотрение

**Зарубежные и отечественные системы анализа текста**

| Название системы | Основные функции | Используемые методы |
|---|---|---|
| InterMedia Text | Тематическая классификация документов на английском языке, автоматическое выделение главных тем из документов, реферирование | Лингвистические и математические |
| RCO | Статистический (извлекаются основные понятия) и семантический (устанавливаются смысловые связи между понятиями) анализ документов на русском языке. Определение ключевых тем документа, автоматического реферирования, функция нечеткого поиска (расширение запроса при опечатках и ошибках) | Лингвистические и математические |
| Intelligent Miner for Text | Набор независимых утилит, предназначенных для построения приложений управления знаниями:<br>— Language Identification Tool — утилита определения языка.<br>— Categorisation Tool — утилита классификации.<br>— Clusterisation Tool — утилита кластеризации.<br>— Feature Extraction Tool — утилита определения новых ключевых слов.<br>— Annotation Tool — утилита составления аннотаций | Лингвистические и математические |
| Text-Analyst | Построение семантической сети большого текста, подготовка резюме текста, поиск по тексту, автоматическая классификация и кластеризация текстов. | Лингвистические и математические |
| SemioMap | Индексирования массивов неструктурированного текста, извлечение ключевые фраз (понятия); кластеризация понятий; графическое отображение и навигация, визуализация карт связей | Лингвистические и математические |
| Autonomy Knowledge Server | Контекстный анализ и извлечение смысла для решения задач автоматической классификации и организации перекрестных ссылок | Математические |
| Galaktika-ZOOM | Поиск информации в больших информационных массивах; выявление ключевых слов и словосочетаний документа, отражающих его смысл; обнаружение сходства, различия, аномалий изучаемых объектов | Лингвистические |
| Стрингер-LT | Тематическая классификация, реферирование, аннотирование | Лингвистические |
| ИАС "Астарта" | Непрерывный мониторинг; рубрицирование; тематическая фильтрация сообщений; полнотекстовая индексация рубрицированных материалов; составление дайджестов, аннотирование; статистический анализ по времени и тематике | Лингвистические и математические |
| XFiles | Выявление фактов из документов; автоматическое выявление прямых и косвенных связей объекта; построение карты связей объектов для различных типов связей, визуализация и фильтрация связей; поиск оптимальных связей между заданными объектами; построение многомерных частотных распределений фактов | Лингвистические и математические |
| ИАС "Семантический архив" | Инструмент для создания интегрированного хранилища информации с возможностью создания досье. Автоматическое выделение понятий, поиск дублей, мониторинг объектов, классификация, развитый поиск и визуализация | Лингвистические и математические |
| ИАС "СОНЕТ" | Автоматизация обработки текстовой информации: классификация, составление тезаурусов, контент-анализ информационных потоков | Лингвистические и математические |
| ЛАК "SEMANTIX" | Автоматизированная обработка неструктурированных потоков текста на естественных языках с извлечением интересующих объектов и анализом взаимосвязей этих объектов | Лингвистические и математические |
| Аналитический курьер | Многоязычный поиск с расширенными возможностями; аннотирование; рубрицирование; определение тональности документов; кластерный анализ публикаций; выявление ключевых тем, построение семантической сети; построение дайджеста (обзора); частотный анализ рубрик и публикаций | Лингвистические и математические |
| Search-Inform | Поиск документов, похожих по содержанию, в текстовых файлах практически любых форматов, базах данных и информационных системах | Лингвистические и математические |


**Рис. 2. Семантическая фильтрация на основе многоаспектного рассмотрения текстов**

текста, представленное на рис. 2 и 3, с использованием следующих показателей:

- $F_1$, рассчитываемого путем выделения из текста множества ключевых слов (тем), а также связей между ними [8];
- $F_2$, рассчитываемого на основе учета вероятностных связей между парами слов внутри текста [9];
- $F_3$, рассчитываемого на основе анализа гипертекстовых связей между различными текстами информационного массива [10].

Здесь под *семантической фильтрацией* понимается процесс отбора из массива текстовых публикаций таких документов, содержание которых *подобно* относительно выбранного критерия некоторому эталону текста. В свою очередь, семантическая фильтрация представляет один из видов семантической обработки текстовой информации, наряду с такими процедурами, как аннотирование, классификация, кластеризация, ответы на запросы и т. д.

Под *структурным подобием* понимается степень совпадения ключевых слов и словосочетаний документов и связей между ними (рис. 4) (структура текста — его внутреннее устройство, т. е. совокупность ключевых слов и связей между ними, отражающих содержание документа).

При *контекстном подобии* рассматриваются связи между цепочками слов внутри документов (рис. 5). При этом под контекстом понимается относительно законченный отрывок текста, общий смысл которого позволяет уточнить значение отдельных входящих в его состав слов (В более широком значении контекст — это среда, в которой существует объект.). В лингвистике различают левый и правый контексты. Левый контекст — те высказывания, которые находятся слева от данного слова, правый контекст — то, что находится справа.

*Внеконтекстное подобие* отражает ссылочные связи между текстами внутри массива документов (рис. 6).

Семантическая фильтрация на основе многоаспектного рассмотрения текста реализована в виде отдельных алгоритмов и подробно описана в работах [8—11].

## Комплексный алгоритм семантической фильтрации текстовой информации

В целях эффективного использования разработанных алгоритмов семантической фильтрации текстовой информации предлагается комплексный алгоритм, представленный на рис. 7. Алгоритм включает в себя следующие этапы.

В блоке 1 осуществляется загрузка текстов для анализа. В блоке 2 определяются следующие свойства загруженных текстов:

1. $R_1 = \dfrac{r_1^{\text{общ}}}{r_1}$ — доля общих ключевых слов ко всем ключевым словам массива текстов.


**Рис. 3. Аспекты семантической фильтрации текстовой информации**


**Рис. 4. Пример множества ключевых слов и словосочетаний документа**

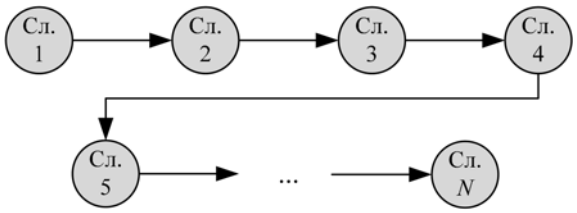



**Рис. 5. Пример связей между цепочками слов внутри текста**

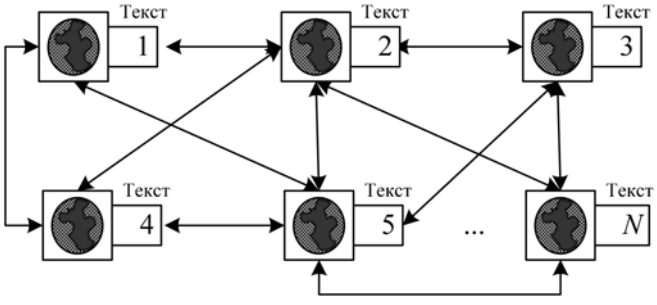



**Рис. 6. Пример ссылочных связей между текстами**

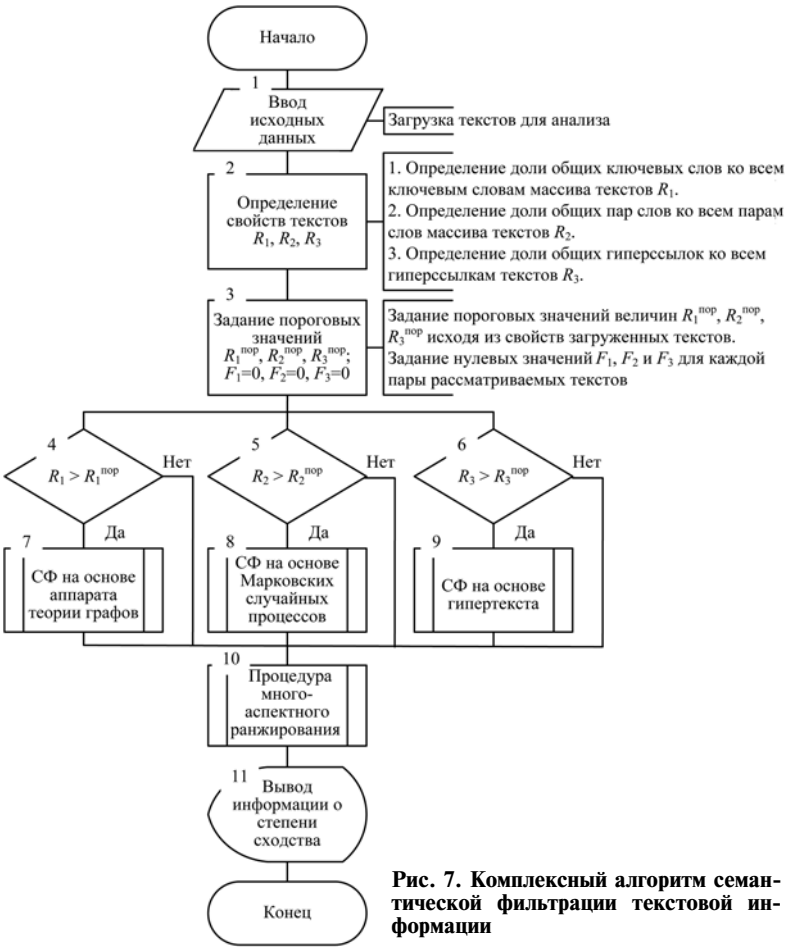



**Рис. 7. Комплексный алгоритм семантической фильтрации текстовой информации**

2. $R_2 = \dfrac{r_2^{общ}}{r_2}$ — доля общих пар слов ко всем парам слов массива текстов.

3. $R_3 = \dfrac{r_3^{общ}}{r_3}$ — доля общих гиперссылок ко всем гиперссылкам текстов.

Указанные свойства характеризуют качество массивов эталонных и произвольных текстов. Формирование эталонных текстов желательно осуществлять таким образом, чтобы значения показателей были максимальны. При осуществлении семантической фильтрации это обеспечит выбор из массива наиболее релевантных и пертинентных документов.

Задание пороговых значений $R_1^{пор}$, $R_2^{пор}$, $R_3^{пор}$ осуществляется в блоке 3. Рекомендуемые пороговые значения: $R_1^{пор} = 0{,}3$; $R_2^{пор} = 0{,}3$; $R_3^{пор} = 0{,}3$.

Кроме того, устанавливаются нулевые значения множества оценок показателей $F_1$, $F_2$ и $F_3$ для каждой пары рассматриваемых текстов ($Y_j$, $X_i$).

В блоках 4, 5 и 6 проверяется, больше ли значения показателей свойств текстов пороговых величин. В случае, если соответствующий показатель больше порогового значения, выполняется соответствующий вид семантической фильтрации (блок 7, 8 или 9). В противном случае выполнять семантическую фильтрацию нецелесообразно, так как рассматриваемые пары текстов будут существенно отличаться друг от друга и при выполнении семантической фильтрации не будет найдено сходных документов.

В блоке 10 выполняется процедура многоаспектного ранжирования загруженных текстов. В качестве такой процедуры в работе использован кластерный анализ *k*-средних, позволяющий осуществлять разбиение исходного массива текстов на необходимое пользователю число классов. Для проведения многоаспектного ранжирования можно использовать и другие методы, например метод опорных векторов, предназначенный для решения задач бинарной классификации (число классов — 2); байесовская классификация и т. д.

В блоке 11 осуществляется визуализация полученных результатов.

## Заключение

На основании анализа современных методов обработки текстов в статье показано, что функция отбора информации, соответствующей потребностям пользователя, на базе использования семантических эталонов практически не реализована на практике, что требует значительных затрат времени при осуществлении информационного поиска. В ходе работы решена актуальная научная задача, заключающаяся в разработке комплексного алгоритма семантической фильтрации текстовой информации, обеспечивающего повышение алгоритмической релевантности (точности и полноты) результатов выполнения поисковых запросов в ИАС обработки текстовой информации, а также существенное сокращение времени отбора полезной информации.

### Список литературы


1. **Меркулов Ф. Г.** Энциклопедия деловой разведки и контрразведки. — М.: Русь-Олимп, 2007. 428 с.
2. **Чубукова И. А.** Data Mining: учеб. пособие // Основы информационных технологий. — М.: БИНОМ. Лаборатория знаний. Интернет-университет информационных технологий. 2006. 382 с.
3. **Ланде Д. В.** Добыча знаний. [Электронный ресурс] / Глубинный анализ текстов. Технология эффективного анализа текстовых данных. — М.: Персональный сайт Дмитрия Ландэ, 2010. URL: http://dwl.kiev.ua/art/dz/index.html, свободный.
4. **Data Mining Community's Top Resource** [Электронный ресурс] / Data Mining and Analytics Resources. — Boston: KDnuggets, 2010. URL: http://www.kdnuggets.com, свободный.
5. **Р-техно.** Экономическая разведка. — М.: ООО "Р-Техно", 2010. URL: http://www.r-techno.com, свободный.

6. **CNews**. Издание о высоких технологиях. — М.: Холдинг РБК, 2010. URL: http://www.cnews.ru, свободный.

7. **Беляев К. В., Босов А. В., Краюшкин Д. В.** Обзор и сравнительный анализ информационно-аналитических систем. — М.: ИПИ РАН, 2008. 136 с.

8. **Бочков М. В., Бородащенко А. Ю., Бочков М. В., Салбиев А. Л.** Алгоритм оценки массива текстов на семантическое сходство с эталоном // Информационные технологии. 2008. № 12. С. 8—11.

9. **Бородащенко А. Ю., Яковлев В. А.** Алгоритм фильтрации текстовой информации на основе марковской модели. // Информационные технологии. 2011. № 5. С. 2—5.

10. **Бородащенко А. Ю., Рябцев А. О.** Алгоритм оценки массива гипертекстовых документов на семантическое сходство с эталоном // Информационные технологии. 2011. № 6. С. 7—12.

11. **Программа** семантической фильтрации текстов: Свид. о государственной регистрации программы для ЭВМ № 2009612007 от 20.04.2009 г. / А. Ю. Бородащенко, М. В. Бочков, А. Л. Салбиев. — М.: ФГУ ФИПС, 2009.

УДК 519.876.2, 004.942

**В. В. Борисов**[1], д-р техн. наук, проф.,
**В. В. Сысков**[2], канд. техн. наук, специалист
[1]Филиал ФГБОУ ВПО
"Национальный исследовательский
университет «МЭИ»" в г. Смоленске,
[2]ФГУП "Центральный
научно-исследовательский институт
экономики, информатики и систем управления"
e-mail: smolenskcity@mail.ru

# Мультиагентное моделирование сложных организационно-технических систем в условиях противоборства

*Предложена мультиагентная модель сложной организационно-технической системы, функционирующей в условиях противоборства. Определены типы агентов, моделирующих различные объекты системы и среды, — агенты-приемники информации из среды, агенты управления системой, агенты-исполнители и противодействующие им агенты внешней среды. Предложено формализованное описание агентов управления системой, включающее задание множеств данных, действий и поведений, описание модели управления и итогового поведения агента, а также описание взаимодействия агентов. Для описания функционирования агентов модели использована алгебра поведений. Для решения сложноформализуемых задач применены способы нечеткого вывода и реализована модель нечеткого управления. По результатам экспериментальных исследований выполнена оценка достоверности результатов мультиагентного моделирования сложной организационно-технической системы в условиях противоборства.*

***Ключевые слова:*** *сложная система, мультиагентное моделирование, мультиагентная модель, агент, противоборство, поведение*

## Введение

В настоящее время для исследования сложных организационно-технических систем (СОТС) применяются различные методы моделирования: аналитическое, аналитическое моделирование с элементами дискретно-событийного моделирования, системная динамика, дискретно-событийное моделирование с элементами аналитического и агентного моделирования. Ограничениями этих методов при исследовании сложных систем, функционирующих в условиях противоборства, являются недостаточная гибкость моделирования ввиду сложностей при формализации и расширении факторного пространства модели, при учете неоднородности системы, при определении различных вариантов функционирования элементов системы и среды, при модернизации модели, а также трудности при описании и реализации сложноформализуемых задач в среде моделирования.

Использование теории мультиагентных систем (МАС) и мультиагентного подхода позволяет повысить гибкость моделирования сложных систем в условиях противоборства.

Основы теории МАС предложены в работах Н. П. Амосова, М. М. Бонгарда, В. А. Лефевра и развиты в работах С. Рассела, П. Норвига, К. Хьюита, М. Вулдриджа, В. Б. Тарасова, А. А. Летичевского и других исследователей. В их работах рассматриваются различные направления МАС, к которым относятся распределенный искусственный интеллект (РИИ), децентрализованный искусственный интеллект, искусственная жизнь.

В РИИ рассматриваются МАС, характеризующиеся централизованным управлением и координацией действий между агентами, представляющими различные объекты системы и/или внешней среды, поведение которых определяет функционирование исследуемой системы в целом. Под отдельным агентом при этом понимается объект, который имеет направленные на достижение своей цели автономное поведение и активное влияние на других агентов, может принимать частные и обобщенные решения в соответствии с некоторым набором правил, взаимодействовать с окружающей средой и другими агентами, а также изменяться [1]. Коллективное интеллектуальное поведение при этом основывается на интеллектуальном поведении отдельных агентов и предполагает согласование их целей, координацию действий и разрешение конфликтов между ними [2].

С точки зрения моделирования МАС относятся к самоорганизующимся системам, в которых с помощью агентов выполняется поиск оптимального решения задачи, на которое необходимо потратить наименьшее количество одних ресурсов системы в условиях ограничения других.

При этом мультиагентная модель системы может быть дополнена и модифицирована без существенной ее переработки.

### Мультиагентная модель сложной системы, функционирующей в условиях противоборства

К особенностям СОТС, функционирующей в условиях противоборства, относятся иерархичность, наличие различных целей системы и среды, а также распределенность задач [3, 4].

Мультиагентные модели таких СОТС состоят из следующих основных компонентов: агентов различных типов; действий, описывающих поведение агентов; среды функционирования; отношений между агентами; коммуникативных актов взаимодействия агентов со средой; целей функционирования системы [1, 2, 5].

Исходя из этого мультиагентная модель сложной системы может быть представлена следующим образом:

$$MM_{\text{СОТС}} = (A,\ E,\ R,\ P,\ I,\ Z,\ C),$$

где $A$ — множество агентов системы; $E$ — множество агентов внешней среды; $R$ — множество отношений между агентами модели; $P$ — множество актов взаимодействия между агентами модели; $I$ — множество показателей оценки эффективности системы; $Z$ — множество целей системы; $C$ — множество целей внешней среды.

Особенностью описания мультиагентной модели СОТС в условиях противоборства является представление не относящихся к системе объектов внешней среды в качестве агентов, стремящихся к достижению целей, не совпадающих с целями системы или противоположных им.

Множество $A$ представляет собой совокупность агентов-приемников информации из среды ($A_{\text{П}}$), агентов управления системой ($A_{\text{У}}$) и агентов-исполнителей ($A_{\text{И}}$). Множество $E$ представляет собой совокупность агентов внешней среды ($E_{\text{С}}$).

Между агентами модели возможны три вида отношений, во-первых, иерархии ($R_{hrh}$), определяющих их соподчиненность, во-вторых, кооперации ($R_{cpr}$), характеризующих сотрудничество между агентами одного уровня иерархии, и, в-третьих, противодействия ($R_{res}$), определяющих конфронтацию между агентами $A_{\text{П}}$, $A_{\text{У}}$, $A_{\text{И}}$ рассматриваемой системы и агентами $E_{\text{С}}$ среды.

Взаимодействие между агентами модели задается с помощью актов взаимодействия, реализующих генерацию агентов ($P_{gen}$), управление одних агентов другими ($P_{con}$), передачу информации ($P_{trn}$), а также воздействие ($P_{frc}$) и восприятие ($P_{per}$) между агентами.

В качестве показателей $I$ выступают показатели производительности системы ($I_{pr}$), потерь системы ($I_{lss}$), расхода ресурсов системы ($I_{cons}$).

Множество целей $Z$ системы включает цели максимизации производительности системы при условии сохранения допустимых потерь и расхода ресурсов ($Z_{pr.\,\max}$), минимизации потерь системы при условии сохранения требуемой производительности ($Z_{lss.\,\min}$), минимизации расхода ресурсов системы при условии сохранения требуемой производительности системы ($Z_{cons.\,\min}$).

Множество целей $C$ внешней среды включает цели минимизации производительности системы ($C_{pr.\,\min}$), максимизации потерь системы ($C_{lss.\,\max}$), максимизации расхода ресурсов системы ($C_{cons.\,\max}$), которые противоположны соответствующим целям из множества $Z$.

Каждая цель из множеств $Z$ или $C$ характеризуется достижением показателями оценки эффективности определенного уровня. Например, цель $Z_{iss.\,\min}$ определяется критерием минимизации потерь системы:

$$Cr_{lss.\,\min} = \begin{cases} \dfrac{I_{lss}(У)}{I_{lss}.\text{opt}} \to \min, \text{ если } I_{lss.\,\text{opt}} \neq 0, \\ I_{lss}(У) \to \min, \text{ если } I_{lss.\,\text{opt}} = 0, \end{cases}$$

и соответствующими ему ограничениями требуемой производительности и расхода ресурсов системы:

$$I_{pr}(У) \geqslant I_{pr.\,\text{dem}},\ I_{cons}(У) \leqslant I_{cons.\,\text{acc}},$$

где $I_{pr.\,\text{dem}}$ — требуемая (минимальная) производительность системы; $I_{lss.\,\text{opt}}$ — оптимальное значение потерь системы; $I_{cons.\,\text{acc}}$ — допустимое значение расхода ресурсов системы.

Взаимосвязь различных типов агентов модели описывается с помощью актов взаимодействия $P$, задаваемых на множестве отношений $R$.

Для задания соподчиненности между агентами модели определяются отношения иерархии между: агентами-исполнителями и агентами управления — $r^1_{hrh}: A_{\text{И}} \to A_{\text{У}}$; агентами управления различных уровней — $r^2_{hrh}: A_{\text{У}} \to A_{\text{У}}$; агентами среды — $r^3_{hrh}: E_{\text{С}} \to E_{\text{С}}$.

Для реализации взаимодействия между агентами одного уровня определяются: отношения кооперации между агентами-приемниками и агентами управления — $r^1_{cpr}: A_{\text{П}} \to A_{\text{У}}$; отношения $r^2_{cpr}: E_{\text{С}} \to A_{\text{П}}$, характеризующие восприятие агентами-приемниками информации от агентов среды; отношения $r^3_{cpr}: A_{\text{П}} \to E_{\text{С}}$, $r^4_{cpr}: A_{\text{У}} \to E_{\text{С}}$ и $r^5_{cpr}: A_{\text{И}} \to E_{\text{С}}$, характеризующие восприятие агентами среды информации от агентов СОТС.

Для реализации противоборства между агентами модели определяются отношения, характери-

зующие противодействие: агентов-исполнителей агентам среды — $r^1_{res}: A_{\text{И}} \to E_{\text{С}}$; агентов среды агентам-приемникам, агентам управления и агентам-исполнителям — $r^2_{res}: E_{\text{С}} \to A_{\text{П}}$, $r^3_{res}: E_{\text{С}} \to A_{\text{У}}$ и $r^4_{res}: E_{\text{С}} \to A_{\text{И}}$.

Представленное описание отношений между агентами модели СОТС позволяет предложить следующий способ описания актов взаимодействия между любыми двумя агентами модели:

$$p_{\text{act}}: p\hat{}r_{rel}(Z', Z''),$$

где $p_{\text{act}}$ — какой-либо акт взаимодействия между двумя агентами $A'$ и $A''$, заданный на некотором отношении $r_{rel}$: $A' \to A''$ с помощью операции приписывания "^" отношения к акту взаимодействия. Акт $p_{\text{act}}$ направлен на достижение взаимодействующими агентами целей $Z'$ и $Z''$ системы (среды) соответственно, причем цели $Z'$ и $Z''$ могут совпадать. Акт взаимодействия является действием между агентами, приводящим к изменению состояний агентов.

Для реализации в модели функций управления одних агентов другими определяются акты управления агентами-исполнителями — $p^1_{con}: p\hat{}\bar{r}^1_{hrh}(Z', Z'')$ и агентами управления нижестоящего уровня — $p^2_{con}: p\hat{}\bar{r}^2_{hrh}(Z', Z'')$ для достижения ими целей из множества $Z$. Акты управления определяются через обратные отношения соответствующих отношений иерархии, например, $p^1_{con}$ — через обратное отношение $\bar{r}^1_{hrh}: A_{\text{У}} \to A_{\text{И}}$.

Для реализации в модели возможности генерации агентов внешней среды определяются акты $p^1_{gen}: p\hat{}r^3_{hrh}(C', C'')$, направленные на достижение агентами среды целей из множества $C$.

Для реализации в модели функций информационного взаимодействия между агентами модели определяются акты передачи информации между ними — $p^1_{trn}: p\hat{}r^1_{hrh}(Z', Z'')$, $p^2_{trn}: p\hat{}\bar{r}^1_{hrh}(Z', Z'')$, $p^3_{trn}: p\hat{}r^2_{hrh}(Z', Z'')$, $p^4_{trn}: p\hat{}\bar{r}^2_{hrh}(Z', Z'')$, $p^5_{trn}: p\hat{}r^1_{cpr}(Z', Z'')$, $p^6_{trn}: p\hat{}\bar{r}^1_{cpr}(Z', Z'')$.

Для реализации в модели функций воздействия агентов друг на друга определяются акты воздействия агентов исполнителей на агентов среды — $p^1_{frc}: p\hat{}r^1_{res}(Z', \bar{C}'')$, а также агентов среды на агентов СОТС — $p^2_{frc}: p\hat{}r^2_{res}(C', \bar{Z}'')$, $p^3_{frc}: p\hat{}r^3_{res}(C', \bar{Z}'')$ и $p^4_{frc}: p\hat{}r^4_{res}(C', \bar{Z}'')$.

Для реализации в модели функций восприятия агентами данных о других агентах определяются акты восприятия агентами-приемниками данных об агентах среды — $p^1_{per}: p\hat{}r^2_{cpr}(C', Z'')$, а также агентами среды данных об агентах СОТС — $p^2_{per}: p\hat{}r^3_{cpr}(Z'', C')$, $p^3_{per}: p\hat{}r^4_{cpr}(Z'', C')$, $p^4_{per}: p\hat{}r^5_{cpr}(Z'', C')$.

Исходя из этого, взаимодействие различных типов агентов модели можно представить с помощью схемы, представленной на рис. 1.

Предложенное описание взаимодействия между агентами модели, помимо основных свойств МАС (автономность, ограниченность представления и децентрализация) [1, 5], позволяет отображать свойства распределенного решения задач, противоборства и иерархичности системы.

Таким образом, под мультиагентной моделью СОТС понимается совокупность взаимодействующих агентов, моделирующих функционирование элементов рассматриваемой системы и объектов внешней среды. Предложенное описание позволяет обеспечить адекватность мультиагентной модели логике функционирования сложной системы в условиях противоборства.

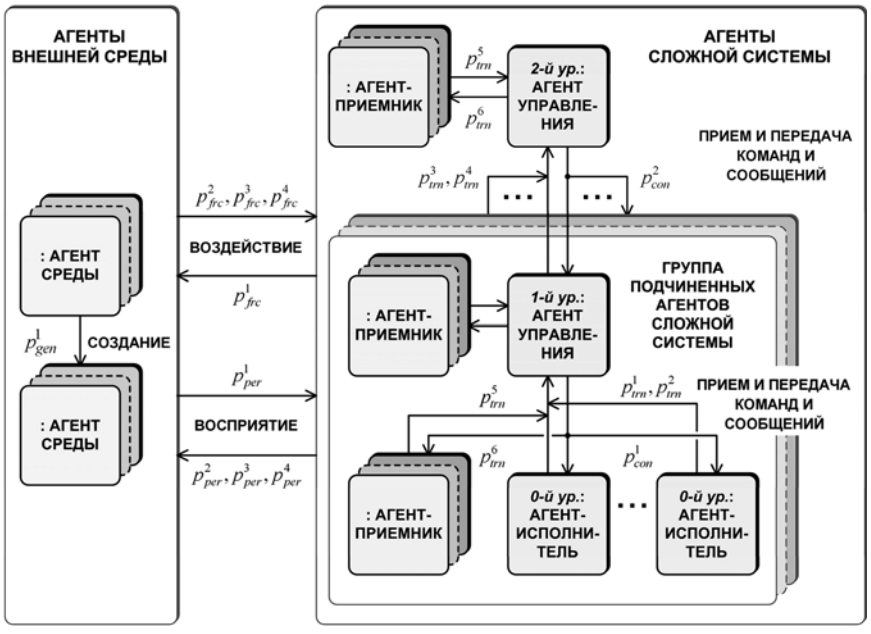



**Рис. 1. Схема взаимодействия между агентами модели**

### Построение агентов модели сложной системы

Для построения агентов модели СОТС выполняется задание структур и формализованное описание соответствующих типов агентов.

Исходя из вариантов описания агентов различных мультиагентных моделей [1, 2, 6, 7] для построения агентов $A_{\text{П}}$, $A_{\text{У}}$, $A_{\text{И}}$ и $E_{\text{С}}$ предлагается использовать следующий набор элементов — данные, действия и поведение, а описание агента выполнять следующим образом:

$$A = (D,\ S,\ U,\ Cn,\ Bh), \qquad (1)$$

где $D$ — множество данных агента; $S$ — множество детерминированных и составных действий агента; $U$ — множество детерминированных и составных различных поведений агента; $Cn$ — модель нечеткого управления агента для достижения различных целей системы; $Bh$ — итоговое поведение агента, которое задается через элементы множеств $D$, $S$, $U$, $Cn$.

Под данными агента понимается совокупность значений характеристик агента, а также показателей, характеризующих других агентов системы и агентов среды (тип агента, время действия агента и т. д.) и решения. Данные агента могут быть заданы в числовом, нечетком и лингвистическом виде, а также в виде структурированных данных и данных, описывающих нечеткие ситуации.

Элемент "действия" включает набор действий, выполнение которых обеспечивает изменение характеристик самого агента или среды. С их помощью агент воздействует на внешнюю среду. Элемент "поведение" позволяет проводить оценку и определение характеристик агента, пошагово или одновременно выполнять действия или наборы действий в виде реакции на воздействие внешней среды.

Элемент "модель нечеткого управления" в структуре агента сложной системы необходим для реализации оценки результатов собственного функционирования при решении задач управления и, по сути, определяет стратегию поведения, являющуюся совокупностью поведений и действий агента системы, направленных на достижение конкретной цели в различных ситуациях [6].

Предложенный набор элементов для описания агента модели позволяет выполнить задание обобщенной структуры агента модели СОТС, определяющей взаимосвязь ее элементов.

В соответствии с предложенной обобщенной структурой агента модели, представленной на рис. 2, элемент "данные" содержит множества значений характеристик, которые воспринимаются агентом модели из внешней среды, а другие значения определяются предварительно или во время его функционирования. С помощью действий агент модели не только воздействует на среду, но и модифицирует собственные данные. Исходя из этого данные используются как входные характеристики для действий и поведения агента.

Отличительной особенностью МАС от моделей, где функционирование их элементов определяется общим алгоритмом функционирования, является то, что агенты имеют структуру, позволяющую каждому из них действовать самостоятельно для достижения определенной цели системы.

### Описание поведения агентов модели на основе алгебры поведений

Описание поведения агентов модели предлагается рассмотреть на примере агентов управления системой, которые осуществляют выработку решений по управлению СОТС.

Исходя из определенного в выражении (1) набора элементов, формализованное описание агента управления системой заключается в задании множеств данных, действий, поведений, описании модели управления и итогового поведения, определяющих функционирование агента $A_{\text{У}}$ при решении задач управления системой.

Задание детерминированных и составных действий и поведения агентов модели предлагается осуществлять на основе алгебры поведения $U^a(S^a)$:

$$u = \sum_{g \in G} s_g \circ u_g + \varepsilon, \qquad (2)$$

где $\circ$ — операция префиксинга, задающая поведение $u_g$ на действии $s_g$; $g$ — некоторое множество индексов поведения и действий; $\varepsilon$ — одна из констант поведения [7, 8]. При этом алгебра поведения $U^a(S^a)$ задается на множестве операций $S^a$: $s_1 + s_2$ — недетерминированного выбора действий (поведения); $s_1 \bullet s_2$ — последовательной композиции действий (поведения); $s_1 \times s_2$ — комбинации действий (поведения); а также следую-

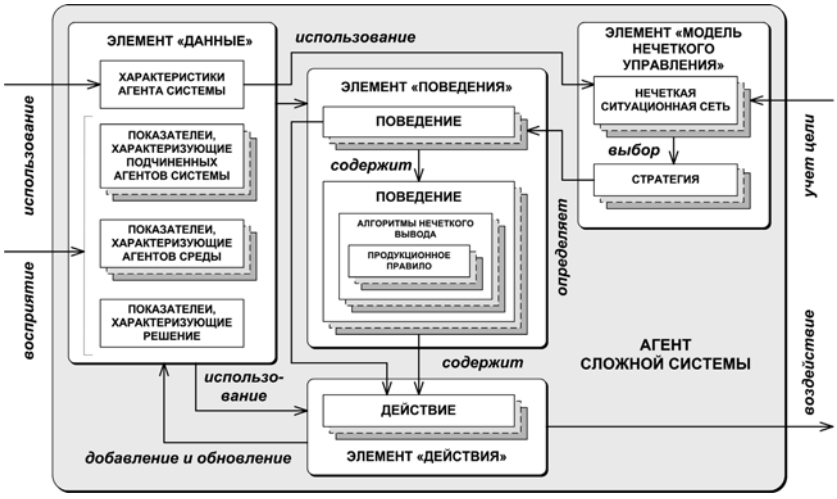



**Рис. 2. Обобщенная структура агента модели**

щих констант: ϕ — нулевого элемента (невозможного действия и поведения); δ — пустого действия; Δ — пустого поведения.

В соответствии со схемой функционирования агента управления системой, представленной на рис. 3, итоговое поведение агента $A_Y$ включает комбинацию его поведений, осуществляемых при обмене информационными сообщениями, обработке данных о других агентах, выработке решений по противодействию агентам среды и управлению агентами сложной системы, а также оценке результатов функционирования системы.


**Рис. 3. Схема функционирования агента управления системой**

Используя процедуру (2), можно предложить следующее описание поведения агента $A_Y$ при обмене информационными сообщениями с вышестоящими и подчиненными агентами модели СОТС:

$$u_{com} = u_{tran} \times u_{rec},$$

где $u_{rec}$ — передача информационных сообщений вышестоящему агенту; $u_{tran}$ — получение информационных сообщений от подчиненных агентов в соответствии с отношениями $r^1_{hrh}$, $r^2_{hrh}$, $r^1_{cpr}$ между рассматриваемым агентом $A_Y$ и взаимодействующими агентами модели. Сообщения, передаваемые между агентами модели, характеризуются типом информации и могут содержать информацию об объектах среды, информацию о характеристиках агента и команды управления. В процессе передачи сообщения другим агентам модели проводится подготовка соответствующих данных.

При функционировании агент $A_Y$ передает информационные сообщения об обрабатываемых им агентах среды, о своем состоянии и подчиненных агентах вышестоящему агенту:

$$u_{tran} = ((s_{prep.env} \bullet s_{tran}(Q_{id.main})) \circ \Delta) \times \\ \times ((s_{prep.sub} \bullet s_{tran}(Q_{id.main})) \circ \Delta),$$

где $s_{prep.env}$ — подготовка данных о рассматриваемых агентах среды; $s_{tran}$ — передача сообщения указанному агенту; $s_{prep.sub}$ — подготовка для передачи данных о своем состоянии и состоянии подчиненных агентов. Передача подчиненному агенту команд управления (информационных сообщений, содержащих команды) осуществляется при выработке соответствующих решений и поэтому не включается в поведение $u_{rec}$.

В ходе функционирования агент модели также получает информационные сообщения от других агентов об обрабатываемых ими агентах среды, их состоянии и состоянии их подчиненных агентов. По типу информации агент $A_Y$ определяет дальнейшие действия по отношению к ней, например, данные об агенте среды сохраняет в вектор $D_{vec.env}$ данных об агентах среды.

Для поведения агента $A_Y$ при получении информационных сообщений от другого агента можно предложить следующее описание:

$$u_{rec} = \left( \left( ((s_{rec} \bullet \prod_{j=1}^{J} s_{sav.sub}(Q^j_{id.sub})) \circ \Delta \right) \times \right. \\ \left. \times \left( \left( s_{rec} \bullet \prod_{n=1}^{N} s_{sav.env}(Q^n_{id.env}) \right) \circ \Delta \right) \right),$$

где $s_{rec}$ — получение информационных сообщений от другого агента; $J$ — множество индексов подчиненных агентов; $s_{sav.sub}$ — сохранение данных $D_{dat.ag}$ о подчиненном агенте в вектор данных $D_{vec.ag}$; $N$ — множество индексов рассматриваемых агентов среды; $s_{sav.env}$ — сохранение данных $D_{dat.env}$ об агенте среды в вектор данных $D_{vec.env} = \{ D^n_{dat.env} \mid k = \overline{1, N} \}$, каждый элемент которого является множеством структурированных данных об агенте среды и содержит $k$ характеристик о нем, например:

$$D_{dat.env} = \\ = \{ \tilde{d}_{env.type}, d_{env.time}, \tilde{d}_{env.time}, \tilde{Q}_{est.env}, Q_{est.env} \},$$

где $\tilde{d}_{env.type}$ — тип агента среды; $d_{env.time}$ — время действия агента среды; $\tilde{d}_{env.time}$ — время действия агента среды в лингвистическом виде; $\tilde{Q}_{est.env}$ — показатель важности агента среды в лингвистическом виде; $Q_{est.env}$ — числовой показатель важности агента среды.

Для решения задач агентами модели реализуются различные способы нечеткого вывода. Например, для описания поведения агента $A_Y$ при определении важности агента среды для решения задачи и выбора способа противодействия ему используется нечеткий вывод характеристик агентов. Поведение агента при этом можно представить следующим образом:

$$u_{est.env.all} = \prod_{n=1}^{N} (s^n_{calc.est.env} \circ u^n_{est.env}),$$

где $s_{calc.est.env}$ — расчет характеристик для определения важности рассматриваемого агента среды; $u_{est.env}$ — определение важности агента среды.

Поведение $u_{est.env}$ при определении важности агента среды реализуется с помощью процедуры нечеткого вывода по схеме Мамдани. Правила вывода, которые используются при этом, учитывают различные сочетания показателей $\tilde{d}_{env.type}$ и $d_{env.time}$ и позволяют определить значение показателя $\tilde{Q}_{est.env}$, для описания которого используются пять термов (лингвистических значений):

$$\tilde{Q}_{est.env} = \left\{ \begin{array}{c} \langle\text{"очень низкая"}/\mu_{ver.low}\rangle, \langle\text{"низкая"}/\mu_{low}\rangle, \\ \langle\text{"средняя"}/\mu_{mid}\rangle, \langle\text{"высокая"}/\mu_{high}\rangle, \\ \langle\text{"очень высокая"}/\mu_{ver.high}\rangle \end{array} \right\},$$

где $\mu_r$ — функция принадлежности терма, характеризующего $r$-е лингвистическое значение важности агента среды, $r = \overline{1,5}$.

Каждое правило, используемое при оценке важности агента среды, соответствует одному из лингвистических значений показателя $\tilde{Q}_{est.env}$ и объединяет все комбинации значений входных переменных, определяющих конкретное значение выходной характеристики в соответствии с рис. 4.

Например, представленное на схеме правило $Pr_{ver.low}$, определяющее "очень низкую" важность агента среды, имеет вид

*ЕСЛИ (((тип агента среды = первый)*
*И (время действия агента среды = малое))*
*ИЛИ (...)*
*ИЛИ ((тип агента среды = первый)*
*И (время действия агента среды = среднее))),*
*ТО (важность агента среды = очень низкая).*

Также для решения агентами модели других задач, связанных с управлением и распределением ресурсов системы, реализуются алгоритмы нечеткого вывода, сформированные на основе правил, а также результатов экспертного опроса специалистов предметной области.

Для описания поведения агентов $A_y$, участвующих в выработке решений, реализуется модель нечеткого управления *Cn* вида "ситуация — стратегия управления — действие". Модель *Cn* задается на основе нечетких ситуационных сетей (НСС), каждая из которых соответствует одной из целей системы [9]. Например, для достижения цели $Z_{lss.min}$ определяется НСС

$$\tilde{G}_{lss,\,\min} = (B^1,\ \tilde{V}^1),$$

где $B^1 = \{\tilde{B}_1^1,\ \tilde{B}_7^1\}$ — множество нечетких эталонных ситуаций сети для цели $Z_{lss.\min}$; $\tilde{V}^1 = \{\langle V_1^1(\tilde{B}_6^1, \tilde{B}_1^1)/1.0\rangle, \langle V_2^1(\tilde{B}_3^1,\ \tilde{B}_6^1)/0.7\rangle, \langle V_{21}^1(\tilde{B}_3^1,\ \tilde{B}_5^1)/0.0\rangle\}$ — множество всех возможных переходов между ситуациями сети со значениями степеней их предпочтений. Граф НСС $\tilde{G}_{lss,\,\min}$ для цели $Z_{lss.\min}$ представлен на рис. 5.

В каждой нечеткой ситуации агент $A_y$ осуществляет наиболее подходящее для данной ситуации поведение на основе правил нечеткого вывода, тем

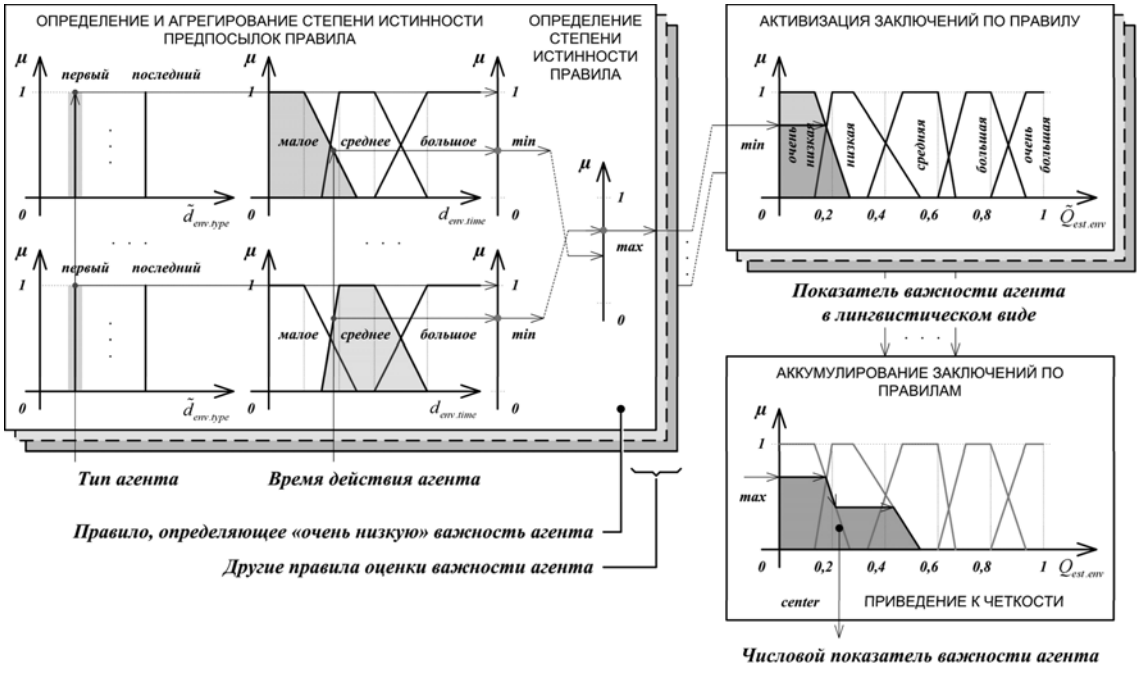



**Рис. 4. Схема вывода показателя важности агента среды**

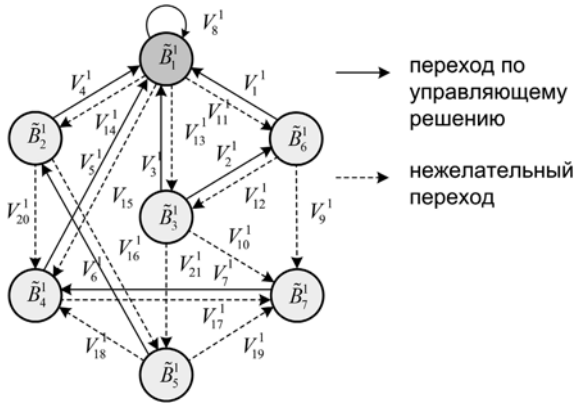



**Рис. 5. Граф нечеткой ситуационной сети**

самым, выполняя переход в целевую ситуацию, в данном случае — $\tilde{B}_1^1$.

Каждая нечеткая ситуация при этом рассматривается в зависимости от целей и определяется через нечеткие признаки ситуации — лингвистические показатели текущих производительности $\tilde{Q}_{pr}$, потерь системы $\tilde{Q}_{lss}$ и расхода ресурсов $\tilde{Q}_{cons}$.

Применение НСС, реализующих модель нечеткого управления, для построения агентов модели сложной системы позволяет им в зависимости от целей системы классифицировать входную ситуацию, корректировать характеристики агентов и реализовывать поведение и действия, необходимые для их перехода в целевую ситуацию.

Модель нечеткого управления агента на основе НСС позволяет обеспечить представление процессов управления и других процессов в мультиагентной модели СОТС с требуемой степенью детализации, а применение нечеткого вывода — учитывать ряд дополнительных (качественных) характеристик и в условиях неполной и неточной исходной информации формировать рациональные решения.

Таким образом, применение мультиагентного подхода позволяет моделируемым элементам сложной системы, функционирующей в условиях противоборства, отображать внешнюю среду в полной мере, принимать решения, изменяющие среду, оценивать результаты действий и использовать различные варианты поведения для достижения различных целей системы.

## Оценка достоверности результатов мультиагентного моделирования сложной системы в условиях противоборства

Для оценки возможности применения предложенных решений выполнена разработка и программная реализация мультиагентной модели СОТС, в которой имееется возможность оценки эффективности функционирования системы, работающей в условиях противоборства. В качестве примера подобной сложной системы на данном этапе была рассмотрена и смоделирована система управления огнем подразделений войсковой противовоздушной обороны (ПВО) [10], выполняющая управление огневыми и другими средствами ПВО при отражении удара воздушного противника.

В ходе исследования выполнены оценка адекватности и чувствительности моделирования данной системы. Экспериментальные исследования показали, что реализованная мультиагентная модель [11] адекватна и чувствительна к существенным факторам, учитываемым при моделировании исследуемой системы, под влиянием которых ее эффективность может существенно изменяться.

Для оценки достоверности моделирования СОТС различными способами использована методика [12], позволяющая оценить достоверность через относительное сокращение погрешности определения искомого результата в зависимости от учитываемых факторов и способов их учета конкретным методом моделирования. Результаты сравнительной оценки достоверности моделирования СОТС различными методами представлены на рис. 6.

К отличительным факторам, учитываемым при мультиагентном моделировании СОТС, относятся следующие: характер действий элементов системы и среды; противоборство системы и среды; решение сложно-формализуемых задач в системе; учет условий выполнения задач; учет целей функционирования системы и среды.

Результаты исследований позволяют сделать вывод о повышении достоверности результатов мультиагентного моделирования СОТС в условиях противоборства на 8—10 % за счет гибкости предложенной модели, непосредственного учета неоднородности системы, возможности задания различных вариантов поведения элементов сложной системы и среды, в том числе группового поведения, при которых рассматриваются распределенное и децентрализованное взаимодействие агентов различных групп, и решения сложно-формализуемых задач в среде моделирования.

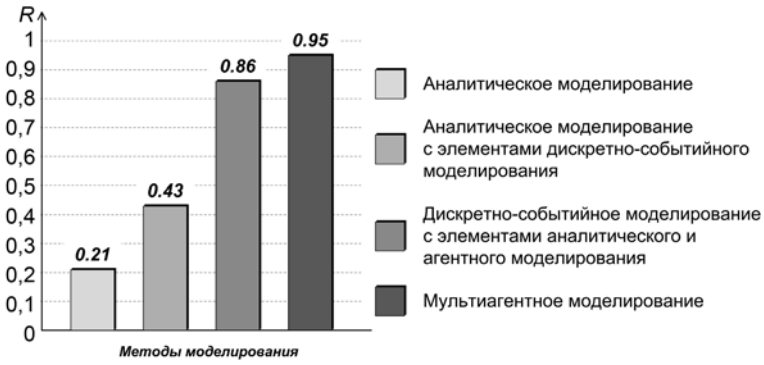



**Рис. 6. Результаты сравнительной оценки достоверности моделирования СОТС различными методами**

*Работа поддержана грантом Президента РФ для ведущих научных школ НШ-7139.2010.9, а также грантом РФФИ проект № 10-07-97506-р_центр_а.*

**Список литературы**

1. **Тарасов В. Б.** От многоагентных систем к интеллектуальным организациям: философия, психология, информатика. — М.: Эдиториал УРСС, 2002. 352 с.

2. **Пэранек Г. В.** Распределенный искусственный интеллект // Искусственный интеллект: применение в интегрированных производственных системах / Под. ред. Э. Кьюсиака. — М.: Машиностроение, 1991. С. 238—267.

3. **Балашов О. В., Борисов В. В., Круглов В. В.** Вопросы управления организационно-техническими системами военного назначения. — Смоленск: ВА ВПВО ВС РФ, 2005. 62 с.

4. **Рассел С., Норвиг П.** Искусственный интеллект: современный подход: [пер. с англ.]. 2-е изд. — М.: Вильямс, 2006. 1408 с.

5. **Кузнецов Н. А., Кульба В. В., Ковалевский С. С., Косяченко С. А.** Методы анализа и синтеза модульных информационно-управляющих систем. — М.: Физматлит, 2002. 800 с.

6. **От моделей** поведения к искусственному интеллекту / Под ред. В. Г. Редько. — М.: КомКнига, 2006. 456 с.

7. **Борисов Е. С.** Агенториентированная технология построения систем искусственного интеллекта / Сайт Е. С. Борисова. 2005. URL: http://mechanoid.narod.ru/misc/agents, свободный.

8. **Лаврищева Е. М.** Современные методы программирования: возможности и инструменты // Проблеми програмування. Теоретичні і методологічні основи програмування. — Киев: Укрпрог, 2006. № 2. С. 60—74.

9. **Мелихов А. Н., Берштейн А. С., Коровин С. Я.** Ситуационные советующиеся системы с нечеткой логикой. — М.: Наука. Физматлит, 1990. 272 с.

10. **Сысков В. В.** Способ построения модели системы управления огнем подразделений войсковой ПВО на основе теории интеллектуальных агентов. — Смоленск: ВА ВПВО ВС РФ, 2010. 24 с.

11. **Сысков В. В., Борисов В. В.** Мультиагентная модель сложной организационно-технической системы. Описание электронного ресурса № 15819 // Объединенный фонд электронных ресурсов "Наука и образование". — М.: ИНИМ РАО, 2010. 11 с.

12. **Моделирование** боевых действий соединений и частей войсковой противовоздушной обороны: учеб. пособие для адъюнктов. — Смоленск: ВА ВПВО ВС РФ, 2010. 294 с.

---

УДК 004.89

**И. П. Норенков,** д-р техн. наук, проф.,
norenkov@wwwcdl.bmdtu.ru,
**М. Ю. Уваров**, инженер,
МГТУ им. Н. Э. Баумана

# Извлечение знаний из текстовых документов на основе концептно-ориентированной типизации запросов*

*Задачи извлечения знаний из текстовых документов основаны на информационном поиске. Эффективность поиска зависит от структуры запросов и словарного заполнения слотов. Поэтому для многих приложений повышение точности и полноты поиска связано с увеличением длины запросов. В статье предложен подход к полуавтоматическому формированию запросов на основе типизации их структуры и ролевой кластеризации прикладных онтологий.*

***Ключевые слова:*** *извлечение информации, язык запросов, структура запроса, типизация запросов, ролевая кластеризация онтологий*

## Введение

Извлечение знаний (IE — *Information Extraction or Knowledge Extraction*) из документов — одна из важнейших функций интеллектуальных систем, связанных с поиском и обработкой знаний. Извлечение знаний требуется при решении ряда задач управления знаниями. Это кластеризация (рубрикация) документов, их аннотирование, поддержка принятия решений и др.

Системы IE классифицируют по нескольким признакам.

По составу входной информации, характеризующей информационные потребности пользователей, различают системы KBS (*keyword-based systems*), ориентированные на ключевые слова, и системы CBS (*concept-based systems*), ориентированные на концепты. Первоначально в большинстве информационно-поисковых систем (ИПС) применяли только поиск, ориентированный на ключевые слова, который по-прежнему широко используют, в частности, в специализированных системах, например в библиотечных [1]. Однако при работе с большими массивами документов точность и полнота поиска в таких системах остаются невысокими. Поэтому наряду с KBS применяют CBS, в которых ориентация на концепты подразумевает использование тезаурусов или онтологий [2]. Если в KBS релевантность запроса и документа определяется по совпадению имеющихся в них слов, то в CBS сопоставляются не отдельные слова, а концепты, т. е. учитываются синонимические связи и, возможно, окрестности концептов в семантических сетях.

В зависимости от масштабов области применения системы IE могут быть специализированными или независимыми от конкретных приложений. Большинство существующих систем относится к специализированным. В них извлечение нужной информации связано с поиском в текстах фраз или упоминаний фактов (NE — *named entities*), характер-

* Работа выполнена при поддержке РФФИ (12-07-002222а).

ных для приложения. В качестве фактов используются имена собственные, названия организаций, географических мест, даты событий и т. п.

Так, в системе Artequakt [3], предназначенной для создания онтологий в области искусства и генерирования биографий артистов, извлечение знаний основано на автоматическом поиске фактов. Прикладная онтология формируется с использованием уже существующих общих онтологий WordNet и CRM с добавлением специфических понятий. Выполняется анализ текстов, выявляются сведения, преобразуемые в триплеты (субъект — предикат — объект), которые представляются средствами RDF. На их основе генерируются тексты биографий.

В системе SOPHIA [4] выполняется кластеризация текстовых документов, порожденных определенными профессиональными группами, на основе выявления специфичных терминов и фраз, характерных для различных профессиональных сообществ, и определения вероятностей совместного употребления слов в одних и тех же документах.

Ориентация на факты широко используется и в системах, инвариантных к приложениям. В системе управления досье "X-Files" [5] компании Ай-Теко осуществляется поиск и аналитическая обработка сведений о различных фактах, что позволяет кластеризовать факты, находить связи между ними, и тем самым преобразовывать неструктурированную информацию в структурированную.

Другой системой извлечения знаний из текстов той же компании является "Аналитический курьер" [6]. В этой системе кластеризация документов основана на выделении общих признаков (сущности, темы) и формировании частотной матрицы "понятие/документ". Кластеры определяются по преобладанию в документах общих понятий.

К числу систем, основанных на извлечении из текстов сведений о фактах, относится также известная система TextRunner [7]. Сведения из текстов извлекаются в виде триплетов $r(x, y)$, где $r$ — отношение между объектами $x$, $y$. Другой используемый в системе подход — семантический разбор, в соответствии с которым лингвистические конструкции преобразуются в логические формулы, используемые, например, в режиме работы "вопрос — ответ".

В научно-образовательной сфере извлечение знаний из баз документов требуется, прежде всего, для поддержки принятия решений и формирования электронных учебных материалов, оптимизированных под запросы пользователя. Остается актуальной проблема представления запросов пользователя в форме, позволяющей добиться адекватного выражения информационных потребностей пользователей. Для решения этой проблемы совершенствуются языки запросов и средства их реализации в ИПС.

В языках запросов известных ИПС имеются возможности создания запросов в виде сложных высказываний, образованных из ключевых слов, т. е. основу формирования запросов составляет подход KBS. Эффективность поиска и извлечения знаний из текстовых документов может быть повышена при использовании подхода CBS.

Статья посвящена описанию метода формирования сложных запросов в системах извлечения знаний из текстовых документов с применением концептно-ориентированного подхода и прикладных онтологий.

## Типизация запросов

Языки запросов в развитых ИПС позволяют формировать сложные запросы из ключевых слов и связок И, ИЛИ, НЕ с ограничениями на последовательность слов и на расстояния между ними в составе документа или отдельного предложения. Формирование запроса на поиск прецедентов принятия решений (ПР), как и для поисковых задач в большинстве других приложений, оказывается многовариантной задачей, причем полезность результатов поиска существенно зависит от выбранного варианта.

Рассмотрим примеры, характеризующие нетривиальность проблемы формирования запросов.

Пусть требуется найти в Веб-пространстве прецеденты решения задачи проектирования объекта $X$, в качестве которого выберем "ПЛИС". На запрос

разработка / + 2 ПЛИС

в системе Яндекс получено 830 ответов. Это избыточно большой объем информации. Поскольку точность поиска можно регулировать изменением числа конъюнктивно связанных элементов запроса, используем запрос

(подход | метод | методика | маршрут) /
+ 1 разработка / + 2 ПЛИС (1)

в ответе на который получены ссылки на два документа, что является недостаточным результатом. На уточненный вариант запроса

(подход | метод | методика | маршрут | идея |
концепция | алгоритм | последовательность) /
+ 1 ((принятие / + 1 решений) | проектирование |
разработка | создание | реализация |
моделирование | выбор) / + 2 ПЛИС (2)

было получено уже 24 ссылки.

*Примечание* 1. В приведенных запросах | — символ связки "ИЛИ", а конструкция / + $n$ означает, что в релевантных документах выбранные слова должны следовать в заданном порядке, причем между каждой парой последовательных слов должно быть не более $n - 1$ посторонних слов.

*Примечание* 2. Части запроса, разделенные символом /, будем далее называть слотами запроса, а термины в словосочетаниях, являющихся релевантными ответами поисковой системы, — слотами ответов (найденных сниппетов).

Рассмотрение подобных примеров позволяет сделать следующий вывод: для увеличения полноты поиска прецедентов необходимо расширение словарного наполнения слотов запроса, что следует из сравнения результатов поиска по запросам (1) и (2). Однако формирование многословных запросов вручную весьма затруднительно, кроме того, в ИПС имеются ограничения на длину запросов.

Решением проблемы может стать полуавтоматическое формирование запросов, основанное на их типизации для конкретных классов задач. Так, одним из основных типов запроса для задач принятия проектных решений (ПР) являются структуры (1) и (2) с соответствующим наполнением слотов и с выделением определенных ролей для каждого слота. Для слотов структур (1) и (2) это роли "средство", "действие", "объект". При этом состав слотов "средство" и "действие" в значительной мере оказывается инвариантным к приложениям и, следовательно, типичным для задач ПР, а прикладная направленность запроса определяется составом слота "объект".

### Ролевая кластеризация онтологий

Использование ИПС типа Яндекс и соответствующего языка запросов для реализации поиска по сложным запросам оказывается возможным лишь при сравнительно малом числе слов в запросе. Однако в общем случае желаемое число слов в каждом слоте может составлять десятки и более. Поэтому для решения задач извлечения знаний из документов целесообразно иметь специализированные системы CBS, в которых слоты типовых запросов формируются из концептов предварительно разработанных прикладных онтологий [8].

| "Средство" | "Свойство" | "Действие" |
|---|---|---|
| Алгоритм<br>Дедукция<br>Диакоптика<br>Допущение<br>Идея<br>Индукция<br>Мера<br>Метод<br>Методика<br>Методология<br>Модель<br>Подход<br>Предположение<br>Проблема<br>Задача<br>Способ<br>Средство<br>Шаблон<br>Эвристика<br>и др. | Архитектура<br>Аспект<br>Быстродействие<br>Вероятность<br>Виртуализация<br>График<br>Неисправность<br>Допуск<br>Достижимость<br>Идемпотентность<br>Идентификатор<br>Интенсивность<br>Интенсионал<br>Интерактивность<br>Интероперабельность<br>Качество<br>Ключ<br>Когерентность данных<br>Концепция<br>Меню<br>Метаданные<br>Надежность<br>Наследование<br>Категоризация<br>Обусловленность<br>Отказоустойчивость<br>Отношение<br>Параметр<br>и др. | Абстракция<br>Автоматизация<br>Агрегация<br>Алгебраизация<br>Анализ<br>Анимация<br>Бифуркация<br>Вейвлет-преобразование<br>Верификация<br>Визуализация<br>Восстановление<br>Выбор<br>Выделение<br>Вычисление<br>Герменевтика<br>Декодирование<br>Декомпозиция<br>Дельта-модуляция<br>Демодуляция<br>Дефаззификация<br>Документооборот<br>Запоминание<br>Идентификация<br>Изучение<br>Инсталляция<br>Использование<br>Исследование<br>Квантификация<br>Кластеризация<br>Кодирование<br>и др. |

Одной из таких систем является система Precedent. Наполнением базы знаний этой системы являются прикладные онтологии. Концепты конкретной онтологии распределены по кластерам. Кластеры соответствуют ролям, которые выполняют концепты в различных словосочетаниях, типичных для прикладной онтологии.

В варианте кластеризации, ориентированной на поиск прецедентов ПР, выделены роли "объект", "свойство", "действие", "средство". Примеры концептов, входящих в кластеры "средство", "свойство", "действие", приведены в таблице. При этом концепты, относящиеся к категории "объект", распределены по нескольким кластерам в соответствии с разделением прикладной онтологии на более частные задачные онтологии.

Кластеризованная база знаний используется для полуавтоматического формирования запросов. Слоты запросов автоматически заполняются концептами определенных кластеров. Пользователь в большинстве случаев лишь выбирает задачную онтологию или заполняет слот "объект" и при необходимости вручную корректирует другие слоты запроса.

При решении задач поиска прецедентов ПР с помощью системы Precedent возможно использование различных структур запросов. Типичными являются структуры "средство — действие — объект", "действие — средство — действие — объект", "средство — действие — действие — объект". Так, при задании этих структур и выборе в качестве объекта концепта "программное обеспечение" были получены ссылки на ряд документов с выдачей сниппетов. Приведенные далее примеры сниппетов позволяют судить о перспективности найденных документов в качестве источников запрошенной информации:

— "Формализован процесс разработки программного обеспечения КИС в виде диаграмм UML, что позволило автоматизировать управление этим процессом";

— "Поставлена задача выбора эффективных вариантов моделей разработки программного обеспечения, которая сводится к задаче гипервекторного ранжирования";

— "Шаблоны проектирования — это многократно используемые решения широко распространенных проблем, возникающих при разработке программного обеспечения";

— "Одной из ключевых проблем, возникающих в процессе разработки программного обеспечения, является проблема реализации унифицированного подхода к фильтрации данных и формирования логического выражения (условия), которому должен соответствовать результирующий набор данных";

— "...описывает процесс разработки программного обеспечения в форме метаданных, которые используются совместно с активами".

Для задач других типов как состав концептов, так и их распределение по кластерам могут иметь определенные отличия. Примером слотов запроса для задач поиска документов, описывающих принципы функционирования объекта *X*, может служить запись

(рассмотрен | представлен | изложен | описан | показан | раскрыт) / + 1 (устройство | структура | конструкция | конфигурация | принцип | условие | основа | работа | функционирование | действие) / + 1 (*X*)

с соответствующим заполнением слота *X*.

### Заключение

Эффективность решения задач извлечения знаний из документов зависит от структуры и словарного содержания запросов к информационной системе. Показано, что повышение точности и полноты поиска может быть достигнуто с помощью структуризации и заполнения слотов запросов в соответствии с распределением концептов онтологий по ролевым кластерам.

### Список литературы


1. **Joten Dingh R. K.** A robast information retrieval technique for a bibliographical database // Annals of Library and Information Studies. 2008. V. 55. P. 135—140.
2. **Haav H., Lubi T.** A Survey of concept-based information retrieval tools on the Web // In Proc. of 5th East-European Conf., Vilnus: Technika, 2001. P. 29—41.
3. **Alani H., Kim S., Millard D. E., Weal M. J., Hall W., Lewis P. H., Shadbolt N.** Web based Knowledge Extraction and Consolidation for Automatic Ontology Instantiation // 2nd Int. Conf. Knowledge Capture (K-Cap'03). Workshop on Knowledge Markup and Semantic Annotation. Sanibel Island, USA, 2003.
4. **Добрынин В. Ю.** Анализ коллекции нормативных документов 2007 года средствами системы SOPHIA // Российский семинар по оценке методов информационного поиска: Труды РОМИП, 2010. С. 164—171.
5. **Компания** Ай-Теко. URL: http://www.i-teco.ru/xfiles.html
6. **Система** извлечения знаний из текстов "Аналитический курьер". URL: http://www.i-teco.ru/ac.html
7. **Yates A., Banko M., Broadhead M., Cafarella M., Etzioni O., Soderland S.** TextRunner: Open Information Extraction on the Web // Procs. of Human Language Technologies, NJ, USA, 2007. P. 25—26.
8. **Норенков И. П., Уваров М. Ю.** Задачи обработки знаний на основе ролевой кластеризации онтологий // Информационные технологии. 2012. № 2. С. 19—24.


# МОДЕЛИРОВАНИЕ И ОПТИМИЗАЦИЯ


УДК 62-50:519.7/8



**В. И. Левин**, д-р техн. наук, проф.,
Пензенская государственная технологическая академия,
e-mail: levin@pgta.ru


## Методы оптимизации систем в условиях интервальной неопределенности параметров


*Рассмотрены существующие подходы к оптимизации систем при неопределенности. Дана точная постановка задачи условной оптимизации в случае интервальной неопределенности параметров целевой функции и ограничений. В связи с этим изложена математическая теория сравнения интервалов. На основе данной теории сформулирован и обоснован метод детерминизации, позволяющий решить поставленную задачу путем ее сведения к двум полностью определенным задачам условной оптимизации того же типа.*



***Ключевые слова:*** *оптимизация систем, неопределенность, детерминированная оптимизация, интервальная оптимизация, сведение интервальной задачи, сравнение интервалов*


### Введение

Задачи оптимизации имеют большое прикладное значение: на их основе строятся методы оптимального проектирования систем — технических, экономических, социальных и т. д., обеспечивающие достижение наилучшего, в определенном смысле, результата работы создаваемой системы. В связи с этим к настоящему времени создано огромное число методов решения задач оптимизации, как универсальных, рассчитанных на применение к задачам различных классов, так и специализированных, позволяющих эффективно решать лишь отдельные узкие классы задач [1—6]. Однако при всем различии существующих методов, все они имеют одно общее свойство — применимость только к тем задачам оптимизации, в которых оптимизируемая функция известна точно (детерминирована). Между тем встречающиеся на практике задачи оптимизации таковы, что их оптимизируемые функции обычно известны не точно, а с той или иной степенью неопределенности (недетерминированы). Это вызвано следующими факторами: 1) многим реальным процессам свойственна естественная неопределенность; 2) параметры большинства систем вслед-

ствие погрешности их вычислений или измерений известны неточно; 3) параметры многих систем изменяются во времени.

Исходя из этого, возникает проблема оптимизации неполностью определенных (недетерминированных) функций. Эта проблема является более сложной, чем традиционная оптимизация полностью определенных (детерминированных) функций, поскольку для нее еще необходимо:

- обобщить понятие экстремума функции;
- выяснить условия существования экстремума функции, связанные с ее недетерминированностью;
- разработать специальные методы поиска экстремума таких функций.

Существуют различные подходы к нахождению оптимума неполностью определенных (недетерминированных) функций, различающиеся достоинствами и недостатками.

Первый подход состоит в решении задачи оптимизации для определенных значений параметров оптимизируемой функции, взятых внутри заданных областей их неопределенности. Так, можно взять наихудшее сочетание значений этих параметров (пессимистический подход) или их наилучшее сочетание (оптимистический подход) и др. Достоинство данного подхода — простота интерпретации полученного решения, недостаток — ориентировка на какое-то одно определенное сочетание значений параметров, которое на практике реализуется редко, что может обернуться неоправданной сложностью решения.

Второй подход заключается в решении задачи оптимизации для усредненных значений параметров оптимизируемой функции, что предполагает задание вероятностных распределений этих параметров внутри областей их неопределенности. Достоинство указанного подхода — ориентировка получаемого решения хотя и на одно, но зато наиболее часто встречающееся сочетание значений параметров функции, недостаток — необходимость знания вероятностных распределений параметров функции, что не всегда возможно.

Третий подход идейно близок второму, но вместо вероятностных распределений параметров функции, являющихся объективными характеристиками, используются нечеткие распределения параметров, получаемые экспертным путем, т. е. субъективно.

В наших работах [7—14] был предложен и детально описан применительно к различным оптимизационным задачам детерминизационный подход к нахождению оптимума неполностью определенных функций. Этот подход принципиально отличен от предыдущих тем, что оптимизация неполностью определенной функции проводится с учетом всего множества возможных значений недетерминированных параметров функции. Этот подход позволяет для любой функции, неопределенность которой выражается в том, что ее параметры известны лишь с точностью до интервалов возможных значений, свести нахождение оптимума этой функции к нахождению одноименных оптимумов двух детерминированных функций. Таким образом, для нахождения оптимума неполностью определенных (недетерминированных) функций становится возможным применять многочисленные хорошо известные и эффективные методы нахождения оптимума полностью определенных (детерминированных) функций. Второй причиной выбора неопределенности именно интервального типа было то, что интервальные оценки неизвестных параметров наиболее просты и доступны для получения. В этом состоит основное достоинство предложенного метода оптимизации неполностью определенных функций — метода детерминизации.

В настоящей работе детерминизационный подход к оптимизации неполностью определенных функций обосновывается в общем виде, не зависящем от особенностей оптимизируемых функций.

## Постановка задачи

Согласно сказанному выше, пусть имеется некоторая произвольная непрерывная функция $n$ переменных

$$y = F(x_1, \ldots, x_n), \tag{1}$$

причем все параметры (коэффициенты) ее явного представления известны точно. Будем рассматривать функцию (1) в ограниченной области, определяемой следующей системой ограничений:

$$\Phi_i(x_1, \ldots, x_n) \leqslant b_i,\ i = \overline{1, m}\ \}. \tag{2}$$

Тогда относительно функции (1) сформулируем полностью определенную задачу условной оптимизации

$$F(x_1, \ldots, x_n) = \max,$$

$$\text{при } \Phi_i(x_1, \ldots, x_n) \leqslant b_i,\ i = \overline{1, m}\ \}. \tag{3}$$

В современном математическом программировании имеется множество различных методов эффективного решения оптимизационных задач вида (3), ориентирующихся на тип функций $F$ и $\Phi_i$, $i = \overline{1, m}$. Литературу по этим методам можно найти, в частности в работах [1—5].

Пусть теперь параметры $p_k$, $k = \overline{1, l}$, явного представления функции $F$ известны не точно, а с точностью до интервалов значений, т. е. имеют вид интервалов $\tilde{p}_k = [p_{k1}, p_{k2}]$. Пусть далее таким же образом заданы параметры $q_s$ явного представления функций $\Phi_i$ в левых частях ограничений, а также параметры $b_i$ в правых частях, т. е. $\tilde{q}_{si} = [q_{si1}, q_{si2}]$,

$s = \overline{1, t}$ , $\tilde{b}_i = [b_{i1}, b_{i2}]$, $i = \overline{1, m}$ . Тогда функции $F$ и $\Phi_i$, $i = \overline{1, m}$ , также становятся интервальными (т. е. принимающими вид интервалов $\tilde{F}$ и $\tilde{\Phi}_i$, $i = \overline{1, m}$ ), определяемыми с точностью до интервалов возможных значений, равно как и параметры $b_i$, $i = \overline{1, m}$ (т. е. принимающие вид интервалов $\tilde{b}_i$, $i = \overline{1, m}$ ). В результате полностью определенная задача условной оптимизации (3) переходит в интервальную (неполностью определенную) задачу условной оптимизации вида

$$\tilde{F}(x_1, ..., x_n) = \max,$$

$$\text{при } \tilde{\Phi}(x_1, ..., x_n) \leqslant \tilde{b}_i,\ i = \overline{1, m}\ \}. \qquad (4)$$

Таким образом, поставленная задача сводится к необходимости разработки методики решения оптимизационной задачи вида (4).

### Математика сравнения интервалов

В основе решения поставленной выше интервальной задачи условной оптимизации (4) лежит математическая теория сравнения интервалов.

Рассмотрим два интервала $\tilde{a} = [a_1, a_2]$ и $\tilde{b} = [b_1, b_2]$. Попытаемся сравнить эти интервалы по величине, рассматривая их как интервальные числа. Можно сравнивать интервалы $\tilde{a}$ и $\tilde{b}$ на основе отношений в отдельных парах вещественных чисел $(a_i, b_j)$, где $a_i \in \tilde{a}$, $b_j \in \tilde{b}$. Однако такой подход приводит к провалу, поскольку в общем случае при произвольных интервалах $\tilde{a}$ и $\tilde{b}$ некоторые пары чисел $(a_i, b_j)$ будут находиться в отношении $a_i > b_j$, в то время как другие — в противоположном отношении $a_i < b_j$. Поэтому единственное, что остается, — реализовать сравнение интервалов на теоретико-множественном уровне, рассматривая их как единое целое, не подлежащее дроблению на части. Этот путь был реализован автором в 1990-е годы. Ниже приводится краткое изложение полученных результатов [15—18].

Согласно высказанному выше подходу к интервалам, операцию взятия максимума $\vee$ и минимума $\wedge$ двух интервалов $\tilde{a} = [a_1, a_2]$ и $\tilde{b} = [b_1, b_2]$ введем в виде следующих теоретико-множественных конструкций:

$$\begin{aligned} \tilde{a} \vee \tilde{b} &= \{a \vee b \mid a \in \tilde{a},\ b \in \tilde{b}\}, \\ \tilde{a} \wedge \tilde{b} &= \{a \wedge b \mid a \in \tilde{a},\ b \in \tilde{b}\}, \end{aligned} \qquad (5)$$

Таким образом, взятие максимума (минимума) двух интервалов $\tilde{a}$ и $\tilde{b}$ определяется, согласно выражениям (5), как нахождение максимума (минимума) двух точечных величин $a$ и $b$, при условии, что конкретные значения этих величин пробегают все возможные значения соответственно из интервалов $\tilde{a}$ и $\tilde{b}$. Теперь для того чтобы интервалы $\tilde{a}$ и $\tilde{b}$ можно было сравнить по величине и установить их отношение — $\tilde{a} \geqslant \tilde{b}$ или $\tilde{a} \leqslant \tilde{b}$, необходимо, во-первых, чтобы введенные операции $\vee$, $\wedge$ над этими интервалами существовали, во-вторых, чтобы эти операции давали в результате один из операндов — $\tilde{a}$ или $\tilde{b}$, и, в-третьих, чтобы эти две операции были согласованы, в том смысле, что если большим (меньшим) является один из интервалов, то меньшим (большим) является другой. Сформулированное условие сравнимости двух интервалов по величине является, как этот хорошо видно, не только необходимым, но и достаточным.

Нетрудно доказать, что условие согласованности операций $\vee$ и $\wedge$ над интервалами выполняется всегда, т. е. для любой пары интервалов $(\tilde{a}, \tilde{b})$. Очевидно также, что всегда выполняется условие существования введенных операций взятия максимума $\vee$ и минимума $\wedge$ двух интервалов, причем результатом операции оказывается некоторый, вообще говоря, новый интервал. Таким образом, необходимым и достаточным условием сравнимости интервалов $\tilde{a}$ и $\tilde{b}$ оказывается условие, по которому операции $\tilde{a} \vee \tilde{b}$ и $\tilde{a} \wedge \tilde{b}$ должны иметь своим результатом один из интервалов — $\tilde{a}$ или $\tilde{b}$. Последняя формулировка условия сравнимости интервалов открывает возможность получения его в конструктивной форме, пригодной для практического применения. Основной результат здесь формулируется следующим образом.

**Теорема 1.** Для того чтобы два интервала $\tilde{a} = [a_1, a_2]$ и $\tilde{b} = [b_1, b_2]$ были сравнимы по величине (отношению $\geqslant$) и находились в отношении $\tilde{a} \geqslant \tilde{b}$, необходимо и достаточно, чтобы их границы подчинялись условиям

$$a_1 \geqslant b_1,\ a_2 \geqslant b_2, \qquad (6)$$

а чтобы они были сравнимы по величине (отношению $\leqslant$) и находились в отношении $\tilde{a} \leqslant \tilde{b}$, необходимо и достаточно, чтобы выполнялись условия

$$a_1 \leqslant b_1,\ a_2 \leqslant b_2. \qquad (7)$$

Из утверждения теоремы 1 можно получить следствие: интервалы $\tilde{a}$ и $\tilde{b}$ являются сравнимыми по отношению $\geqslant$ и $\leqslant$ (и находятся именно в этом отношении) только в случае, когда в таком же отношении находятся их одноименные границы $a_1$, $b_1$ и $a_2$, $b_2$. Другими словами, два интервала $\tilde{a}$ и $\tilde{b}$ находятся в отношении $\tilde{a} \geqslant \tilde{b}$ только тогда, когда

$\tilde{a}$ сдвинут обеими границами вправо относительно $\tilde{b}$, и находятся в отношении $\tilde{a} \leqslant \tilde{b}$ только тогда, когда $\tilde{a}$ сдвинут обеими границами влево относительно $\tilde{b}$.

Значение теоремы 1 в том, что она сводит сравнение двух интервалов и выбор большего (меньшего) из них к сравнению границ этих интервалов, которые являются вещественными числами. Таким образом разрешается проблема сравнения интервалов.

**Теорема 2.** Для того чтобы два интервала $\tilde{a} = [a_1, a_2]$ и $\tilde{b} = [b_1, b_2]$ были несравнимы по величине (по отношению $\geqslant$ и $\leqslant$), т. е. не находились в отношении $\tilde{a} \geqslant \tilde{b}$ или $\tilde{a} \leqslant \tilde{b}$, необходимо и достаточно выполнения условий

$$(a_1 < b_1,\ a_2 > b_2) \text{ или } (b_1 < a_1,\ b_2 > a_2). \qquad (8)$$

Эта теорема показывает, что интервалы $\tilde{a}$ и $\tilde{b}$ несравнимы по отношению $\geqslant$ и $\leqslant$ только тогда, когда один из них полностью "накрывает" другой.

Значение теоремы 2 в том, что она показывает существование определенных случаев несравнимости интервалов по отношениям $\geqslant$ и $\leqslant$, в отличие от вещественных чисел, которые всегда сравнимы по этим отношениям. Несравнимость величин некоторых интервалов — естественный результат того, что интервальные числа, в отличие от вещественных чисел, задаются не точно, а с неопределенностью (известно, что число принимает некоторое значение в заданном интервале, но не уточняется, какое именно это значение). На основе теорем 1 и 2 можно доказать следующие положения.

**Теорема 3.** Для того чтобы в некоторой системе интервалов числовой оси $\tilde{a}(1) = [a_1(1),\ a_2(1)]$, $\tilde{a}(2) = [a_1(2), a_2(2)]$, ... существовал максимальный интервал (который находится со всеми остальными интервалами в отношении $\geqslant$), необходимо и достаточно, чтобы границы данного интервала были расположены относительно одноименных границ всех остальных интервалов согласно таким условиям (системе неравенств):

$$\left.\begin{array}{l} a_1(1) \geqslant a_1(2),\ a_1(1) \geqslant a_1(3),\ \ldots \\ a_2(1) \geqslant a_2(2),\ a_2(1) \geqslant a_2(3),\ \ldots . \end{array}\right\}. \qquad (9)$$

Условия (9) даны в случае, когда максимальным является интервал $\tilde{a}(1)$, что, очевидно, не ограничивает общности.

**Теорема 4.** Для того чтобы в некоторой системе интервалов числовой оси $\tilde{a}(1) = [a_1(1),\ a_2(1)]$, $\tilde{a}(2) = [a_1(2), a_2(2)]$, ... существовал минимальный интервал (который находится со всеми остальными интервалами в отношении $\leqslant$), необходимо и достаточно, чтобы границы данного интервала были расположены относительно одноименных границ всех остальных интервалов согласно следующей системе неравенств

$$\left.\begin{array}{l} a_1(1) \leqslant a_1(2),\ a_1(1) \leqslant a_1(3),\ \ldots \\ a_2(1) \leqslant a_2(2),\ a_2(1) \leqslant a_2(3),\ \ldots . \end{array}\right\}. \qquad (10)$$

Аналогично теореме 3 условия (10) записаны для случая, когда минимальным является интервал $\tilde{a}(1)$, что не ограничивает общности.

Теоремы 3 и 4 показывают, что интервал является максимальным (минимальным) среди множества имеющихся интервалов только тогда, когда максимальны (минимальны) его нижняя граница — среди нижних границ всех интервалов, и верхняя граница — среди верхних границ всех интервалов.

### Идея решения

В задаче (4) целевая функция $\tilde{F}(x_1, \ldots, x_n)$, функции $\tilde{\Phi}_i(x_1, \ldots, x_n)$, $i = \overline{1, n}$, в левых частях ограничений и параметры $\tilde{b}_i$, $i = \overline{1, m}$, в их правых частях являются интервальными и поэтому могут быть записаны в виде интервалов

$$\begin{array}{l} \tilde{F}(x_1, \ldots, x_n) = [F_1(x_1, \ldots, x_n),\ F_2(x_1, \ldots, x_n)]; \\ \tilde{\Phi}_i(x_1, \ldots, x_n) = [\tilde{\Phi}_{i1}(x_1, \ldots, x_n),\ \tilde{\Phi}_{i2}(x_1, \ldots, x_n)], \\ i = \overline{1, m}, \\ \tilde{b}_i = [b_{i1}, b_{i2}],\ i = \overline{1, m}. \end{array} \qquad (11)$$

После этого задачу (4) можно переписать в явном интервальном виде:

$$\begin{array}{l} [F_1(x_1, \ldots, x_n),\ F_2(x_1, \ldots, x_n)] = \max, \\ [\Phi_{i1}(x_1, \ldots, x_n),\ \Phi_{i2}(x_1, \ldots, x_n)] \leqslant [b_{i1}, b_{i2}], \\ i = \overline{1, m}, \end{array} \qquad (12)$$

который уже поддается решению. Действительно, согласно теореме 3 интервальное уравнение в (12) можно записать в виде эквивалентной пары обычных (детерминированных) уравнений:

$$F_1(x_1, \ldots, x_n) = \max,\ F_2(x_1, \ldots, x_n)] = \max. \qquad (13)$$

Далее, по теореме 1 систему интервальных неравенств в (12) можно записать в виде эквивалентной системы детерминированных неравенств

$$\Phi_{i1}(x_1, \ldots, x_n) \leqslant b_{i1},\ \Phi_{i2}(x_1, \ldots, x_n) \leqslant b_{i2},\ i = \overline{1, m}. \qquad (14)$$

Соединяя пару уравнений (13) с системой неравенств-ограничений (14), получим две детерминированные задачи условной оптимизации вида (3):

$$\left.\begin{aligned} &F_1(x_1, ..., x_n) = \max, \\ &\Phi_{i1}(x_1, ..., x_n) \leqslant b_{i1},\ i = \overline{1, m}\ , \\ &\Phi_{i2}(x_1, ..., x_n) \leqslant b_{i2},\ i = \overline{1, m}\ , \end{aligned}\right\} \quad (15)$$

$$\left.\begin{aligned} &F_2(x_1, ..., x_n) = \max, \\ &\Phi_{i1}(x_1, ..., x_n) \leqslant b_{i1},\ i = \overline{1, m}\ , \\ &\Phi_{i2}(x_1, ..., x_n) \leqslant b_{i2},\ i = \overline{1, m}\ . \end{aligned}\right\} \quad (16)$$

При этом задачу (15) назовем нижней граничной задачей исходной интервальной задачи (4), а задачу (16) — ее верхней граничной задачей.

Из выполненного нами построения следует, что пара детерминированных задач условной оптимизации (15), (16), рассматриваемых в совокупности, эквивалентна исходной интервальной задаче (4). Таким образом, для получения решения интервальной задачи условной оптимизации (4) надо решить ее нижнюю (15) и верхнюю (16) граничные задачи. В общем случае решения нижней и верхней граничных задач имеют вид $\{M_н(x), F_{1\,max}\}$, $\{M_в(x), F_{2\,max}\}$, где $M_н(x)$ $\{M_в(x)$ — множества точек решений $x = (x_1, ..., x_n)$ нижней и верхней задачи; $F_{1,\,max}$, $F_{2,\,max}$ — полученные максимальные значения целевых функций этих задач. Решение интервальной задачи (4) составляется из решений ее нижней и верхней граничных задач в виде:

$$\{x^* \in M_н(x) \cap M_в(x),\ \tilde{F}_{max} = [F_{1max}, F_{2max}]\}. \quad (17)$$

Таким образом, в качестве точки решения $x^*$ в (17) берется любая точка из пересечения множеств точек решения нижней и верхней граничных задач, а в качестве максимума целевой функции $\tilde{F}_{max}$ — интервал от максимального значения целевой функции нижней задачи $F_{1\,max}$ до максимального значения целевой функции верхней задачи $F_{2max}$.

Преимущество нашего подхода к решению интервальной задачи условной оптимизации заключается в возможности использования для этого традиционных, хорошо разработанных методов решения детерминированных задач такой оптимизации. Основанный на этом подходе метод решения интервальной задачи условной оптимизации можно назвать методом детерминизации, поскольку он сводит решение недетерминированной оптимизационной задачи (4) к решению двух детерминированных задач (15) и (16).

## Алгоритм решения

Для решения интервальной задачи (4) методом детерминизации необходимо действовать по следующему алгоритму.

**Шаг 1.** Используя формулы интервальной математики, выражающие результаты элементарных преобразований интервалов [18]

$$\begin{aligned} &[a_1, a_2] + [b_1, b_2] = [a_1 + b_1, a_2 + b_2]; \\ &[a_1, a_2] - [b_1, b_2] = [a_1 - b_2, a_2 - b_1]; \\ &[a_1, a_2]/[b_1, b_2] = [a_1, a_2] \cdot [1/b_2, 1/b_1], \\ &k[a_1, a_2] = \begin{cases} [ka_1, ka_2],\ k > 0, \\ [ka_2, ka_1],\ k < 0; \end{cases} \\ &[a_1, a_2] \cdot [b_1, b_2] = [\min_{i,\,j}(a_i \cdot b_j),\ \max_{i,\,j}(a_i \cdot b_j)], \end{aligned} \quad (18)$$

представляем целевую функцию $\tilde{F}$ и функции ограничений $\tilde{\Phi}_i$ задачи (4) в интервальной форме. Так же представляем параметры $b_i$ в ограничениях. Полученные представления имеют вид (11).

**Шаг 2.** Взяв за основу полученные на шаге 1 представления, формируем нижнюю (15) и верхнюю (16) граничные задачи интервальной задачи (4).

**Шаг 3.** С помощью известных методов решения детерминированных задач условной оптимизации получаем решения нижней $\{M_н(x), F_{1\,max}\}$ и верхней $\{M_в(x), F_{2\,max}\}$ граничных задач. В приведенных формулах $M_н(x)$ — множество точек решения $x = (x_1, ..., x_n)$ нижней граничной задачи, где ее целевая функция $F_1$ достигает максимума $\tilde{F}_{1max}$, а $M_в(x)$ — множество точек решения $x = (x_1, ..., x_n)$, соответственно, верхней граничной задачи, в которых ее целевая функция $F_2$ достигает максимума $F_{2\,max}$.

**Шаг 4.** Выбирая в качестве точки решения интервальной задачи (4) любую точку $x^*$ из пересечения множеств $M_н(x)$ и $M_в(x)$ точек решения нижней и верхней граничных задач и беря в качестве нижней границы максимума $\tilde{F}_{max}$ интервальной целевой функции $\tilde{F}$ задачи (4) максимум $\tilde{F}_{1max}$ целевой функции нижней граничной задачи, а в качестве верхней границы максимума целевой функции $\tilde{F}$ задачи (4) максимум $\tilde{F}_{2\,max}$ целевой функции верхней граничной задачи, получаем полное решение интервальной задачи условной оптимизации (4) в виде (17).

## Заключение

В настоящей статье показано, что проблема оптимизации неполностью определенных (недетерминированных) функций достаточно просто разрешима, если указанную неопределенность задавать в интервальной форме и использовать при этом конструктивную теорию сравнения величин интервалов, сводящую это сравнение к сравнению одноименных границ интервалов. Тем самым нахождение оптимума неполностью определенной функции сводится к отысканию одноименного оптимума двух полностью определенных (детерминированных) функций. Наш подход (его естественно назвать де-

терминизацией) примечателен тем, что позволяет свести оптимизацию неполностью определенных функций к хорошо известным и эффективным методам оптимизации полностью определенных функций.


**Список литературы**

1. **Юдин Д. Б., Гольдштейн Е. Г.** Задачи и методы линейного программирования. — М.: Советское радио, 1961. 494 с.
2. **Вентцель Е. С.** Введение в исследование операций. — М.: Советское радио, 1964. 388 с.
3. **Уайлд Д. Дж.** Методы поиска экстремума. — М.: Наука, 1967. 266 с.
4. **Корбут А. А., Финкельштейн Ю. Ю.** Дискретное программирование. — М.: Наука, 1969. 370 с.
5. **Моисеев Н. Н., Иванилов Ю. П., Столярова Е. М.** Методы оптимизации. — М.: Наука, 1978. 352 с.
6. **Левин В. И.** Структурно-логические методы исследования сложных систем. — М.: Наука, 1987. 304 с.
7. **Левин В. И.** Дискретная оптимизация в условиях интервальной неопределенности // Автоматика и телемеханика. 1992. № 7.
8. **Левин В. И.** Булево линейное программирование с интервальными коэффициентами // Автоматика и телемеханика. 1994. № 7.
9. **Левин В. И.** Интервальное дискретное программирование // Кибернетика и системный анализ. 1994. № 6.
10. **Левин В. И.** Оптимизация расписаний в системах с неопределенными временами обработки. I, II // Автоматика и телемеханика. 1995. № 2, 3.
11. **Левин В. И.** Задача трех станков с неопределенными временами обработки // Автоматика и телемеханика. 1996. № 1.
12. **Левин В. И.** Интервальная модель общей задачи линейного программирования. I, II // Вестник Тамбовского университета. Серия: Естественные и технические науки. 1998. Т. 3, № 4; 1999. Т. 4. № 1.
13. **Левин В. И.** Нелинейная оптимизация в условиях интервальной неопределенности // Кибернетика и системный анализ. 1999. № 2.
14. **Левин В. И.** Антагонистические игры с интервальными параметрами // Кибернетика и системный анализ. 1999. № 3.
15. **Левин В. И.** О недетерминистской дискретной оптимизации // Принятие решений в условиях неопределенности: Сб. статей. — Уфа: Изд-во Уфимского авиационного института, 1999.
16. **Левин В. И.** Математическая теория сравнения интервальных величин и ее применение в задачах измерения // Измерительная техника. 1998. № 5.
17. **Левин В. И.** Математическая теория сравнения интервальных величин и ее применение в задачах измерения, контроля и управления // Измерительная техника. 1998. № 9.
18. **Левин В. И.** Интервальная математика и исследование систем в условиях неопределенности. — Пенза: Изд-во Пензенского технологического института, 1998.


---


УДК 519.176



**А. Р. Ураков**, канд. физ.-мат. наук, доц.,
e-mail: urakov@ufanet.ru,
**Т. В. Тимеряев**, магистр,
e-mail: timeryaev@yandex.ru,
Уфимский государственный
авиационный технический университет


# Многоуровневый алгоритм разбиения графов по критерию средней длины


*Рассматривается задача разбиения взвешенного графа на ограниченное число подграфов с минимизацией максимума средних перемещений по подграфам в условиях равной вероятности перемещений между вершинами графа. Для решения задачи предлагается многоуровневый алгоритм разбиения. Проводится сравнение предложенного алгоритма с экспертными разбиениями и другими алгоритмами на графах реальных дорожных сетей.*



***Ключевые слова:*** *разбиение графа, декомпозиция графа, многоуровневый алгоритм*


## Введение

Задача массового обслуживания в некоторых случаях ставится в распределенно-территориальном виде — это означает, что прежде чем выполнить обслуживание заявки, требуется выполнить физическое перемещение заявки к обслуживающему прибору или прибора к заявке. В этом случае обслуживаемая территория, как и в транспортной логистике, моделируется в виде взвешенного графа. Узлам, из которых поступают заявки на обслуживание, как и узлам, в которых находятся обслуживающие приборы, соответствуют вершины графа. Тогда вес ребра представляет собой расход на перемещение (заявки или обслуживающего прибора) между связанными узлами.

Одна из основных задач, из тех, что приходится здесь решать, — это взаимное попарное распределение заявок и приборов таким образом, чтобы уменьшить затраты на перемещение.

Очевидное решение заключается в том, чтобы каждую новую заявку назначать на ближайший к ней прибор. Однако на практике при достаточно больших объемах обслуживания решение оказывается неудачным, так как приводит к большим избыточным перемещениям.

Перемещения можно значительно сократить, если создавать буфер (или очередь) из ожидающих заявок, а дальнейшее распределение проводить через решение транспортной задачи и задачи коммивояжера. Проблема тут заключается в том, что обе указанные подзадачи имеют комбинаторную сложность и при большом числе узлов крайне сложны для решения, тем более, не могут быть решены качественно и оперативно (в режиме реального времени).

Популярный и широко известный способ решения заключается в следующем. Территория разби-

вается на части (секторы), обслуживающие приборы распределяются между секторами, новая заявка может быть обслужена только теми приборами, которые принадлежат тому же сектору, в котором заявка появилась. Этот способ показал высокую эффективность, так как даже при неудачном выборе секторов позволяет значительно сократить перемещения. Кроме того, малое число узлов в секторах позволяет применять для них способ с буфером ожидающих заявок, приведенный выше. Способ настолько очевиден, прост и эффективен, что получил массовое применение. Разбиение территории при этом обычно происходит вручную, наобум, без применения каких-либо математических методов. Соответственно, основной недостаток этого способа заключается в некачественном ручном разбиении, которое приводит к неравномерной нагрузке на приборы и дополнительным перемещениям, которые можно было бы избежать, разбив территорию правильнее.

В данной статье рассматривается один из возможных алгоритмов разбиения графа на части.

## Постановка задачи

Поставим следующее условие: граф должен быть разбит на секторы таким образом, чтобы получить минимум перемещений в каждом из полученных секторов (минимаксный критерий — требуется добиться минимума перемещений в том секторе, перемещения в котором максимальны). Для оценки объема перемещений внутри выбранного сектора предлагается использовать понятие среднего перемещения, под которым будем понимать среднеарифметический размер всех возможных перемещений (матожидание размера случайного перемещения). В этой статье оценка среднего перемещения сделана при следующих условиях:

- появление заявки равновероятно для всех узлов;
- прибор перемещается между заявками;
- обслуживание заявок происходит в порядке их появления.

Число возможных вариантов перемещения $S_z$ из одной вершины в другую в подграфе $G_z$ с учетом перемещения из вершины в себя же равно квадрату числа вершин в подграфе $S_z = |V_z|^2$. Для определения длины каждого перемещения используем матрицу достижимости **M** с элементами $m_{ij}$. Тогда, с учетом равной вероятности перемещений между вершинами, среднее перемещений $A_z$ по подграфу $G_z$ будет определяться выражением

$$A_z = \sum_{i,j \in V_z} m_{ij}/|V_z|^2, \qquad (1)$$

где $m_{ij}$ — элементы матрицы достижимости **M**.

В результате задача получает следующий вид. Дан взвешенный граф $G = (V, E, w)$, $w$: $E \to R$, $1 < k \leqslant k_{\max}$, $k \in N$, требуется найти разбиение множества вершин графа $V_1, V_2, ..., V_k$: $V_i \cap V_j = \varnothing$ $\forall\, i \neq j$, $\bigcup_{i=1}^{k} V_i = V$ с критерием оптимальности $A = \max_{i = \overline{1, k}} A_i \to \min$. В постановке задачи $E$ — множество ребер графа $G$, $R$ — множество вещественных чисел, а $N$ — множество натуральных чисел.

## Решение

**Выбор числа подграфов разбиения.** Задача, поставленная в таком виде, имеет тривиальное, но бессмысленное решение — число секторов $k$ должно быть равно числу узлов (в этом случае $A_z = 0$). На практике число допустимых секторов ограничено сверху некоторой величиной: $k \leqslant k_{\max}$. Отсюда появляется задача выбора числа секторов, на которое следует разбивать граф, т. е. при каком числе $k_o$ секторов может быть получено наименьшее значение целевой функции.

Вводим дополнительное условие: матрица достижимости **M** является симметричной. Тогда число подграфов разбиения $k_o$ можно определить следующим образом. Для подграфа $G_m$, состоящего из одной вершины, величина (1) полагается равной нулю: $A_m = 0$, а приводимые ниже формулы (2), (3) определяют, соответственно, величину среднего перемещения для подграфа из $n$ вершин и величину среднего перемещения при удалении одной из $n$ вершин:

$$A^n = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}}{n^2} = \frac{S_n}{n^2}; \qquad (2)$$

$$A^{n-1} = \frac{S_n - 2d_x}{(n-1)^2}, \qquad (3)$$

где $d_x$ — сумма длин перемещений от удаленной вершины $v_x$ до всех остальных: $d_x = \sum_{i=1}^{n} m_{xi}$, $m_{ij}$ — элементы матрицы достижимости **M**. Так как существует вершина $v_x$ такая, что $d_x \geqslant \frac{S_n}{n}$, то из неравенства $A^n \geqslant A^{n-1}$ следует $2n \geqslant 2n - 1$. То есть при разбиении на большее число подграфов всегда можно выделить вершину с максимальным значением $d_x$ в новый подграф, что не приведет к увеличению значения целевой функции $A$.

Другими словами, в случае с симметричной матрицей достижимости **M** разбиение следует проводить на максимально возможное число подграфов $k_{\max}$. Это же означает, что если разбиение $k_1$ подграфов дает лучшее значение целевой функции $A$, чем разбиение на $k_2$, но при этом $k_1 < k_2$, то раз-

биение на $k_2$ подграфов является не оптимальным и может быть улучшено.

**Многоуровневый алгоритм разбиения.** Для решения задачи используется многоуровневый алгоритм разбиения. Алгоритм состоит из трех основных этапов:

1) *огрубление*. Из исходного графа $G^0 = (V^0, E^0)$ строится последовательность графов меньшей размерности $G^1, G^2, ..., G^m$: $|V^i| > |V^j| \ \forall \ i < j$;

2) *разбиение*. Проводится разбиение на меньшем графе $G^m$ из полученной последовательности;

3) *уточнение*. Разбиение меньшего графа $G^m$ проецируется через все графы полученной последовательности обратно на исходный граф $G^0$.

Использование многоуровневого алгоритма позволяет получать хорошие решения на графах большой размерности за приемлемое время.

**Огрубление графа.** Для сокращения числа итераций алгоритма на этапе проектирования и улучшения разбиения желательно, чтобы огрубленные графы были близки к исходному по интересующим, при данной задаче разбиения, характеристикам. Для достижения этой цели изменяется модель и постановка задачи. Для каждой вершины $v_i$ графа $G$ вводятся два числа. Вес вершины — $a_i$ и степень вершины — $b_i$. Элементы матрицы достижимости $\mathbf{M}^{i+1}$ меньшего графа $G^{i+1}$ полагаются равными $m_{kj}^{i+1} = \sum\limits_{h, l} m_{hl}^{i}$, $h \in V^{i,k}$, $l \in V^{i,j}$. Здесь $V^{i,j}$ — множество вершин графа $G^i$, объединенных в вершину $v_j$ графа $G^{i+1}$. Характеристики вершин нового меньшего графа $G^{i+1}$ вычисляются по формулам $a_j^{i+1} = \sum\limits_{h, l} m_{hl}^{i} + a_h^i + a_l^i$, $h, l \in V^{i,j}$ и $b_j^{i+1} = \sum\limits_{k \in V^{i,j}} b_k^i$. Формула расчета средних перемещений по подграфу (1) изменяется следующим образом:

$$A_z = \frac{\sum\limits_{i,j \in V_z} m_{ij} + \sum\limits_{i \in V_z} a_i}{\left(\sum\limits_{i \in V_z} b_i\right)^2}. \qquad (4)$$

Для исходного графа $G^0$ величины $a_i$ и $b_i$ полагаются равными $a_i = 0$, $b_i = 1$, $\forall v_i \in V^0$. При такой постановке величины (4) совпадают с величинами (1) из начальной постановки задачи.

Для получения графа меньшей размерности на каждом шаге ищется максимальное паросочетание и пары смежных вершин объединяются. Ребра паросочетания определяются последовательно. Каждый раз к паросочетанию добавляется ребро, пара вершин которого обладает минимумом величины $\frac{m_{ij} + m_{ji} + a_i + a_j}{(b_i + b_j)^2}$. Вершины, не вошедшие в паросочетание, просто переносятся в меньший граф без объединения.

**Разбиение графа.** Этап разбиения графа наименьшей размерности состоит их двух основных стадий: первоначального разбиения и улучшения разбиения.

Для получения первоначального разбиения используется метод роста регионов с неслучайным выбором начальных вершин. Выбор начальных $k$ вершин осуществляется по максимальной удаленности от уже выбранных $l$ вершин, определяемой по формуле $\max\limits_{i \notin V_i, ..., V_l} \left( \max\limits_{j \in V_i, ..., V_l} (m_{ij}) \right)$. Остальные $n - k$ вершин приращиваются к подграфам, образованным начальными вершинами. Вершина приращивается к подграфу, добавление которой к нему даст наименьшее увеличение целевой функции. Для этого для каждой не присоединенной вершины рассчитывается выигрыш

$$g_i = \max_{j = \overline{1, k}} (A_j) - \max_{j = \overline{1, k}} (A_j^i)$$

для подграфов $V_j$, таких, что добавление вершины $v_i$ не нарушит связность. $A_j^i$ здесь — значение (4) для подграфа $G_j$ после добавления к нему вершины $v_i$. Вершина с максимальным выигрышем $g_i$ приращивается к подграфу, на котором этот максимум достигается. Процесс продолжается до распределения всех вершин по подграфам.

Для улучшения разбиения используется модификация FM-алгоритма [1]. Одна итерация модифицированного алгоритма для бисекции выглядит следующим образом. Для всех вершин вычисляются выигрыши при перемещении в другой подграф:

$$g_i = \max(A_1, A_2) - \max(A_1', A_2'),$$

где $A_j'$ — значение (4) для $G_j$ после перемещения вершины $v_i$. Вершина $v_e$ с максимальным выигрышем $g_e = \max\limits_{j = \overline{1, k}} (g_i)$ перемещается, для не перемещенных вершин выполняется пересчет выигрышей. После перемещения всех вершин определяется последовательность перемещений $s_1, s_2, ..., s_k$, при которой достигается минимум величины $A' = \max(A_1', A_2')$. Если этот минимум меньше значения целевой функции до начала итерации $A' \leqslant A^0$, то разбиение с последовательностью перемещений $s_1, s_2, ..., s_k$ принимается на следующей итерации за исходное. В противном случае достигнут локальный минимум и дальнейшее улучшение невозможно.

**Уточнение разбиения.** На каждом шаге этого этапа выполняется проецирование разбиения более грубого графа $G^m$ на более точный $G^{m-1}$. Проециро-

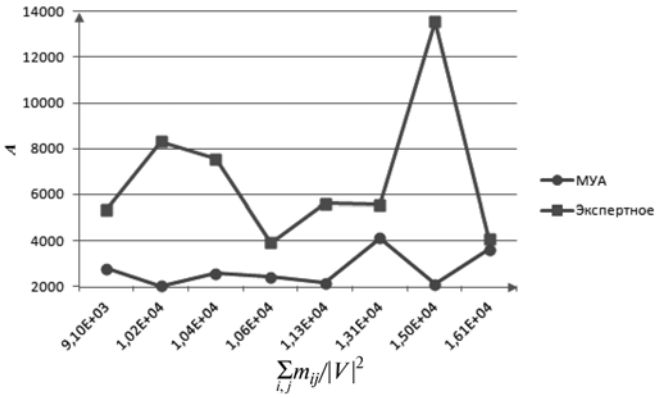



**Рис. 1. Сравнение экспертных разбиений с разбиениями разработанным многоуровневым алгоритмом (МУА)**

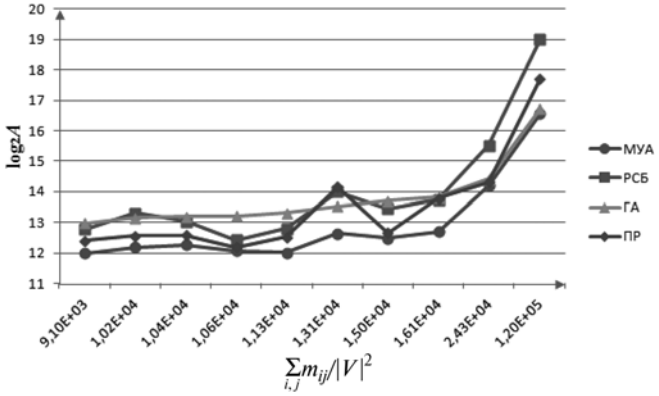



**Рис. 2. Результаты разбиения на 8 подграфов**

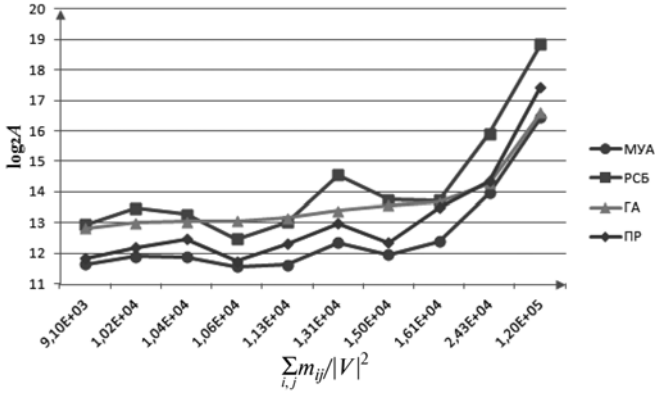



**Рис. 3. Результаты разбиения на 16 подграфов**

вание проводится путем назначения вершинам $v_j$: $j \in V^{i,k}$ графа $G^i$, образующим вершину $v_k$ графа $G^{i+1}$, подграфа вершины $v_k$, $P_i(j) = P_{i+1}(k)$. После каждого проецирования выполняется процедура улучшение разбиения, описанная выше.

**Улучшение разбиения на *k* подграфов.** При улучшении разбиения на $k$ подграфов используется техника бисекции, чтобы сделать разбиение попарно оптимальным. Пара подграфов выбирается из множества пар подграфов, в котором хотя бы один из пары подграфов был изменен с момента последнего выбора. Каждый раз выбирается пара, в которой первый подграф имеет наибольшее значение (4), а второй соединен с первым и обладает минимальным значением (4).

## Результаты

Эффективность разработанного многоуровневого алгоритма (МУА) была оценена сравнением с экспертными разбиениями, проведенными в ходе эксплуатации на графах реальных дорожных сетей. На рис. 1 изображена зависимость $A$ от $\sum_{i,j} m_{ij}/|V|^2$ для этого сравнения. Разбиения разработанным алгоритмом обладают значением целевой функции $A$, меньшим (от 11 до 545 %), чем экспертные разбиения.

Разработанный многоуровневый алгоритм (МУА) также был сравнен с алгоритмом рекурсивной спектральной бисекции (РСБ), генетическим алгоритмом (ГА) и алгоритмом приращения регионов (ПР). Зависимость $\log_2(A)$ от $\sum_{i,j} m_{ij}/|V|^2$ при разбиении на 8 и 16 подграфов представлена на рис. 2 и 3 соответственно. Разбиения, полученные разработанным алгоритмом, обладают в среднем на 30 % меньшим значением целевой функции, чем лучший из сравниваемых алгоритмов.

### Список литературы

1. **Fiduccia C. M., Mattheyses R. M.** A Linear Time Heuristic for Improving Network Partitions // 19th Design Automation Conference, IEEE Press Piscataway, NJ, USA. 1982. P. 175—181.

2. **Karypis G., Kumar V.** A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs // SIAM Journal on Scientific Computing. 1999. Vol. 20 (1). P. 359—392.

3. **Kernighan B. W., Lin S.** An Efficient Heuristic Procedure for Partitioning Graphs // The Bell System Technical Journal. 1970. Vol. 49 (1). P. 291—307.

УДК 004.7

**С. В. Минухин**, канд. техн. наук, проф.,
e-mail: ms_vl@mail.ru,
**С. В. Знахур**, канд. экон. наук, доц.,
e-mail: sergznakhur@mail.ru,
Харьковский национальный
экономический университет

# Оптимизация энергопотребления вычислительных ресурсов двухуровневого Grid на основе балансировки их загрузки

*Рассмотрен подход к оптимизации энергопотребления узлами вычислительного кластера в условиях равномерной загрузки (балансировки) кластеров метапланировщиком двухуровневой Grid-системы. Приведены результаты моделирования в пакете GridSim и примеры, иллюстрирующие и обосновывающие возможность оптимизации энергопотребления в случае изменения интенсивности потока заданий.*

***Ключевые слова:*** *Grid-система, кластер, метапланировщик, моделирование, оптимизация, узел, энергопотребление*

## Введение

Развитие современных Grid-систем предполагает не только увеличение количества и производительности вычислительных ресурсов (кластеров), но и их рациональное использование в соответствии с динамикой потока и характеристик заданий пользователей. В настоящее время проблема рационального использования ресурсов является актуальной, поскольку ресурсы кластеров используются менее чем на 30 %, что обусловлено инертностью перехода пользователей от традиционного программирования к использованию сред и языков для программирования потоков и параллельного выполнения заданий. Отметим, что увеличение коэффициента использования кластеров (ресурсов) не может решить проблему их рационального использования, поскольку встроенные в Grid брокеры планирования загрузки ресурсов используют, как правило, алгоритм FCFS [1], который, в первую очередь, загружает первые ресурсы в их общем списке (рис. 1) (по оси абсцисс приведены номера ресурсов с их производительностью). Это приводит к тому, что при увеличении общего коэффициента использования ресурсов отдельных кластеров первые ресурсы загружаются до 90...100 %, а остальные ресурсы имеют низкий коэффициент загрузки (в случае существенной гетерогенности ресурсов Grid коэффициенты использования первых ресурсов могут колебаться в пределах от 60 до 100 %). С точки зрения экономической эффективности для владельца ресурса, как для одного из участников функционирования Grid-среды, реализация данного подхода представляется достаточно затратной, поскольку бóльшую часть ресурсов (если не все) необходимо поддерживать в актуальном и рабочем состоянии, а именно, нести текущие эксплуатационные затраты и, в первую очередь, затраты на энергопотребление. Действительно, существующие технологии динамического управления энергопотреблением современных процессоров позволяют снизить их энергопо-

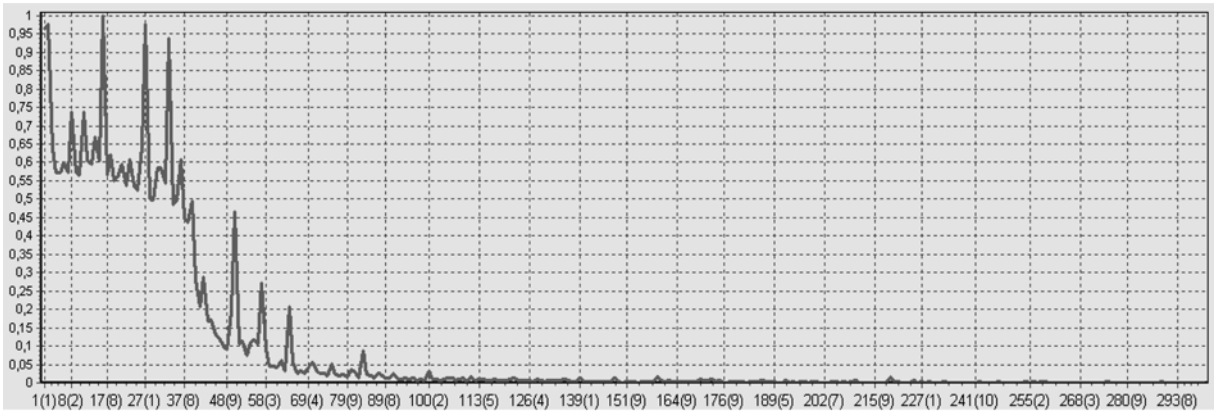



**Рис. 1. Коэффициент использования гетерогенных ресурсов при потоке неоднородных заданий в случае использования метода FCFS**

требление (если это поддерживается операционной системой) за счет уменьшения напряжения и частоты, но, как правило, выигрыш при этом составляет не более 50 % от штатного режима работы процессора и, что особо важно, устройство автоматически включается в штатный режим, если на него было направлено задание [3].

Поэтому одним из важнейших условий реализации режимов энергосбережения является обеспечение равномерной, сбалансированной загрузки кластеров, эффективные методы реализации которой предложены в работе [1]. Целью данной статьи является разработка и исследование подхода к оптимизации энергопотребления вычислительного кластера Grid, в основе которого лежит управление процессами включения и отключения его вычислительных ресурсов при обеспечении высокой загрузки и балансировки загрузки активных (оставшихся) ресурсов в условиях изменения интенсивности внешнего по отношению к вычислительному кластеру потока заданий [1].

В статье используется понятие Grid с неотчуждаемыми ресурсами, причем интенсивность внутреннего потока заданий предполагается не влияющей на решение заданий внешнего потока.

В данной работе решаются следующие задачи:

- разработка модели оптимизации энергопотребления кластера в зависимости от величины поступающего потока заданий и коэффициента его использования;
- моделирование работы Grid в условиях изменения количества доступных ресурсов и определения влияния их количества на средний коэффициент загрузки кластера при стационарном потоке внешних заданий;
- разработка эффективной политики снижения энергопотребления при эксплуатации кластера.

## Оптимизация энергопотребления вычислительного кластера

Предлагаемым в данной работе направлением оптимизации работы вычислительного кластера является уменьшение (оптимизация) энергопотребления как основной статьи затрат текущих эксплуатационных расходов вычислительного кластера [2]. Оптимизация энергопотребления предполагает достижение баланса между производительностью и функциональностью системы и ее энергопотреблением.

Оптимизацию энергопотребления кластера (процессоров) можно реализовать, используя возможности отключения/включения ресурсов в условиях изменения интенсивности потока поступающих заданий. При этом изменения загрузки ресурса определяют изменения его энергопотребления, что приводит к необходимости выбора такого режима, при котором в условиях соблюдения штатной (нормативной) производительности ресурса его энергопотребление было бы минимальным.

Экономический смысл оптимизации планирования загрузки кластера заключается в том, что метапланировщик Grid-системы должен обеспечивать максимальную загрузку ресурсов при выполнении ограничений на время выполнения заданий в условиях изменения их интенсивности. Это обеспечит снижение эксплуатационных затрат: например, снижение затрат на энергопотребление и обслуживание отключенных ресурсов при одновременном увеличении загрузки ресурсов, оставшихся неотключенными, находится в пределах от штатного уровня до возможно максимального.

Рассмотрим два основных сценария управления работой ресурсов в целях оптимизации их энергопотребления.

*Сценарий 1.* Отключение ресурсов в случае снижения интенсивности потока заданий (коэффициент использования ресурсов также снижается) — минимизация энергопотребления.

*Сценарий 2.* Включение ресурсов в случае увеличения интенсивности потока заданий при условии максимальной загрузки уже работающих ресурсов. В данном случае работающие ресурсы имеют коэффициент использования на уровне 90...100 %, и дальнейшее увеличение интенсивности потока заданий приводит к тому, что время решения задания увеличивается за счет простоя задания в очереди к ресурсу.

Для первого и второго сценариев важным является "упаковка" каждого из ресурсов до рекомендуемого значения коэффициента использования ресурса.

В качестве целевой функции предлагается использовать следующую:

$$\sum_{j=1}^{n} TaP_j \to \min,$$

где $P_j$ — потребляемая мощность $j$-го вычислительного ресурса; $n$ — количество ресурсов; $a$ — параметр, определяющий режим включения (1) или отключения (0) ресурса; $T$ — период загрузки ресурса, который определяется как максимальное время решения $i$-го задания очереди для ресурса $j$:

$$\max_{i \in \overline{(1, m)}}(t_{ij}) < T,$$

где $t_{ij}$ — директивный срок решения $i$-го задания на $j$-м ресурсе.

Как показали результаты моделирования работы Grid, полученные в [1], выбор метода (стратегии) планирования ресурсов существенно влияет на эффективность (равномерность) загрузки вычислительных кластеров в Grid. В большинстве работ эффективность загрузки вычислительных кластеров рассматривается на основе следующих метрик [1, 5]:

$K_{\text{испj}}$ — коэффициент использования (загрузки) $j$-го ресурса;

$T_{\text{простоя}}$ — суммарное время простоя $M$ заданий очереди при пакетной обработке.

При высокой интенсивности заданий значение $K_{\text{исп}j}$ ресурса стремится к 1 (100 %), однако, в зависимости от выбранного метода планирования существенно различаются значения $T_{\text{простоя}}$ и $T_{\text{решения}}$. Таким образом, выбор метода планирования в условиях изменяющейся интенсивности может обеспечить равномерную загрузку и оптимизировать время решения всех заданий очереди в заданный срок и минимизировать штрафы за их возможное невыполнение.

Для расчета энергопотребления, как правило, используют упрощенную модель вида [6—8]:

$$P_j = (P_{\max} - P_{\min})K_{\text{исп}j} + P_{\min}, \qquad (1)$$

где $P_j$ — потребляемая мощность вычислительного ресурса; $P_{\max}$, $P_{\min}$ — максимальное и минимальное значения энергопотребления вычислительного ресурса соответственно.

Данная модель не учитывает технологических и физических особенностей работы вычислительных устройств — процессоров. Для решения задачи оптимизации энергопотребления кластера предлагается использовать следующую зависимость между энергопотреблением и частотой (производительностью) процессора [6—8], но уже применительно к коэффициенту использования ресурса:

$$P_j \approx V_j^2 f_j + P_{\min} \approx f_j^3 + P_{\min}, \qquad (2)$$

$$f_j = f_{\max} K_{\text{исп}j}, \qquad (3)$$

где $V_j$ — регулируемое напряжение процессора $j$; $f_j$ — регулируемая тактовая частота процессора $j$; $f_{\max}$ — максимальная тактовая частота процессора $j$.

Отметим, что для каждого процессора существует некоторый оптимальный рабочий (штатный) частотный режим, при котором обеспечивается оптимальное соотношение его производительности и энергопотребления.

Согласно формуле (2), процессор, работающий на частоте 800 МГц, потребляет в 16 раз меньше электроэнергии, чем процессор, работающий на штатной частоте 3200 МГц. Следовательно, при наличии динамики изменения интенсивности входного потока заданий планирования загрузки ресурса на период $T$ при условии управления тактовой частотой процессоров динамически регулируемая частота даст эффект экономии энергопотребления при прочих равных условиях, который определяется выражением

$$\textit{Эффект\_ энергопотребления\_за\_период\_}T = \sum_{j=1}^{N} T(f_{\max} - f_j)^3.$$

Данный расчет справедлив при условии, что установленная частота $f_j$ определяется штатной загрузкой процессора и обеспечивает требуемый директивный срок выполнения заданий.

Рассмотрим подробнее оптимизацию энергопотребления кластера на основе балансировки загрузки ресурсов на примере планирования загрузки ресурсов 1...5, загрузка которых представлена в виде пронумерованных гистограмм (рис. 2). Под ресурсами в данном случае будем понимать вычислительные узлы самого кластера. В данном примере рассмотрим управление отключением/включением ресурсов при изменении потока заданий на вычислительные ресурсы гомогенного кластера.

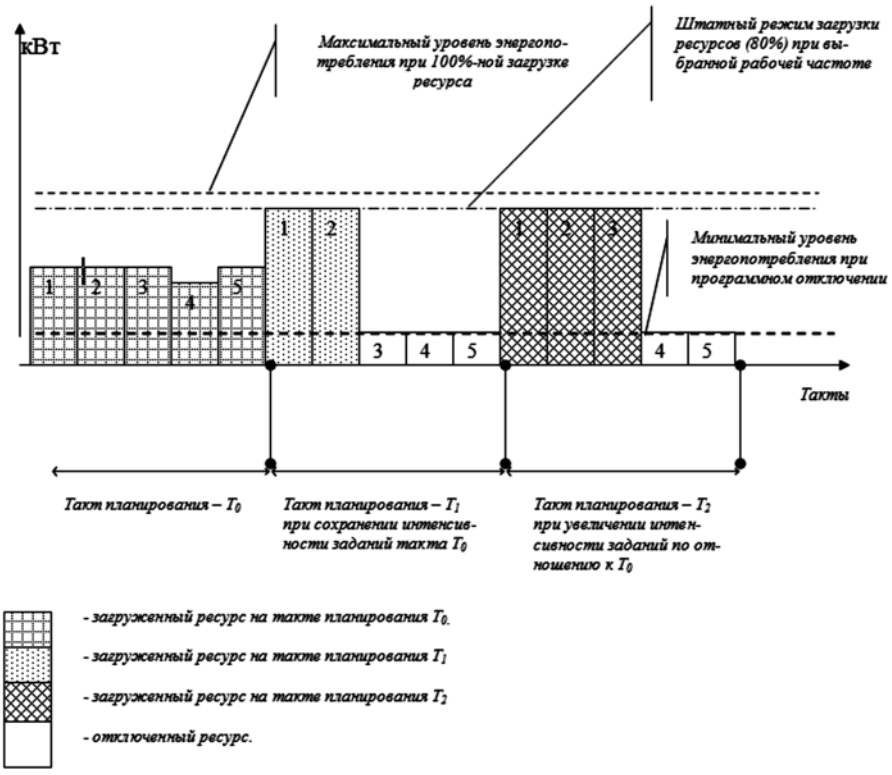



**Рис. 2. Результаты балансировки загрузки при отключении процессоров кластера**

На рис. 2 приведены результаты планирования загрузки ресурсов на каждом такте планирования при изменении интенсивности потока заданий для отключения/включения ресурсов. Предполагается, что время решения задач на ресурсах находится в пределах директивного. В данном примере на такте планирования $T_1$ принимается решение об отключении ресурсов на основе мониторинга их загрузки после такта планирования $T_0$ и текущей интенсивности потока заданий (т. е. плановая загрузка ресурсов на такте планирования $T_1$

в режиме без отключения может находиться в интервале от 1 до 79 %). Верхний предел 79...80 % определен опытным путем на основе анализа результатов максимальной загрузки ресурсов при соблюдении условия решения заданий в директивное время [1].

Максимальное количество отключаемых ресурсов определяется при выполнении следующих условий (ограничений): загрузка оставшихся ресурсов не должна превышать 90 % (10 % выделяются для ОС, сервисов и резерва для случая непланового отключения других ресурсов), при этом общее время решения заданий очереди на кластере не должно увеличиваться по сравнению с режимом, когда все ресурсы включены.

Таким образом, на такте планирования $T_1$ были определены ресурсы 1 и 2, которые максимально загружены при той же интенсивности потока заданий, что и на такте $T_0$, и, соответственно, ресурсы 3...5 могут отключаться на время решения текущей очереди заданий для периода планирования $T_1$. В случае увеличения интенсивности потока заданий на момент времени планирования $T_2$ для обеспечения директивного срока выполнения заданий возможно включение ресурсов.

Так, на рис. 2 видно, что для разрешения очереди заданий, размер которой больше, чем на такте $T_1$, на такте $T_2$ включается ресурс 3 при условии, что 1-й и 2-й ресурсы имеют максимальную загрузку.

### Пример расчета экономического эффекта от оптимизации энергопотребления кластера

Выигрыш экономии энергопотребления в зависимости от выбора метода планирования и равномерности загрузки ресурсов проиллюстрируем на следующем примере. Рассмотрим одну и ту же очередь заданий $\lambda$ для тактов $T_0$ и $T_1$. Сумма затрат на электропотребление для такта $T_0$ будет всегда больше, чем для $T_1$, при прочих равных условиях (времени разрешения заданий очереди), поскольку все ресурсы 1...5 работают в штатном режиме энергопотребления за период $T$, а в случае такта $T_1$ работают только ресурсы 1 и 2.

В случае отсутствия режима изменения напряжения процессоров расчет для примера на рис. 2 имеет следующий вид: для такта $T_0$ при штатной мощности энергопотребления узла 0,8 кВт общее энергопотребление равно $0{,}8 \cdot 5 = 4$ кВт, для такта $T_1$ энергопотребление составляет $0{,}8 \cdot 2 = 1{,}6$ кВт. Таким образом, выигрыш составляет 2,4 кВт/ч. Эффект достигается также и при динамическом изменении напряжения процессоров в зависимости от загрузки ресурсов. Например, энергопотребление ресурсов 1...5 для такта $T_0$ рассчитаем при средней загрузке ресурсов 18 %, штатной мощности 0,8 кВт и минимальной 0,3 кВт следующим образом: $(0{,}8 \text{ кВт} \cdot 0{,}18 + 0{,}3 \text{ кВт}) \cdot 5 = 2{,}24$ кВт, а для такта $T_1$ энергопотребление составляет $(0{,}8 \text{ кВт} \cdot 0{,}9\ \% + 0{,}3 \text{ кВт}) \cdot 2 = 2{,}04$ кВт. Таким образом, выигрыш по энергопотреблению составил 0,2 кВт/ч.

Рассмотренный механизм оптимизации загрузки процессоров вычислительных узлов возможно реализовать средствами операционной системы кластера и технологиями, реализованными на уровне инструкций процессоров [3].

Для сценария 2, т. е. для случая увеличения интенсивности потока заданий, предложенный метод реализуется в обратной последовательности: по мере увеличения нагрузки на кластер его узлы начинают последовательно включаться с учетом динамики увеличения загрузки — все используемые ресурсы загружаются по "жадному" алгоритму: полностью загружаются до уровня 95 % уже частично загруженные, а потом начинают последовательно загружаться остальные незадействованные узлы кластера.

### Анализ результатов экспериментальных исследований

Исследование зависимости коэффициента использования (загрузки) ресурсов от их количества было проведено в работе [1] с использованием программного продукта GridSim, предназначенного для имитационного и полунатурного моделирования работы Grid [10]. В эксперименте были использованы следующие параметры:

- кластеры Grid — гомогенные, число кластеров изменялось от 2 до 15 с шагом 1;
- число процессоров в каждом кластере — 10;
- размер общей очереди заданий — 2000 (использовались реальные данные [11]);
- методы планирования ресурсов представляют собой модификации метода FCFS: FCFS_RING, FCFS_RANDOM, FCFS_RoundPobin (как показано в работе [1], использование FCFS в чистом виде не позволяет достичь балансировки загрузки ресурсов, поскольку в первую очередь загружаются первые свободные ресурсы и с увеличением количества используемых ресурсов и числа процессорных элементов эта ситуация ухудшается).

В качестве экспериментальных данных выполняемых заданий использовалась реальная статистическая выборка данных работы сегмента Grid [11].

В данном эксперименте для случая низкой интенсивности, который моделируется реальным потоком заданий в Grid (число заданий по отношению к количеству ресурсов за выбранный период), было использовано всего 15 кластеров по 10 процессоров в каждом. Результаты эксперимента показали, что предложенные методы планирования обеспечивают близкие характеристики загрузки заданий (балансировку) на ресурсы (отклонения среднего значения коэффициента использования — межкластерной балансировки — для алгоритмов составляют менее 1 %) и их можно использовать в сценариях отключения/включения. На рис. 3...5

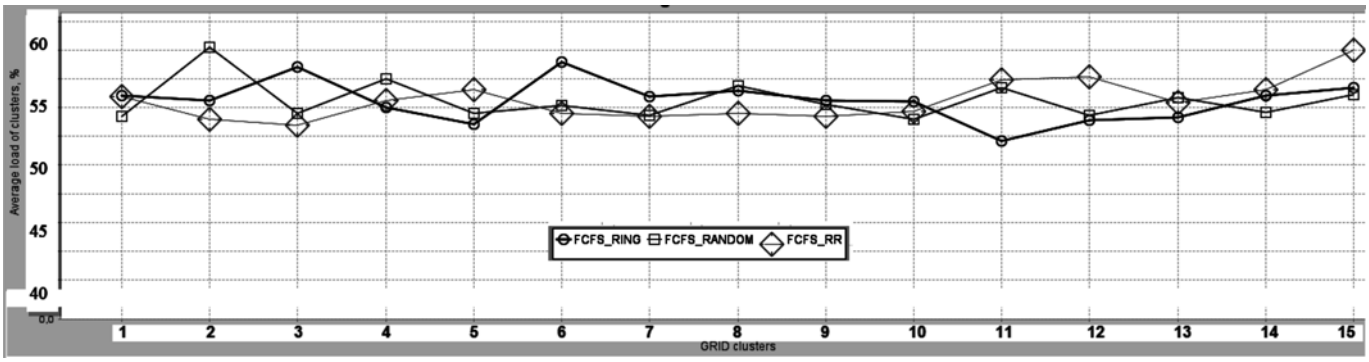



**Рис. 3. Значения среднего коэффициента использования ресурса для 15 кластеров**

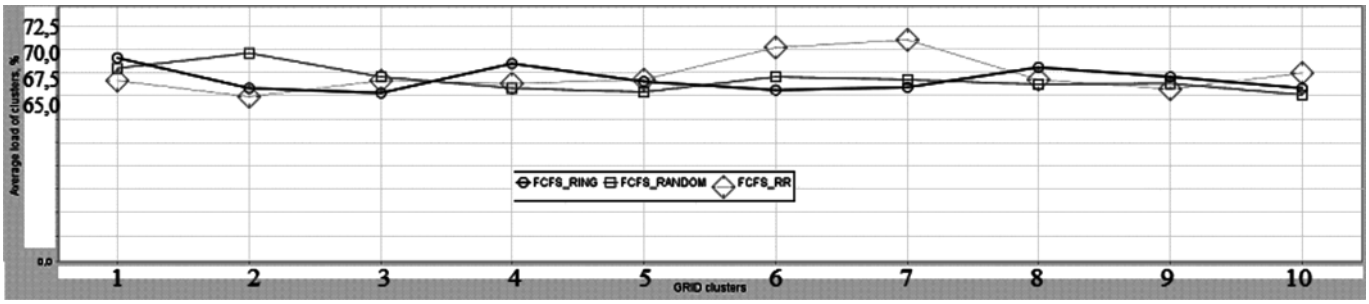



**Рис. 4. Значения среднего коэффициента использования ресурса для 10 кластеров**

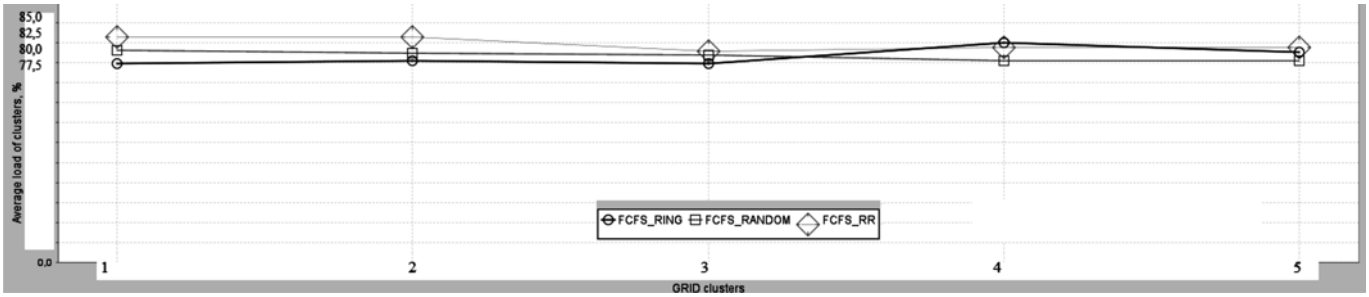



**Рис. 5. Значения среднего коэффициента использования ресурса для 5 кластеров**

приведены результаты моделирования соответственно следующих сценариев:

- ресурсы не отключаются;
- отключается 33 % ресурсов (используется 10 кластеров);
- отключается 66 % ресурсов (используется 5 кластеров).

Таким образом, при условии постоянного времени решения всех заданий очереди отключение некоторых ресурсов (кластеров) позволяет пропорционально увеличивать средний коэффициент их использования.

Для определения характера зависимости между средним коэффициентом загрузки (использования) кластера и числом используемых кластеров для моделируемых сценариев отключения/включения ресурсов в пакете GridSim была получена выборка средних значений коэффициентов загрузки кластеров при пошаговом их отключении в интервале от 15 до 2 при условии постоянного числа процессоров в одном кластере (10) и величины входного потока заданий. Проведенный статистический анализ коэффициента загрузки показал наличие закономерностей, которые для всех используемых методов планирования описываются полиномами второго порядка с коэффициентами детерминации не ниже 0,99 (рис. 6).

Экономический эффект, получаемый после оптимизации загрузки вычислительных ресурсов кластера, рассчитаем по следующей формуле:

$$ЭЭ = \sum_{i}^{t}\sum_{j}^{n} N_{ij} P_j T_{ij} C_i,$$

где $T_{ij}$ — период времени использования $j$-го ресурса в штатном режиме для интервала периода

планирования $i$; $\sum_{j}^{n} N_{ij}$ — количество используемых ресурсов за период времени $i$; $P_j$ — энергопотребление ресурса $j$-го в единицу времени (мощность); $C_i$ — стоимость 1 кВт/ч для периода времени $i$.

В качестве примера проведем расчеты для 100 узлов кластера — серверов Intel Xeon E5540, с двумя процессорами Intel 5-й серии по 4 ядра со штатным потреблением электроэнергии 80 Вт/ч.

Максимальная экономия за год при условии отключения 50 % узлов может составить:

$$0{,}5 \cdot 100 \text{ узлов} \cdot 0{,}8 \text{ кВт} \cdot 24 \text{ ч} \cdot 365 \text{ дней} \times \\ \times 4 \text{ руб./кВт/ч} = 1\,401\,600 \text{ руб.}$$

Таким образом, для приведенного примера даже в случае минимального уменьшения энергопотребления в размере 5 % от штатного для всех ресурсов (узлов) кластера экономия денежных средств может составить 140 160 руб., что является весьма существенным с точки зрения экономии.

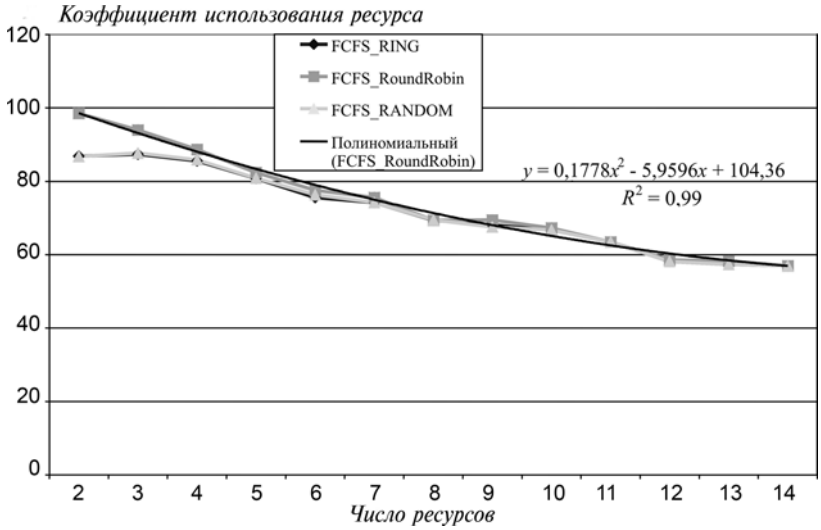



**Рис. 6. Зависимость коэффициента использования ресурса от количества доступных кластеров при использовании планировщиков FCFS_RING, FCFS_RANDOM, FCFS_RoundRobin**

## Выводы

В статье предложены направления оптимизации экономических результатов владельцев ресурсов при условии равномерной загрузки их ресурсов в двухуровневой Grid-системе. Равномерная загрузка кластеров обеспечивает "справедливое" распределение заданий, что дает возможность максимально эффективно использовать все кластеры Grid и является универсальным средством для всех ресурсов Grid. Для снижения эксплуатационных затрат на содержание вычислительного кластера предложен подход к динамическому уменьшению (изменению) энергопотребления вычислительного ресурса (узла) на основе использования методов планирования, обеспечивающих эффективную балансировку загрузки ресурсов кластера.

Получены результаты моделирования загрузки ресурсов вычислительных кластеров (для сценария 1) в условиях их поэтапного отключения, которые показали, что при низкой интенсивности заданий в Grid они могут быть решены в директивный срок на значительно меньшем количестве ресурсов Grid (до 50 % от их общего числа), причем их загрузку можно обеспечить на штатном уровне — в диапазоне от 80 до 90 % максимальной загрузки. При этом остальные ресурсы (кластеры) Grid на время решения заданий очереди можно отключать или переводить в режим минимального энергопотребления. В случае увеличения интенсивности потока заданий (сценарий 2) необходимо подключать дополнительные ресурсы при условии, что уже работающие загружены (имеют коэффициент использования) до уровня 90 %.

### Список литературы


1. **Минухин С. В., Коровин А. В.** Моделирование планирования ресурсов GRID средствами пакета GridSim // Системи обробки інформації. Інформаційні технології та комп'ютерна інженерія. 2011. Вип. 3 (93). С. 62—68.
2. **Минухин С. В., Знахур С. В.** Методика выбора и расчета затрат совокупной стоимости владения вычислительным кластером // Радіоелектронні і комп'ютерні системи. 2011. № 1 (49). С. 90—96.
3. **Richling J., Werner M., Mühl G.** Event-Driven Processor Power Management // Proceedings of the 1st International Conference on Energy-Efficient Computing and Networking, 2010. URL: http://www.e-energy.uni-passau.de/fileadmin/user_upload/presentation_final/Session2/ 04_Jan_Schoenherr_EventDrivenProcessorPowerManagement.pdf
4. **Rusu C., Ferreira A., Scordino C., Watson A.** Energy efficient real-time heterogeneous server clusters // Proceedings of the Real-Time and Embedded Technology and Applications Symposium, 2006. P. 418—428.
5. **Сериков Д. А.** Математическое и программное обеспечение эффективного разделения ресурсов для решения задач в распределенной вычислительной среде. URL: http://seminar.s2s.msu.ru/files/20090324_Сериков.pdf
6. **Lee Y. C., Zomaya A. Y.** Energy efficient utilization of resources in cloud computing systems. URL: http://www.lifl.fr/~derbel/cgc/papers/energytask.pdf
7. **Liu Y., Zhu H. A.** Survey of the Research on Power Management Techniques for High Performance Systems. URL: http://cms.brookes.ac.uk/staff/HongZhu/Publications/Power_Mgt-final.pdf.
8. **Hays R.** Active/Idle Toggling with Low-Power Idle. In IEEE 802.3az Task Force Group Meeting. 2008. URL: http://www.ieee802.org/3/az/public/jan08/hays 01 0108.pdf
9. **Nedevschi S.** Reducing Network Energy Consumption via Sleeping and Rate-Adaptation. URL: http://berkeley.intel-research.net/sylvia/power_nsdi08.pdf
10. **Buyya R., Murshed M.** Gridsim: a toolkit for the modeling and simulation of distributed resource management and scheduling for grid computing // Concurrency and computation: practice and experience. 2002. Vol. 14. P. 1175—1220.
11. **The Grid** Workloads Archive. URL: http://gwa.ewi.tudelft.nl/pmwiki/pmwiki.php?n = Workloads.Gwa-t-4

УДК 004.7

**А. Э. Саак,** канд. техн. наук, доц.,
Технологический институт
Южного федерального университета,
г. Таганрог,
e-mail: saak@tti.sfedu.ru

# Диспетчеризация в Grid-системах на основе однородной квадратичной типизации массивов заявок пользователей

*Протяженные массивы линейных полиэдралей координатных ресурсных прямоугольников, представляющих заявки пользователей на компьютерное обслуживание в Grid-системах и многопроцессорных вычислительных системах, требуют локализации в ресурсную рамку по правилам ориентации, аддитивности, целостности. Согласованность с параметрами рамки осуществляется кольцевой локализацией, упорядочивающей линейную полиэдраль большой протяженности в планарный полиэдр кольцевой структуры с последующим перераспределением по возможно минимальному числу ресурсных рамок. Кольцевая локализация строится с учетом кругового, гиперболического, параболического квадратичного типа первоначального массива ресурсных прямоугольников.*

*Более полная информация о свойствах однородности, монотонности и некоторых других параметрах в пределах предыдущих квадратичных типов массива индуцирует углубленную классификацию линейных полиэдралей и приводит к акселерации алгоритмов кольцевой локализации, предложенных в предыдущих авторских публикациях настоящего журнала. Указанным результатам посвящена предлагаемая статья.*

***Ключевые слова:*** *Grid-система, многопроцессорная вычислительная система, диспетчирование, однородный квадратичный тип массива требований пользователей, линейные полиэдрали со свойством монотонности, целевые критерии симметризации ресурсной оболочки и ресурсной меры оболочки заявок пользователей*

## Введение

В работах [1—3] исследовалось диспетчирование множественного компьютерного обслуживания в Grid-системах [4, 5], многопроцессорных вычислительных системах (МВС) [6—8], базирующееся на условно-аналитических алгоритмах. Введен и изучался класс массивов триодной горизонтально-вертикальной структуры с факторизацией массива на подмножества трех квадратичных типизаций.

Сопоставление каждого квадратичного типа и адекватного упорядочивания первоначально протяженного массива спроса на обслуживание в ресурсную рамку $M \times M$ операционного поля МВС приводит к акселерированному диспетчированию множества заявок пользователей, требующему оценки качества по сравнению с оптимизацией наилучшего выбора.

В настоящей статье понятия кругового, гиперболического, параболического типов линейных и триодных полиэдралей координатных ресурсных прямоугольников — заявок пользователей — углубляются свойством однородности квадратичной типизации, характеризующим взаимосвязь транспонированно-симметричных элементов, подобную связи элементов арифметической прогрессии, впервые замеченную Гауссом. Замена локальных массивов координатных прямоугольников ресурсными оболочками приводит к необходимости выделения класса линейных полиэдралей со свойством монотонности, воспроизводящего структуру исходного массива по отношению к массиву локальных оболочек. Найденные свойства используются для рамочного упорядочивания соответствующих протяженных линейных полиэдралей согласно однородной или монотонной типизации.

## 1. Классификация однородных полиэдралей

Рассмотрим массивы линейных полиэдралей координатных ресурсных прямоугольников (со сторонами $a$ и $b$), квадратичный тип которых обладает свойством однородности, определяемым ниже.

Пусть грани линейной полиэдрали имеют горизонтальную форму

$$\bigcup_{j_1=1}^{k} [(a(j_1),\ b(j_1))],\ a(j_1) \geqslant b(j_1),$$

медленно убывающие высоты

$$b(j_1)\downarrow,\ j_1\uparrow$$

и парную транспонированную униаддитивность измерений 2-го рода

$$b(1) + b(k) = b(2) + b(k-1) = \ldots = b(m) + b(m+1) = \text{const},\ k = 2m.$$

По совокупности указанных свойств линейная полиэдраль относится к однородно-круговому квадратичному типу.

Пусть имеется линейная полиэдраль последовательно несравнимых координатных ресурсных прямоугольников

$$\Delta a(j_1)\Delta b(j_1) < 0$$

с медленным убыванием высот и ростом оснований $b(j_1)\downarrow$, $a(j_1)\uparrow$, $j_1\uparrow$.

Выделением центрального элемента $j_{1,*}$ с минимальной асимметрией измерений

$$|b(j_{1,*}) - a(j_{1,*})| \leqslant |b(j_1) - a(j_1)|$$

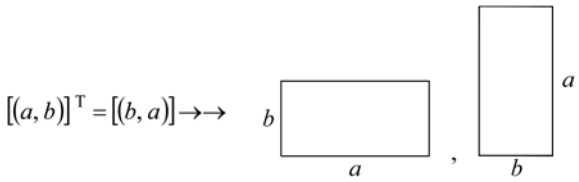



**Рис. 1. Транспонированная пара ресурсных прямоугольников**

заданной линейной полиэдрали разбиваем последнюю на подмножества граней, левее и правее центрального элемента

$$\bigcup_{j_1=1}^{j_{1,*}-1}[(a(j_1),\ b(j_1))],\ j_1 = \bigcup_{j_{1,*}+1}^{k}[(a(j_1),\ b(j_1))].$$

В случае транспонированности элементов (рис. 1) симметричного расположения относительно центра

$$[(a(j_{1,*} + \Delta j_{1,*}),\ b(j_{1,*} + \Delta j_{1,*}))] =$$
$$= [(a(j_{1,*} - \Delta j_{1,*}),\ b(j_{1,*} - \Delta j_{1,*}))]^{\mathrm{T}}$$

предыдущая линейная полиэдраль относится к однородно-гиперболическому квадратичному типу.

В статье [1] приведено определение однородно-параболической линейной полиэдрали и предложен алгоритм локализации.

Квадратично-однородную типизацию линейных полиэдралей обобщаем на триодные полиэдральные массивы, принимая вертикальные слои в качестве граней-блоков.

## 2. Рамочное упорядочивание однородно-круговых полиэдралей

Введем метрирование линейных полиэдралей на основе однородно-квадратичного упорядочивания в ресурсную оболочку по критерию ресурсной меры.

С этой целью в линейной полиэдрали кругового типа со свойством однородности квадратичной типизации проведем попарный транспонировано-симметричный синтез граней $[(a(j_1),\ b(j_1))] \cup \cup [(a(k-j_1),\ b(k-j_1))]$ суперпозицией вдоль вертикали с образованием ресурсной оболочки (рис. 2)

$$\max\{a(j_1);\ a(k-j_1)\} \times [b(j_1) + b(k-j_1)] =$$
$$= \max\{a(j_1);\ a(k-j_1)\} \times H,$$

в которой $H = \mathrm{const}$ — общее значение аддитивности транспонировано-симметричных измерений 2-го рода.

Для построенных оболочек применяем суперпозицию последовательно по горизонтали (рис. 3) в полосу уровня $H$ и протяженности

$$L = \sum_{j_1=1}^{m} \max\{a(j_1);\ a(k-j_1)\},\ 2m = k.$$

Для рамочного $M \times M$ распределения синтезированных блоков общего уровня $H$ и протяженностей $\max\{a(j_1);\ a(k-j_1)\}$ применяем горизонтальную суперпозицию наилучшего приближения измерения $M$ с недостатком последовательными отрезками $j_1$-индексации указанных оснований. Получаем некоторый массив ресурсных оболочек $(M-0) \times H$. В одной рамке-стадии размещаем $M/H$ ресурсных оболочек (рис. 4).

Для натуральных ресурсных квадратов

$$\bigcup_{j_1=1}^{k} j_1 \times j_1,\ k = 2m,$$

парным вертикальным транспонированным синтезом получаем ресурсные оболочки (рис. 5)

$$(j_1 + m) \times [j_1 + (k - j_1 + 1)],\ j_1 = 1,\ 2,\ \ldots,\ m,$$

общего уровня $(k+1)$.

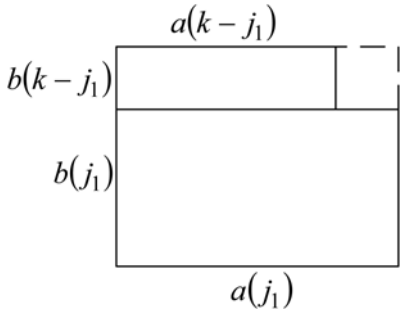



**Рис. 2. Парный транспонированно-симметричный синтез**

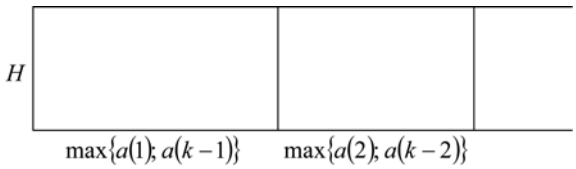



**Рис. 3. Суперпозиция ресурсных оболочек в полосу**

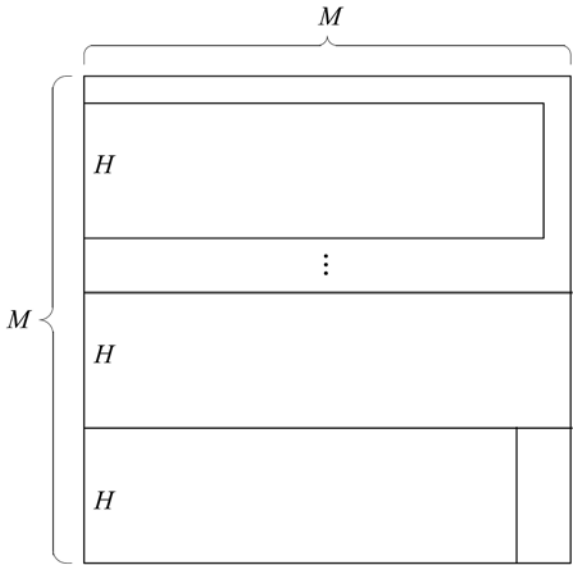



**Рис. 4. Заполнение рамки ресурсными оболочками**

**Сравнение условно- аналитического и оптимального алгоритмов**

| $k$ | $L \times H$ | $L_{opt} \times H_{opt}$ | $\Delta$, % | $k$ | $L \times H$ | $L_{opt} \times H_{opt}$ | $\Delta$, % |
|---|---|---|---|---|---|---|---|
| 5 | 12 × 6 | 5 × 12 | 20 | 19 | 145 × 20 | 47 × 53 | 16,4 |
| 6 | 15 × 7 | 9 × 11 | 6 | 20 | 155 × 21 | 34 × 85 | 12,6 |
| 7 | 22 × 8 | 11 × 14 | 14,3 | 21 | 176 × 22 | 38 × 85 | 15,8 |
| 8 | 26 × 9 | 14 × 15 | 11,4 | 22 | 187 × 23 | 39 × 98 | 12,5 |
| 9 | 35 × 10 | 15 × 20 | 16,7 | 23 | 210 × 24 | 64 × 68 | 15,8 |
| 10 | 40 × 11 | 15 × 27 | 8,6 | 24 | 222 × 25 | 56 × 88 | 12,6 |
| 11 | 51 × 12 | 19 × 27 | 19,3 | 25 | 247 × 26 | 43 × 129 | 15,8 |
| 12 | 57 × 13 | 23 × 29 | 11,1 | 26 | 260 × 27 | 70 × 89 | 12,7 |
| 13 | 70 × 14 | 22 × 38 | 17,2 | 27 | 287 × 28 | 47 × 148 | 15,5 |
| 14 | 77 × 15 | 23 × 45 | 11,6 | 28 | 301 × 29 | 63 × 123 | 12,6 |
| 15 | 92 × 16 | 23 × 55 | 16,4 | 29 | 330 × 30 | 81 × 106 | 15,3 |
| 16 | 100 × 17 | 28 × 54 | 12,4 | 30 | 345 × 31 | 51 × 186 | 12,7 |
| 17 | 117 × 18 | 39 × 46 | 17,4 | 31 | 376 × 32 | 91 × 110 | 20,2 |
| 18 | 126 × 19 | 31 × 69 | 11,9 | 32 | 392 × 33 | 85 × 135 | 12,7 |

Горизонтальная суперпозиция данных блоков $\bigcup_{j_1=1}^{k} [(j_1 + m, k + 1)]$ образует полосу (рис. 6) уровня $(k + 1)$ и протяженности

$$L = \sum_{j_1=1}^{m} (j_1 + m) = m^2 + \frac{m(m+1)}{2},\ m = \frac{k}{2}.$$

Сравним качество приведенного условно-аналитического алгоритма акселерированного (ускоренного) диспетчирования с оптимальной укладкой в объемлющий прямоугольник минимальной площади, полученной в работах [9—12]. В таблице приведены результаты размещения условно-аналитическим и оптимальным алгоритмами для последовательности натуральных ресурсных квадратов от $1 \times 1$ до $k \times k$. Здесь $L$ — горизонтальное и $H$ — вертикальное измерения объемлющего прямоугольника условно-аналитического алгоритма; $L_{opt}$, $H_{opt}$ — соответствующие измерения оптимального алгоритма; $\Delta$ — погрешность площади ресурсной оболочки в % относительно оптимального значения.

Видим, что погрешность не превосходит 21 %, причем на массивах с четным числом натуральных ресурсных квадратов не более 13 %, что является подтверждением целесообразности использования предложенного алгоритма при диспетчировании процессорно-временными ресурсами.

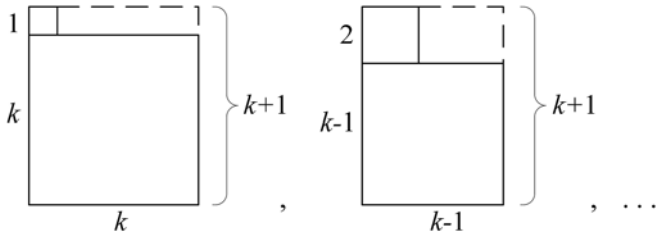



**Рис. 5. Ресурсные оболочки натуральных квадратов**

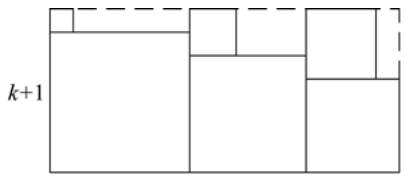



**Рис. 6. Горизонтальная суперпозиция блоков ресурсных квадратов**

## 3. Рамочное упорядочивание однородно-гиперболических полиэдралей

Натуральным квадратам сопоставляем натуральные гиперболические ресурсные прямоугольники

$$j_1 \times j_1 \to j_1 \times (k - j_1)$$

с убыванием высот и ростом оснований (рис. 7).

При $k = 2m$ элемент с индексацией

$$j_1 = m \to j_1 \times (k - j_1) = m \times m$$

является центральным. Грани симметричного расположения относительно центрального квадрата

$$j_1 \times (k - j_1),\ (k - j_1) \times [k - (k - j_1)] = (k - j_1) \times j_1$$

образуют транспонированную пару. Поэтому вертикальный синтез транспонированных пар указанных граней (рис. 8) дает ресурсные оболочки общего уровня

$$k - j_1 + j_1 = k$$

и протяженности $k - j_1 \geqslant m$ для $j_1 \leqslant m$.

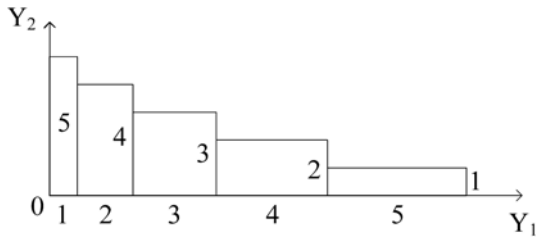



**Рис. 7. Гиперболические ресурсные прямоугольники**

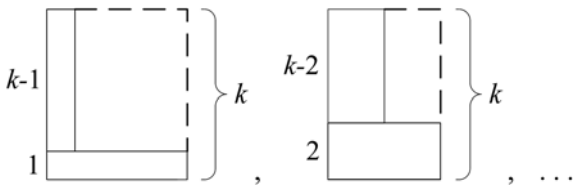



**Рис. 8. Ресурсные оболочки транспонированных пар**

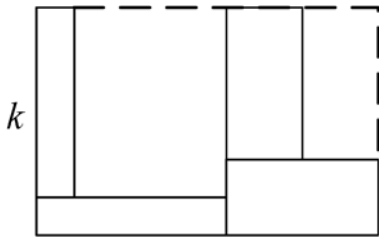



**Рис. 9. Горизонтальная суперпозиция блоков гиперболических прямоугольников**

Горизонтальная суперпозиция данных блоков $\bigcup_{j_1=1}^{k}[(k-j_1,\ k)]$ образует полосу (рис. 9) уровня $k$ и протяженности

$$L=\sum_{j_1=1}^{m}(k-j_1)=2m^2-\frac{m(m+1)}{2}=\frac{3}{2}m^2-\frac{m}{2}.$$

Факторизация данной полосы на допустимые протяженности наилучшего приближения рамочного измерения $M$, как и в предыдущем разделе 2, позволит построить требуемое рамочное распределение указанного массива ресурсных гиперболических прямоугольников.

### 4. Полиэдрали монотонной составности

Рассмотрим $k$-модульную линейную полиэдраль координатных ресурсных прямоугольников

$$\bigcup_{i'=1}^{k}\bigcup_{j'=1}^{k}[(a_{i'}(j'),\ b_{i'}(j'))],\ [(a_{i'}(j'+1),\ b_{i'}(j'+1))]\subseteq$$
$$\subseteq[(a_{i'}(j'),\ b_{i'}(j'))],$$

с упорядочиванием граней по вложению в пределах каждого $i'$-го модуля (рис. 10).

Выполнив кольцевой синтез кругового массива каждого модуля, придем к линейной полиэдрали из $k$ ресурсных оболочек аддитивной графики граней первоначальных модулей (рис. 11).

Исходный массив относим к классу $k$-модульных линейных полиэдралей монотонной круговой составности, если построенное множество оболочек в свою очередь упорядочено по вложению:

$$\bigcup_{i'=1}^{k}[(a(i'),\ b(i'))],\ [(a(i'+1),\ b(i'+1))]\subseteq$$
$$\subseteq[(a(i'),\ b(i'))],\ i'=1,\ 2,\ ...,\ k-1.$$

Дополнительное свойство монотонной составности $k$-модульной линейной полиэдрали кругового квадратичного типа обеспечивает кольцевой синтез указанного массива в две стадии: вначале помодульно, затем — кольцевым синтезом ресурсных оболочек, полученных на первой стадии.

Введем понятие *монотонной составности* для $k$-модульной линейной полиэдрали из модулей гиперболического квадратичного типа, содержащих $k$ несравнимых граней с медленным убыванием высот и медленно растущими основаниями.

С этой целью применим центрально-угловой синтез граней каждого $i'$-го гиперболического модуля в угловую полиэдраль (рис. 12).

В случае, когда $(i'+1)$-я угловая полиэдраль допускает размещение внутри угловой координатной области предыдущей $i'$-й угловой полиэдрали для $i'=1,\ 2,\ ...,\ k-1$ (рис. 13), исходный массив относим в класс $k$-модульных линейных гиперболических полиэдралей монотонной угловой составности.

Данное дополнительное свойство гиперболической монотонности позволяет на втором этапе центрально-углового синтеза гиперболических модулей получить координатную ресурсную оболочку аддитивной графики граней исходного массива с вычислением измерений оболочки на первой стадии синтеза.

Введем далее понятие *параболической монотонной составности*. Пусть $k$-модульная линейная полиэдраль координатных ресурсных прямоугольников состоит из модулей параболического типа с ресурсными оболочками в виде координатных трапеций.

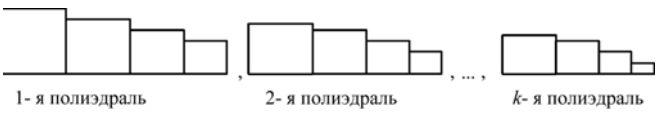



**Рис. 10. $k$-модульная линейная полиэдраль ресурсных прямоугольников**

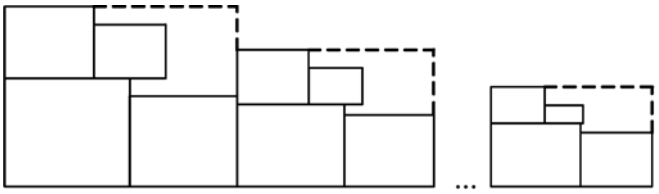


**Рис. 11. Линейная полиэдраль из $k$ ресурсных оболочек**

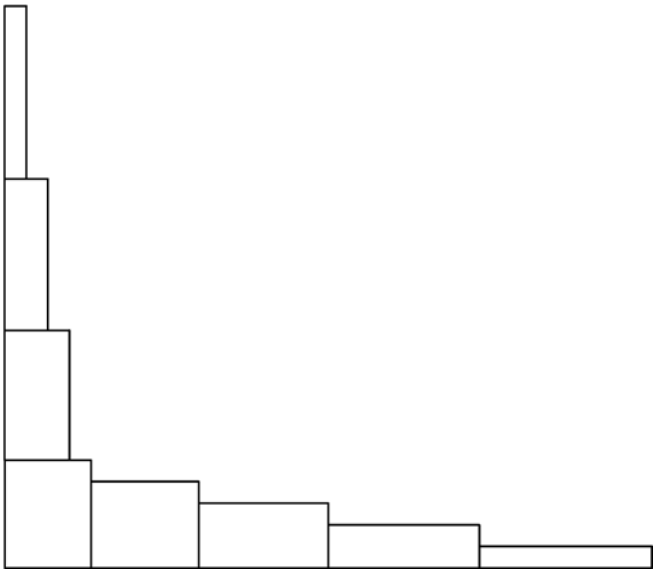


**Рис. 12. Центрально-угловой синтез граней гиперболического модуля**

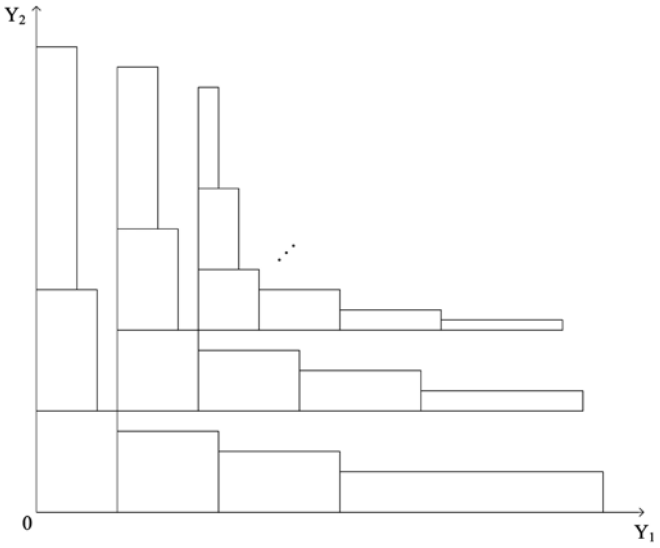



**Рис. 13. Вершинно-диагональное соединение центрально-углового синтеза граней *k*-модульной гиперболической полиэдрали**

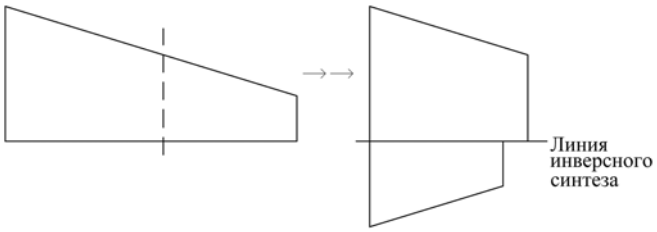



**Рис. 14. Центрально-инверсный синтез факторов параболической ресурсной оболочки**

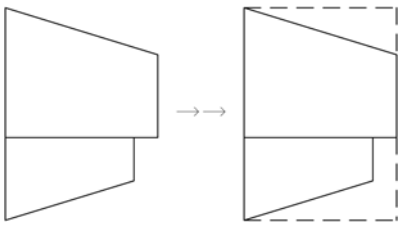


**Рис. 15. Оболочка блока инверсного синтеза**

Подвергнем последние центрально-инверсному синтезу посредством факторизации оболочки вертикалью через середину основания с точностью до ближайшей координатной грани, с тем чтобы не нарушить целостности планарных ресурсных элементов.

Правую часть подвергаем инверсии изменения ориентации по вертикали и выполняем суперпозицию с левой частью по линии основания одинаковой протяженности с указанной выше точностью расположения ближайшей грани к линии факторизации модуля (рис. 14).

Образованные *k* блоков центрального инверсного синтеза первоначальных оболочек (трапеций) принимаем в качестве линейной полиэдрали координатных граней-оболочек блоков (рис. 15).

Свойство упорядоченности линейной полиэдрали вновь образованных граней по вложению определяет искомую монотонную параболическую составность *k*-модульной линейной полиэдрали в указанных условиях.

Прикладное значение выделенного класса параболической монотонности состоит в возможности кольцевого синтеза оболочек на второй стадии локализации протяженного массива координатных ресурсных прямоугольников.

## Заключение

Реальностью синтеза Grid-технологий глобального массива компьютерной техники является ограниченность как массива спроса, так и ресурса предложений со стороны диспетчирования. В статье дана модель рамочного распределения (локальной части координатного квадранта) протяженного массива координатных ресурсных прямоугольников (заявок пользователей) в качестве основной задачи множественного диспетчирования.

Выделение свойств однородности и монотонности квадратичной типизации множества заявок пользователей используется для рамочного упорядочивания соответствующих протяженных линейных массивов. Приводятся результаты сравнения условных аналитических алгоритмов решения рамочной задачи с оптимальными алгоритмами размещения и показывается приемлемость значений погрешности. Условные аналитические алгоритмы могут использоваться в диспетчере как МВС, так и центра Grid-технологий.

### Список литературы

1. **Саак А. Э.** Алгоритмы диспетчеризации в Grid-системах на основе квадратичной типизации массивов заявок // Информационные технологии. 2011. № 1. С. 9—13.

2. **Саак А. Э.** Локально-оптимальные ресурсные распределения // Информационные технологии. 2011. № 2. С. 28—34.

3. **Саак А. Э.** Локально-оптимальный синтез расписаний для Grid-технологий // Информационные технологии. 2010. № 12. С. 16—20.

4. **Барский А. Б.** Параллельные информационные технологии. М.: ИНТУИТ; БИНОМ. Лаборатория знаний, 2007. 503 с.

5. **Барский А. Б.** Параллельные информационные технологии в основе Grid-системы // Информационные технологии. 2006. № 12. С. 54—60.

6. **Хорошевский В. Г.** Архитектура вычислительных систем. М.: Изд-во МГТУ им. Н. Э. Баумана, 2005. 512 с.

7. **Воеводин В. В., Воеводин Вл. В.** Параллельные вычисления. СПб.: БХВ-Петербург, 2002. 608 с.

8. **Каляев И. А., Левин И. И., Семерников Е. А., Шмойлов В. И.** Реконфигурируемые мультиконвейерные вычислительные структуры. Изд. 2-е, перераб. и доп. / Под общ. ред. И. А. Каляева. Ростов н/Д.: Изд-во ЮНЦ РАН, 2009. 344 с.

9. **Korf R.** Optimal rectangle packing: Initial results // Proc. of the thirteenth international conference on automated planning and scheduling (ICAPS 2003). 2003. P. 287—295. Trento: AAAI Press.

10. **Korf R.** Optimal rectangle packing: New results // Proceedings of the fourteenth international conference on automated planning and scheduling (ICAPS 2004). Whistler: AAAI Press. 2004. P. 142—149.

11. **Korf R., Moffitt M., Pollack M.** Optimal rectangle packing // Annals of Operations Research. 2010. Vol. 179, N 1. P. 261—295.

12. **Korf R., Huang E.** New Improvements in Optimal Rectangle Packing // Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI 2009). Pasadena, California, USA, July 11—17, 2009. P. 511—516.

УДК 004.3

**А. Н. Каленик**[1], вед. инж.-программист,
e-mail: andreik@gmail.com,
**А. А. Коляда**[2], д-р физ.-мат. наук, гл. науч. сотр.,
e-mail: razan@tut.by,
**Н. А. Коляда**[2], науч. сотр.,
**А. Ф. Чернявский**[2],
акад. НАН Беларуси, зав. отделом,
**Е. В. Шабинская**[2], канд. техн. наук, науч. сотр.,
e-mail: shabinskaya@rambler.ru
[1] ИООО "Софтек Девелопмент", г. Минск
[2] "Институт прикладных физических проблем им. А. Н. Севченко" Белорусского государственного университета, г. Минск

# Умножение и возведение в степень по большим модулям с использованием минимально избыточной модулярной арифметики

*Предлагаются новые быстрые алгоритмы умножения и возведения в степень по большому модулю, основанные на минимально избыточной модулярной схеме Монтгомери. Главной отличительной особенностью разработанной схемы является использование интервально-индексных характеристик и интервально-модулярной формы чисел в базовых процедурах расширения кода. Достигаемая за счет этого оптимизация синтезированных мультипликативных алгоритмов обеспечивает (3,5—3,6)-кратное повышение производительности (в сравнении с наиболее близким модулярным аналогом) при выполнении на однопроцессорной ЭВМ. В случае мультипроцессорной реализации получаемый выигрыш в быстродействии является (7—8)-кратным. Созданные алгоритмы предназначены для применения в криптосистемах с открытым ключом.*

***Ключевые слова:*** *криптосистема, умножение и возведение в степень по большому модулю, мультипликативная схема Монтгомери, модулярная система счисления, минимально избыточная модулярная арифметика, интервальный индекс, интервально-модулярная форма, расширение модулярного кода*

## Введение

Как известно [1—11], мультипликативные операции, определенные на кольцах вычетов по большим модулям, составляют эффективную основу для построения систем криптографической защиты информации. В частности, их широко применяют в системах электронной цифровой подписи, а также в криптосистемах с открытым ключом, базирующихся на схемах RSA, Рабина и т. д. В свете сказанного особую важность приобретают разработки по внедрению в практику новых вычислительных технологий, которые обеспечивают высокую производительность при оперировании в диапазонах больших чисел (ДБЧ) и, прежде всего, при выполнении операций умножения и возведения в степень по большим модулям. В этом отношении значительный интерес представляет модулярная вычислительная технология (МВТ).

В настоящее время арифметику модулярных систем счисления (МСС) — модулярную арифметику (МА) широко применяют в системах параллельной обработки для решения задач, требующих быстрых точных вычислений. Внутренний (кодовый) параллелизм модулярных вычислительных структур (МВС) обеспечивает ей ряд существенных преимуществ над позиционными структурами при проведении расчетов в ДБЧ. К таким преимуществам относятся:

- независимость длительности модульных операций при их параллельной реализации от числа оснований, а значит, от длины кода МСС;
- идеальная приспособленность алгоритмов МА к конвейеризации и табличным вычислениям;
- простота организации на базе инструментальных платформ позиционного типа многомашинного и мультипроцессорного режимов обработки данных;
- гибкость табличного механизма реконфигурации МВС и др.

С повышением уровня модульности выполняемых вычислительных процессов продуктивность МСС значительно возрастает, причем на ДБЧ влияние данного фактора особенно ощутимо. Весьма показательным в этом отношении примером служат мультипликативные МА-процедуры, основанные на схеме Монтгомери [4—15]. В рамках метода Монтгомери используется операция деления нацело, а не операция общего деления. Поэтому модулярные конфигурации алгоритма Монтгомери для умножения по большим модулям отличаются высокой производительностью.

Наиболее трудоемкую часть МА-процедур Монтгомери составляют операции расширения модулярного кода (МК). Оптимизация данных операций является ключевым направлением развития МВТ на ДБЧ для криптографических приложений. Эффективной основой для решения обозначенной оптимизационной проблемы могут служить минимально избыточные МСС (МИМСС) [16—19].

Представляемая разработка нацелена на реализацию фундаментальных преимуществ табличной конфигурации компьютерной арифметики МИМСС — минимально избыточной МА (МИМА) на ДБЧ в части оптимизации алгоритмов умножения и возведения в степень по большому модулю, базирующихся на схеме Монтгомери.

## 1. Компьютерно-арифметическая база модулярных мультипликативных процедур на основе схемы Монтгомери

Введем обозначения:

- $\mathbf{Z}$ — множество целых чисел (ЦЧ);
- $\lfloor x \rfloor$ и $\lceil x \rceil$ — наибольшее и наименьшее ЦЧ соответственно, не большее и не меньшее вещественной величины $x$;
- $\mathbf{Z}_m = \{0, 1, ..., m-1\}$ и $\mathbf{Z}_m^- = \{-\lfloor m/2 \rfloor, -\lfloor m/2 \rfloor + 1, ..., \lceil m/2 \rceil - 1\}$ — множества наименьших неотрицательных и абсолютно наименьших вычетов по натуральному модулю $m > 1$ соответственно;
- $|a|_m$ — элемент множества $\mathbf{Z}_m$, сравнимый с $a$ (в общем случае рациональной величиной) по модулю $m$;
- $\mathrm{sgn}(x)$ — знаковая функция вида $\mathrm{sgn}(x) = \begin{cases} 0, \text{ если } x \geqslant 0, \\ 1, \text{ если } x < 0; \end{cases}$
- $M_n = \prod_{j=1}^{n} m_j$, $M_{i,n} = M_n / m_i$ $(i = \overline{1, n})$, где $m_1, m_2, ..., m_n$ — натуральные модули ($n \geqslant 1$);
- $p$ — рабочий модуль (большое ЦЧ) для мультипликативных операций.

На множестве $\mathbf{Z}$ МСС определяется посредством набора попарно простых модулей (оснований) — $m_1, m_2, ..., m_k$ ($k > 1$). Число $X \in \mathbf{Z}$ в данной МСС представляется в виде $X = (\chi_1, \chi_2, ..., \chi_k)$ $(\chi_i = |X|_{m_i}$ $(i = \overline{1, k}))$. В неизбыточной МСС с основаниями $m_1, m_2, ..., m_k$ можно закодировать не более $M_k$ ЦЧ. При этом в качестве диапазона используют множества $\mathbf{Z}_{M_k}$ или $\mathbf{Z}_{M_k}^-$.

Декодирующее отображение для МСС с диапазоном $\mathbf{Z}_{M_k}$, ставящее в соответствие коду $(\chi_1, \chi_2, ..., \chi_k)$ единственный элемент $X$ из $\mathbf{Z}_{M_k}^-$, может быть реализовано [16] с помощью соотношения

$$X = \sum_{i=1}^{k-1} M_{i,k-1} \left| M_{i,k-1}^{-1} \chi_i \right|_{m_i} + M_{k-1} I_k(X), \quad (1)$$

где $I_k(X)$ — интервальный индекс (ИИ) числа $X$ относительно модулей $m_1, m_2, ..., m_k$. Выражение (1) называется интервально-модулярной формой (ИМФ) ЦЧ $X$.

Справедливо [16] следующая теорема.

**Теорема 1.** Для ИИ $I_l(X)$ ЦЧ $X \in \mathbf{Z}_{M_l}$ в МСС с попарно простыми основаниями $m_1, m_2, ..., m_{l-1}$, $m_l \geqslant l - 2$ ($l \geqslant 2$) имеет место формула

$$I_l(X) = \hat{I}_l(X) - m_l \Theta_l(X), \quad (2)$$

где

$$\hat{I}_l(X) = |I_l(X)|_{m_l} = \left| \sum_{i=1}^{l} R_{i,l}(\chi_i) \right|_{m_l}; \quad (3)$$

$$R_{i,l}(\chi_i) = \left| -m_i^{-1} \left| M_{i,l-1}^{-1} \chi_i \right|_{m_i} \right|_{m_l} \ (i = \overline{1, l-1}),$$

$$R_{l,l}(\chi_l) = \left| M_{l-1}^{-} \chi_l \right|_{m_l}; \Theta_l(X) \in \{0, 1\}. \quad (4)$$

Величина $\Theta_l(X)$ называется минимальной интегральной характеристикой МК (ИХМК), отвечающая ЦЧ $X$ в МСС с основаниями $m_1, m_2, ..., m_l$.

Из-за наличия в формуле (2) $\Theta_l(X)$ в классической (неизбыточной) МСС вычисление интервально-индексной характеристики $I_l(X)$ требует применения общего алгоритма формирования ИХМК [16, 20], который является довольно трудоемким.

Арифметические свойства МСС удается значительно улучшить за счет избыточного кодирования элементов рабочего диапазона. Предложенное в работе [16] так называемое минимально-избыточное модулярное кодирование $\Phi_{\text{МИМСС}}$: $(\mathbf{Z}_{m_1} \times \mathbf{Z}_{m_2} \times ... \times \mathbf{Z}_{m_k}) \to \mathbf{D}$ предусматривает использование диапазона $\mathbf{D}$ с мощностью $|\mathbf{D}| < M_k$. Сущность реализуемого принципа раскрывает нижеследующая теорема.

**Теорема 2.** Для того чтобы в МСС с попарно простыми основаниями $m_1, m_2, ..., m_k$ ИИ $I(X) = I_k(X)$ каждого элемента $X$ диапазона $\mathbf{D} = \mathbf{Z}_{2M}^- = \{-M, -M+1, ..., M-1\}$ ($M = m_0 M_{k-1}$; $m_0$ — вспомогательный модуль) однозначно определялся компьютерным ИИ — вычетом $\hat{I}_k(X) = |I(X)|_{m_k}$, необходимо и достаточно, чтобы $k$-е основание МСС удовлетворяло условию $m_k \geqslant 2m_0 + k - 2$ ($m_0 \geqslant k - 2$). При этом для $I(X)$ верна формула

$$I(X) = \begin{cases} \hat{I}_k(X), \text{ если } \hat{I}_k(X) < m_0; \\ \hat{I}_k(X) - m_k, \text{ если} \\ \hat{I}_k(X) \geqslant m_k - m_0 - k + 2. \end{cases} \quad (5)$$

Компьютерный ИИ $\hat{I}_k(X)$ вычисляется согласно формулам (3), (4) при $l = k$.

Из теоремы 2 видно, что при минимально избыточном модулярном кодировании ИИ и отвечающая ему ИМФ позволяют достичь принципиально нового, в сравнении с традиционными конфигурациями МА, уровня оптимизации немодульных про-

цедур по таким, в частности, характеристикам, как быстродействие и объем реализационных затрат. Весьма показательными в этом отношении являются операции расширения МК, выполняемые в рамках МА-алгоритмов Монтгомери для умножения по большим модулям $p$ [17—19].

Пусть, например, некоторый МК $(\chi_{l+1}, \chi_{l+2}, ..., \chi_k)$ по набору оснований $\{m_{l+1}, m_{l+2}, ..., m_k\}$ $(1 < l < k)$ необходимо расширить на модули $m_j$ $(j = \overline{1, l})$. Если на $m_k$ наложить условие

$$m_k \geqslant 2m_0 + k - l - 2 (m_0 \geqslant k - l - 2), \qquad (6)$$

то согласно теореме 2 МСС с основаниями $m_{l+1}$, $m_{l+2}$, ..., $m_k$ и диапазоном $\mathbf{D}' = \mathbf{Z}^{-}_{2M'} = \{-M', -M'+ + 1, ..., M' - 1\}$ $(M' = M/M_l)$ будет минимально избыточной. Предположим, что коду $(\chi_{l+1}, \chi_{l+2}, ..., \chi_k)$ отвечает ЦЧ $X \in \mathbf{D}'$. Тогда требуемая операция расширения сводится к расчету ИИ $I'_k(X)$ числа $X = (\chi_{l+1}, \chi_{l+2}, ..., \chi_k)$ относительно $m_{l+1}, m_{l+2}, ..., m_k$ по формулам типа (3)—(5):

$$I'_k(X) = \begin{cases} \hat{I}'_k(X), \text{ если } \hat{I}'_k(X) < m_0; \\ \hat{I}'_k(X) - m_k, \text{ если} \\ \hat{I}'_k(X) \geqslant m_k - m_0 - k + l + 2; \end{cases} \qquad (7)$$

$$\hat{I}'_k(X) = |I'_k(X)|_{m_k} = \left| \sum_{i=l+1}^{k} |R'_{i,k}(\chi_i) \right|_{m_k}; \qquad (8)$$

$$R'_{i,k}(\chi_i) = \left| -m_i^{-1} \left| \frac{M_l m_i}{M_{k-1}} \chi_i \right|_{m_i} \right|_{m_k} \quad (i = \overline{l+1, k-1}),$$

$$R'_{k,k}(\chi_k) = \left| \frac{M_l}{M_{k-1}} \chi_k \right|_{m_k} \qquad (9)$$

и применению к ИМФ

$$X = \sum_{i=l+1}^{k-1} \frac{M_{k-1}}{M_l m_i} \left| \frac{M_l m_i}{M_{k-1}} \chi_i \right|_{m_i} + \frac{M_{k-1}}{M_l} I'_k(X) \qquad (10)$$

(см. (1)) операции приведения к остатку по модулю $m_j$. Результирующее расчетное соотношение имеет вид

$$\chi_j = |X|_{m_j} = \left| \sum_{i=l+1}^{k-1} \left| \frac{M_{k-1}}{M_l m_i} \left| \frac{M_l m_i}{M_{k-1}} \chi_i \right|_{m_i} \right|_{m_j} + \right.$$

$$+ \left. \left| \frac{M_{k-1}}{M_l} I'_k(X) \right|_{m_j} \right|_{m_j} \quad (j = \overline{1, l}). \qquad (11)$$

Для операции расширения МК ЦЧ $X$ по набору $\mathbf{M}_1$ оснований на основания набора $\mathbf{M}_2$, где $\mathbf{M}_1$, $\mathbf{M}_2 \subset \mathbf{M}$ $\{m_1, m_2, ..., m_k\}$ будем употреблять условное обозначение ЕС($X$; $\mathbf{M}_1$, $\mathbf{M}_2$).

Что касается модульных операций над произвольными ЦЧ $A$ и $B$, заданными своими МК:

$$A = (\alpha_1, \alpha_2, ..., \alpha_k), \; B = (\beta_1, \beta_2, ..., \beta_k)$$

$$(\alpha_i = |A|_{m_i}, \; \beta_i = |B|_{m_i} \; (i = \overline{1, k})),$$

то в МСС с основаниями $m_1, m_2, ..., m_k$ они выполняются независимо по каждому из оснований, т. е. по правилу

$$A \circ B = (\alpha_1, \alpha_2, ..., \alpha_k) \circ (\beta_1, \beta_2, ..., \beta_k) =$$

$$= (|\alpha_1 \circ \beta_1|_{m_1}, |\alpha_2 \circ \beta_2|_{m_2}, ..., |\alpha_k \circ \beta_k|_{m_k})$$

$$(\circ \in \{+, -, \cdot\}). \qquad (12)$$

В свойстве (12) заключается главное фундаментальное преимущество МА над арифметикой позиционных систем счисления (ПСС).

### Метод Монтгомери для умножения по большому модулю

Пусть $A$, $B$ — операнды подлежащей выполнению операции умножения по некоторому большому модулю $p$. Сущность основополагающей идеи, выдвинутой Монтгомери [21] для построения требуемой мультипликативной схемы, состоит в аддитивной вариации произведения $C = AB$, которая обеспечивает деление без остатка значения результирующего выражения на специально выбираемый вспомогательный модуль $S$. Для достижения указанной цели предложено варьирующее соотношение вида

$$\tilde{C} = C + |-Cp^{-1}|_S p. \qquad (13)$$

При $S$, взаимно простом с $p$, из формулы (13) следует, что

$$|\tilde{C}|_S = |C + |-Cp^{-1}|_S p|_S = |C - Cp^{-1}p|_S = 0. \qquad (14)$$

Таким образом, число

$$\tilde{C}/S = (C + p|-Cp^{-1}|_S/S) \qquad (15)$$

является целым. Переход в формуле (15) к остаткам по модулю $p$ дает

$$|\tilde{C}/S|_p = |C/S|_p = |AB/S|_p. \qquad (16)$$

В соответствии с выражением (16) по методу Монтгомери в качестве искомого произведения операндов $A$ и $B$ принимается ЦЧ

$$\tilde{\gamma} = |AB/S|_p = (\tilde{C}/S) - Qp, \qquad (17)$$

где $Q$ — однозначно определяемый (для заданных $A$ и $B$) целочисленный коэффициент.

Изложенное позволяет заключить, что базовая вычислительная схема для метода Монтгомери сводится к операционной последовательности:

$$< C = AB; \; D = |CF|_S \; (F = |-p^{-1}|_S);$$

$$\tilde{C} = C + Dp; \; \tilde{\gamma} = (\tilde{C}/S) - Qp >. \qquad (18)$$

Как видно из формулы (18), трудоемкость ПСС-версий метода Монтгомери определяется главным образом сложностью операций умножения больших чисел: $AB$, $|CF|_S$ и $Dp$. Сказанное относится и к операции мультипликативной инверсии: $F = |-p^{-1}|_S$. Однако, являясь параметром долговременного использования, величина $F$ может быть получена на этапе предварительных вычислений. Поэтому сложность операции определения $F$ принципиального значения не имеет. В МСС все указанные операции относятся к разряду модульных (см. формулу (12)) и реализуются значительно проще, чем в ПСС. Именно этим обстоятельством, обусловленным кодовым параллелизмом МВС, в первую очередь, и продиктована целесообразность применения МА в криптосистемах.

### Минимально избыточная модулярная схема Монтгомери для умножения по большим модулям

Пусть операнды $A$, $B$ и модуль $p$ заданы в МСС с основаниями $m_1, m_2, ..., m_k$: $A = (\alpha_1, \alpha_2, ..., \alpha_k)$, $B = (\beta_1, \beta_2, ..., \beta_k)$, $p = (\pi_1, \pi_2, ..., \pi_k)$ ($\alpha_i = |A|_{m_i}$, $\beta_i = |B|_{m_i}$, $\pi_i = |p|_{m_i}$ ($i = \overline{1, k}$)) и пусть $S = M_l$ ($1 < l < k$). Так как мультипликативная инверсия $F \in \mathbf{Z}_{M_l}$, то она однозначно определяется своим МК $(\varphi_1, \varphi_2, ..., \varphi_l)$, цифры которого находятся согласно равенствам $\pi_i = |-1/\varphi_i|_{m_i}$ ($i = \overline{1, l}$). Реализация данных равенств осуществляется на стандартных компьютерных диапазонах, причем в ходе предварительных вычислений. Число $D$ также определяется кодом МСС с основаниями $m_1, m_2, ..., m_l$:

$$D = |CF|_{M_l} = (\delta_1, \delta_2, ..., \delta_l) =$$
$$= (|\gamma_1\varphi_1)_{m_1}, |\gamma_2\varphi_2)_{m_2}, ..., |\gamma_l\varphi_l)_{m_l})$$
$$(\delta_i = |D|_{m_i}, \gamma_i = |C|_{m_i} = |\alpha_i\beta_i|_{m_i} \ (i = \overline{1, l}). \quad (19)$$

Поскольку ЦЧ $D$, полученное согласно (19), участвует в дальнейших вычислениях при получении $\tilde{C}$ по полной системе модулей — $m_1, m_2, ..., m_k$, то МК $(\delta_1, \delta_2, ..., \delta_l)$ должен быть расширен на остальные модули: $m_{l+1}, m_{l+2}, ..., m_k$. В рамках неизбыточного модулярного кодирования, а именно таковым является кодовое пространство МСС с основаниями $m_1, m_2, ..., m_l$ и диапазоном $\mathbf{Z}_{M_l}$, данная операция требует использования сложно вычисляемых ИХМК, например ранга [8]. В целях устранения отмеченного негативного фактора заменим $D$ на число

$$\hat{D} = \sum_{i=1}^{l-1} M_{i,l-1} |M_{i,l-1}^{-1} \delta_i|_{m_i} + M_{l-1} \hat{I}_l(D), \quad (20)$$

где $\hat{I}_l(D) = |I_l(D)|_{m_l}$ — компьютерный ИИ ЦЧ $D$ и $\hat{D}$ относительно модулей $m_1, m_2, ..., m_l$, который вычисляется по формулам (3), (4) при $X = D$, $(\chi_1, \chi_2, ..., \chi_l) = (\delta_1, \delta_2, ..., \delta_l)$.

Применяя формулы (1) и (2), запишем $\hat{D}$ в виде

$$\hat{D} = \sum_{i=1}^{l-1} M_{i,l-1} |M_{i,l-1}^{-1} \delta_i|_{m_i} +$$
$$+ M_{l-1}(\hat{I}_l(D) - m_l\Theta_l(D) + m_l\Theta_l(D)) =$$
$$= \sum_{i=1}^{l-1} M_{i,l-1} |M_{i,l-1}^{-1} \delta_i|_{m_i} + M_{l-1} I_l(D) + M_l\Theta_l(D) =$$
$$= D + M_l\Theta_l(D), \quad (21)$$

где $\Theta_l(D)$ — двузначная минимальная ИХМК (см. теорему 1). Из формулы (21) следует равенство

$$|\hat{D}|_{M_l} = D. \quad (22)$$

В соответствии с изложенным искомая модификация базового соотношения для вариации произведения $C$ имеет вид

$$\hat{C} = C + \hat{D}p =$$
$$= C + \left( \sum_{i=1}^{l-1} M_{i,l-1} |M_{i,l-1}^{-1} \delta_i|_{m_i} + M_{l-1} \hat{I}_l(D) \right) p. \quad (23)$$

С учетом формул (20), (22) и (14) из выражения (23) следует, что

$$|C|_{M_l} = |C + |\hat{D}p|_{M_l}|_{M_l} = |C + |Dp|_{M_l}|_{M_l} = 0.$$

Таким образом, ЦЧ (23) без остатка делится на $M_l$, т. е. число

$$\hat{\gamma} = \hat{C}/M_l = \left( C + \left( \sum_{i=1}^{l-1} M_{i,l-1} |M_{i,l-1}^{-1} \delta_i|_{m_i} + \right. \right.$$
$$\left. \left. + M_{l-1} \hat{I}_l(D)p \right) \right) / M_l \quad (24)$$

является целым.

Процесс реализации (24) в МСС с основаниями модулей $m_1, m_2, ..., m_k$, т. е. получения кода $(\hat{\gamma}_1, \hat{\gamma}_2, ..., \hat{\gamma}_k)$ ЦЧ $\hat{\gamma}$ ($\hat{\gamma}_i = |\hat{\gamma}|_{m_i}$ ($i = \overline{1, k}$)) состоит из двух шагов. На первом шаге $\hat{\gamma}$ вычисляется по набору модулей $\{m_{l+1}, m_{l+2}, ..., m_k\}$, а на втором шаге сформированный усеченный МК $(\hat{\gamma}_{l+1}, \hat{\gamma}_{l+2}, ..., \hat{\gamma}_k)$ расширяется на модули $m_1, m_2, ..., m_l$. При этом данная операция ЕС($\hat{\gamma}$; $\{m_{l+1}, m_{l+2}, ..., m_k\}$, $\{m_1, m_2, ..., m_l\}$), естественно, должна выполняться по упрощенной минимально избыточной процедуре

расширения согласно формулам (7)—(9), (11) при $X = \hat{\gamma}$, $(\chi_{l+1}, \chi_{l+2}, ..., \chi_k) = (\hat{\gamma}_{l+1}, \hat{\gamma}_{l+2}, ..., \hat{\gamma}_k)$.

Необходимым и достаточным условием корректности двухшагового процесса вычисления величины $\hat{\gamma}$ служит принадлежность ЦЧ $\hat{C}$ и $\hat{\gamma}$ соответственно неотрицательным компонентам $\{0, 1, ..., \lceil M_k/2 \rceil - 1\}$ диапазона $\mathbf{Z}^-_{M_k}$ МСС с основаниями $m_1, m_2, ..., m_k$ и $\mathbf{Z}_{M'}$ диапазона $\mathbf{D}' = \mathbf{Z}^-_{2M'}$ МИМСС с основаниями $m_{l+1}, m_{l+2}, ..., m_{k-1}, m_k \geqslant 2m_0 + k - l - 2$ (см. формулу (6)). Поскольку из соотношения $0 \leqslant \hat{\gamma} = \hat{C}/M_l < M' = m_0 M_{k-1}/M_l$ вытекает неравенство $0 \leqslant \hat{C} < m_0 M_{k-1} \leqslant M_{k-1}((m_k - k + l + 2)/2) < M_k/2$, то условие $\hat{\gamma} \in \mathbf{Z}_{M'}$ гарантирует выполнение и условия $\hat{\mathbf{C}} \in \mathbf{Z}^-_{M_k}$. При этом МСС с основаниями $m_1, m_2, ..., m_k$ необязательно должна быть минимально избыточной.

Найдем ограничения на модули $m_1, m_2, ..., m_k$ и $p$, гарантирующие выполнение условия $\hat{\gamma} \in \mathbf{D}'$. Получим сначала верхнюю оценку для $\hat{D}$.

Пусть $m\_\max = \max\{m_1, m_2, ..., m_k\}$. Из формулы (20) имеем:

$$\hat{D} = M_{l-1}\left(\sum_{i=1}^{l-1} m_i^{-1} |M_{i,l-1}^{-1} \delta_i|_{m_i} + \hat{I}_l(\hat{D})\right) \leqslant$$

$$\leqslant M_{l-1}\left(\sum_{i=1}^{l-1} m_i^{-1}|m_i - 1) + m_l - 1\right) =$$

$$= M_{l-1}\left(l - 1 - \sum_{i=1}^{l-1} \frac{1}{m_i}\right) + M_l - M_{l-1} <$$

$$< M_l + M_{l-1}\left(l - 2 - \sum_{i=1}^{l-1} \frac{1}{m\_\max}\right) <$$

$$< M_l + M_{l-1}\left(l - 2 - \left\lfloor \frac{l-1}{m\_\max} \right\rfloor\right) =$$

$$= M_l + M_{l-1}(l-2). \qquad (25)$$

Пусть $A, B \in \mathbf{Z}_p$. Тогда с учетом (25) из (24) получим:

$$\hat{\gamma} < (p^2 + (M_l + M_{l-1}(l-2))p)/M_l =$$
$$= p(1 + (p + M_{l-1}(l-2))/M_l). \qquad (26)$$

Из формулы (26) видно, что при $p + M_{l-1}(l-2) < M_l$ ЦЧ $\hat{\gamma} \in \mathbf{Z}_{2p}$. Такого же результата, т. е. принадлежности $\hat{\gamma}$ множеству $\mathbf{Z}_{2p}$, можно достичь и в случае, когда $A, B \in \mathbf{Z}_{2p}$. Соответствующее ограничение на основания МСС и модуль $p$ вытекает из неравенства $\hat{\gamma} < (4p^2 + (M_l + M_{l-1}(l-2))p)/M_l < 2p$ и имеет вид

$$4p + M_{l-1}(l-2) < M_l. \qquad (27)$$

В рамках данного условия, обеспечивающего $\hat{\gamma} \in \mathbf{Z}_{2p}$ при $A, B \in \mathbf{Z}_{2p}$, допускается режим многократного обращения к процедуре умножения по модулю $p$ с использованием в качестве операндов результатов уже выполненных операций умножения. Это, в частности, необходимо для реализации в криптосистемах операций возведения в степень по модулю $p$.

Так как на первом шаге применяемой мультипликативной схемы $\hat{\gamma}$ вычисляется в МИМСС с основаниями $m_{l+1}, m_{l+2}, ..., m_k$, то ее диапазон $\mathbf{D}'$ ввиду $\hat{\gamma} \in \mathbf{Z}_{2p}$ должен удовлетворять требованию

$$2p < M' = m_0 M_{k-1}/M_l = M/M_l. \qquad (28)$$

Таким образом, приведенные оценочные выкладки позволяют заключить, что при $A, B, \hat{\gamma} \in \mathbf{Z}_{2p}$ корректность предлагаемой вычислительной МИМА-схемы метода Монтгомери (см. формулы (23), (24)) обеспечивается в рамках условий (27) и (28).

Переход в формуле (24) к остаткам по модулю $p$ дает $\tilde{\gamma} = |\hat{\gamma}|_p = |\hat{C} M_l^{-1}|_p = = |ABM_l^{-1}|_p$. Поскольку $\hat{\gamma} < 2p$, то искомое произведение ЦЧ $A$ и $B$ по модулю $p$ можно получить по $\hat{\gamma}$ с использованием равенства $\tilde{\gamma} = |ABM_l^{-1}|_p = \hat{\gamma} - Qp$ $(Q \in \{0, 1\})$. При этом для величины $Q$ справедлива формула $Q = 1 - \mathrm{sgn}(\hat{\gamma} - p)$.

С учетом вышеизложенного результирующую мультипликативную схему, основанную на МИМА, можно записать в виде операционной последовательности:

$$\begin{aligned}
&\langle C = AB = (\gamma_1, \gamma_2, ..., \gamma_k,);\ D = |CF|_{M_l} = \\
&= (\delta_1, \delta_2, ..., \delta_l)\ (F = |-p^{-1}|_{M_l} = (\varphi_1, \varphi_2, ..., \varphi_l)); \\
&(\hat{\delta}_{l+1}, \hat{\delta}_{l+2}, ..., \hat{\delta}_k) = \\
&= EC(\hat{\mathbf{D}}; \{m_1, m_2, ..., m_l\}, \{m_{l+1}, m_{l+2}, ..., m_k\}); \\
&\hat{\gamma} = \hat{C}/M_l = (\hat{\gamma}_{l+1}, \hat{\gamma}_{l+2}, ..., \hat{\gamma}_k)(\hat{\gamma}_j = \\
&= |(\gamma_j + |\hat{\delta}_j \pi_j|_{m_j}) M_l^{-1}|_{m_j})\ (j = \overline{l+1, k})); \\
&(\hat{\gamma}_1, \hat{\gamma}_2, ..., \hat{\gamma}_l) = \\
&= EC(\hat{\gamma}; \{m_{l+1}, m_{l+2}, ..., m_k\}\{m_1, m_2, ..., m_l\}); \\
&\tilde{\gamma} = \hat{\gamma} - (1 - \mathrm{sgn}(\hat{\gamma} - p))p\rangle.
\end{aligned} \qquad (29)$$

Отметим, что расчет цифр МК $(\hat{\delta}_{l+1}, \hat{\delta}_{l+2}, ..., \hat{\delta}_k)$ числа $\hat{D}$ по набору модулей $\{m_{l+1}, m_{l+2}, ..., m_k\}$

осуществляется с помощью указанной в (29) операции расширения, которая в соответствии с (20) выполняется по правилу

$$\hat{\delta}_j = |\sum_{i=1}^{l-1} |M_{i,l-1}| M_{i,l-1}^{-1} \delta_i|_{m_i}|_{m_j} + |M_{l-1}\hat{I}_l(D)|_{m_j}|_{m_j}$$
$$(j = \overline{l+1, k}) \qquad (30)$$

с применением формул (3), (4). Поскольку $|\hat{C}|_{M_l} = 0$, то согласно (23) для МК числа $\hat{C}$ по набору оснований $\{m_1, m_2, ..., m_k\}$ верна формула

$$\hat{C} = (0, 0, ..., 0, |\gamma_{l+1} +$$
$$+ |\hat{\delta}_{l+1}\pi_{l+1}|_{m_{l+1}}|_{m_{l+1}}, ..., |\gamma_k + |\hat{\delta}_k\pi_k|_{m_k}|_{m_k}, \quad (31)$$

а для ЦЧ (24), вычисляемого в МИМСС с модулями $m_{l+1}, m_{l+2}, ..., m_k$, — формула

$$\tilde{\gamma} = \hat{C}/M_l = (\hat{\gamma}_{l+1}, \hat{\gamma}_{l+2}, ..., \hat{\gamma}_k)(\hat{\gamma}_j =$$
$$= |M_l^{-1}(\gamma_j + |\hat{\delta}_j\pi_j|_{m_j})|_{m_j} \ (j = \overline{l+1, k})). \qquad (32)$$

Из изложенного вытекает нижеследующая теорема.

**Теорема 3.** Пусть наборы оснований (простых чисел): $\{m_1, m_2, ..., m_l\}$ и $\{m_{l+1}, m_{l+2}, ..., m_k\}$ неизбыточной и минимально избыточной МСС соответственно с диапазонами $\mathbf{Z}_{M_l}$ и $\mathbf{D}' = \mathbf{Z}_{2M'}^{-}$ ($M' = M/M_l$, $M = m_0M_{k-1}$, $m_0 \geqslant k - l - 2$, $m_k \geqslant 2m_0 + k - l - 2$, $1 < l < k$) совместно с модулем $p$, взаимно простым с $M_l$, удовлетворяют условию

$$\begin{cases} 4p + M_{l-1}(l-2) < M_l; \\ 2p < M/M_l \end{cases} \qquad (33)$$

и пусть операнды $A$ и $B$ мультипликативной операции $\tilde{\gamma} = |ABM_l^{-1}|_p$ принадлежат множеству $\mathbf{Z}_{2p} = \{0, 1, ..., 2p-1\}$. Тогда величина $\hat{\gamma}$, вычисляемая в рамках схемы (29), также является элементом множества $\mathbf{Z}_{2p}$, при этом $\hat{\gamma} \equiv \tilde{\gamma} (\mathrm{mod} p)$ и для $\tilde{\gamma}$ верна формула

$$\tilde{\gamma} = \hat{\gamma} - (1 - \mathrm{sgn}(\hat{\gamma} - p))p. \qquad (34)$$

## 4. Алгоритмы умножения и возведения в степень по большому модулю на основе МИМА-схемы Монтгомери

На базе вычислительной МИМА-схемы (29) типа Монтгомери синтезированы алгоритмы модульного умножения и возведения в степень для криптосистем. Ключевой отличительной особенностью данных алгоритмов является широкое применение таблиц. При этом необходимый комплект рабочих таблиц генерируется на этапе предварительных вычислений с обеспечением минимизации трудоемкости процесса, реализуемого в реальном времени.

***Алгоритм умножения по большому модулю p***

**Параметры алгоритма:** определяющие основания МСС — $m_0, m_1, ..., m_k$ и модуль $p = (\pi_1, \pi_2, ..., \pi_k))$, которые удовлетворяют условиям теоремы 3.

**Входные данные:** операнды $A$ и $B$ ($A, B \in \mathbf{Z}_{2p}$), представленные в МСС — $A = (\alpha_1, \alpha_2, ..., \alpha_k)$, $B = (\beta_1, \beta_2, ..., \beta_k)$.

**Выходные данные:** МК $(\hat{\gamma}_1, \hat{\gamma}_2, ..., \hat{\gamma}_k)$ аналога $\hat{\gamma} \in \mathbf{Z}_{2p}$ произведения Монтгомери $\tilde{\gamma} = |ABM_l^{-1}|_p$ ($\hat{\gamma} \equiv \tilde{\gamma} (\mathrm{mod}\ p)$).

**Предварительно вычисляемые данные:**

- код $(\varphi_1, \varphi_2, ..., \varphi_l)$ противоположного значения $F = |-p^{-1}|_{M_l}$ мультипликативной инверсии $|-p^{-1}|_{M_l}$ модуля $p$ в МСС с основаниями $m_1, m_2, ..., m_l$, получаемый с помощью равенств $\varphi_i = |-1/\pi_i|_{m_i}$ $(i = \overline{1, l})$;
- таблицы ИИ — *TIIi* и *TII_i*, которые формируются согласно (4) и (9):

$$TIIi\,[\chi] = R_{i,l}(\chi) \ (\chi \in \mathbf{Z}_{m_i};\ i = \overline{1, l});$$

$$TII\_i\,[\chi] = R'_{i,k}(\chi) \ (\chi \in \mathbf{Z}_{m_i};\ i = \overline{l+1, k});$$

- таблицы расширения МК — *TEi_j* и *TE_i_j*, генерируемые в соответствии с (11) и (30) по формулам

$$TEi\_j[\chi] =$$
$$\begin{cases} |M_{i,l-1}|M_{i,l-1}^{-1}\chi|_{m_i}|_{m_j} \ (\chi \in \mathbf{Z}_{m_i}) \text{ при } i = \overline{1, l-1}, \\ |M_{l-1}\chi|_{m_j} \ (\chi \in \mathbf{Z}_{m_l}) \text{ при } i = l, \end{cases}$$
$$(j = \overline{l+1, k});$$
$$TE\_i\_j[\chi] =$$
$$= \begin{cases} \left|\frac{M_{k-1}}{M_l m_i}\left|\frac{M_l m_i}{M_{k-1}}\chi\right|_{m_i}\right|_{m_j} \ (\chi \in \mathbf{Z}_{m_i}) \text{ при } i = \overline{l+1, k-1}, \\ \left|\frac{M_{k-1}}{M_l}\chi\right|_{m_j} \ (\chi \in \mathbf{Z}_{m_k}) \text{ при } i = k \text{ и } \chi < m_0, \\ \left|\frac{M_{k-1}}{M_l}(\chi - m_k)\right|_{m_j} \ (\chi \in \mathbf{Z}_{m_k}) \text{ при } i = k \text{ и и } \chi \geqslant m_0 \end{cases}$$
$$(j = \overline{1, l});$$

- таблицы *TMPli* умножения на константу $M_l^{-1}$, которые согласно (32) рассчитываются по формуле: $TMPli\,[\chi] = |M_l^{-1}\chi|_{m_i}$ $(\chi \in \mathbf{Z}_{2m_i - 1};\ i = \overline{l+1, k})$.

***Тело алгоритма***

**УМ.М1.** Найти произведение $C = AB = (\gamma_1, \gamma_2, ..., \gamma_k) = (|\alpha_1\beta_1|_{m_1}, |\alpha_2\beta_2|_{m_2}, ..., |\alpha_k\beta_k|_{m_k})$.

**УМ.М2.** В МСС с основаниями $m_1, m_2, ..., m_l$ сформировать код числа $D = |CF|_{M_l}$: $(\delta_1, \delta_2, ..., \delta_l) = (|\gamma_1\varphi_1|_{m_1}, |\gamma_2\varphi_2|_{m_2}, ..., |\gamma_l\varphi_l|_{m_l})$.

**УМ.М3.** Вычислить интервально-индексную характеристику $\hat{I}_l(\hat{D}) = \hat{I}_l(D)$ числа $\hat{D} = \sum_{i=1}^{l-1} M_{i,l-1} \times \times |M_{i,l-1}^{-1}\delta_l|_{m_i} + M_{l-1}\hat{I}_l(D)$ по расчетному соотношению $\hat{I}_l(\hat{D}) = \eta_l = \left|\sum_{i=1}^{l} TIIi[\delta_i]\right|_{m_l}$.

**УМ.М4.** Определить цифры МК $(\hat{\delta}_{l+1}, \hat{\delta}_{l+2}, ..., \hat{\delta}_k)$ ЧЧ $\hat{D}$ в МСС с основаниями $m_{l+1}, m_{l+2}, ..., m_k$ по правилу

$$\hat{\delta}_j = \left|\sum_{i=1}^{l-1} TEi\_j[\delta_i] + TEl\_j[\eta_l]\right|_{m_j} \quad (j = \overline{l+1, k}).$$

**УМ.М5.** Получить код числа $\hat{C} = C + \hat{D}p$ в МИМСС с модулями $m_{l+1}, m_{l+2}, ..., m_k$:

$$(\gamma'_{l+1}, \gamma'_{l+2}, ..., \gamma'_k) = (\gamma_{l+1} + |\hat{\delta}_{l+1}\pi_{l+1}|_{m_{l+1}}, \gamma_{l+2} + |\hat{\delta}_{l+2}\pi_{l+2}|_{m_{l+2}}, ..., \gamma_k + |\hat{\delta}_k\pi_k|_{m_k}.$$

**УМ.М6.** В МИМСС с основаниями $m_{l+1}, m_{l+2}, ..., m_k$ сформировать код $(\hat{\gamma}_{l+1}, \hat{\gamma}_{l+2}, ..., \hat{\gamma}_k)$ ЧЧ $\hat{\gamma} = \hat{C}M_l^{-1}$ по правилу $\hat{\gamma}_i = TMPli[\gamma'_i]$ $(i = \overline{l+1, k})$.

**УМ.М7.** Рассчитать интервально-индексную характеристику $\hat{I}'_k(\hat{\gamma})$ ЧЧ $\hat{\gamma}$: $\hat{I}'_k(\hat{\gamma}) = \eta'_k = \left|\sum_{i=l+1}^{k} TII\_i[\hat{\gamma}_i]\right|_{m_k}$.

**УМ.М8.** Расширить МИМК $(\hat{\gamma}_{l+1}, \hat{\gamma}_{l+2}, ..., \hat{\gamma}_k)$ на модули $m_1, m_2, ..., m_l$ с применением соотношения

$$\hat{\gamma}_j = \left|\sum_{i=l+1}^{k-1} TE\_i\_j[\hat{\gamma}_i] + TE\_k\_j[\eta'_k]\right|_{m_j} \quad (j = \overline{1, l}).$$

**УМ.М9.** Число $\hat{\gamma} = \hat{C}M_l^{-1} = (\hat{\gamma}_1, \hat{\gamma}_2, ..., \hat{\gamma}_k)$ зафиксировать в качестве искомого аналога нормированного произведения $\tilde{\gamma} = |ABM_l^{-1}|_p$ операндов $A$ и $B$ по модулю $p$ и завершить работу алгоритма.

Мультипликативные МИМА-процедуры Монтгомери УМ.М1—УМ.М9 предназначены в первую очередь для вычисления степеней натуральных чисел по большим модулям $p$ — целочисленных величин вида $Y = |X^E|_p$, где $X$ и $E$ заданы соответственно модулярным и двоичным кодами: $X = (\chi_1, \chi_2, ..., \chi_k)$ и $E = (e_{s-1}\, e_{s-2}, ..., e_0)_2$ $(e_{s-1} = 1; s$ — разрядность ЧЧ $E)$.

Примем в качестве основы для расчета степеней традиционно применяемый метод умножения с возведением в квадрат (*square multiply method*) [1, 2, 8], который использует мультипликативную декомпозицию функции $Y$:

$$Y = \left|X^{\sum_{j=0}^{s-1} e_j 2^j}\right|_p =$$

$$= |X^{e_0}(X^{e_1}(X^{e_2}(...(X^{e_{s-2}}(X^{e_{s-1}})^2)^2...)^2)^2)^2|_p. \quad (35)$$

Введем для операции умножения по модулю $p$, выполняемой согласно процедурам УМ.М1—УМ.М9, обозначение ММ$(A, B)$ ($A$ и $B$ — операнды, представленные в МСС с основаниями $m_1, m_2, ..., m_k$). Тогда на базе формулы (35) можно сформулировать нижеследующий алгоритм возведения в степень.

**Входные данные:** $X = (\chi_1, \chi_2, ..., \chi_k)$ $(\chi_i = |X|_{m_i})$ $(i = \overline{1, k})$, $X \in \mathbf{Z}_{2p}$; $E = (e_{s-1}\, e_{s-2}, ...\, e_0)_2$ $(e_{s-1} = 1; s \geqslant 1)$.

**Выходные данные:** $Y = (\xi_1, \xi_2, ..., \xi_k)$ $(\xi_i = |Y|_{m_i}$ $(i = \overline{1, k})$, $Y \equiv X^E (\mathrm{mod}\, p)$, $Y \in \mathbf{Z}_{2p}$.

**Предварительно вычисляемые данные:**

$$N = |M_l^2)_p = (\nu_1, \nu_2, ..., \nu_k) \; (\nu_i = |N|_{m_i} \; (i = \overline{1, k})),$$

$$M_l = \prod_{i=1}^{l} m_i \; (1 < l < k).$$

***Тело алгоритма***

**ВС1.** Получить МК $(\chi'_1, \chi'_2, ..., \chi'_k)$ ЧЧ $X' =$ $= \mathrm{MM}(X, N)$.

**ВС2.** Присвоить переменной $Y = (\xi_1, \xi_2, ..., \xi_k)$ начальное значение $Y = X'$.

**ВС3.** Для всех $j = s - 2, s - 3, ..., 0$ выполнить:

а) $Y = \mathrm{MM}(Y, Y)$;

б) если $e_j = 1$, то найти $Y = \mathrm{MM}(Y, X')$.

**ВС4.** Определить МК $(\chi_1, \chi_2, ..., \chi_k)$ искомого значения степени: $Y = MM(Y, 1)$ и завершить работу алгоритма.

Используемая в алгоритме ВС1—ВС4 константа $N$ обеспечивает отсутствие в конечном результате $Y$ коэффициента $M_l^{-1}$ произведений Монтгомери. Требуемый МК этой константы можно получить с помощью синтезированного в работе [22] МИМА-алгоритма деления по схеме Ферма.

В таблице приведены времена выполнения алгоритма УМ.М1—УМ.М9 на ПЭВМ и мультипроцессорным кластером (МПК). Представленные данные получены в предположении, что основания МСС являются 16-битовыми. По сравнению с наиболее близким модулярным аналогом — разработкой фир-

**Времена выполнения МИМА-алгоритма умножения по модулю на основе метода Монтгомери с использованием процессоров Intel Pentium 4 (3ГГц)**

| Параметры алгоритма и базовой МИМСС | | | | Временные затраты на реализацию алгоритма, нс | |
|---|---|---|---|---|---|
| $\lceil \log_2 p \rceil$ | l | k | M | в ПЭВМ | в МПК |
| 64 | 5 | 10 | $1{,}000183 \cdot 2^{132}$ | 295,46 | 62,8 |
| 128 | 9 | 18 | $1{,}000427 \cdot 2^{260}$ | 668,98 | 78,8 |
| 256 | 17 | 34 | $1{,}000916 \cdot 2^{516}$ | 1800,02 | 110,8 |
| 512 | 33 | 67 | $1{,}001893 \cdot 2^{1028}$ | 5750,36 | 176,8 |
| 1024 | 66 | 133 | $1{,}003910 \cdot 2^{2052}$ | 20183,9 | 308,8 |
| 2462 | 157 | 315 | $1{,}009480 \cdot 2^{4928}$ | 105121,48 | 672,8 |

мы Toshiba [8], однопроцессорная программная версия предложенного алгоритма обеспечивает повышение производительности в 3,5—3,6 раз при $p$ разрядностью 1024—2462 бит. В случае мультипроцессорной реализации достигаемый выигрыш в быстродействии, как минимум, 8-кратный. Временные затраты на выполнение операции модульного возведения в степень прямо пропорциональны соответствующим характеристикам применяемых процедур умножения с коэффициентом, примерно составляющим 1,5 $s$ ($s$ — разрядность показателя степени).

## Заключение

Представленная разработка по оптимизации модулярной схемы Монтгомери для умножения по большим модулям показывает, что для решения данной проблемы табличная МИМА обеспечивает принципиально новые возможности. Наиболее важные результаты выполненных исследований состоят в нижеследующем.

- В мультипликативной МА-схеме Монтгомери применен новый способ аддитивной вариации произведения операндов, обеспечивающий сокращение реализационных затрат при расчете базовой ИХМК на первом каскаде схеме в $l/2$ раз.
- Для мощностей диапазонов используемых усеченных МСС получены устанавливаемые в соответствии с разрядностью рабочего модуля $p$ условия, которые гарантируют корректность режима многократного обращения к МИМА-процедуре умножения без выхода результатов за пределы кольца $Z_{2p}$, в том числе и в рамках созданного алгоритма модульного возведения в степень. При этом также достигается уменьшение затрат при формировании ИХМК на втором каскаде схемы в $(k - l)/2$ раз.
- На базе МИМА-схемы Монтгомери синтезирован алгоритм умножения по модулю $p$, имеющий сумматорно-табличную конфигурацию. Для его выполнения требуются лишь операции извлечения вычетов из таблиц и суммирование ЦЧ на позиционных сумматорах стандартной разрядности.
- Для предложенного МИМА-алгоритма умножения по большим модулям приведены оценки минимальных временных затрат на его реализацию как в мультипроцессорном кластере, так и в ПЭВМ. Однопроцессорная программная версия алгоритма при (1024—2462)-битовых $p$ превосходит адекватный вариант наиболее близкого модулярного аналога [8] по производительности в 3,5—3,6 раз. Это относится и к синтезированной процедуре модульного возведения в степень.

### Список литературы


1. **Харин Ю. С., Агиевич С. В.** Компьютерный практикум по математическим методам защиты информации. Мн.: БГУ, 2001. 190 с.
2. **Харин Ю. С.** Математические и компьютерные основы криптологии / Ю. С. Харин и др. // Мн.: Новое знание, 2003. 382 с.
3. **Инютин С. А.** Основы модулярной арифметики. Ханты-Мансийск: Полиграфист, 2008. 208 с.
4. **Posch K. S., Posch R.** Modulo reduction in residue number system // IEEE Trans. on parallel and distributed syst. 1995. Vol. 6. N 5. P. 449—454.
5. **Schwemmlein J., Posch K. S., Posch R.** RNS-modulo reduction upon a restricted base value set and its applicability to RSA cryptography // Comput. and security. 1998. Vol. 17. N 7. P. 637—650.
6. **Bajart J.-C., Didier L.-S., Kornerup P.** An RNS montgomery modular multiplication algorithm // IEEE Trans. Comput. 1998. Vol. 47. N 7. P. 766—776.
7. **Hiasat A. A.** New efficient structure for a modular multiplier for RNS // IEEE Trans. Comput. 2000. Vol. 49. N 2. P. 170—174.
8. **Kawamura S., Koike M., Sano F., Shimbo A.** Cox-Rower architecture for fast parallel Montgomery multiplication // Eurocrypt 2000, LNCS. Berlin, 2000. Vol. 1807. P. 523—538.
9. **Nozaki H., Motoyama M., Shimbo A., Kawamura S.** Implementation of RSA Algorithm Based on RNS Montgomery Multiplication // Proc. Cryptographic Hardware and Embedded Systems (CHES 2001). Sept., 2001. P. 364—376.
10. **Bajard J.-C., Imbert L.** A Full RNS Implementation of RSA // IEEE Trans. Comp. 2004. Vol. 53. N 6. P. 769—774.
11. **Амербаев В. М., Дьячков В. Н.** Модулярная арифметика как криптографический примитив // Юбил. международная науч.-техн. конференция "50 лет модулярной арифметики" (В рамках 5-й Международной научно-технической конференции "Электроника и информатика — 2005"). Зеленоград, РФ 23—25 нояб., 2005. Сб. науч. тр. М., Зеленоград: НИЭТ; М., Зеленоград: АНГСТРЕМ, 2006. С. 187—193.
12. **Lim Z., Phillips B. J.** An RNS-Enhanced microprocessor implementation of public key cryptography // Signals, Systems and Computers. 2007. ACSSC 2007. Conf. Rec. of the forte-first Asilomar Conf. 4—7 nov. 2007. P. 1430—1434.
13. **Shien M.-D., Chen J.-H., Wu H.-S., Lin W.-C.** An new modular exponentiation architecture for efficient design of RSA cryptosystem // IEEE Trans. Very Large Scale Integr. (VLSI) Syst. 2008. Vol. 16. N 9. P. 1151—1161.
14. **Lee K.-J., Yoo K.-J.** Systolic multiplier for Montgomery's algorithm // Integration. 2002. Vol. 32. N 1—2. P. 99—109.
15. **RSA** speedup with residue number system immune against hardware fault cryptanalysis / S.-M. Yen [et al.] // Lect. Notes Comput. Sci. 2002. Vol. 2288. P. 297—413.
16. **Коляда А. А., Пак И. Т.** Модулярные структуры конвейерной обработки цифровой информации. Мн.: БГУ. 1992. 256 с.
17. **Чернявский А. Ф., Коляда А. А., Коляда Н. А., Шабинская Е. В.** Интервально-индексная технология расширения модулярного кода // Электроника инфо. 2010. № 6. С. 72—77.
18. **Коляда А. А., Чернявский А. Ф.** Умножение по большим модулям с использованием минимально избыточной модулярной схемы Монтгомери // Информатика. 2010. № 3. С. 31—48.
19. **Чернявский А. Ф., Коляда А. А., Коляда Н. А., Шабинская Е. В.** Умножение по большим модулям методом Монтгомери с применением минимально избыточной модулярной арифметики // Матер. Всерос. науч. конф. с элементами научной школы для молодежи "Параллельная компьютерная алгебра", Ставрополь, 11—15 окт. 2010 г. // Нейрокомпьютеры: разработка, применение. 2010. № 9. С. 3—8.
20. **Коляда А. А., Чернявский А. Ф.** Общая технология вычисления интегральных характеристик модулярного кода // Доклады НАН Беларуси. 2008. Т. 52. № 4. С. 38—44.
21. **Montgomery P. L.** Modular multiplication without trial division // Mathematics of Computation. 1985. Vol. 170. N 44. P. 519—521.
22. **Коляда А. А., Коляда Н. А., Ревинский В. В., Чернявский А. Ф., Шабинская Е. В.** Умножение по большому модулю в минимально избыточной модулярной системе счисления с применением операций масштабирования // Информатика. 2009. № 4. С. 49—65.

УДК 004.75 + 004.056.5

**И. В. Крупнов**, аспирант, мл. науч. сотр.,
e-mail: krupnov_iv@mail.ru
Институт точной механики и вычислительной техники им. С. А. Лебедева РАН

# Анализ проблем обеспечения информационной безопасности системы электронного голосования в условиях российского информационного пространства

*Представлен анализ проблем обеспечения информационной безопасности информационной системы Интернет-голосования (ОИСИГ), предназначенной для использования в реальных условиях Российской Федерации. Анализируются потенциально слабые звенья системы, предлагаются дополнительные механизмы повышения уровня надежности и защищенности ОИСИГ. Проводится анализ алгоритма контроля состояния голоса, поданного избирателем. По результатам анализа системы предлагаются пути ее модернизации, позволяющие существенно повысить защищенность системы без кардинального изменения используемой программно-аппаратной базы.*

***Ключевые слова:*** *информационная безопасность, электронные выборы, электронное правительство*

## Введение

Идея использования современных информационных технологий для проведения электронных выборов становится все более популярной наравне с другими современными механизмами повышения гражданской активности. Новое поколение избирателей не всегда охотно посещает избирательные участки и с большим желанием готово воспользоваться преимуществами удаленного голосования через Интернет [1].

В работе [2] нами был проведен анализ реализованных к настоящему моменту систем удаленного голосования. Затем по результатам анализа была предложена модель Оптимальной информационной системы Интернет-голосования (ОИСИГ), ориентированная на особенности информационного пространства РФ [3].

В данной статье будет проведен анализ проблем, связанных с обеспечением информационной безопасности при использовании ОИСИГ. По результатам анализа будут предложены те направления развития системы, которые позволят существенно повысить защищенность системы без кардинального изменения предложенной архитектуры и программно-аппаратной базы.

## Средства анализа

ОИСИГ, являясь распределенной информационной системой со многими участниками, должна удовлетворять как общим требованиям безопасности, предъявляемым к информационным системам, так и требованиям, уникальным для систем тайного голосования. Для анализа общих требований мы используем модель угроз STRIDE (spoofing, tampering, repudiation, information disclosure, denial of service, elevation of privilege) и средство анализа SDL (secure development lifecycle) [4]. Уникальные требования тайного голосования наиболее полно сформулированы Брюсом Шнайером [5]. Подход к анализу системы аналогичен ранее использованному автором в работе [2].

Предложенная модель ОИСИГ имеет достаточно много общего с наиболее известной эстонской системой Интернет-голосования (далее ЭС). Напомним отличительные черты ОИСИГ.

1. Для аутентификации избирателя можно использовать документ, идентифицирующий личность с электронным носителем, например паспорт нового поколения (ПНП).

2. Новые избиратели, регистрирующиеся в системе, получают активный криптографический токен.

3. Для усиления защиты при аутентификации с использованием ПНП применяется предварительная SMS (short message service)-регистрация.

4. Для контроля состояния голоса избирателем может быть применен альтернативный используемому в ЭС алгоритм **A1**, основанный на применении времени подачи голоса [3].

5. В рабочем режиме системы устранена зависимость от внешней PKI (public key infrastructure) системы.

Анализ системы будет сосредоточен на этих особенностях.

## Соответствие требованиям общей информационной безопасности

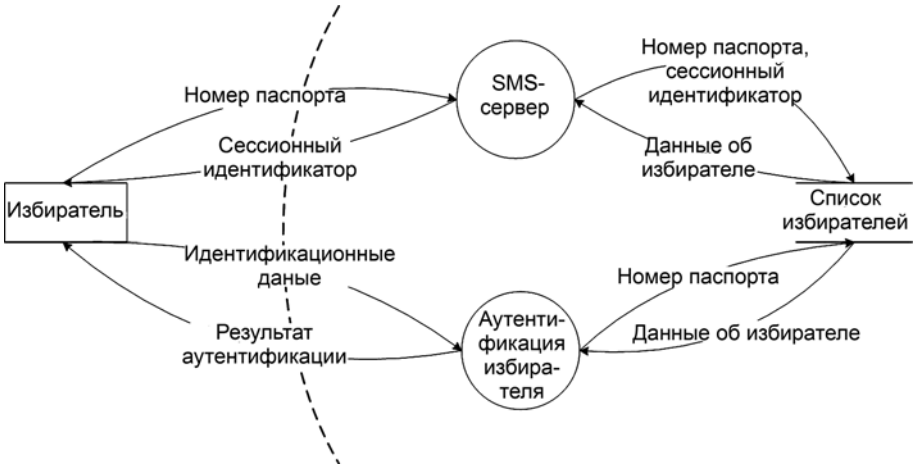
Проанализируем систему с точки зрения общих угроз информационной безопасности. Для данного анализа используем программу SDL Threat Modeling Tool v3.14 [4].

Приведем контекстную диаграмму и анализ процедуры аутентификации избирателя с использованием паспорта (см. рисунок).

*Внешние роли:* избиратель.

*Процессы:* SMS регистрация избирателя, осуществляемая SMS-сервером; аутентификация избирателя, реализуемая сервером перенаправления голосов.

*Хранилища данных:* список избирателей — заранее собранная и полученная в рабочем режиме системы информация по избирателям, имеющим право принимать участие в голосовании; информация включает хеш файла, содержащего фотографию владельца (файл с фотографией) из паспорта, полу-




**Аутентификация с использованием паспорта**

ченную заранее из баз данных системы оформления паспортов, а также сессионный идентификатор, номер паспорта и номер телефона избирателя, зарегистрированные через SMS-сервис.

*Информационные потоки:* номер паспорта; сессионный идентификатор; данные об избирателе — вся информация об избирателе, хранящаяся в системе; идентификационные данные — сессионный идентификатор, номер паспорта, хеш файла с фотографией; результат аутентификации.

*Границы доверия:* граница между внешним участником и внутренней составляющей системы.

*Основные угрозы:* угрозы, специфичные для ОИСИГ проявляются при SMS-регистрации избирателя. Наибольший интерес представляет взаимодействие избирателя и SMS-сервера. Роль "Избиратель" может быть подвержена атаке типа спуфинга (*spoofing*). Нелегальный пользователь может выдать себя за произвольного избирателя, отправив случайный номер паспорта. При этом он будет зарегистрирован вместо легального избирателя и получит его сессионный идентификатор. С учетом логики построения системы проголосовать злоумышленник не сможет, так как не имеет хеша файла с фотографией. Однако в этом случае и легальный пользователь не сможет проголосовать, так как не получит уникальный идентификатор ввиду существующего правила однократной регистрации. Угроза не может быть полностью устранена, но может быть уменьшена. Для этого в SMS-сообщении избирателя SMS-сервису необходимо указать информацию, относящуюся к избирателю, отличную от номера его паспорта, например ФИО. Таким образом, для проведения атаки злоумышленнику придется каким-либо образом получить данную пару: ФИО и номер электронного документа (например паспорта нового поколения). Таким образом, атака не может быть проведена в массовом порядке при отсутствии утечек больших объемов персональной информации избирателей. Помимо этого возможно предусмотреть процедуру повторной SMS-регистрации избирателя при отсутствии поданного избирателем голоса. В любом случае за избирателем всегда остается возможность проголосовать на избирательном участке. В указанных условиях эта угроза несет малые риски для ОИСИГ.

Информационные потоки "Номер паспорта" и "Сессионный идентификатор" подвержены атаке раскрытия информации (*information disclosure*). С помощью доступного по цене оборудования возможно осуществлять прослушивание GSM (Global System for Mobile Communications) сетей [6], в том числе SMS-сообщений. Сессионный идентификатор не представляет интереса для злоумышленника, так как для голосования помимо этого необходим хеш файла фотографии избирателя. Номер паспорта и ФИО (в исправленной версии протокола обмена) могут быть использованы для повторной SMS-регистрации злоумышленника (если данная процедура будет принята). Максимальный вред, который злоумышленник может при этом нанести ОИСИГ, — это неудобство для избирателя при повторной SMS-регистрации. Так как атака не может быть проведена в массовом порядке, угроза ее реализации в системе невысока.

SMS-сервер также может быть подвержен атаке спуфинга. Злоумышленники могут попытаться направить избирателя на нелегальный SMS-сервис для регистрации, в связи с чем персональная информация избирателя может быть скомпрометирована.

Другим вариантом угрозы может стать попытка провести атаку "человек посередине" с использованием поддельного SMS-сервиса. Для предотвращения атаки на легальной стороне необходимо обеспечить защиту от SMS-спуфинга. В этой области существуют готовые решения [7]. Данная защита и контроль номеров, с которых осуществляется регистрация избирателей, ослабят атаку. Она не сможет быть проведена в массовом порядке. Дополнительными способами подавления угрозы могут являться повышение осведомленности избирателей и мониторинг схожих SMS-номеров в целях выявления серверов злоумышленников.

### Анализ алгоритма генерации суррогатного идентификатора

Проведем анализ предлагаемого нами алгоритма генерации суррогатного идентификатора, позволяющего отследить состояние голоса. Напомним, что в ЭС используется алгоритм генерации идентификатора с подмешиванием к выбору избирателя случайного значения, что делает идентификатор уникальным и устраняет возможность восстановления по нему голоса избирателя. Для обеспечения

доверия алгоритму со стороны избирателей необходима публикация исходных кодов системы с возможностью восстановления из них бинарных модулей приложения голосования. В случае невозможности публикации, а также для упрощения контроля предложен алгоритм **A1**, описанный в работе [3]. Алгоритм является одной из реализаций отказуемого шифрования (*deniable encryption*) [8].

Для формального подтверждения корректности данного алгоритма воспользуемся аппаратом БАН-логики [9], применяемой для подтверждения корректности протоколов аутентификации, а также проверки подлинности информации, участвующей в обмене между участниками произвольного протокола.

Опишем нотацию, используемую нами при доказательстве (более полное описание БАН-логики дано в работе [9]).

Символы $P_1$, $P_2$ и $S$ используем здесь для обозначения определенных участников протокола, $X$ и $Y$ — для обозначения утверждений.

*Базовая нотация:*

$P \models X$ — Участник $P$ верит участнику $X$. То есть участник протокола $P$ может действовать так, как если бы утверждение $X$ было истинно. Данная конструкция является центральной для логики.

$P \lhd X$ — Участник $P$ видит $X$. Некто послал $P$ сообщение, содержащее $X$. $P$ может прочитать или повторить (отправить) $X$.

$P \mid\sim X$ — $P$ однажды сказал $X$. Участник протокола $P$ в неопределенный момент времени в прошлом отправил сообщение, включающее утверждение $X$.

$P \Rightarrow X$ — $P$ обладает юрисдикцией над $X$. Участник протокола $P$ имеет полномочия над $X$. Например, часто серверу выделяют юрисдикцию (право) на создание сессионных ключей.

$\#X$ — утверждение $X$ свежее. $X$ не было отправлено ни в одном сообщении до начала текущей сессии протокола.

$P_1 \overset{Y}{\rightleftharpoons} P_2$ — формула $Y$ — секрет, известный только $P_1$ и $P_2$ и, возможно, участникам протокола, которым они верят. Примером секрета может служить пароль.

$\langle X \rangle_Y$ читается, как $X$, объединенное с $Y$. Подразумевается, что $Y$ — секрет, и его присутствие идентифицирует того, кто создал $\langle X \rangle_Y$. Чаще всего используется простая конкатенация $X$ и пароля $Y$.

*Логические постулаты.*

Правило интерпретации сообщения для распределенных секретов:

$$\frac{P_2 \models P_1 \overset{Y}{\rightleftharpoons} P_2,\ P_2 \lhd \langle X \rangle_Y}{P_2 \models P_1 \mid\sim X}.$$

Правило означает, что если $P_2$ верит, что секрет $Y$ известен только ему и $P_1$, и видит $\langle X \rangle_Y$, то $P_2$ верит, что $P_1$ однажды сказал $X$.

Правило верификации свежести сообщения.

Данное правило показывает, что сообщение было создано в текущей сессии протокола и значит отправитель верит ему:

$$\frac{P_2 \models P_1 \mid\sim X,\ P_2 \models \#X}{P_2 \models P_1 \models X}.$$

Правило означает, что если $P_2$ верит, что $P_1$ однажды сказал $X$, и $P_2$ верит, что $X$ свежее, то $P_2$ верит, что $P_1$ верит $X$.

Правило юрисдикции:

$$\frac{P_2 \models P_1 \Rightarrow X,\ P_2 \models P_1 \models X}{P_2 \models X}.$$

Правило означает, что если $P_2$ верит, что $P_1$ имеет юрисдикцию над $X$, и $P_2$ верит, что $P_1$ верит $X$, то $P_2$ верит $X$.

Процедура верификации протокола состоит из трех фаз.

1. Формирование базовых предположений с использованием ранее описанной нотации.

2. Идеализация протокола — запись каждого шага протокола в виде формулы.

3. Анализ протокола. Пошаговый вывод определенного утверждения из базовых предположений с использованием формализованного протокола.

В нашем случае необходимо показать, что при определенных базовых условиях избиратель может проверить состояние своего голоса, но не может убедить злоумышленника в том, за кого он проголосовал.

*Базовые обозначения:*

$P_1$ — избиратель в момент голосования; $P_2$ — тот же самый избиратель в момент проверки поданного им голоса; $S$ — злоумышленник; $Y$ — время голосования; $X$ — поданный голос; $\langle X \rangle_Y$ — суррогатное число, опубликованное на сайте для отслеживания состояния голоса, $\langle X \rangle_Y$ получается простым комбинированием (например, "побитовым или") поданного избирателем голоса ($X$) и времени голосования ($Y$).

***Базовые предположения:***

$P_2 \models P_1 \overset{Y}{\rightleftharpoons} P_2$ — участник $P_2$ верит, что он использовал секрет $Y$ при голосовании и может пользоваться им при проверке.

$P_1 \models \#X$ — $P_1$ верит в свежесть $X$.

$P_2 \models P_1 \models \#X$ — более того, $P_2$ верит в то, что $P_1$ верит в свежесть $X$.

$P_2 \models P_1 \Rightarrow \#X$ — $P_2$ верит, что $P_1$ обладает юрисдикцией на свежесть $X$.

$P_2 \models P_1 \Rightarrow X$ — $P_2$ верит, что $P_1$ обладает юрисдикцией на $X$, т. е. верит, что $P_1$ корректно выбирает $X$.

***Идеализация протокола:***

$P_2 \lhd \langle X \rangle_Y$ — $P_2$ видит $\langle X \rangle_Y$.

***Анализ протокола:***

$$\frac{P_2 \models P_1 \Rightarrow \#X,\ P_2 \models P_1 \models \#X}{P_2 \models \#X},$$

$$\frac{P_2 \models P_1 \overset{Y}{\rightleftharpoons} P_2,\ P_2 \lhd \langle X \rangle_Y}{P_2 \models P_1 |\sim X},$$

$$\frac{P_2 \models P_1 |\sim X,\ P_2 \models \#X}{P_2 \models P_1 \models X},$$

$$\frac{P_2 \models P_1 \Rightarrow X,\ P_2 \models P_1 \models X}{P_2 \models X},$$

это показывает, что проверяющий избиратель доверяет поданному голосу. Что и требовалось доказать.

Попытка избирателя сделать свой голос предметом продажи не увенчается успехом, так как для заинтересованного лица ($S$) не будет выполнено базовое предположение $S \models P_1 \overset{Y}{\rightleftharpoons} S$. Такой результат возможен, поскольку избиратель не может указать (доказательно) момент своего голосования. Следовательно, доказательство утверждения $S \models X$ становится невозможно. Таким образом, требуемое свойство получено.

### Анализ полноты реализации специальных требований

Проанализируем систему на соответствие уникальным требованиям, впервые сформулированным Брюсом Шнайером [5].

1. Участвовать в выборах могут только граждане, имеющие право голоса. "Легитимность голоса" — поддерживается за счет надежной аутентификации пользователя с использованием активного криптографического токена или электронного документа (паспорта нового поколения). Надежность данных способов аутентификации доказана выше.

2. Каждый избиратель может голосовать только один раз. "Единственность выбора" — поддерживается за счет процедуры отмены повторяющихся голосов. Данная функциональность аналогична используемой в ЭС [2].

3. Никто не может установить, за кого проголосовал каждый избиратель. "Тайна голосования" — поддерживается за счет разделения электронных бюллетеней. Аналогично ЭС [2].

4. Никто не может сделать дубликат бюллетеня с волеизлиянием любого избирателя. "Корректность подсчета голосов" — поддерживается за счет процедуры отмены повторяющихся голосов и HTTPS (Hypertext Transfer Protocol Secure) соединений.

5. Никто не может изменить результат голосования любого избирателя. "Защита волеизлияния" — поддерживается за счет HTTPS-соединений и надежной аутентификации избирателя. Для голосования используются подписанные системой ОИСИГ приложения. Драйверы, используемые для работы активного токена, должны быть также подписаны.

6. Каждый избиратель может проверить, что его бюллетень учтен при подведении итогов голосования. "Прозрачность голосования" — достигается за счет использования аналогичного применяемому в ЭС алгоритму контроля состояния голоса избирателем или предложенного ранее альтернативного механизма A1.

7. Всем известно, кто участвовал в голосовании, а кто нет. "Публичность голосования" — может быть достигнута за счет сохранения подписей выделенных из электронных конвертов и публикации общего реестра избирателей.

### Направления модернизации ОИСИГ

Как видно из проведенного анализа, основные угрозы для ОИСИГ будут сосредоточены в области SMS-регистрации пользователя. Для уменьшения угроз нами предложено:

1. Использование в SMS-сообщении избирателя SMS-сервису информации, относящейся к избирателю, но отличной от номера его паспорта, например ФИО.

2. Добавление возможности повторной SMS-регистрации не проголосовавшего избирателя.

3. Повышение осведомленности избирателей.

4. Мониторинг схожих SMS-серверов.

5. Использование специального программно-аппаратного комплекса в целях защиты от SMS-спуфинга.

Данные улучшения позволяют утверждать, что процедура SMS-регистрации для ОИСИГ является надежной в смысле защиты от возможных угроз.

Стоит обратить отдельное внимание на предварительную подготовку информации об избирателе: сбор хешей файлов фотографий избирателей. Эта операция должна быть проведена до начала самого голосования в целях устранения внешней зависимости в рабочем режиме ОИСИГ. Слабое звено системы — это наличие SMS-регистрации. Данная процедура будет неизбежна при отсутствии ключевой пары в электронном удостоверении личности (паспорте нового поколения). При наличии ключевой пары необходимость SMS-регистрации пропадает и ОИСИГ в основном следует ЭС. В то же время при выдаче зарегистрированным избирателям активного криптографического токена нельзя пренебрегать биометрическими данными. Биометрическая информация об избирателе позволит значительно улучшить надежность системы. Например, в ЭС возможна следующая ситуация: глава семейства собирает электронные паспорта со всей семьи и голосует один за каждого члена семьи. Для традиционных выборов это невозможно. Решением этой проблемы

может являться использование биометрической информации, например применение алгоритмов определения подмены пользователя (*liveness detection*) [10] при аутентификации пользователя.

В связи с запуском программы ввода в обращение универсальных электронных карт для граждан РФ рассмотрение особенности реализации ОИСИГ переходит в сугубо практическую плоскость. Если универсальная карта владельца будет содержать пару ключей (открытый/закрытый ключ) и соответствующий им сертификат, зарегистрированный в уполномоченном удостоверяющем центре, а также биометрические данные владельца, то она может быть идеально использована для проведения удаленных Интернет-выборов с применением ОИСИГ.

Публикация исходных кодов системы, независимо от того, какой алгоритм контроля состояния голоса используется, безусловно, повысит доверие ко всей системе. Описываемый здесь алгоритм контроля будет более понятен рядовому избирателю. Отображение времени в приложении для голосования позволит избирателю проголосовать в любой выбранный момент, что повысит степень его доверия к итоговому контрольному числу.

Отдельное внимание стоит уделить проблемным местам системы.

1. Подача избирателем заведомо неверного голоса.

2. Контроль использования закрытого ключа системы, применяемого для расшифровывания поданных обезличенных бюллетеней.

Первая ситуация возможно в случае, когда злоумышленник является законным избирателем и использует стороннее, например написанное им самим, ПО (программное обеспечение) для подачи голоса. Злоумышленник может указать неверное значение в бюллетене, в связи с чем, его голос будет зарегистрирован, но отвергнут при подсчете. Это не является большой проблемой, так как в этом случае будет утерян единственный голос избирателя по его собственной вине. Для полного устранения этой угрозы можно предложить использование алгоритма доказательства с нулевым разглашением.

Для контроля использования основного закрытого ключа системы мы предлагаем рассмотреть стандартную пороговую схему разделения секрета. Алгоритм позволит разделить ключ между несколькими участниками подсчета результатов голосования и уменьшить вероятность его несанкционированного использования.

## Заключение

В данной работе проведен детальный анализ модели ОИСИГ, предложенной в работе [3]. Объектами анализа являлись как возможность выполнения общих требований по безопасности, предъявляемых к распределенным информационным системам, так и специфических требований к системам тайного голосования. Анализ показал, что предложенные в данной работе модификации ОИСИГ снижают общий уровень угроз до допустимого уровня.

Соответствие системы уникальным для тайного голосования требованиям в основном унаследовано от ЭС. Новым элементом является использование алгоритма А1 контроля состояния голоса избирателя. Проведен формальный анализ алгоритма с использованием БАН-логики и доказана его корректность.

Итогом работы стал набор рекомендаций по подбору программно-аппаратной базы для использования в ОИСИГ. В частности показано, что использование электронного документа, удостоверяющего личность (например, паспорта нового поколения), содержащего как ключевую пару, так и биометрическую информацию об избирателе, существенно повышает надежность системы.

### Список литературы


1. **Gerlach J., Gasser U.** Three Case Studies from Switzerland: E-Voting. March 2009.
2. **Крупнов И. В.** Анализ решений в области автоматизации процедуры выборов // Качество, Инновации, Образование, 2011. № 3. С. 50—55.
3. **Крупнов И. В.** Определение оптимальной в рамках РФ информационной системы Интернет голосования // Качество. Инновации. Образование. 2011. № 5. С. 62—68.
4. **SDL** Threat Modeling Tool. URL: http://www.microsoft.com/security/sdl/adopt/threatmodeling.aspx (дата обращения 12.06.2011)
5. **Шнайер Б.** Прикладная криптография. М.: Триумф, 2002. 151 с.
6. **Nohl K., Munaut S.** GSM Sniffing. Chaos Computer Club Congress. 2010.
7. **Global** Networks. URL: http://www.gni.ch/index.php (дата обращения: 12.06.2011)
8. **Canetti R., Dwork C., Naor M., Ostrovsky R.** Deniable Encryption // Lecture Notes in Computer Science. 1997. Vol. 1294. P. 90—104.
9. **Burrows M., Abadi M., Needham R.** A Logic of Authentication // Digital Equipment Corporation. 1989. V. 426, N 12. P. 233—271.
10. **Schuckers S., Hornak L., Norman T.** et al. Issues for Liveness Detection in Biometrics. West Virginia University. 2003.

УДК 519.24 + 519.25

**А. В. Антонов,** д-р техн. наук, проф., декан,
**С. В. Соколов,** ассистент,
**В. А. Чепурко,** канд. физ.-мат. наук, доц.,
e-mail: v.a.chepurko@mail.ru,
Обнинский институт атомной энергетики
НИЯУ МИФИ

# Бутстреп-метод оценки характеристик надежности восстанавливаемых объектов по специфическим данным об отказах

*Предложено применение бутстреп-метода для построения интервальных оценок показателей надежности. Оценивание проводится в предположении неоднородного пуассоновского потока событий. Построенная модель позволяет, к примеру, посчитать интервальные оценки для показателей надежности стареющего оборудования. Разобран пример построения оценки.*

***Ключевые слова:*** *метод "складного ножа", бутстреп-метод, неоднородный пуассоновский поток, ведущая функция потока, среднее прямое остаточное время*

## Введение

В задачах непараметрического оценивания статистических характеристик надежности разрабатываются так называемые перестановочные методы. К ним относятся метод "складного ножа" и бутстреп-метод, являющийся развитием первого из указанных методов. Метод "складного ножа" был первоначально предложен М. Кенуэем [1] для снижения смещения оценок, получаемых по малым выборкам. В дальнейшем метод стал эффективно применяться для построения доверительных интервалов в случае, когда нельзя использовать результаты асимптотической теории [3].

Схема получения оценок с помощью метода "складного ножа" довольно проста. Она состоит в последовательном исключении из зафиксированной выборки одного из наблюдений и оценивании требуемой характеристики на основании так называемых "проколотых" данных. Пусть в результате наблюдений за функционированием группы однотипных объектов получена выборка объема $n$ независимых одинаково распределенных случайных величин: $T_1, T_2, ..., T_n$ с функцией распределения $F(\theta, t)$. В результате последовательного удаления из выборки одного из наблюдений $T_i$, $i = \overline{1, n}$, вычисляем $n$ оценок искомого параметра, например $\hat{\theta}_i$, $i = \overline{1, n}$. На основании полученных оценок $\hat{\theta}_i$ можно рассчитывать искомый показатель надежности $\theta$ и выполнять доверительное оценивание данного показателя. Изложенная схема вычислительной процедуры подробно описана в работе [4].

Суть бутстреп-метода заключается в том, что теоретическое распределение генеральной совокупности заменяется выборочным. Далее происходит переход от одновыборочной схемы наблюдений к схеме многих выборок того же объема, которые извлекаются из первоначальной выборки и имеют распределение, совпадающее с выборочным. Эта процедура позволяет проводить построение выборочного распределения оцениваемого параметра без каких-либо дополнительных предположений и строить непараметрические доверительные интервалы. Бутстреп-метод разрабатывался Б. Эфроном [2, 5]. Поясним суть бутстреп-метода на простейшем примере.

Пусть случайная выборка объема $n$ извлекается из неизвестного распределения $F$: $T_1, T_2, ..., T_n \sim F(\theta, t)$, где $F(\theta, t)$ — функция распределения. Имеется параметр $\theta$, для которого необходимо найти оценку $\hat{\theta}_n = \hat{\theta}_n(T)$, зависящую от выборки $T = (T_1, ..., T_n)$. На основании зафиксированных данных (реализации выборки) $T_1, T_2, ..., T_n$ необходимо выполнить оценивание показателя $\theta$.

Согласно [1] решение проводится следующим образом.

**Шаг 1.** Подбирают непараметрическую гистограммную оценку для $F$. Каждая из зафиксированных случайных величин $T_i$, $i = \overline{1, n}$, реализуется с одинаковой вероятностью. Следовательно, $\hat{F}$ представляет собой ступенчатую возрастающую функцию со скачком $1/n$ в точках $T_i$, $i = \overline{1, n}$.

**Шаг 2.** При фиксированном $\hat{F}$ из первоначальной выборки извлекают случайную выборку $T_1^*$, $T_2^*, ..., T_n^* \sim F$ объема $n$. Полученная выборка называется бутстреп-выборкой. Здесь следует заметить, что на этом шаге не получается перестановочного

распределения, поскольку значения $T_i^*$ отбираются из множества $T_1$, $T_2$, ..., $T_n$ с возвращением. На основании сформированной бутстреп-выборки проводят оценивание

$$\hat{\theta}_1^* = \hat{\theta}(T_1^*, T_2^*, ..., T_n^*).$$

**Шаг 3.** Независимо многократно повторяется шаг 2. При этом получают бутстреп-повторения $\hat{\theta}_1^*$, $\hat{\theta}_2^*$, ..., $\hat{\theta}_r^*$ и определяют требуемые характеристики

$$\hat{\theta}^* = \frac{\sum_{i=1}^{r} \hat{\theta}_i^*}{r}; \quad \hat{\sigma}_\theta^2 = \frac{\sum_{i=1}^{r} (\hat{\theta}_i^* - \hat{\theta}^*)^2}{r-1},$$

где $\hat{\theta}^*$ и $\hat{\sigma}_\theta^2$ — соответственно выборочная средняя и выборочная дисперсия бутстреп-оценок $\hat{\theta}_1^*$, $\hat{\theta}_2^*$, ..., $\hat{\theta}_r^*$; $r$ — число повторений.

Достоинство бутстреп-метода состоит в том, что с его помощью удается весьма эффективно выполнить доверительное оценивание показателей надежности, выражаемых в виде функционала от плотности распределения наработки до отказа.

В ряде работ приводятся процедуры для оценки параметров с учетом цензурированных данных. Так, в работе [7] излагается методика оценки показателей в случае, когда имеет место выборка, содержащая как полные наработки, так и наблюдения, цензурированные справа. В работе [4] описывается бутстреп-метод для оценки характеристик надежности по комбинированной выборке, когда наряду с полными наработками присутствуют цензурированные данные типа цензурированных справа и слева. Применение бутстреп-метода для анализа характеристик надежности можно найти также в работе [6].

## Постановка задачи

Рассмотрим применение бутстреп-метода для случая оценки характеристик надежности восстанавливаемых объектов, когда зафиксированная информация представлена в весьма специфическом виде. Будем рассматривать ситуацию, согласно которой в процессе сбора эксплуатационной информации исследователь не имеет данных о значениях наработок отказавших объектов, имеются лишь сведения о том, что отказы произошли на некоторых интервалах работоспособности. Неизвестно также, какой именно объект из группы однотипных элементов отказал. Пусть под наблюдением находятся $m$ однотипных восстанавливаемых объектов. Будем считать, что известным является лишь число отказавших объектов из некоторой совокупности однотипных устройств (в количестве $m$ изделий), отказы которых распределены по интервалам работоспособности.

В противоположность обычному бутстреп-методу в нашем случае данные об отказах являются сгруппированными по интервалам работоспособности (например, по годам эксплуатации) и наработки объектов до отказа неизвестны. В случае классического применения бутстреп-метода моделируется случайная равномерно распределенная на оси вероятностей (на интервале [0, 1]) величина и затем она проецируется на ось времени с известными наработками. В предлагаемом для рассмотрения случае смоделированную равномерно распределенную случайную величину на оси вероятностей необходимо отобразить на ось временных интервалов с известным числом попаданий наблюдаемой случайной величины в каждый из интервалов, т. е. моделируется полиномиальная выборка с вероятностями, пропорциональными наблюдаемым частотам.

Рассмотрим последовательность действий, выполняемых при проведении доверительного оценивания с помощью бутстреп-метода.

**Этап 1.** Подготовка исходных данных для бутстреп-метода. Имеется $n$ отказов, распределенных по $k$ временным интервалам. Перенумеруем события, определяющие отказы объектов из группы однотипных устройств, в порядке возрастания времени их реализации (проранжируем отказы), т. е. каждому отказу поставим в соответствие индекс $i$. Определим временной интервал, на который попало данное событие. Если на один интервал времени приходится несколько отказов, порядок нумерации событий внутри интервала произвольный. В рассматриваемом в последующем разделе примере проведена демонстрация ранжирования представленных данных.

**Этап 2.** Ось вероятностей делим на $n$ равных непересекающихся интервалов $[0, y_1)$, $[y_1, y_2)$, ..., $[y_n - 1, y_n]$. Моделируем непрерывную равномерно распределенную случайную величину $U_C[0, 1]$ на оси вероятностей. Далее определяем, какой из зафиксированных отказов соответствует смоделированной переменной. Если смоделированная случайная величина $U_C[0, 1]$ принадлежит интервалу $[y_i - 1, y_i)$, это означает, что реализовалось событие с индексом $i$. Повторяем данную процедуру моделирования $n$ раз, получаем выборку номеров событий отказов. Далее наблюдаемое событие должно быть поставлено в соответствие временному интервалу наблюдения, согласно описанной на первом этапе процедуре. Используя полученную с помощью бутстреп-процедуры выборку, вычисляем оцениваемый параметр $\hat{\theta}_1^*$.

**Этап 3.** Повторяем многократно этап 2. Получаем множество оценок $\hat{\theta}_j^*$, $j = 1, ..., r$, где $r$ — число бутстреп-повторений третьего этапа. Упорядочиваем их.

**Этап 4.** Непараметрические оценки границ доверительного интервала определяем следующим образом. Вначале задаем уровень значимости $\alpha$, соответствующий доверительной вероятности $1 - 2\alpha$. Далее определяем границы интервалов, удовлетворяющие следующим соотношениям для заданных $\alpha$:

$$\alpha = \frac{\#(\hat{\theta}_i \leqslant \theta_{low})}{r}; \quad (1)$$

$$1 - \alpha = \frac{\#(\hat{\theta}_i \leqslant \theta_{high})}{r}, \quad (2)$$

где $\alpha$ — уровень значимости; $r$ — число бутстреп-повторений; $\#(A)$ — частота появления события $A$. Таким образом, $\#(\hat{\theta}_i \leqslant \theta_{low})$ и $\#(\hat{\theta}_i \leqslant \theta_{high})$ определяют порядковый номер рассчитанного показателя надежности, для которого впервые выполнится условие $\{\hat{\theta}_i \leqslant \theta_{low}\}$, и номер рассчитанного показателя, для которого впервые выполнится условие $\{\hat{\theta}_i \leqslant \theta_{high}\}$, соответственно. В этом случае значения $\theta_{low}$ и $\theta_{high}$, определенные выражениями (1) и (2), будут характеризовать приблизительные границы доверительного интервала, соответствующие доверительной вероятности $1 - 2\alpha$.

## Математическое обоснование применения бутстреп-процедуры для восстанавливаемых объектов

В теории восстановления широкое практическое применение имеет функция восстановления или ведущая функция потока $\Omega(t)$. Известно, что все числовые и функциональные характеристики потока восстановления выражаются через эту функцию. По сути, $\Omega(t)$ представляет собой математическое ожидание числа отказов системы к моменту времени $t$. Вместе с тем, она является бесконечной суммой функций распределения моментов отказов $T_i = \sum_{j=1}^{i} \Delta_j$, где $\Delta_j$ — независимые одинаково распределенные случайные времена между двумя последовательными точками потока:

$$\Omega(t) = MN(t) = \sum_{j=1}^{\infty} F_{T_i}(t),$$

где $T_i$ — моменты отказов; $F_{T_i}$ — их функция распределения.

В следующей теореме доказывается, что если применить строго монотонное, дифференцируемое преобразование $\Omega(t)$ к рекуррентному потоку с ведущей функцией $\Omega(t)$, то получающийся поток будет простейшим.

**Теорема**. Пусть $\Omega(t)$ — строго монотонная, дифференцируемая ведущая функция потока (ВФП) для неоднородного пуассоновского потока отказов $\{T_i;\ i = 1, 2, ...\}$, где $T_i$ — моменты отказов. Тогда $\{\Omega(T_i);\ i = 1, 2, ...\}$ будет простейшим потоком отказов.

**Следствие.** Пусть $n_t$ — число отказов к моменту времени $t$ рекуррентного потока $\{T_i;\ i = 1, 2, ...\}$, т. е. $n_t = \sum_i I\{T_i \leqslant t\}$ и $\Omega(t)$ удовлетворяет условиям теоремы. Тогда $n_{\Omega^{-1}(t)} = k_t$ — число отказов к моменту времени $t$ простейшего потока $\{\Omega(T_i);\ i = 1, 2, ...\}$. Следовательно, процесс $n_{\Omega^{-1}(t)}$ будет однородным пуассоновским и будет обладать всеми свойствами пуассоновского процесса.

В этом случае $n_{\Omega^{-1}(t_1)}$, $n_{\Omega^{-1}(t_2)} - n_{\Omega^{-1}(t_1)}$, ..., $n_{\Omega^{-1}(t_n)} - n_{\Omega^{-1}(t_{n-1})}$ будут являться независимыми приращениями пуассоновского процесса, и распределение вероятностей будет определяться формулами

$$P(n_{\Omega^{-1}(t_i)} = k_i,\ i = 1, ..., s) =$$

$$= \frac{t_1^{k_1}}{k_1!} \frac{(t_2 - t_1)^{k_2 - k_1}}{(k_2 - k_1)!} \cdots \frac{(t_s - t_{s-1})^{k_s - k_{s-1}}}{(k_s - k_{s-1})!} \mathbf{e}^{-t_s},$$

$$P(n_{\Omega^{-1}(t_s)} = k_s = n) = \frac{t_s^n}{n!} \mathbf{e}^{-t_s}.$$

Тогда условные вероятности частот при условии, что произошло $n$ отказов, будут равны

$$P(n_{\Omega^{-1}(t_i)} = k_i,\ i = 1, ..., s | n_{\Omega^{-1}(t_s)} = n) =$$

$$= \frac{n!}{\prod_{j=1}^{s} (k_i - k_{i-1})!} \prod_{j=1}^{s} \left(\frac{t_j - t_{j-1}}{t_s}\right)^{k_j - k_{j-1}}$$

при $k_0 = 0$ и будут подчиняться полиномиальному распределению.

Заменив $\Omega^{-1}(t_i) = u_i$, получим:

$$P(n_{u_i} = k_i,\ i = 1, ..., s) =$$

$$= \frac{\Omega(u_1)^{k_1}}{k_1!} \frac{(\Omega(u_2) - \Omega(u_1))^{k_2 - k_1}}{(k_2 - k_1)!} \cdots$$

$$\cdots \frac{(\Omega(u_s) - \Omega(u_{s-1}))^{k_S - k_{s-1}}}{(k_s - k_{s-1})!} \mathbf{e}^{-\Omega(u_s)};$$

$$P(n_{u_s} = n) = \frac{\Omega(u_s)^n}{n!} \mathbf{e}^{-\Omega(u_s)}$$

и условные вероятности

$$P(n_{u_1} = k_i,\ i = 1,\ ...,\ s | n_{u_s} = n) =$$

$$= \frac{n!}{\prod\limits_{j=1}^{s} (k_i - k_{i-1})!} \prod_{j=1}^{s} \frac{\Omega(u_j) - \Omega(u_{j-1})}{\Omega(u_s)},$$

где $\Omega(u_0) = 0$.

Окончательно, условное распределение частот возникновения отказов также будет полиномиальным. Таким образом, доказана следующая теорема.

**Теорема**. Условное распределение частот возникновения отказов в неоднородном пуассоновском процессе будет полиномиальным:

$$P(\nu_i = m_i,\ i = 1,\ ...,\ s | n_{u_s} = n) = \frac{n!}{\prod\limits_{j=1}^{s} m_j!} \prod_{j=1}^{s} (p_j)^{m_j},$$

где $p_i = \dfrac{\Omega(u_i) - \Omega(u_{i-1})}{\Omega(u_s)}$ — отношение приращения функции восстановления на промежутке времени $(u_{i-1};\ u_i]$ (причем $u_0 = 0$) к функции восстановления в момент времени $u_s$.

Естественной оценкой вероятностей будет отношение наблюденного числа отказов $\#((u_{i-1};\ u_i])$ на промежутке времени $(u_{i-1};\ u_i]$ к общему числу отказов $n$

$$\hat{p}_i = \frac{\#((u_{i-1};\ u_i])}{\#((0;\ u_s])} = \frac{\#((u_{i-1};\ u_i])}{n}.$$

Оценив вероятности, можно смоделировать полиномиальную выборку, имеющую примерно такое же распределение частот возникновения отказов, как и частоты имеющегося группированного ряда наблюдений отказов восстанавливаемых объектов.

Таким образом, по предложенной методике можно построить доверительный интервал для характеристик надежности, определяемых через плотность распределения наработки.

### Пример оценки прямого остаточного времени при группированных данных об отказах

Покажем возможность проведения процедуры доверительного оценивания бутстреп-методом на примере оценки среднего прямого остаточного времени на основании исходной информации, когда отказы группы однотипных элементов распределены по интервалам работоспособности. Исходную информацию будем записывать в следующем виде: период наблюдения за функционированием группы однотипных объектов будем представлять в виде его конечного разбиения — массива временных промежутков $LR = \{[l_1,\ r_1];\ (l_2,\ r_2];\ ...;\ (l_{s1},\ r_s]\}$, (где $l_i$ и $r_i$ — левая и правая границы $i$-го интервала), на которых имело место случайное число отказов $\nu = (\nu_1,\ \nu_2,\ ...,\ \nu_s)$.

Формула для определения оценки среднего прямого остаточного времени имеет вид [8]

$$MV(t) = \int_t^{\infty} (1 - F_\xi(x))dx + \int_0^t f_\xi(t-u) MV(u) du, \quad (3)$$

где $F_\xi(x)$ — функция и $f_\xi(x)$ — плотность распределения наработки до отказа.

Для восстанавливаемых объектов определение плотности распределения наработки осуществляется путем решения уравнения Вольтерра 2-го рода

$$f_\xi(x) = \omega(t) - \int_0^t f_\xi(u)\omega(t-u)du, \quad (4)$$

где $\omega(t)$ — параметр потока отказов. Таким образом, задача сводится к оценке параметра потока отказов на основании имеющихся данных с последующим применением бутстреп-процедуры.

Рассмотрим числовой **пример**. Предположим, что под наблюдением находятся $m = 4$ однотипных объекта. Информация о функционировании объектов представлена в виде числа отказов, распределенных по годам эксплуатации $\nu = (1, 9, 3, 4, 3, 2, 1, 0, 3, 0, 3, 0, 0, 0, 0, 1, 0, 6, 1, 0, 1, 2, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0)$. Число лет наблюдения, а соответственно, и число интервалов равно $k = 34$, число зафиксированных отказов $n = 42$.

Оценим параметр потока отказов, используя метод ядерного оценивания [9], по формуле

$$\hat{\omega}(t;\ h) = \sum_{i=1}^{k} \frac{v_i}{m(r_i - l_i)} \left( G\left(\frac{t - l_i}{h(i)}\right) - G\left(\frac{t - r_i}{h(i)}\right)\right) + \varepsilon(t),$$

где $G(x) = \dfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} \exp\left(-\dfrac{u^2}{2}\right) du$ — гауссовское ядро; $m$ — число однотипных объектов, формирующих поток; $k$ — объем конечного разбиения (число промежутков); $h(i) = \sigma_1 \sqrt{-0{,}5 v_i + \sum\limits_{i=1}^{k} v_j}$ — параметр локальности (мера, зависящая, в частности, от $\sigma_1$ — среднего квадратического отклонения случайной величины наработки до отказа); $\varepsilon(t)$ — систематическая ошибка оценки параметра потока отказов (ППО), которая определяется выражением [9]

$$\varepsilon(t,\ n,\ m) \approx$$

$$\approx \frac{1}{2a}\left[\operatorname{erfc}\left(\frac{an/m - t}{\sqrt{2n/m}\sigma_2}\right) + \mathbf{e}^{\frac{2at}{\sigma_2^2}} \operatorname{erfc}\left(\frac{an/m + t}{\sqrt{2n/m}\sigma_2}\right)\right],$$

где $n$ — число зафиксированных отказов; $a = \dfrac{m}{n} r_k$ — оценка математического ожидания наработки до отказа; $\sigma_2$ — среднее квадратическое отклонение случайной наработки до отказа.

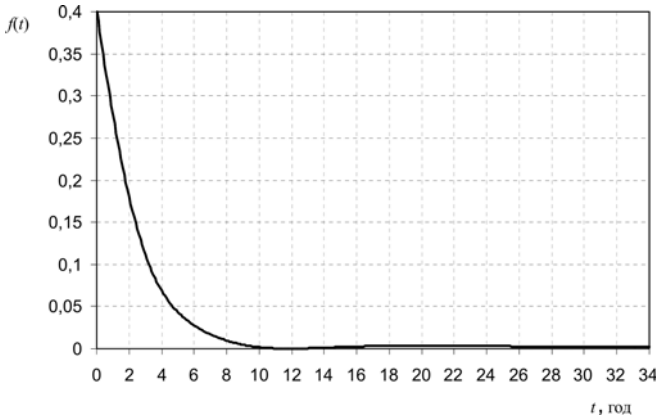



**Рис. 1. Ядерная оценка плотности распределения $f(t)$**

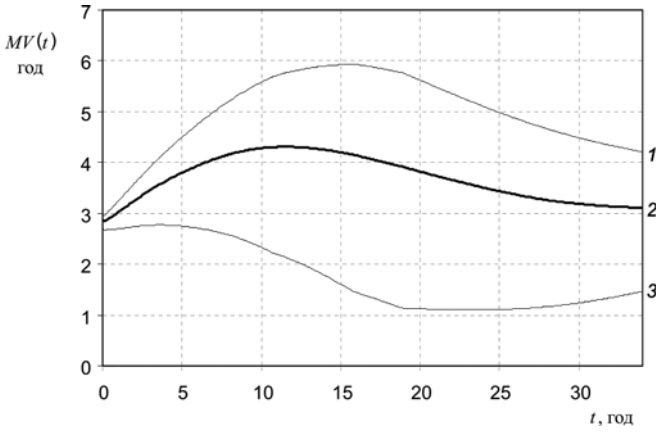



**Рис. 2. Значения доверительного интервала, полученные для среднего прямого остаточного времени:**
*1* — верхняя доверительная граница $MV_{high}(t)$; *2* — среднее прямое остаточное время $MV(t)$; *3* — нижняя граница доверительного интервала $MV_{low}(t)$

В идеальном случае настроечные параметры $\sigma_1$ и $\sigma_2$ равны и оцениваются итерационно.

Решая уравнение (4), определим плотность распределения наработки. При решении задачи необходимо учесть условия, накладываемые на плотность распределения случайной величины, а именно:

- оценка плотности распределения наработки до отказа не должна иметь отрицательных значений;
- должно выполняться условие нормировки — интеграл от плотности распределения должен быть равен единице.

Результат вычисления плотности распределения приведен на рис. 1.

Далее, подставив полученную оценку плотности распределения наработки до отказа в формулу (3), можно оценить математическое ожидание прямого остаточного времени. Результат вычисления представлен на рис. 2.

На последнем шаге проведем вычисление нижней и верхней границы для оценки среднего прямого остаточного времени. Для этого последовательно выполним этапы 1—4, описанные выше. При выполнении этапа 2 при проведении расчетов в данном примере число бутстреп-повторений $r = 1000$. В результате расчетов получим множество оценок $MV^j(t)$, $j = 1, ..., 1000$. Зададим уровень значимости $\alpha = 0{,}01$ и построим доверительный интервал, воспользовавшись формулами (1) и (2). Результаты вычислений верхней и нижней границ доверительного интервала для прямого остаточного времени ($MV_{low}(t)$ и $MV_{high}(t)$) также показаны на рис. 2.

Предложенный метод дает возможность построения доверительного интервала для показателей надежности на основании информации малого объема специфической формы. В рассмотренном примере информация об отказах распределена по годам эксплуатации, причем не было возможности идентифицировать отказавший объект. Известным было только то, что зафиксирован отказ объекта из группы однотипных изделий. Такое представление статистических данных об отказах характерно для большинства компонентов систем ядерных энергетических установок, функционирующих в настоящее время в России.

## Заключение

Рассмотрен бутстреп-метод оценки показателей надежности для случая, когда зафиксированная информация представлена в весьма специфическом виде. А именно, в процессе сбора эксплуатационной информации исследователь не имеет данных о значениях наработок отказавших объектов, имеются лишь сведения о числе объектов, находящихся под наблюдением, и о том, что отказы произошли на некоторых интервалах работоспособности. Доказана теорема, обосновывающая возможность применения бутстреп-процедуры для восстанавливаемых объектов. Показано, что по предлагаемой методике можно построить доверительный интервал для характеристик надежности, функционально связанных с параметром потока отказов или плотностью распределения наработки до отказа. Приведен пример оценки прямого остаточного времени реального объекта, информация об отказах которого представлена в виде отказов распределенных по интервалам работоспособности.

### Список литературы

1. **Quenouille M.** Approximate tests of correlation in time series // J. Roy. Statist. Soc. Ser. 1949. B. 11. P. 18—84.
2. **Эфрон Б.** Нетрадиционные методы многомерного статистического анализа. М.: Финансы и статистика, 1988. 261 с.
3. **Мостеллер Ф., Тьюки Дж.** Анализ данных и регрессия. М.: Финансы и статистика, 1982. Вып. 1.
4. **Антонов А. В., Острейковский В. А.** Оценивание характеристик надежности элементов и систем ЯЭУ комбинированными методами. М.: Энергоатомиздат, 1993. 368 с.
5. **Efron B.** Censored data and bootstrap // J.A.S.A. 1976. P. 312—319.
6. **Belyaev Yu. K.** Resampling and bootstrap methods in analysis of reliability data // Safety & Reliability. ESREL. 2001. P. 1877—1882.
7. **Grabski F., Zaleska-Fornal A.** Bootstrap methods for the censored data in empirical Bayes estimation of the reliability parameters // RT & A, #2 (17). 2010. Vol. 1. P. 115—121.
8. **Antonov A., Sokolov S.** Assessment of residual lifetime of NPP equipment based on operational information specific type // Third International Conference on Accelerated Life Testing. ALT'2010. Clermont-Ferrand. 2010. P. 85—90.
9. **Чепурко В. А.** Ядерная оценка параметра потока отказов // Диагностика и прогнозирование состояния сложных систем. Сб. науч. тр. Обнинск: ИАТЭ. 2004. С. 19—31.

# ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ЭКОНОМИКЕ И УПРАВЛЕНИИ

УДК 15.519.876

**Р. А. Караев**, проф., руководитель лаборатории,
Институт кибернетики НАН Азербайджана,
e-mail: karayevr@rambler.ru,
**Р. Г. Гюльмамедов**, доц.,
Азербайджанский государственный
экономический университет,
**Н. Ю. Садыхова,** науч. сотр.,
Институт кибернетики НАН Азербайджана,
**М. А. Нагиев**, ген. директор,
Консалтинговая компания USTAD LLC

# Индикаторы состояния и факторы развития ИКТ-сектора регионов

*Отмечается важность своевременного и точного представления данных о состоянии ИКТ-сектора регионов для принятия директивных и инвестиционных решений в условиях современной информационной экономики. Рассматриваются приоритетные задачи и методы мониторинга ИКТ-сектора. Предметное обсуждение вопроса дается на примере проекта мониторинга ИКТ-сектора Республики Азербайджан.*

***Ключевые слова**: информационно-коммуникационные технологии, индикаторы состояния, факторы развития, регион*

## Введение

Высокий потенциал информационных и коммуникационных технологий (ИКТ) как катализатора социально-экономического развития сегодня неоспорим. Международная практика [1—5] убедительно свидетельствует о том, что ИКТ способствуют повышению производительности и экономическому росту, однако как именно и в какой степени, остается по-прежнему предметом дискуссий. В решении этого актуального вопроса наметились несколько подходов, но первоочередными на данном этапе являются два:

- разработка системы индикаторов состояния и факторов развития ИКТ;
- анализ качества динамики этих показателей [3, 4].

Своевременные и точные сведения о состоянии ИКТ необходимы для принятия взвешенных решений в приоритетных областях экономики и соответствующего распределения ресурсов. В условиях, когда важность включения проблематики ИКТ в национальные стратегии социально-экономического развития получила широкое признание, дефицит данных об ИКТ является серьезной преградой для эффективного планирования. Информация о распространении и использовании ИКТ помогает в принятии продуманных деловых и инвестиционных решений. В докладе ЮНКТАД [4] отмечается, что на национальном уровне важно поддерживать проекты по сбору данных об ИКТ, с тем чтобы обеспечить долгосрочное международное сотрудничество в деле количественного анализа ИКТ для целей соответствующей директивной деятельности.

В настоящей статье приводятся результаты разработки такого проекта, осуществленного в Республике Азербайджан. Формат исследований, выполненных в составе проекта, а также некоторые из полученных оценок могут, на наш взгляд, представить интерес для коллег, занятых в данной области.

## Общая характеристика проекта мониторинга

Отличительной особенностью Республики Азербайджан является большое внимание, уделяемое ИКТ-сектору высшим руководством страны, наличие развитой ИКТ-инфраструктуры и высокие темпы внедрения ИКТ во все сферы общественно-политической жизни. В настоящее время ИКТ-сектор страны теснейшим образом взаимодействует с лидерами мирового рынка информационных технологий (ИТ) — компаниями Microsoft, HP, ACER, APC, Eaton, TrippLite, Gemalto, Nokia-Siemens, Cisco и др. По оценкам Международного телекоммуникационного союза (ITU), представленным на Международном экономическом форуме в Давосе, страна является одним из лидеров в сфере ИКТ среди стран СНГ. В рамках программы "Электронный Азербайджан" и "Электронное правительство" планируется компьютеризация всех государственных органов страны. Азербайджан выступил инициатором проекта "Трансевразийская суперинформационная магистраль", предусматривающего создание современной региональной высокоскоростной сети между Востоком и Западом, охватывающей более 20 стран Европы и Азии. На фоне динамично развивающейся экономики страны, улучшения уровня жизни населения и роста спроса на информационные услуги вопрос оценки состояния ИКТ-сектора и определения ключевых факторов его развития при-

обретает важное государственное значение. Вместе с тем было ясно, что решение вопроса не может ограничиться общими рекомендациями, содержащимися в рамочных документах МСЭ, Евростат и ОЭСР [6], и необходимо детальное изучение местных условий, определяющих текущий потенциал ИКТ-сектора и возможности его дальнейшего развития.

Цель проекта состояла в исследовании ключевых индикаторов, отражающих общее состояние ИКТ-сектора, а также в выявлении основных факторов макро- и микросреды, определяющих возможности развития ИКТ в различных отраслях.

Методическую базу проекта составили методы экспертного опроса (интервьюирование, анкетирование, метод круглого стола), а также широко известные методы стратегического анализа: SWOT-анализ и PEST-анализ.

**SWOT-анализ** направлен на выявление сильных и слабых сторон ИКТ-проектов предприятий в их взаимодействии с угрозами и возможностями внешней среды (*Strengths* — сильные стороны, *Weaknesses* — слабые стороны, *Opportunities* — возможности, *Threats* — угрозы). SWOT-анализ дает возможность установить внутренние факторы ("слабые стороны" предприятий) и внешние факторы ("угрозы" внешней среды), препятствующие достижению максимальных результатов от использования ИКТ.

**PEST-анализ** предназначен для выявления ведущих политических, экономических, социальных и технологических факторов макросреды (*Policy* — политика, *Economy* — экономика, *Society* — общество, *Technology* — технология), которые потенциально могут оказывать влияние на эффективность ИКТ-проектов предприятий. Результаты идентификации этих факторов и оценка их значимости для предприятий составляют важную характеристику макроэкономической среды конкретного региона, определяющую возможности реализации ИКТ-потенциала региона.

В качестве источников использовались аналитические обзоры Всемирного банка [5], Комитета ОСЭР по ИКТ [3], информационный документ "Реализация преимуществ ИКТ и экономический рост в Европе", подготовленный группой Economist Intelligence Unit (http://www.eiu.com).

В процессе проведения исследований наряду с количественными и качественными показателями, дающими общую характеристику ИКТ-сектора [1], нас интересовал ряд ключевых вопросов, критически важных сегодня для большинства развивающихся стран:

- мнение менеджмента предприятий (в первую очередь, малых и средних) относительно роли и возможностей ИКТ, а также степень использования ИКТ в текущей деятельности предприятий;
- степень важности взаимоотношений между финансовыми и ИКТ-подразделениями предприятия для достижения бизнес-целей;
- основные внутренние и внешние факторы, препятствующие достижению максимальных результатов от внедрения ИКТ на предприятиях;
- степень использования ИКТ в торговых операциях, проводимых предприятиями;
- степень соответствия ИКТ-проектов бизнес-целям предприятий;
- субрегиональные показатели ИКТ, отражающие степень "цифровой асимметрии" страны.

Значительное внимание в проекте уделялось вопросу оценки "качества бизнес-среды" — "условий ведения бизнеса". Экономическая практика [1], а также исследования Комитета ОЭСР по ИКТ [3] и группы Economist Intelligence Unit показывают, что "условия ведения бизнеса" играют чрезвычайно важную роль при реализации потенциальных возможностей ИКТ. Сводные перечни таких условий представлены в указанных отчетах. Наряду с этим группой Economist Intelligence Unit предложена рейтинговая модель "условий ведения бизнеса", в основе которой лежит совокупный индекс (со шкалой от 1 до 5), основанный на показателях, сгруппированных по следующим категориям:

1) политика по отношению к частным предприятиям;
2) условия финансирования;
3) налоговый режим;
4) макроэкономическая обстановка;
5) рынок труда.

Однако эти показатели носят весьма общий (справочный) характер. Международная практика показывает, что при обследовании конкретных регионов повсеместно возникает необходимость в корректировке этих показателей с учетом местных условий, которые могут быть определены с помощью PEST-анализа. Результаты идентификации условий ведения бизнеса, полученные с помощью PEST-анализа, и оценка значимости этих условий для различных отраслей составляют важную характеристику макросреды, определяющую возможности развития ИКТ-сектора конкретного региона.

В качестве интегральных индикаторов состояния ИКТ-сектора были приняты:

- широко используемые в международной практике сводные индексы развития ИКТ, а также динамика изменения этих индексов за установленный период времени;
- индексы развития ИКТ по отдельным экономическим районам региона, сопоставительный анализ которых дает представление о степени "цифровой асимметрии" информационного пространства региона.

В выполненном проекте индексы развития ИКТ определялись по методике Economist Intelligence Unit, ориентированной на прогнозные оценки. Индекс ИКТ для составления прогнозов отличается от индекса ИКТ для оценки экономического роста. Переменные ИКТ зависят от периода времени, поэтому их непросто сравнивать из-за технологиче-

ских перемен. По этой причине группой Economist Intelligence Unit предложен более сложный показатель использования инфраструктуры ИКТ на период прогноза. Для этого в индексе ИКТ используются некоторые из полученных группой качественных показателей "готовности к электронным взаимодействиям" (*e-readyness*).

Сводный индекс развития ИКТ вычисляется как среднее арифметическое комплекса первичных показателей: распространенности фиксированных телефонных линий (число линий на 100 человек); распространенности мобильных телефонов (на 100 человек); числа персональных компьютеров (на 100 человек); числа пользователей сети Интернет (на 100 человек); числа Интернет-серверов на 1 млн человек; распространенности широкополосного доступа (на 10 000 человек), и четырех качественных показателей из рейтингов "готовности к электронным взаимодействиям": качество Интернет-подключений; развитость электронного бизнеса; развитость онлайновой коммерции; знакомство населения с Интернетом ("Интернет граммотность").

Каждый показатель преобразовывался в балльную оценку (по шкале от 1 до 100) исходя из минимального и максимального значений этого показателя в выборке региона.

## Результаты мониторинга

Как показал опрос, проведенный среди 72 руководителей и сотрудников финансовых подразделений предприятий из различных отраслей экономики (табл. 1, рис. 1), только 24 % (A) опрошенных (из 72) уверены, что затраты на ИКТ положительно влияют на производительность, и это влияние может быть измерено; 42 % (B) признают положительное влияние, но считают, что это влияние не может быть измерено; 10 % (C) не находят связи между затратами на ИКТ и производительностью; остальные 24 % (D) не уверены в ответе.

Опрос показал также, что взаимоотношения между финансовыми и ИКТ-подразделениями предприятия для достижения бизнес-целей "очень важны" с точки зрения 25 % (четверти) опрошенных; еще 70 % оценивают их как "важные"; 4 % частично признают эту точку зрения.

Оценка соответствия ИКТ-проектов бизнес-целям предприятий показала, что для большинства из них разрабатываемые и внедряемые ИКТ-проекты напрямую не отвечают коммерческим целям руководства и владельцев предприятий (рис. 2).

Все еще низкой остается степень использования ИКТ в торговых операциях, проводимых предприятиями (рис. 3).

Результаты SWOT-анализа, направленного на выявление сильных и слабых сторон ИКТ-проектов предприятий в их взаимодействии с угрозами и возможностями внешней среды, позволили установить внутренние факторы ("слабые стороны" предприятий) и внешние факторы ("угрозы" внеш-

Таблица 1

**К какой отрасли экономики принадлежит ваше предприятие? (в процентах респондентов)**

| | |
|---|---|
| Финансовые услуги (банки, страховые компании) | 7 |
| Услуги в сфере связи и телекоммуникаций | 5 |
| Услуги профессионального характера (врачи, адвокаты, брокеры по операциям с недвижимостью) | 8 |
| Производство стройматериалов | 9 |
| Здравоохранение | 4 |
| Торговля и общественное питание | 11 |
| Наука, образование и обучение | 8 |
| Транспорт и транспортные услуги | 8 |
| Индустрия спорта, туризма, отдыха, реклама, СМИ, издательская деятельность | 6 |
| Машиностроение | 4 |
| Химические и нефтехимические предприятия | 7 |
| Топливная промышленность (добыча нефти и газа, нефтепереработка, торговля нефтепродуктами, АЗС) | 10 |
| Электроэнергетика (ТЭЦ, ГЭС, торговля электроэнергией) | 7 |
| Прочее | 9 |

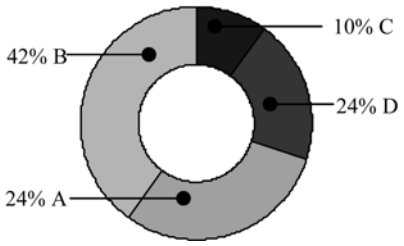



**Рис. 1. Мнение менеджмента предприятий относительно роли ИКТ**

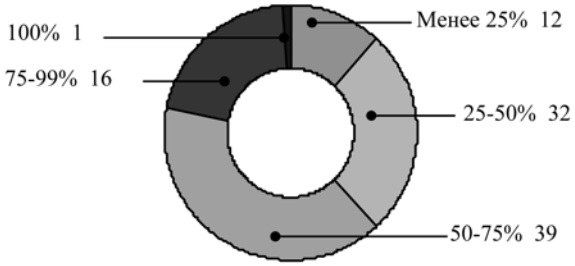



**Рис. 2. Диаграмма соответствия ИКТ-проектов бизнес-целям предприятий**


**Рис. 3. Диаграмма использования сети Интернет в торговле**

ней среды), препятствующие достижению максимальных результатов от использования ИКТ. Эти результаты приведены в табл. 2 и 3.

Результаты идентификации условий ведения бизнеса ("качество бизнес-среды"), выполненной с помощью PEST-анализа, и полученные оценки значимости этих условий для предприятий-респондентов, приведены в табл. 4.

В заключительной части статьи мы приводим:

- ключевые индикаторы состояния ИКТ в Республике Азербайджан, широко используемые в международной практике, а также динамику изменения этих индексов за период 2005—2010 гг. (рис. 4);
- сводные индексы развития ИКТ по отдельным экономическим районам Республики Азербайджан (рис. 5, см. третью сторону обложки).

При расчете этих индексов, наряду с авторскими данными [6], использовались также отчеты Госкомстата Азербайджана (Статистические ежегодники) [7, 8], пресс-релизы и доклады Министерства связи и информационных технологий Азербайджана.

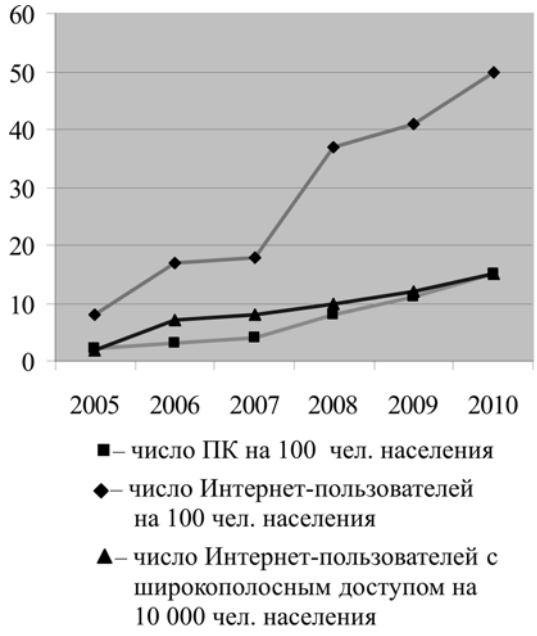



**Рис. 4. Ключевые индикаторы состояния ИКТ**

Представленное на рис. 5 распределение свидетельствует о существенной "цифровой асимметрии" национального информационного пространства. Оно должно быть преодолено на второй фазе Государственной программы "E-Азербайджан" [6]. Программа осуществляется под пристальным вниманием и при всемерной поддержке со стороны высшего руководства страны и продиктована растущим спросом на ИКТ, динамичным развитием экономики страны и улучшением уровня жизни населения.

Таблица 2

**Каковы основные внутренние факторы, препятствующие получению максимальных результатов от использования ИКТ? (в процентах; респонденты могли указать не более двух ответов)**

| | |
|---|---|
| Недостаток ИКТ-навыков у менеджеров высшего звена | 37 |
| Отсутствие эффективного взаимодействия руководителей коммерческих и ИТ-подразделений | 32 |
| Недостаточная интеграция различных технологий в бизнесе | 31 |
| Ценовые ограничения | 39 |
| Неспособность эффективно работать с данными и использовать их в бизнес-проектах | 23 |
| Недостатки в планировании и/или реализации ИКТ-проектов | 26 |
| Сопротивление сотрудников внедрению ИКТ | 22 |
| Прочее | 7 |

Таблица 3

**Каковы основные внешние факторы, препятствующие получению максимальных результатов от использования ИКТ? (в процентах респондентов)**

| | |
|---|---|
| ИКТ слишком часто не соответствуют потребностям бизнеса | 36 |
| Нехватка единых технических стандартов | 33 |
| Плохая ИКТ-инфраструктура | 31 |
| Недостаток прозрачности стоимости владения для информационных технологий | 26 |
| Недостаток профессиональной квалификации в сфере ИКТ | 28 |
| Высокая скорость устаревания ИКТ | 19 |
| Плохое послепродажное обслуживание со стороны поставщиков ИКТ | 15 |
| Недостаток стимулов, поощряющих инновации и инвестиции в ИКТ | 24 |
| Ограничивающие методы работы | 14 |
| Последствия автоматизации рабочих мест в глазах общественности | 8 |
| Прочее | 3 |

Таблица 4

**Насколько важны перечисленные «условия ведения бизнеса» для того, чтобы ваше предприятие могло получить пользу от ИКТ? (в процентах респондентов)**

| Условия ведения бизнеса | Очень важно | Важно | Средне | Неважно | Очень неважно |
|---|---|---|---|---|---|
| Свободный от регулирования и конкурентный рынок ИКТ | 40 | 30 | 19 | 8 | 3 |
| Государственная политика, способствующая распространению ИКТ среди потребителей | 27 | 35 | 25 | 11 | 2 |
| Инвестиционное и налоговое стимулирование предприятий, активно использующих ИКТ | 30 | 35 | 27 | 6 | 2 |
| Действенная законодательная база и правоприменение, защищающие электронную торговлю | 41 | 25 | 27 | 6 | 1 |
| Эффективные законы, защищающие интеллектуальную собственность и сетевые ресурсы юридических и физических лиц | 45 | 26 | 15 | 3 | 2 |

## Заключение

Международный союз электросвязи оценил Азербайджан как одну из быстроразвивающихся стран по ИКТ (http://www.regnum.ru/news/1133149.html?forprint). За последние три года темпы доходного роста в ИКТ-секторе страны примерно в 2,5—3 раза опережают темпы всемирного развития и в среднем ежегодно составляют 30—35 % (http://www.bakutel.az/2011/?p = index). В настоящее время по темпам роста сектор занимает в Республике Азербайджан второе место после нефтегазового сектора. Ожидается, что такая тенденция будет продолжена, и в течение нескольких лет доход в ИКТ-секторе приблизится к доходам от нефтегазового сектора, а в 2018—2020 гг. уже и превысят их (http://news.day.az/hitech/165073_print.html). Это обстоятельство делает проблему мониторинга ИКТ-сектора еще более актуальной и ставит вопрос долгосрочного оптимального управления сектором в контексте стратегических приоритетов страны, связанных с поэтапным переходом от ресурсно-экспортной экономики к ресурсно-инновационной и затем к инновационной экономике. Эта стратегическая линия является общей для стран СНГ, являющихся крупными экспортерами углеводородного сырья, — России, Казахстана, Азербайджана.

## Список литературы


1. **Штрик А. А.** Использование информационно-коммуникационных технологий для экономического развития и государственного управления в странах современного мира // Информационные технологии. Приложение. 2009. № 6. 32 с.
2. **Международная** конференция "ЮНЕСКО между двумя этапами Всемирного саммита по информационному обществу". 2005. URL: http://confifap.cpic.ru/conf2005/rus/info/progr_ru.htm
3. **Организация** экономического сотрудничества и развития (ОЭСР). Наука, технологии и промышленность: перспективы 2008. URL: http://www.oecd.org/html
4. **ЮНКТАД.** Конференция ООН по торговле и развитию. Доклад об информационной экономике. 2010. URL: http://www.unctad.org/Templates/webflyer.asp?docid = 13912&intItemID = 3594&lang = 1.
5. **IC4D.** Information and Communications for Development 2009: Extending Reach and Increasing Impact**.** (IC4D is a regular publication of the World Bank on the critical role of information and communication technology (ICT) in economic development). URL: http://go.worldbank.org/DMY979SNP0.
6. **ООН.** Экономический и Социальный Совет. Статистическая комиссия. Доклад Партнерства для статистического измерения информационно-коммуникационных технологий в интересах развития: статистика информационно-коммуникационных технологий. 2009. URL: http://www.itu.int/ITU-D/ict/partnership/material/2009-19-ICT-R.pdf
7. **Gulmamedov R. H.** The National ICT Strategy and Key Indicators on Information Society of Azerbaijan Republic // Proceedings of International Conference on e-Government Sharing Experiences. Antalya. Turkey, 8—11 December 2009. P. 373—381.
8. **Телекоммуникация** и связь в Азербайджане. Статистический ежегодник. Госкомстат Республики Азербайджан. 2010. 356 с.
9. **Информационное** общество в Азербайджане. Информационные и коммуникационные технологии. Статистический ежегодник. Госкомстат Республики Азербайджан. 2010. 147 с.


---


УДК 004.94



**Д. В. Капулин**, канд. техн. наук, доц.,
e-mail: kapulin@gmail.com,
Институт космических и информационных технологий Сибирского федерального университета, г. Красноярск


# Прикладное решение по подготовке информации о бизнес-процессах для платформы 1С: Предприятие с использованием *ERwin Process Modeler*


*Предложено универсальное решение по преобразованию моделей бизнес-процессов, созданных с помощью ERwin Process Modeler в модели бизнес-процессов 1С: Предприятие. Применение предложенного решения позволяет осуществлять поддержку и обновление бизнес-процессов, не изменяя конфигурации ERP-системы.*



***Ключевые слова:*** *проектирование информационных систем, процессный подход, структурный анализ и проектирование, XML*


## Введение

Своевременная обработка информации способствует совершенствованию организации производства, оперативному и долгосрочному планированию, прогнозированию и анализу хозяйственной деятельности. Каждая организация стремится минимизировать затраты времени, материальных, трудовых ресурсов в ходе своей деятельности и упростить процесс обработки информации. Эти задачи можно решить с использованием автоматизированных информационных систем.

Широкое распространение в управлении производством получили комплексные информационные системы класса ERP (*Enterprise Resource Planning System* — система управления ресурсами предприятия). В основе ERP-системы лежит принцип создания единого хранилища данных, содержащего всю деловую информацию, накопленную организацией в процессе ведения бизнеса, в частности, финансовую информацию, данные, связанные с производством, управлением персоналом и др. Стандарт ERP позволяет объединить все ресурсы предприятия и повысить эффективность управления ими.

Таким образом, использование комплексных информационных систем становится неотъемлемой составляющей функционирования организаций.

В связи с этим большую актуальность приобретает освоение принципов построения и эффективного применения соответствующих технологий и программных продуктов: систем управления базами данных, CASE-средств проектирования и др.

## Постановка задачи

Сегодня большинство российских предприятий использует ERP-системы и продолжает работать по принципам структурного подхода [1]. В этом случае организация и управление деятельностью осуществляется по структурным элементам (отделы, департаменты и т. п.), а взаимодействие структурных элементов — через должностных лиц и структурные подразделения более высокого уровня. Применение структурного подхода к управлению деятельностью предприятия на протяжении многих лет выявило ряд существенных недостатков, которые заметно влияют на рост его экономических показателей. В настоящее время стало очевидным, что для оптимизации работы предприятия в современных условиях, для повышения его финансовых, качественных и внутренних показателей эффективности применение структурно-функционального подхода не является достаточным условием. Повышение эффективности бизнеса требует "интеллектуального" управления, которое может быть достижимо использованием процессного подхода [2, 3].

Под процессом понимается совокупность взаимосвязанных или взаимодействующих видов деятельности, которые преобразуют входы в выходы [3]. В [3] также дается пояснение, что любая деятельность или совокупность видов деятельности, которая использует ресурсы для преобразования входов в выходы, может рассматриваться как процесс. Процессный подход позволяет серьезным образом повысить конкурентоспособность предприятия, сделать его более адекватным к изменениям на рынке, принципиально улучшить качество продуктов и услуг. Он заставляет устранить фрагментарность в работе, организационные и информационные разрывы, дублирование функций, нерациональное использование материальных и людских ресурсов, а также значительно сократить операционные издержки.

Успешное внедрение процессного подхода — задача многошаговая, занимающая достаточное количество времени, трудовых и материальных затрат. Здесь принципиально важно использовать профессиональные инструментальные средства, позволяющие описывать и анализировать бизнес-процессы, делать их более прозрачными и управляемыми. Одним из таких продуктов, реализующих методологию структурного анализа и проектирования, является *ERwin Process Modeler* [4] — инструмент для моделирования, анализа, документирования и оптимизации бизнес-процессов.

В качестве представителя ERP-систем широко внедряются программные продукты, разработанные на платформе *1С: Предприятие*. Эта платформа является универсальной системой автоматизации экономической и организационной деятельности предприятия, а наличие в ней такого важного свойства, как конфигурируемость (т. е. возможность гибкой настройки и модификации прикладных решений под особенности конкретного предприятия и класс решаемых задач), делает платформу *1С: Предприятие* высокоэффективной, расширяемой, конкурентоспособной [5].

Платформа *1С: Предприятие* имеет свой базовый набор инструментов по работе с управлением процессами (объект конфигурации "Бизнес-процессы"), но не имеет достаточных возможностей для описания структуры всех процессов и подпроцессов предприятия и того функционала, который заложен в *ERwin Process Modeler*. Кроме того, работа с бизнес-процессами в *1С: Предприятие* заложена только на уровне программного кода, т. е. возможно запрограммировать действующие процессы на предприятии, но возможность модифицировать и редактировать схемы их функционирования в пользовательском режиме отсутствует. Какие-либо инструментальные средства, позволяющие организовать подготовку информации о бизнес-процессах в *ERwin Process Modeler* и осуществить дальнейшее ее использование в *1С: Предприятие*, также широко не известны.

Таким образом, создание универсального решения, позволяющего импортировать модели бизнес-процессов из *ERwin Process Modeler* в среду *1С: Предприятие* при проектировании информационных систем, является актуальной задачей. Реализация такого решения позволит:

- уменьшить временны́е и иные затраты на определение, внедрение, запуск в работу, мониторинг и оптимизацию бизнес-процессов предприятия;
- обеспечить управление изменениями в существующих моделях бизнес-процессов;
- унифицировать процессы в соответствии со стандартами предприятия и сократить время отклика на изменения в процессах при управлении ими.

## Методика решения и анализ результатов

Комплексное описание организации состоит из множества взаимосвязанных моделей различных типов. Для описания алгоритма выполнения бизнес-процесса, отдельного сценария в виде последовательности процедур, их начальных и конечных событий, а также ссылок на смежные процессы требуется создать модель сценария бизнес-процесса (IDEF3-модель). В этой модели главное внимание уделяется логической последовательности выполнения процедур, составляющих данный сценарий.

Пример модели сценария типового бизнес-процесса, представленного с помощью *ERwin Process Modeler*, приведен на рис. 1. Сценарная модель бизнес-процесса отражает последовательность действий на предприятии в рамках данного процесса и позволяет назначить владельца процесса, распре-

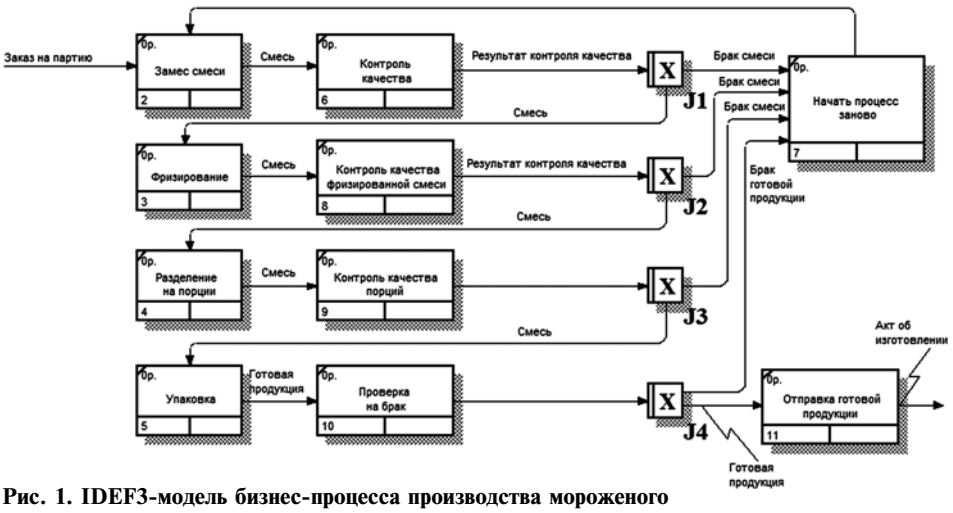
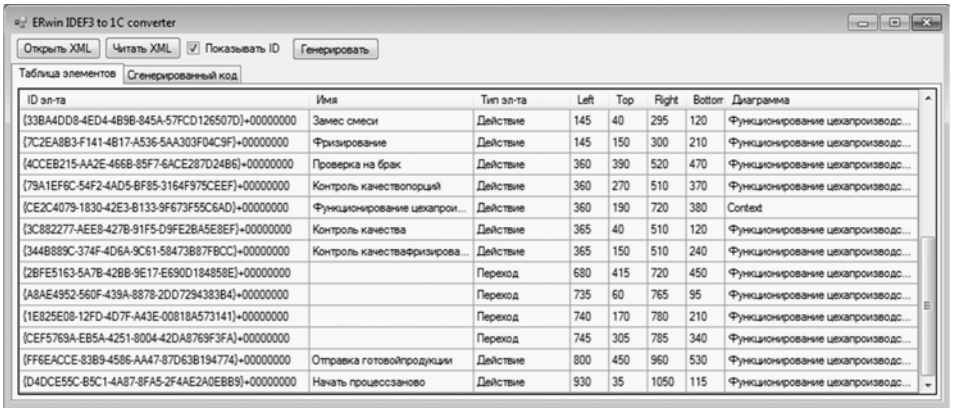



**Рис. 1. IDEF3-модель бизнес-процесса производства мороженого**




| ID эл-та | Имя | Тип эл-та | Left | Top | Right | Bottom | Диаграмма |
|---|---|---|---|---|---|---|---|
| {33BA4DD8-4ED4-4B9B-845A-57FCD126507D}+00000000 | Замес смеси | Действие | 145 | 40 | 295 | 120 | Функционирование цехапроизводс... |
| {7C2EA8B3-F141-4B17-A536-5AA303F04C9F}+00000000 | Фризирование | Действие | 145 | 150 | 300 | 210 | Функционирование цехапроизводс... |
| {4CCEB215-AA2E-466B-85F7-6ACE287D24B6}+00000000 | Проверка на брак | Действие | 360 | 390 | 520 | 470 | Функционирование цехапроизводс... |
| {79A1EF6C-54F2-4AD5-BF85-3164F975CEEF}+00000000 | Контроль качестваПорций | Действие | 360 | 270 | 510 | 370 | Функционирование цехапроизводс... |
| {CE2C4079-1830-42E3-B133-9F673F55C6AD}+00000000 | Функционирование цехапрои... | Действие | 360 | 190 | 720 | 380 | Context |
| {3C882277-AEE8-427B-91F5-D9FE2BA5E8EF}+00000000 | Контроль качества | Действие | 365 | 40 | 510 | 120 | Функционирование цехапроизводс... |
| {344B889C-374F-4D6A-9C61-58473B87FBCC}+00000000 | Контроль качестваФризирова... | Действие | 365 | 150 | 510 | 240 | Функционирование цехапроизводс... |
| {2BFE5163-5A7B-42BB-9E17-E690D184858E}+00000000 | | Переход | 680 | 415 | 720 | 450 | Функционирование цехапроизводс... |
| {A8AE4952-560F-439A-8878-2DD7294383B4}+00000000 | | Переход | 735 | 60 | 765 | 95 | Функционирование цехапроизводс... |
| {1E825E08-12FD-4D7F-A43E-00818A573141}+00000000 | | Переход | 740 | 170 | 780 | 210 | Функционирование цехапроизводс... |
| {CEF5769A-EB5A-4251-8004-42DA8769F3FA}+00000000 | | Переход | 745 | 305 | 785 | 340 | Функционирование цехапроизводс... |
| {FF6EACCE-83B9-4586-AA47-87D63B194774}+00000000 | Отправка готовойпродукции | Действие | 800 | 450 | 960 | 530 | Функционирование цехапроизводс... |
| {D4DCE55C-B5C1-4A87-8FA5-2F4AE2A0EBB9}+00000000 | Начать процессзаново | Действие | 930 | 35 | 1050 | 115 | Функционирование цехапроизводс... |

**Рис. 2. Интерфейс конвертора с загруженным XML-файлом**

делить все операции по исполнителям, а также определить ключевые показатели эффективности, влияющие на достижение целей компании.

Работа с процессами в *1С: Предприятие* реализована в виде объектов конфигурации "Бизнес-процессы" и "Задачи". Бизнес-процессы предназначены для управления последовательностью действий, направленных на достижение цели в контексте автоматизируемой предметной области. Задачи бизнес-процесса предназначены для отражения выдачи и исполнения заданий участниками бизнес-процессов или обычными пользователями системы. Задачи могут применяться самостоятельно или использоваться для обеспечения функционирования бизнес-процессов разного вида.

У бизнес-процесса есть свойство "Карта маршрута". Оно является ключевым свойством бизнес-процесса, наглядно описывает жизненный цикл от старта до завершения, позволяет реализовать визуальное проектирование в терминах предметной области, а также является нотацией, понятной не только специалистам, но и владельцам бизнес-процессов.

Для реализации механизма конвертации моделей были определены соответствия объектов сценарных моделей *ERwin Process Modeler* и карт маршрутов бизнес-процессов *1С: Предприятие*, которые представлены в таблице.

В качестве средства обмена данными между *ERwin Process Modeler* и *1С: Предприятие* использован язык XML. Для переноса модели в ERP-систему из *ERwin Process Modeler* экспортируется XML-файл, содержащий полное описание бизнес-процесса. Выбор формата файла обусловлен его открытостью, легкостью для чтения и пониманияя, возможностью просмотра в любом текстовом редакторе без привлечения специализированного программного обеспечения, а также наличием в языках программирования готовых классов и библиотек для удобной и быстрой работы с файлами этого формата. Сохранение в XML осуществляется стандартными средствами *ERwin Process Modeler*.

Для преобразования XML-файла в файл *1С: Предприятие* разработан конвертор, который позволяет загрузить полученный XML-файл и преобразовать сведения об объектах модели *ERwin Process Modeler* в объекты графической схемы *1С: Предприятие*. На рис. 2 приведен пример считывания информации о бизнес-процессе, приведенном на рис. 1, разработанным конвертором. После открытия XML-файла в окне конвертора происходит

**Соответствие основных объектов сценарных моделей *ERwin Process Modeler* объектам карт маршрутов бизнес-процессов в *1С: Предприятие***

| Название объекта | Описание преобразования *Erwin* — 1С |
|---|---|
| Действие (*Activity*) | Объект соответствует процедуре. Преобразуется в 1С в точку «Действие» |
| Перекресток (*Junction*): AND, XOR, OR | Правила ветвления или соединения процесса: «И», «Исключающее ИЛИ», «ИЛИ». В 1С «И» преобразуется в точки «Слияние» и «Разделение», «Исключающее ИЛИ» — в точку «Выбор варианта», «ИЛИ» — в совокупность указанных точек |
| Вложенный бизнес-процесс | Объект указывает на наличие вложенного подпроцесса. В 1С преобразуется в точку «Вложенный процесс» |

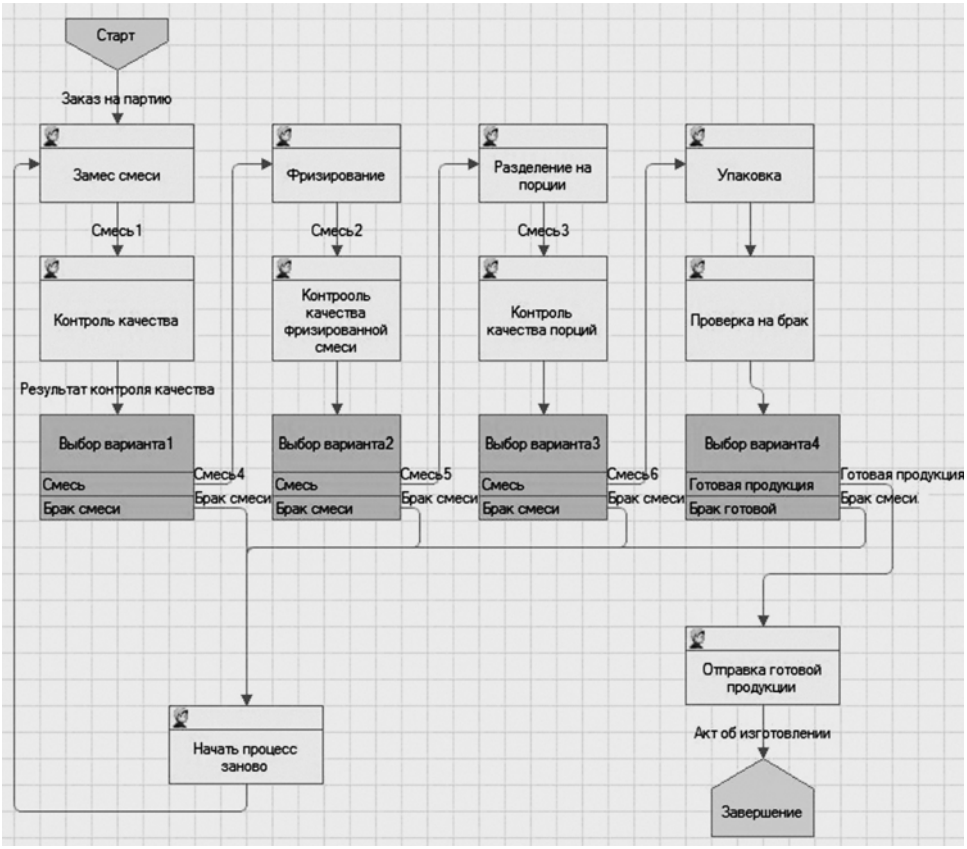



**Рис. 3. Результат работы конвертора в виде карты маршрута бизнес-процесса производства мороженого**

автоматическое заполнение таблицы, содержащей в себе полный список объектов бизнес-процесса.

Для получения списка всех объектов бизнес-процесса требуется поэлементное считывание XML-файла, экспортированного из *ERwin Process Modeler*. Каждый элемент может иметь атрибут (или несколько атрибутов), содержимое или вложенный элемент. Необходимые данные хранятся в следующих элементах:

- *PMArrow_Groups* — список стрелок;
- *PMActivity_Groups* — список действий;
- *PMJunction_Groups* — список перекрестков;
- *PMDiagram_Groups* — список диаграмм.

Данные об объектах включают в себя вложенные элементы, содержащие следующую информацию: имя линии, идентификационный номер объекта *ID*, автор, примечание, описание и т. д.

Графическая схема бизнес-процесса представляет собой совокупность точек различного типа и соединительных линий. Каждая точка карты маршрута в *1С: Предприятие* может быть представлена в виде определенной последовательности символов. Для преобразования данных конвертор из входных параметров, считанных из XML-файла, генерирует текст описания карты маршрута в формате *1С: Предприятие*. Результатом работы является файл с расширением *.GRS, который включает в себя весь сгенерированный текст. Полученный файл можно открыть в *1С: Предприятие* для дальнейшей настройки и создания бизнес-процесса. Пример экспорта результатов работы конвертора (рис. 2) в среду *1С: Предприятие* 8.1 приведен на рис. 3.

Таким образом, с использованием разработанных программных решений возможно осуществить интеграцию CASE-средства проектирования *ERwin Process Modeler* и платформы *1С: Предприятие*. При этом проектирование информационной системы предприятия принимает характер последовательного перехода от одной модели к другой: от модели сценария выполнения бизнес-процесса (IDEF3-модели) к модели процесса в формате *1С: Предприятие*, т. е. файлу в формате *.GRS и карте маршрутов *1С: Предприятие*.

## Заключение

В статье предложено универсальное прикладное решение по конвертации бизнес-процессных моделей *ERwin Process Modeler* в формат *1С: Предприятие*, применение которого позволяет осуществлять экспорт моделей, созданных в *ERwin Process Modeler*, в действующую или проектируемую ERP-систему. Разработанное решение является универсальным: оно не зависит от конфигурации, в которую будет встраиваться конвертируемый бизнес-процесс, — она остается типовой и сохраняется возможность ее поддержки и обновления в стандартизованном виде.

### Список литературы

1. **Репин В. В.** Бизнес-процессы компании. Построение, анализ, регламентация. М.: Стандарты и качество, 2007. 240 с.

2. **Ивлев В. А., Попова Т. В.** Процессная организация деятельности: методы и средства // Управление качеством. 2007. № 1 (3). URL: http://www.iteam.ru/publications/quality/ section_60/ article_750

3. **ГОСТ** Р ИСО 9000—2008. Системы менеджмента качества. Основные положения и словарь. (Взамен ГОСТ Р ИСО 9000—2001; дата введ. 18.12.2008.) М.: Стандартинформ, 2009. 30 с.

4. **Официальный** сайт компании CA Technologies [Электронный ресурс]. — USA, NY: CA Technologies, 2011. URL: http://erwin.com/products/detail/ca_erwin_process_modeler/

5. **Габец А. П., Гончаров Д. И., Козырев Д. В., Кухлевский Д. С., Радченко М. Г.** Профессиональная разработка в системе 1С: Предприятие 8. М.: "1С-Паблишинг"; СПб.: Питер, 2006. 808 с.

УДК 159.9 + 004.3

**Л. С. Куравский**,
д-р техн. наук, проф., зав. каф., декан,
e-mail: l.s.kuravsky@gmail.com,
**Г. А. Юрьев**, аспирант,
e-mail: grinch89@mail.ru,
Московский городской
психолого-педагогический университет

# Применение фильтра Калмана для фильтрации артефактов при адаптивном тестировании

*Представлен метод фильтрации результатов адаптивного тестирования, основанный на использовании обучаемых структур в форме марковских моделей с непрерывным временем. Устранение артефактов, обусловленных различными формами некорректного целенаправленного вмешательства в процедуру испытаний, выполняется на основе сравнения наблюдаемых и прогнозируемых результатов ответов на вопросы с помощью фильтра Калмана, адаптированного для решения рассматриваемой задачи.*

***Ключевые слова:*** *адаптивное тестирование, марковские модели, фильтр Калмана*

## Введение

Компьютерное тестирование в настоящее время широко используется в медицине, психологии и образовании в целях диагностики, определения уровня компетенций и пригодности испытуемых для выполнения тех или иных функций, включая контроль качества обучения. Качество тестирования и достоверность его результатов в значительной степени зависят от технологий проведения тестов, которые в последние десятилетия стали предметом активных научных исследований.

В первое время тесты строились на основе классической теории тестирования [2, 15, 18, 19], в основе которой лежит теория погрешности измерений, заимствованная из физики: полагалось, что измеряемые характеристики имеют некоторые "истинные" значения, искажаемые случайными и систематическими погрешностями. Этот подход получил определенное распространение, однако его практическому применению препятствует ряд существенных недостатков:

- возникают проблемы при сравнении сходных особенностей тестируемых, выявленных с помощью разных методик;
- не решается проблема валидности;
- тестовые баллы становятся недостаточно надежными в областях экстремальных значений;
- технология в целом недостаточна надежна и универсальна.

Для преодоления указанных проблем была разработана новая технология тестирования, основанная на латентно-структурном анализе и названная *теорией ответов на вопросы* (*Item Response Theory — IRT*)[1] [15, 17]. В ней реализована концепция *адаптивного тестирования*, согласно которой тестируемому с определенной текущей расчетной оценкой уровня знаний или способностей на каждом шаге тестирования предлагаются задания определенной сложности. Основная концепция новой теории, предложенная Г. Рашем в 1960 г. [26], предполагает, что вероятность правильного ответа на задание определяется разностью уровня знаний или способностей и трудности теста. В зависимости от условий прикладной задачи на практике используются и другие, более сложные модели, построенные на базе данной концепции [1, 26, 28, 29].

Применение технологии IRT приводит к следующим проблемам:

- "статичность" оценок — игнорирование того факта, что результат тестирования вследствие усталости испытуемых и других факторов может, вообще говоря, существенно изменяться со временем, принимая различные значения в процессе сеанса тестирования;
- невозможность учета времени, затрачиваемого на решение тестовых задач, при построении расчетных оценок;
- необходимость выполнения достаточно большого числа заданий для получения оценок с приемлемой точностью;
- сложность вычисления распределения вероятностей возможных результатов теста, что необходимо для оценки их надежности;
- сравнительно сложная для практической реализации процедура оценки точности результата, связанная с применением метода максимального правдоподобия и расчетом доверительных интервалов.

Указанные проблемы делают актуальной разработку новых технологий тестирования. В этой работе

[1] В русскоязычной литературе также используются и другие варианты ее названия: стохастическая теория тестов, математическая теория измерений, современная теория тестирования, теория латентных черт, теория характеристических кривых заданий, теория моделирования и параметризации педагогических тестов и т. д.

рассматриваются новые аспекты применения разработанного ранее авторами подхода к адаптивному тестированию [4—11, 20—25], построенного на использовании *обучаемых структур в форме марковских моделей* с дискретным и непрерывным временем. Его особенностями, обеспечивающими преимущества перед аналогичными способами тестирования, являются:

- выявление и использование при построении расчетных оценок временно́й динамики изменения способности справляться с заданиями теста;
- возможность учета при построении расчетных оценок времени, затрачиваемого на решение тестовых задач;
- возможность исследования временно́й динамики знаний или способностей как в дискретной, так и в непрерывной временной шкале;
- меньшее по сравнению с другими подходами число заданий, которое следует предъявлять испытуемому для получения оценок знаний или способностей с заданной точностью, что ускоряет процесс тестирования;
- получение распределения вероятностей возможных результатов теста в качестве конечного результата;
- развитая техника идентификации параметров моделей.

Одной из наиболее серьезных проблем, возникающих в процессе тестирования, является появление в истории ответов испытуемого искажающих результаты *артефактов*, обусловленных подсказками, угадыванием и другими формами некорректного целенаправленного вмешательства в процедуру испытаний. Представленная выше технология адаптивного тестирования позволяет бороться с этими явлениями, устраняя артефакты на основе сравнения наблюдаемых и прогнозируемых результатов ответов на вопросы для разных уровней способностей испытуемых. В качестве инструмента для сопоставления в данной работе предлагается использовать *фильтр Калмана* [14, 16] — нестационарную систему с обратной связью, включающую в себя, как составную часть, формирующий фильтр, воспроизводящий идеализированную модель поведения.

Выбор фильтра Калмана для устранения артефактов тестирования среди близких по содержанию подходов является оптимальным решением, поскольку он наилучшим образом согласуется с принятой концепцией адаптивного тестирования и контекстом ее использования. В частности, этот фильтр:

- в отличие от фильтра Винера способен обрабатывать текущую информацию об ответах испытуемого в реальном времени, формируя свои оценки сразу же после получения очередного ответа и не требуя полного протокола тестирования, который недоступен до завершения всей процедуры ответов на вопросы;
- в отличие от фильтра Стратоновича использует только линейные методы оценки, наилучшим образом согласующиеся с применяемой линейной дифференциальной моделью адаптивного тестирования, и не приводит к неоправданному усложнению процесса решения;
- в отличие от фильтра Льюинбергера учитывает ошибки наблюдений и обеспечивает оптимальные оценки.

Далее кратко представлен новый подход к адаптивному тестированию, основанный на использовании марковских моделей, поставлена задача фильтрации артефактов с помощью фильтра Калмана и рассмотрены особенности ее решения.

## 1. Марковские модели адаптивного тестирования

### 1.1. Структура и математическое описание применяемых марковских моделей с непрерывным временем. Процедура оценки знаний или способностей

Оценка вероятностей различных уровней способностей проводится по результатам тестирования с использованием параметрических математических моделей, описываемых *марковскими случайными процессами с дискретными состояниями и непрерывным и дискретным временем* [12, 13]. Дальнейшее изложение относится только к моделям с непрерывным временем. Непосредственно наблюдаемой величиной является трудность выполняемого теста, измеряемая в логитах. Допустимый диапазон значений этой величины делится на несколько интервалов, каждый из которых рассматривается как отдельное состояние $x_i$, $i = 0, 1, ..., n$, в котором тестируемый может находиться с некоторой вероятностью, переходя из одного состояния в другое по определенным правилам. Длина указанных интервалов определяет разрешающую способность оценок, получаемых в процессе тестирования. В свою очередь, число состояний определяется желаемой разрешающей способностью оценок и доступным объемом выборки[2].

Как трудности заданий, так и способности тестируемых измеряются в единой безразмерной *шкале логитов*, выражающей соотношение долей правильных и неправильных ответов. Перевод в шкалу логитов осуществляется по формуле

$$C = \ln \frac{r}{1 - r},$$

где $C$ — значение в шкале логитов; $r$ — вероятность правильного выполнения задания. В случае оценки трудности этот параметр характеризует возможность выполнения определенного задания для всего мно-

[2] Рассматривая непрерывно изменяющуюся характеристику как дискретную величину, мы теряем часть информации (это имеет место при любой идеализации). Однако эти потери несущественны в случае достаточно больших выборок, когда мы имеем возможность устанавливать длину интервалов состояний так, чтобы она не превышала ошибок измерений.

жества тестируемых, а в случае оценки способностей — результаты определенного тестируемого для всего множества допустимых заданий. Статистические приближения указанных величин получаются после замены в приведенной формуле вероятности $r$ на ее выборочные оценки.

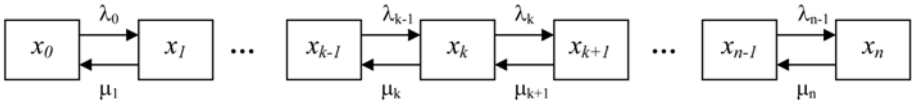



**Рис. 1. Сеть Маркова, представляющая процесс тестирования с непрерывным временем: $x_i$ ($i = 0, 1, \ldots, n$) — состояния; $\lambda_i$ ($i = 0, 1, \ldots, n - 1$) и $\mu_i$ ($i = 1, 2, \ldots, n$) — интенсивности переходов**

Если обозначить верхнюю и нижнюю границы диапазона возможных значений трудности тестов как $D_{bot}$ и $D_{top}$, состояние $x_0$ будет соответствовать интервалу от $D_{bot}$ до $D_{bot} + (D_{top} - D_{bot})/(n + 1)$, состояние $x_1$ — интервалу от $D_{bot} + (D_{top} - D_{bot})/(n + 1)$ до $D_{bot} + 2(D_{top} - D_{bot})/(n + 1)$ и т. д.

Модели для описания динамики этих переходов представляются ориентированными графами, в которых вершины[3] соответствуют состояниям, а дуги[4] — переходам.

В случае моделей с непрерывным временем процесс тестирования может рассматриваться как случайное блуждание по графу с переходами из одного состояния в другое согласно направлениям дуг. Эти переходы мгновенны и происходят в случайные моменты времени.

Предполагается, что для них выполняются следующие два свойства *пуассоновских потоков событий*:

- *ординарность* (поток называется ординарным, если вероятность появления двух и более событий в течение малого интервала времени намного меньше, чем вероятность появления за это же время одного события);
- *независимость приращений* (это свойство означает, что количества событий, попадающих в два непересекающихся интервала, не зависят друг от друга).

Можно показать, что в рассматриваемых потоках число событий $X$, попадающих в любой временной интервал длины $\tau$, начинающийся в момент $t$, распределено согласно *закону Пуассона*:

$$P_{t,\tau}(X = m) = \frac{a(t,\tau)^m}{m!}\,\mathbf{e}^{-a(t,\tau)},$$

где $P_{t,\tau}(X = m)$ — вероятность появления $m$ событий в течение рассматриваемого интервала; $a(t, \tau)$ — среднее число событий, попадающих в интервал длины $\tau$, начинающийся в момент времени $t$. Далее будут рассматриваться только *стационарные потоки* (в которых $a(t, \tau) = \eta\tau$, $\eta = \text{const}$). Параметр $\eta$ называется *интенсивностью стационарного потока*. Он равен среднему числу событий в единицу времени. Средняя продолжительность времени между двумя смежными событиями в этом случае равна $1/\eta$.

Упомянутые выше предположения о свойствах потоков событий обычны для прикладных задач, так как эти потоки (или потоки, близкие к ним по свойствам) часто встречаются на практике благодаря предельным теоремам для потоков событий [12, 13].

Для моделей с непрерывным временем неизвестными (свободными) параметрами модели являются интенсивности потоков событий. Их значения определяются путем сравнения наблюдаемых и прогнозируемых гистограмм, описывающих распределения частот пребывания в состояниях модели, а именно: вычисляются значения, обеспечивающие наилучшее соответствие наблюдаемых и ожидаемых частот попадания в определенное состояние системы в заданные моменты времени. Прогнозируемые вероятности нахождения в состояниях получаются путем численного интегрирования систем уравнений Колмогорова.

Марковские модели с непрерывным временем и свободными параметрами, которые идентифицируются по данным наблюдений, называются *сетями Маркова* [5, 7, 21—23].

Для описания того, как вероятности нахождения в заданных состояниях изменяются со временем, применяются сети Маркова, организованные по так называемой схеме "гибели и размножения"[5] (рис. 1). Эта схема представляет собой конечную цепь из $n + 1$ состояния, в которой переходы из состояния $x_k$ ($k \neq 0$, $k \neq n$) возможны только в предшествующее состояние $x_{k-1}$ или в следующее по порядку состояние $x_{k+1}$. Из состояний $x_0$ и $x_n$ доступны только состояния $x_1$ и $x_{n-1}$ соответственно.

Динамика вероятностей нахождения в различных состояниях указанной схемы описывается следующей системой обыкновенных дифференциальных *уравнений Колмогорова*:

$$\frac{dp_0(t)}{dt} = -\lambda_0 p_0(t) + \mu_1 p_1(t);$$

.....

$$\frac{dp_k(t)}{dt} = -(\lambda_k + \mu_k)p_k(t) + \lambda_{k-1}p_{k-1}(t) + \mu_{k+1}p_{k+1}(t)$$

$(k = 1, 2, \ldots, n - 1)$;

.....

$$\frac{dp_n(t)}{dt} = -\mu_n p_n(t) + \lambda_{n-1}p_{n-1}(t),$$

где $p_*(t)$ есть вероятность нахождения в состоянии $x_*$ в момент времени $t$; $*$ — номер состояния;

[3] Обозначаются как прямоугольники.
[4] Обозначаются как стрелки.

[5] Она была впервые применена в биологии для анализа динамики роста популяций.

$\lambda_i$ ($i = 0, 1, ..., n-1$) и $\mu_i$ ($i = 1, 2, ..., n$) — интенсивности переходов между состояниями, которые определяются отдельно для каждого из рассматриваемых уровней способностей. Для интегрирования указанной системы необходимо задать начальные условия: $p_0(0), p_1(0), ..., p_n(0)$. Нормализующее условие $\sum_{k=0}^{n} p_k(t) = 1$ выполняется в любой момент времени.

Для упрощения задачи, а также для обеспечения приемлемой процедуры идентификации, интенсивности потоков часто полагаются зависящими от индекса $i$ по определенным правилам, включая тривиальный вариант: $\lambda_0 = \lambda_1 = ... = \lambda_{n-1} = \lambda$ и $\mu_1 = \mu_2 = ... = \mu_n = \mu$. Оптимальный выбор подобных зависимостей опирается на технику проверки статистических гипотез. В случае моделей с дискретным временем аналогичные зависимости исследуются для вероятностей переходов.

*Процедура адаптивного тестирования* заключается в последовательном предъявлении испытуемому задач, трудность которых определяется состоянием сети или цепи Маркова, в котором он находится в данный момент. Если испытуемый, находясь в состоянии $x_i$, решает задачу, он переходит в состояние $x_{i+1}$, в противном случае — в состояние $x_{i-1}$. По завершении тестирования он оказывается в одном из состояний $x_*$, наилучшим образом соответствующих его уровню способностей. Принцип выбора очередного теста заключается в выборе задачи, трудность которой примерно соответствует уровню способностей испытуемого. Согласно проведенным наблюдениям и результатам современной теории тестирования это обеспечивает наилучшую *дифференциацию* испытуемых по уровню их способностей.

### 1.2. Идентификация марковских моделей с непрерывным временем

Идентификации марковских моделей проводятся по выборкам испытуемых, отдельно для каждого из рассматриваемых уровней способностей. Каждому уровню способностей $C_i$, $i = 1, ..., I$, при этом ставится в соответствие свой уникальный набор оценок параметров модели, что позволяет в дальнейшем выявлять значение этого показателя, наилучшим образом согласующегося с наблюдениями. Таким образом, вероятности и интенсивности переходов являются функциями двух характеристик: уровня способностей и трудности задачи. Число уровней способностей — это дискретный параметр, который задает разрешающую способность оценки данной характеристики и устанавливается при решении каждой прикладной задачи в зависимости от объема выборки испытуемых, имеющейся у исследователя при решении задачи идентификации, и желаемой точности результата.

С каждой изменяющейся со временем гистограммой пребывания в состояниях модели связывается марковский процесс с дискретными состояниями. *Статистика Пирсона*

$$X^2 = \sum_{k=0}^{n} \frac{(F_k - p_k N)^2}{p_k N},$$

где $N$ — число элементов в выборке; $p_k$ — прогнозируемая вероятность попадания в $k$-е состояние модели; $F_k$ — наблюдаемая частота нахождения в $k$-м состоянии модели, используется как мера соответствия в том смысле, что ее большие значения означают плохое согласование прогнозируемых и наблюдаемых результатов, а малые значения — хорошее согласование. Для идентификации модели минимизируется сумма указанных статистик в те моменты времени, для которых имеются результаты наблюдений. Наблюдаемые числа попаданий в различные интервалы трудностей задач определяются по результатам тестирования группы испытуемых. В качестве искомых оценок свободных параметров моделей используются значения, обеспечивающие наилучшее соответствие наблюдаемых и прогнозируемых частот попадания в определенное состояние системы в заданные моменты времени.

Доказано, что при выполнении ряда общих условий значения статистики Пирсона $X^2$, получаемые при подстановке истинных решений, асимптотически описываются распределением $\chi^2$ с $n-l$ степенями свободы, где $l$ — число определяемых параметров, причем вычисленные значения свободных параметров при увеличении объема выборки сходятся по вероятности к искомому решению [3, с. 462—470]. Это позволяет использовать приведенную статистику для проверки гипотезы о том, что полученный прогноз согласуется с результатами наблюдений. Данный способ идентификации свободных параметров называется *методом минимума* $\chi^2$ [3] и дает решения, близкие к полученным методом максимального правдоподобия [3, с. 461—462].

Используемая процедура вычисления оцениваемых параметров состоит из двух этапов. На подготовительном этапе с помощью электронной таблицы для указанной системы дифференциальных уравнений кодируется численная схема интегрирования, позволяющая вычислять вероятностные функции $p_k$ [5, 7, 21]. Эти функции вычисляются с некоторым заданным временным шагом. Для вычисления решения с приемлемой точностью оказались достаточными *методы Рунге—Кутты* или их эквиваленты.

На заключительном этапе запускается численная процедура многомерной нелинейной оптимизации[6] [5, 7, 21], позволяющая находить искомые значения свободных параметров. Полученные оценки свободных параметров рассматриваются как характе-

[6] В настоящее время предлагается достаточно много программных продуктов для решения задач численной оптимизации. В частности, пользователи электронной таблицы *Excel* могут применять программное обеспечение компании *Frontline Systems, Inc.*

ристики модели, выявленные в результате наблюдений. Рассмотренный критерий также позволяет сравнивать между собой различные варианты марковских моделей, выбирая среди них оптимальные.

### 1.3. Поиск оптимального решения

Зная состояние модели, в котором оказался тестируемый после решения последней предложенной ему задачи, и рассчитав вероятность нахождения в этом состоянии в заданный момент времени для каждого из рассматриваемых уровней способностей с помощью дифференциальных зависимостей (см. раздел 1.1), можно оценить вероятности пребывания в указанном конечном состоянии по формуле Байеса:

$$P(C_i|S) = \frac{P(C_i)P(S|C_i)}{\sum\limits_{k=1}^{I} P(C_k)P(S|C_k)},$$

где $C_i$ — событие, связанное с наличием у тестируемого $i$-го уровня способностей ($i = 1, ..., I$); $S$ — событие, связанное с нахождением в заданном конечном состоянии модели в заданный момент времени; $P(C_i)$ — априорная вероятность появления $i$-го уровня способностей у тестируемого; $P(S|C_i)$ — вероятность нахождения в заданном конечном состоянии модели в заданный момент времени при наличии $i$-го уровня способностей; $P(C_i|S)$ — вероятность $i$-го уровня способностей при условии нахождения в заданном конечном состоянии модели в заданный момент времени.

Уровень способностей, при котором достигается наибольшая условная вероятность

$$P(C_{\max}|S) = \max_i \{P(C_i|S\}_{i=1, ..., I},$$

дает искомую оценку. Распределение вероятностей $\{P(C_i|S\}_{i=1, ..., I}$, которое является результатом решения задачи, позволяет оценить степень надежности этой оценки.

Как указано в разделе 1.1, разрешающая способность полученной оценки определяется длиной интервала между соответствующими смежными уровнями способностей в логитах, которая, в свою очередь, при условии постоянства таких длин задается числом уровней способностей $I$.

## 2. Математическая постановка и решение задачи фильтрации Калмана при адаптивном тестировании с использованием марковских моделей

В случае обсуждаемого варианта адаптивного тестирования наблюдаемый процесс представляет историю пребывания в состояниях марковских моделей. Он выражается вектором $\mathbf{x}(t) = (x_0(t), x_1(t), ..., x_n(t))^{\mathrm{T}}$, в котором в каждый момент времени одна и только одна из компонент $x_i(t)$, $i = 0, ..., n$, соответствующая состоянию, где находится испытуемый, равна единице, а остальные компоненты равны нулю. В свою очередь, исследуемый информационный процесс $\mathbf{P}(t) = (p_0(t), p_1(t), ..., p_n(t))^{\mathrm{T}}$ представляет динамику изменения вероятностей пребывания в состояниях модели.

Уравнения информационного и наблюдаемого процессов, используемые при построении многомерного непрерывного фильтра Калмана для моделей рассматриваемого типа[7], имеют следующий вид [14, 16]:

$$\frac{d\mathbf{P}}{dt} = \mathbf{F}\mathbf{P};$$

$$\mathbf{x}(t) = \mathbf{P}(t) + \mathbf{v}(t),$$

где на случайные ошибки наблюдений $\mathbf{v}(t)$ накладываются условия $\mathbf{E}(\mathbf{v}(t)) = \mathbf{0}$ и $\mathbf{E}(\mathbf{v}(t)\mathbf{v}^{\mathrm{T}}(\tau)) = \mathbf{R}\delta(t-\tau)$, матрица формирующего фильтра $\mathbf{F}$ размерности $(n+1)\times(n+1)$ есть

$$\begin{pmatrix} -\lambda_0 & \mu_1 & . & . & . & 0 & 0 \\ \lambda_0 & -(\lambda_1+\mu_1) & \mu_2 & . & . & 0 & 0 \\ . & . & . & . & . & . & . \\ 0 & . & \lambda_{k-1} & -(\lambda_k+\mu_k) & \mu_{k+1} & 0 & 0 \\ . & . & . & . & . & . & . \\ 0 & 0 & . & . & \lambda_{n-2} & -(\lambda_{n-1}+\mu_{n-1}) & \mu_n \\ 0 & 0 & . & . & . & \lambda_{n-1} & -\mu_n \end{pmatrix},$$

а $\mathbf{R}$ — симметричная положительно определенная матрица, которую мы далее будем полагать не зависящей от времени. При проведении практических расчетов эта матрица может заменяться на одну их своих выборочных оценок $\hat{\mathbf{R}}$, полученных для каждого из рассматриваемых уровней способностей на основе результатов наблюдений.

Дифференциальное уравнение фильтра Калмана, определяющее несмещенную оценку исследуемого процесса[8] $\hat{\mathbf{P}}(t) = (\hat{p}_0(t), \hat{p}_1(t), ..., \hat{p}_n(t))^{\mathrm{T}}$ с минимальным средним квадратом ошибки $\mathbf{e}(t) = \mathbf{P}(t) - \hat{\mathbf{P}}(t)$, представляется в виде

$$\frac{d\hat{\mathbf{P}}(t)}{dt} = \mathbf{F}\hat{\mathbf{P}}(t) + \mathbf{K}_{\mathbf{c}}(t)(\mathbf{x}(t) - \hat{\mathbf{P}}(t)),$$

где $\mathbf{K}(t)$ — матричный коэффициент усиления фильтра Калмана.

В классическом случае этот коэффициент задается уравнением

$$\mathbf{K}_{\mathbf{c}}(t) = \mathbf{U}(t)\mathbf{R}^{-1},$$

[7] Особенностями этих моделей являются: отсутствие информационного шума, равенство размерностей информационного процесса и процесса наблюдений и единичная матрица наблюдений.

[8] Выход фильтра Калмана.

в котором ковариационная матрица ошибок $\mathbf{U}(t) = \mathbf{E}(\mathbf{e}(t)\mathbf{e}^{\mathrm{T}}(t))$ является решением одной из матричных форм уравнения Риккати

$$\frac{d\mathbf{U}(t)}{dt} = \mathbf{F}\mathbf{U}(t) + \mathbf{U}(t)\mathbf{F}^{\mathrm{T}} - \mathbf{U}(t)\mathbf{R}^{-1}\mathbf{U}(t).$$

Однако, поскольку в рассматриваемой задаче компоненты оценки информационного процесса $\hat{\mathbf{P}}(t)$ представляют собой нормированные величины — вероятности пребывания в состояниях сети Маркова с суммой, равной единице, — необходима коррекция коэффициента усиления $\mathbf{K_c}(t)$, обеспечивающая поддержание данного условия.

Если нормализующее условие $\sum_{k=0}^{n} \hat{p}_k(t) = 1$ выполняется в начальный момент времени $t = 0$, а правая часть уравнения фильтра Калмана такова, что при $t \geqslant 0$ обеспечивается равенство $\sum_{k=0}^{n} \frac{d\hat{p}_k(t)}{dt} = 0$, то указанное нормализующее условие выполняется в любой момент времени $t \geqslant 0$. Очевидно, что условие $\sum_{k=0}^{n} \frac{d\hat{p}_k(t)}{dt} = 0$ равносильно равенству нулю суммы компонентов вектора, заданного матричным выражением $\mathbf{F}\hat{\mathbf{P}}(t) + \mathbf{K_c}(t)(\mathbf{x}(t) - \hat{\mathbf{P}}(t))$. Поскольку нулевая сумма компонент вектора $\mathbf{F}\hat{\mathbf{P}}(t)$ обеспечивается приведенной выше структурой матрицы $\mathbf{F}$, то для равенства нулю суммы компонент всего указанного матричного выражения необходимо и достаточно нулевой суммы компонент вектора $\mathbf{K_c}(t)(\mathbf{x}(t) - \hat{\mathbf{P}}(t))$.

Сумма компонент вектора $\mathbf{x}(t) - \hat{\mathbf{P}}(t)$ равна нулю по условиям рассматриваемой задачи, так как эти величины интерпретируются как вероятности. Учитывая данный факт, несложно доказать, что достаточным условием нулевой суммы компонент вектора $\mathbf{K_c}(t)(\mathbf{x}(t) - \hat{\mathbf{P}}(t))$ является равенство сумм элементов матрицы $\mathbf{K_c}(t)$ во всех ее столбцах. Таким образом, если матричный коэффициент усиления $\mathbf{K_c}(t)$ в уравнении фильтра Калмана заменить на близкий к нему нормализованный коэффициент $\mathbf{K_n}(t)$ с равными во всех столбцах суммами элементов, то условие $\sum_{k=0}^{n} \frac{d\hat{p}_k(t)}{dt} = 0$ будет выполнено. Матрицу $\mathbf{K_n}(t)$ можно получить, домножив справа матрицу $\mathbf{K_c}(t)$ на диагональную матрицу $\mathbf{D}$, элементы которой вычисляются по формуле

$$d_{jj} = \frac{\sum_{l,\,m=0}^{n} k_{lm}}{(n+1)k_{*j}},$$

где $d_{jj}$ — $j$-й диагональный элемент матрицы $\mathbf{D}$; $k_{lm}$, $l, m = 0, ..., n$, — элементы матрицы $\mathbf{K_c}(t)$; $k_{*j}$ — сумма элементов в $j$-м столбце матрицы $\mathbf{K_c}(t)$. Такая замена корректна, если $\mathbf{K_n}(t) = \mathbf{U}(t)\mathbf{R}^{-1}\mathbf{D}$ лежит в допустимых границах вариаций коэффициента $\mathbf{K_c}(t)$, обусловленных ошибками выборочных оценок матрицы $\mathbf{R}$, что проверяется с помощью подходящих критериев согласия.

В частности, для этого можно:

- сгенерировать множество выборочных оценок ковариационной матрицы $\mathbf{R}$, соответствующих доверительным интервалам для заданного объема выборки $N$;
- вычислить, используя эти оценки, выборку матриц $\{\mathbf{K_{ni}}(t)\}_{i=1,\,..,\,M}$;
- вычислить выборочное распределение евклидовой нормы разностей $\{\|\mathbf{K_{ni}}(t) - \mathbf{K_c}(t)\|_E\}_{i=1,\,..,\,M}$ классического и нормированного коэффициентов усиления;
- учитывая, что полученное выборочное распределение при достаточно большом числе элементов в матричных коэффициентах усиления приблизительно соответствует нормальному, построить для него выборочные оценки математического ожидания и дисперсии и оценить вероятность $p$ превышения евклидовой нормы разности $\|\mathbf{K_n}(t) - \mathbf{K_c}(t)\|_E$.

Если $p \geqslant 0{,}05$, то использование нормализованного коэффициента $\mathbf{K_n}(t)$ является допустимым. Рассмотренный метод может быть совмещен с процедурой кластеризации, использующей самоорганизующиеся карты Кохонена [9, 25].

В соответствии с представленной выше процедурой адаптивного тестирования фильтрация Калмана выполняется автономно для каждого из уровней способностей, учитываемых при постановке решаемой задачи.

В заключение следует отметить, что существует ряд интересных аналогий между фильтром Калмана и скрытыми марковскими моделями [8, 24], частично рассмотренных в обзоре [27].

### 3. Программная реализация

Рассмотренная процедура фильтрации реализована в среде графического программирования LabVIEW (рис. 2, см. третью сторону обложки). При этом интегрирование матричного уравнения Риккати и уравнения фильтра Калмана выполнено

численными методами[9], а для оценки начального состояния ковариационной матрицы ошибок $\mathbf{U}(0)$, о которой наблюдения дают, как правило, мало полезной информации, использованы следующие предположения:

- $\mathbf{E}(\mathbf{e}(0)) = \mathbf{0}$;
- компоненты вектора ошибок фильтрации $\mathbf{e}(0)$ статистически независимы;
- дисперсии компонент вектора ошибок фильтрации $\mathbf{e}(0)$ пропорциональны соответствующим дисперсиям компонент случайного шума наблюдения $\mathbf{v}(t)$.

## 4. Основные результаты и выводы

- Разработан и программно реализован вероятностный метод фильтрации искажающих результаты артефактов при адаптивном тестировании, построенном на использовании обучаемых структур в форме марковских моделей с непрерывным временем.
- Устранение артефактов, обусловленных различными формами некорректного целенаправленного вмешательства в процедуру испытаний, выполняется на основе сравнения наблюдаемых и прогнозируемых результатов ответов на вопросы для разных уровней способностей испытуемых с помощью фильтра Калмана, адаптированного для задачи адаптивного тестирования.
- Выбор фильтра Калмана для устранения артефактов является оптимальным среди близких по содержанию подходов, поскольку он наилучшим образом согласуется с принятой концепцией адаптивного тестирования и контекстом ее использования.

### Список литературы


1. **Аванесов В. С.** Педагогическое измерение латентных качеств // Педагогическая диагностика. 2003. № 4. С. 69—78.
2. **Карданова Е. Ю.** Моделирование и параметризация тестов: основы теории и приложения. М.: Федеральный центр тестирования, 2008.
3. **Крамер Г.** Математические методы статистики. М.: Мир, 1976. 648 с.
4. **Куравский Л. С., Баранов С. Н.** Синтез сетей Маркова для прогнозирования усталостного разрушения // Нейрокомпьютеры: разработка и применение. 2002. № 11. С. 29—40.
5. **Куравский Л. С., Баранов С. Н.** Применение нейронных сетей для диагностики и прогнозирования усталостного разрушения тонкостенных конструкций // Нейрокомпьютеры: разработка и применение. 2001. № 12. С. 47—63.
6. **Куравский Л. С., Баранов С. Н., Корниенко П. А.** Обучаемые многофакторные сети Маркова и их применение для исследования психологических характеристик // Нейрокомпьютеры: разработка и применение. 2005. № 12. С. 65—76.
7. **Куравский Л. С., Баранов С. Н., Малых С. Б.** Нейронные сети в задачах прогнозирования, диагностики и анализа данных: Учеб. пособие. М.: РУСАВИА, 2003. 100 с.
8. **Куравский Л. С., Баранов С. Н., Юрьев Г. А.** Синтез и идентификация скрытых марковских моделей для диагностики усталостного разрушения // Нейрокомпьютеры: разработка и применение. 2010. № 12. С. 20—36.
9. **Куравский Л. С., Ушаков Д. В., Мармалюк П. А., Панфилова А. С.** Исследование факторных влияний на развитие психологических характеристик с применением нового подхода к оценке адекватности моделей наблюдениям. № 11. Информационные технологии. 2011. № 11. С. 67—77.
10. **Куравский Л. С., Юрьев Г. А.** Адаптивное тестирование как марковский процесс: модели и их идентификация // Нейрокомпьютеры: разработка и применение. 2011. № 2. С. 21—29.
11. **Куравский Л. С., Юрьев Г. А.** Использование марковских моделей при обработке результатов тестирования // Вопросы психологии. 2011. № 2. С. 98—107.
12. **Овчаров Л. А.** Прикладные задачи теории массового обслуживания. М.: Машиностроение, 1969. 324 с.
13. **Саати Т. Л.** Элементы теории массового обслуживания и ее приложения. М.: ЛИБРОКОМ, 2010. 520 с.
14. **Тихонов В. И., Шахтарин Б. И., Сизых В. В.** Случайные процессы. Примеры и задачи. Т. 5. Оценка сигналов, их параметров и спектров. Основы теории информации. М.: Горячая линия—Телеком, 2009. 400 с.
15. **Тюменева Ю. А.** Психологическое измерение. М.: Аспект-Пресс, 2007.
16. **Шахтарин Б. И.** Случайные процессы в радиотехнике. Т. 1. Линейные преобразования. М.: Горячая линия—Телеком, 2010. 520 с.
17. **Baker F. B.** The Basics of Item Response Theory. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD, 2001.
18. **Gregory R. J.** Psychological testing: History, principles, and applications (5th edition). New York: Pearson, 2007.
19. **Gulliksen H.** Theory of Mental Tests. — John Wiley & Sons Inc, 1950.
20. **Kuravsky L. S., Malykh S. B.** Application of Markov models for analysis of development of psychological characteristics // Australian Journal of Educational & Developmental Psychology. 2004. Vol. 2. P. 29—40.
21. **Kuravsky L. S. and Baranov S. N.** Condition monitoring of the structures suffered acoustic fatigue failure and forecasting their service life // Proc. Condition Monitoring 2003, Oxford, United Kingdom. July 2003. P. 256—279.
22. **Kuravsky L. S. and Baranov S. N.** Neural networks in fatigue damage recognition: diagnostics and statistical analysis // Proc. 11th International Congress on Sound and Vibration, St.-Petersburg, Russia. July 2004. P. 2929—2944.
23. **Kuravsky L. S. and Baranov S. N.** The concept of multifactor Markov networks and its application to forecasting and diagnostics of technical systems // In: Proc. Condition Monitoring 2005, Cambridge, United Kingdom. July 2005. P. 111—117.
24. **Kuravsky L. S., Baranov S. N. and Yuryev G. A.** Synthesis and identification of hidden Markov models based on a novel statistical technique in condition monitoring // In: Proc. 7th International Conference on Condition Monitoring & Machinery Failure Prevention Technologies, Stratford-upon-Avon, England. June 2010. 23 p.
25. **Kuravsky L. S., Marmalyuk P. A. and Panfilova A. S.** Estimation of goodness-of-fit measures for identification of unrestricted factor models employing arbitrarily distributed observed data // In: Proc. 8th International Conference on Condition Monitoring & Machinery Failure Prevention Technologies, Cardiff, UK. June 2011. 19 p.
26. **Rasch G.** Probabilistic models for some intelligence and attainment tests // Copenhagen, Danish Institute for Educational Research, expanded edition (1980) with foreword and afterword by B. D. Wright. Chicago: The University of Chicago Press. 1960/1980.
27. **Roweis S. and Ghahramani Z.** A unifying review of linear Gaussian models // Neural Computation. 1999. Vol. 11. N 2. P. 305—345.
28. **Wright B. D., Masters G. N.** Rating scale analysis. Rasch measurements. Chicago: MESA Press, 1982.
29. **Wright B. D., Stone M. N.** Best Test Design. Chicago: MESA Press, 1979.


[9] Следует отметить, что процедура фильтрации, в которой $\mathbf{U}(t)$ определяется путем интегрирования уравнения Риккати, является более корректной, чем используемые в значительном числе приложений аналогичные процедуры, где $\mathbf{U}(t)$ находится как решение уравнения $\mathbf{F}\mathbf{U}(t) + \mathbf{U}(t)\mathbf{F}^{\mathrm{T}} - \mathbf{U}(t)\mathbf{R}^{-1}\mathbf{U}(t) = 0$ для стационарного случая.

УДК 615.47: 681.31

**В. В. Мажуга**, магистрант,
Российский университет дружбы народов,
**В. М. Хачумов**, д-р техн. наук, гл. науч. сотр.,
e-mail: vmh@isa.ru
Федеральное государственное
бюджетное учреждение науки
Институт системного анализа РАН

# Цифровая фильтрация и анализ электрокардиограмм

*Предложены алгоритмы предобработки электрокардиограмм цифровыми фильтрами и построения амплитудного распределения для облегчения их дальнейшего анализа. Разработан алгоритм выделения и определения параметров зубцов. Приведены результаты экспериментальных исследований.*

***Ключевые слова:*** *электрокардиограмма, цифровые фильтры, визуализация, зубцы, алгоритм, время вычисления*

## Введение

В общем случае интеллектуальный анализ данных представляет собой процесс обнаружения закономерностей в потоках данных или изображений. Обычно такие закономерности нельзя обнаружить простым просмотром, поскольку связи слишком сложны и требуют обработки большого объема информации в условиях высокой зашумленности данных. Врачи-кардиологи обладают богатым практическим опытом и высокой квалификацией, что позволяет им визуально интерпретировать биомедицинские сигналы [1]. В то же время они довольно медленно выполняют вычисления, направленные на выявление различных закономерностей в протяженных во времени процессах. Практически невозможной для человека является предобработка данных, направленная на улучшение качества электрокардиограмм (ЭКГ). Должен быть достигнут определенный компромисс в анализе ЭКГ между врачом и компьютерной программой. Программа поддержки принятия решений должна предоставлять эксперту обработанные снимки для визуального анализа и результаты расчетов. При этом наиболее востребованными являются быстрые алгоритмы, написанные для мобильных (переносимых) медицинских устройств. Конечная цель проводимых исследований — создание алгоритмов и программ, способных автоматически анализировать и классифицировать сигналы ЭКГ, оставляя при этом принятие окончательного решения врачу. В настоящей работе представлена часть исследований, охватывающая методы предварительной обработки сигналов, поступающих от кардиографа, распознавания *QRS*-комплекса, а также классификации пиков кардиограммы.

## 1. Постановка задачи

Электрокардиограмма представляет собой запись изменения электрических потенциалов сердца с течением времени [2, 3]. Интерпретация сигнала относится к сложной области распознавания образов. На рис. 1 показан в качестве примера сигнал из базы данных Медицинского центра Массачусетского технологического института (MIT-BIN ECG).

На рис. 1 каждая ячейка сетки имеет размер 0,5 мВ по оси ординат и 0,2 с по оси абсцисс.

На рис. 2 показаны основные этапы, осуществляемые при обработке сигналов, которые заканчиваются классификацией ЭКГ. Существенную часть этого процесса составляет предварительная обработка сигнала.

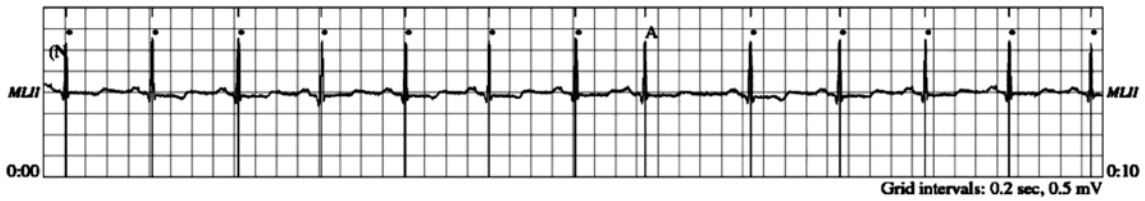



**Рис. 1. Пример сигнала ЭКГ**

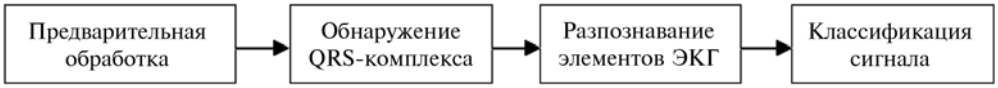



**Рис. 2. Основные этапы анализа ЭКГ**

Прохождение электрического импульса по проводящей системе сердца записывается в виде пиков, называемых зубцами ЭКГ. Зубцы принято обозначать латинскими буквами *P*, *Q*, *R*, *S* и *T*, как это показано на рис. 3.

В статье делается упор на предобработку, выделение и измерение ряда параметров указанных зубцов. Отнесение ЭКГ к классам возможных заболеваний по найденным параметрам зубцов, образующих комплекс, является конечной целью дальнейших исследований.

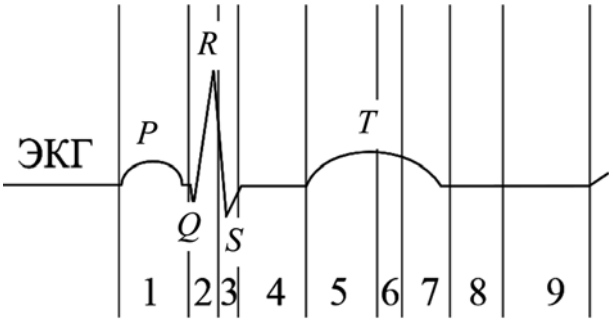



**Рис. 3. Схематическое изображение нормальной ЭКГ**

## 2. Предварительная обработка сигнала ЭКГ

Анализ ЭКГ основан на измерении ее амплитудных и временных параметров. Сигнал, поступающий от кардиографа (рис. 4), содержит многочисленные помехи, поэтому возникает необходимость в его предварительной обработке [1, 4].

Здесь и далее на всех рисунках по оси ординат отложено напряжение в милливольтах, а по оси абсцисс — время в миллисекундах.

Выделяют следующие виды помех, присутствующих в ЭКГ:

- сетевая наводка частотой 50 или 60 Гц [5]);
- дрейф изолинии в результате плохого контакта электродов с кожей [6];
- двигательные артефакты [1]) и др.

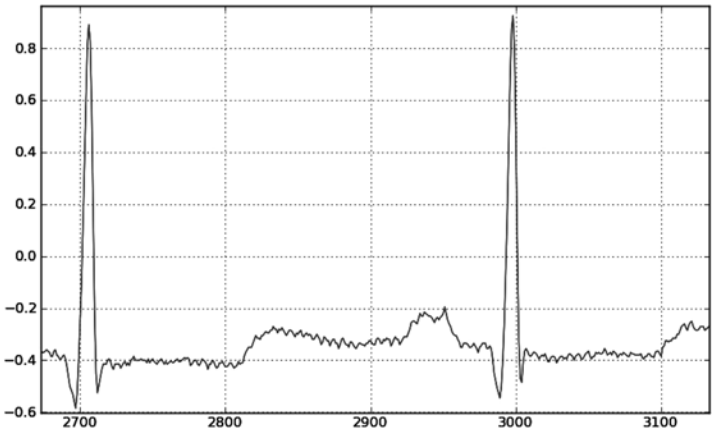



**Рис. 4. Исходный ЭКГ-сигнал (фрагмент)**

Большая часть помех может быть устранена применением цифровых фильтров. На рис. 5 показан сегмент ЭКГ с низкочастотным шумом, а также приведены результаты обработки сигнала с помощью фильтров Баттерворта и Гаусса [7—9].

Как видно из рисунков, низкочастотные артефакты легко удаляются, а высокочастотные компоненты сигнала остаются практически без изменений, что является положительным фактором. Например, обработка фильтром Баттерворта не меняет характеристик *QRS*-комплекса. Это объясняется тем фактом, что коэффициент усиления фильтра при высоких частотах близок к единице.

Для анализа ЭКГ важной информацией является амплитудный спектр сигнала. Результаты вычисления амплитуд для исходного зашумленного ЭКГ-сигнала, а также амплитудный спектр сигнала после фильтрации представлены на рис. 6.

Видно, что применяемые фильтры достаточно хорошо справляются с удалением низкочастотного шума. Часто для фильтрации используют

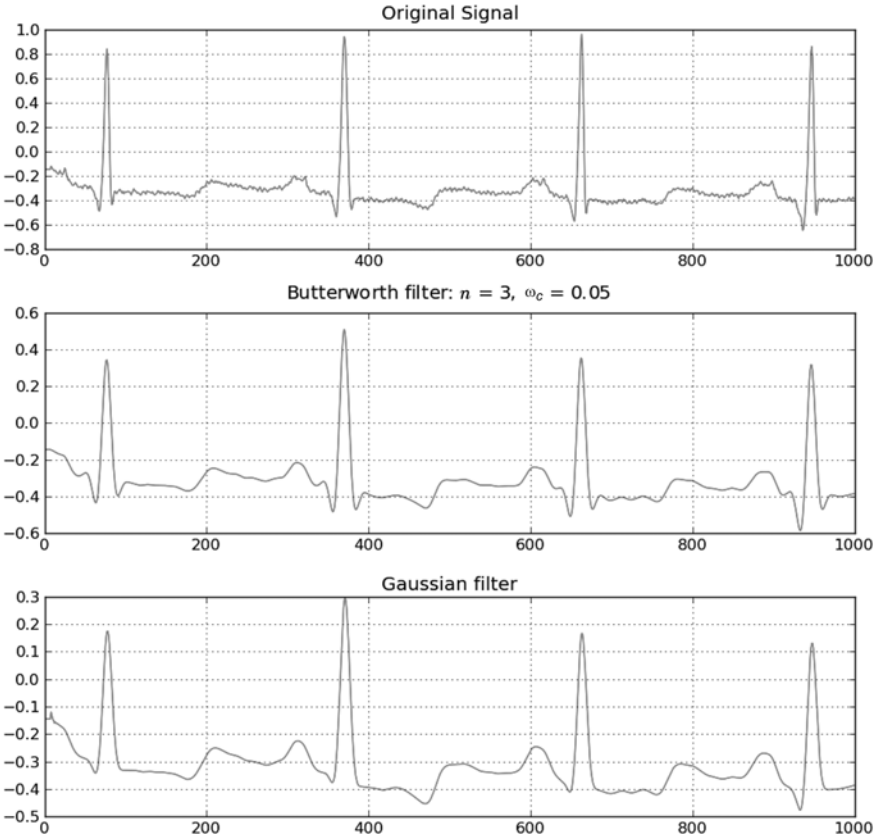



**Рис. 5. Исходный сигнал и результаты фильтрации**

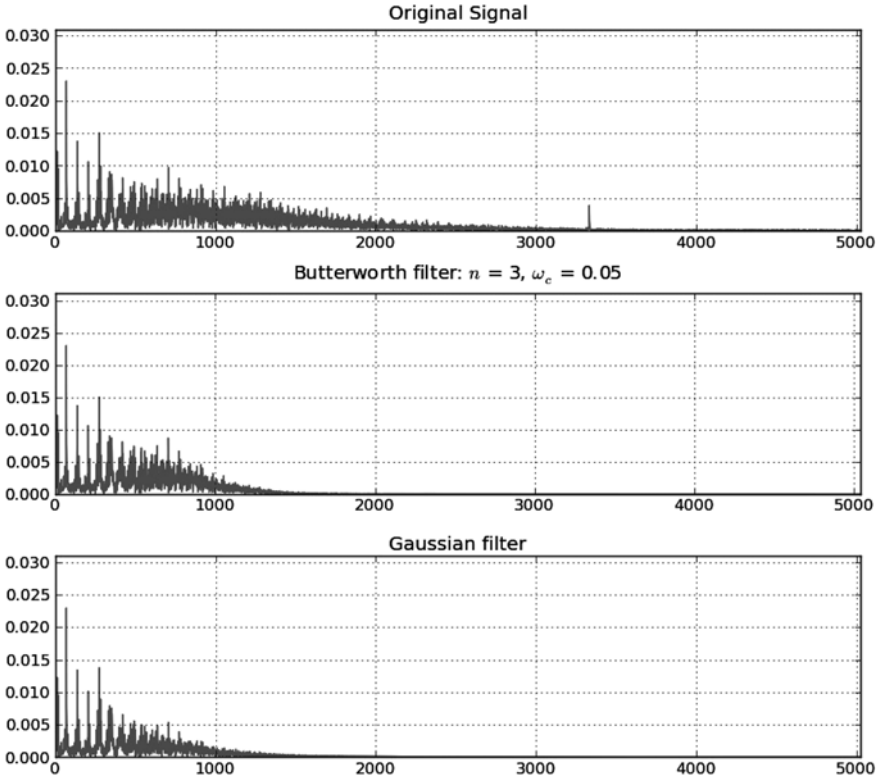



**Рис. 6. Амплитудные спектры ЭКГ до и после фильтрации**

быстрое преобразование Фурье (БПФ) [10]. Воздействуя на коэффициенты прямого Фурье-преобразования и проводя восстановление сигнала, можно добиться требуемого качества.

### 3. Обнаружение *QRS*-комплекса

Распознавание *QRS*-комплекса является отправной точкой при определении сердечного цикла. *QRS*-комплекс образуется в результате сокращений мускулатуры желудочков сердца. *R*-зубец представлен на ЭКГ резкими и высокими пиками. Его амплитуда составляет примерно 1 мВ, а продолжительность 80...100 мс. Поскольку скорость изменения функции определяется производной, будем использовать оператор $d/dt$ для обнаружения зубца *R*.

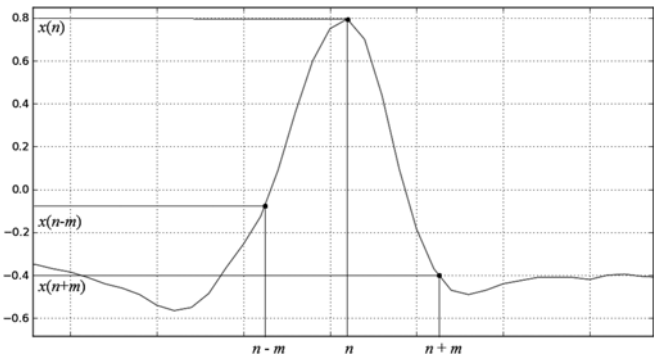



**Рис. 7. Выявление *R*-пика**

Представим исходный ЭКГ-сигнал в виде одномерного массива $x(n)$, $n = 1, ..., N$. Определение *R*-пика осуществляется согласно следующему алгоритму: для каждого отсчета данного сигнала $x(n)$ вычисляются значения левой и правой производных. Критерием наличия пика является выполнение условия

$$(x(n) - x(n + m)) \times \times (x(n - m) - x(n)) < 0,$$

где $2m$ — ширина сканирующего окна (рис. 7).

Точка-претендент на пик проверяется с использованием порога $TH = 0{,}5$ мВ. Зубцы *Q* и *S* определяются как глобальные минимумы в пределах заданного интервала, отсчитываемого от соответствующего *R*-пика. Длина интервала определяется, основываясь на продолжительности *QRS*-комплекса. На рис. 8 представлены результаты работы алгоритма.

Как видно из рисунка, на каждом интервале имеется единственный импульс с амплитудой, превышающей 0,5 мВ, сигнал в местах расположения пиков *P* и *T* значительно ниже, что помогает однозначно определить зубец *R*. Обнаружение зубцов *P* и *T* является трудной задачей в связи с тем, что они довольно малы по амплитуде и имеют изменчивую форму [1, 11, 12].

Сигнал ЭКГ обрабатывается следующим алгоритмом.

1. Обнаруживается *QRS*-комплекс, удаляется и заменяется базовой линией.

2. Результирующий сигнал подвергается фильтрации с удалением низкочастотных компонент.

3. На каждом *SQ*-участке отмечаются точки, удовлетворяющие следующим условиям:

$$(x(n) - x(n + m)) \cdot (x(n - m) - x(n)) < 0$$
$$|x(n) - x(n + m)| < TH_r$$
$$|x(n) - x(n - m)| < TH_l.$$

4. Выбранные на предыдущем шаге точки $x(n_k)$, ..., $x(n_{k+p})$, распределяются по группам согласно следующему правилу:

```
groups = []
peaks = []
p_prev = −1
for x(i) in x(n_k), ..., x(n_{k+p}):
    if p_prev ! = −1:
        # если расстояние между пиками меньше
        заданного значения
        if |p_prev − x(i)| < max_dist:
```

```
            # записываем пик в текущую группу
            peaks.add(x(i))
        # если расстояние между пиками больше
        заданного значения
        else:
            # добавляем текущую группу в список
            всех групп
            groups.add(peaks)
            # создаем новую группу
            peaks = []
            # добавляем найденный
            пик в новую группу
            peaks.add(x(i))
    else:
        peaks.add(x(i))
p_prev = x(i)
```

(Таким образом, множество groups содержит все группы пиков, удаленных друг от друга на расстояние, превышающее константу max_dist (рис. 9)).

5. Для каждого подмножества peaks из множества groups находится максимальный элемент.

6. Из всех значений, найденных на предыдущем шаге, определяются первый и второй по значению глобальные максимумы и отмечаются соответственно как зубцы *P* и *T*.

7. Осуществляется переход к шагу 3 до исчерпания множества интервалов *SQ*.

Результат работы алгоритма представлен на рис. 10.

Как видно из рисунка, в результате работы алгоритма удается обнаруживать все виды зубцов и соответственно определять их параметры.

### 4. Применение полученных результатов

Выходные данные, полученные в результате работы алгоритма выделения зубцов ЭКГ, могут быть использованы для обнаружения различных патологий в работе сердца. Для анализа вариабельности сердечного ритма (ВСР) строится график Пуанкаре (рис. 11), отражающий зависимость между длинами соседних *RR*-интервалов.

Нормальная форма распределения точек Пуанкаре представляет собой эллипс, вытянутый вдоль биссектрисы. Такое расположение эллипса означает, что к значению дыхательной аритмии прибавлено некоторое значение недыхательной аритмии. Форма в виде круга означает отсутствие недыхательных компонентов аритмии.

Узкий овал соответствует преобладанию недыхательных компонентов в общей вариабельности ритма. Этот способ оценки ВСР относится к методам нелинейного анализа и является особенно полезным для случаев, когда на фоне монотонности ритма встречаются редкие и внезапные нарушения. В таблице приведен небольшой фрагмент резуль-

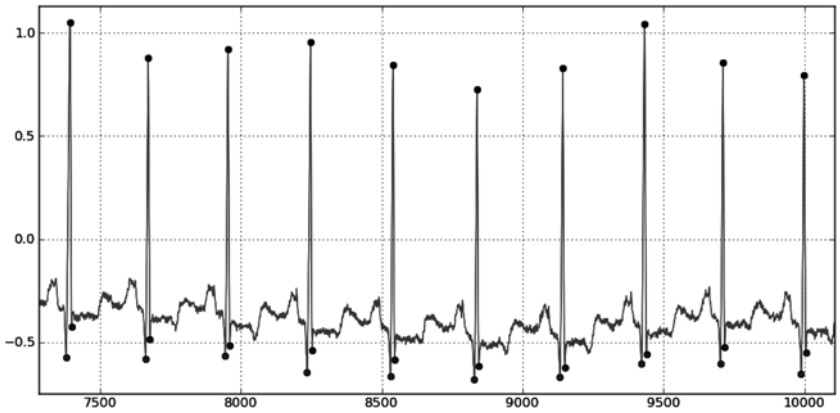



**Рис. 8. Сигнал ЭКГ с размеченным *QRS*-комплексом**

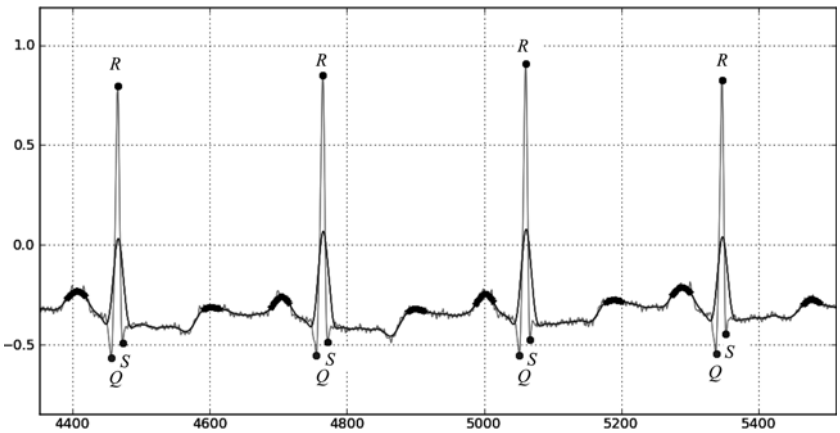



**Рис. 9. Обнаружение зубцов *Q*, *R*, *S***

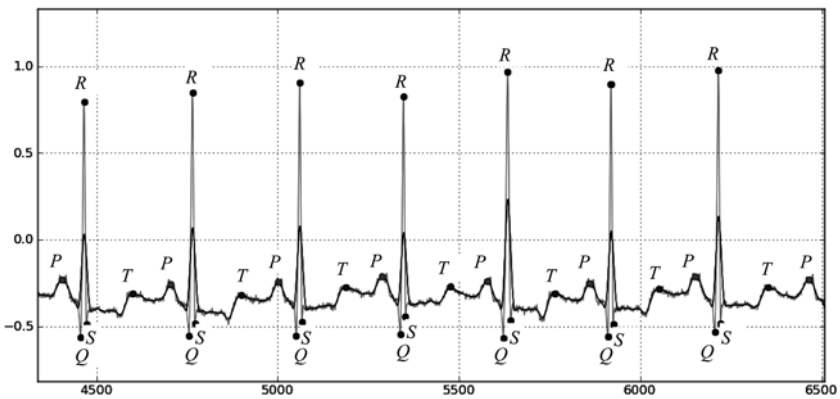



**Рис. 10. Результат работы алгоритма обнаружения зубцов *P* и *T***

**Время обнаружения и амплитуда зубцов *Q, R, S, T* и *P***

| *Q* | | *R* | | *S* | | *T* | | *P* | |
|---|---|---|---|---|---|---|---|---|---|
| с | мВ | с | мВ | с | мВ | с | мВ | с | мВ |
| 1,0 | −0,23 | 1,03 | 0,178 | 1,05 | −0,092 | 1,43 | −0,309 | 0,69 | −0,246 |
| 1,82 | −0,243 | 1,84 | 0,06 | 1,86 | −0,06 | 2,18 | −0,309 | 2,47 | −0,269 |
| 2,6 | −0,373 | 2,63 | 0,026 | 2,65 | −0,079 | 2,97 | −0,289 | 3,25 | −0,255 |
| 3,39 | −0,273 | 3,42 | −0,001 | 3,44 | −0,09 | 3,8 | −0,303 | 4,05 | −0,235 |
| 4,18 | −0,286 | 4,21 | 0,038 | 4,22 | −0,068 | 4,59 | −0,283 | 4,86 | −0,225 |

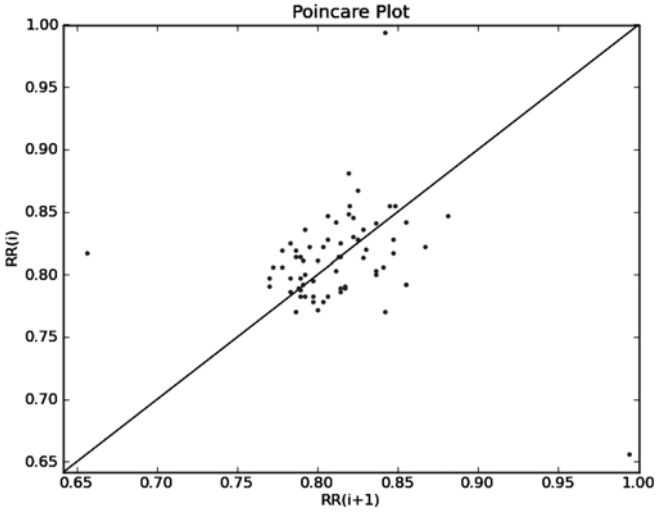



**Рис. 11. График Пуанкаре**

тата работы алгоритма по обнаружению зубцов с фиксацией времени вычисления на ПЭВМ.

На основе данных, представленных в таблице, можно выявить различные патологии с применением специального программного обеспечения. Эктопический предсердный ритм характеризуется частотой сердечных сокращений (ЧСС) от 50 до 100 мин$^{-1}$, при этом зубец *P* отрицателен в отведениях, а интервал $PQ \geqslant 0{,}12$ с. Признаками желудочковой тахикардии служат правильный ритм с частотой 110—250 мин$^{-1}$, комплексом $QRS > 0{,}12$ с, а также разнонаправленностью сегмента *ST* и зубца *T* с комплексом *QRS*. Время анализа ЭКГ можно существенно сократить, достигая практически режима реального времени, используя программно-аппаратные средства параллельной обработки данных [13].

## Заключение

В работе рассмотрен метод компьютерного анализа медицинских данных для выявления характерных участков электрокардиограммы. Реализованы алгоритмы, осуществляющие предварительную обработку ЭКГ-сигнала и выделение пяти основных зубцов: *P*, *Q*, *R*, *S* и *T*. Компьютерный анализ ЭГК позволяет получать информацию об амплитуде зубцов и расстояний между ними, тем самым становится возможным усилить достоверность и точность диагностики, проводимой врачом.

*Работа выполнена при частичной поддержке Государственных контрактов № 02.740.11.0526 и № 07.514.11.4048, предусматривающих интеллектуальную обработку сложных сигналов.*

**Список литературы**

1. **Rangaraj M. R.** Biomedical Signal Analysis: A case study approach. Wiley-Interscience / IEEE Press, 2002. 552 p.

2. **Зубдинов Ю. И.** Азбука ЭКГ и боли в сердце. Ростов н/Д: Феникс, 2006. 240 с.

3. **Азимов А.** Тело человека: строение и функции. М.: Эксмо, 2010. 416 с.

4. **Levkov C., Mihov G., Ivanov R., Daskalov I., Christov I., Dotsinsky I.** Removal of power-line interference from the ECG: a review of the subtraction procedure. — Biomed Eng Online. 2005. URL: http://www.biomedical-engineering-online.com/content/4/1/50

5. **Лазарева Г. Ю.** Справочник фельдшера. М.: Рипол Классик, 2010. 640 с.

6. **Электрокардиограмма**. URL: http://www.happydoctor.ru/info/536

7. **Прэтт У.** Цифровая обработка изображений: В 2 т. М.: Мир, 1982.

8. **MATLAB** & Toolboxes. Обработка сигналов и изображений. URL: http://matlab.exponenta.ru/signalprocess/book1/index.php

9. **Рабинер Л. Р., Гоулд В.** Теория и применение цифровой обработки сигналов. М.: Мир, 1978. 848 с

10. **Залманзон Л. А.** Преобразования Фурье, Уолша, Хаара и их применение в управлении, связи и других областях. М.: Наука, 1989. 496 с.

11. **Clifford G. D., Azuaje F., McSharry P. E.** Advanced Methods and Tools for ECG Analysis. — Boston/London: Artech House Publishing, 2006. URL: http://www.biomedical-engineering-online.com/content/6/1/18

12. **Bhardwaj S., Lee D.-S, Chung W.-Y.** An advanced ECG signal processing for ubiquitous healthcare system // Proc. of International Conference on Control, Automation and Systems, Seoul. 2007. P. 2433—2436. URL: http://alexandria.tue.nl/openaccess/Metis226366.pdf

13. **Параллельная** программная система для распознавания графических образов на основе искусственных нейронных сетей (ИНС ППС). — Свидетельство о государственной регистрации программы для ЭВМ № 2010610208 от 11 января 2010 г.

УДК 004.946

**В. А. Немтинов,** д-р техн. наук, проф., зав. каф.,
e-mail: nemtinov@mail.gaps.tstu.ru,
**Н. В. Пеньшин**, канд. экон. наук, доц., зав. каф.,
**Ю. А. Донских**, магистр,
**К. В. Немтинов**, студент,
**Е. С. Егоров,** студент,
Тамбовский государственный
технический университет, Тамбов,

# Имитационное моделирование динамических процессов при управлении городским пассажирским транспортом

*Рассмотрены вопросы разработки имитационной математической модели функционирования городской маршрутной транспортной сети и информационной системы, позволяющей автоматизировать процесс построения модели и обработку результатов моделирования в среде SIMUL8.*

***Ключевые слова**: имитационная математическая модель, городская маршрутная транспортная сеть, информационная система, среда моделирования SIMUL8*

## Введение

В настоящее время движение транспорта во многих городах РФ представляет собой сложную динамическую систему, характеризующуюся высоким уровнем неопределенности исходной информации и сложностью ее поведения. Для решения многих проблем, связанных с управлением таких систем, можно использовать компьютерное моделирование, реализующее методологию системного анализа, центральной процедурой которого является построение обобщенной модели, отражающей все факторы реальной системы. При этом в качестве методологии исследования выступает вычислительный эксперимент.

Компьютерное моделирование значительно расширяет возможности и эффективность работы лиц, принимающих решения (ЛПР), предоставляя им удобный инструмент и средства для достижения поставленных целей. Оно реализует итерационный характер разработки модели системы, поэтапный характер детализации моделируемых подсистем, что позволяет постепенно увеличивать полноту оценки принимаемых решений по мере выявления новых проблем и получения новой информации.

Современные тенденции в области имитационного моделирования связаны с развитием проблемно-ориентированных систем, созданием встроенных средств для интеграции моделей в единый модельный комплекс. Технологический уровень современных систем моделирования характеризуется большим выбором базовых концепций формализации и структуризации моделируемых систем, развитыми графическими интерфейсами и анимационным выводом результатов [1, 2].

Транспортная система является одной из основных составных частей инфраструктуры города, которая обеспечивает жизненно важные потребности населения. Функционирование всех отраслей городского хозяйства невозможно без рациональной и налаженной работы городской транспортной системы (ГТС). Поэтому рационализация ее развития и планирования является одной из актуальных проблем теории и практики планирования. К группе планирования относятся задачи принятия централизованных решений об использовании ресурсов ГТС. В эту группу входят задачи планирования развития транспортной системы, маршрутизации, составления расписаний и т. д. [3, 4].

Рост концентрации и увеличение доли городского населения — это объективная тенденция развития общества. Быстрые темпы роста городского населения и увеличение его подвижности порождают целый ряд проблем, связанных с развитием транспорта в городах. Роль и масштабность работы ГТС в условиях непрерывного роста городов, концентрации в них населения и насыщенности транспортными средствами требуют проведения широкого круга научных исследований и практических работ, направленных на совершенствование сети городского пассажирского транспорта (ГПТ).

Цель данной работы — разработка математической модели функционирования сети ГПТ и механизма, позволяющего автоматизировать процесс создания имитационной модели на ее основе для проведения дальнейших компьютерных экспериментов. Для выполнения поставленной цели необходимо решить следующий комплекс задач, связанных с формальным описанием структуры сети ГПТ и параметров ее функционирования; разработкой математической модели функционирования сети ГПТ, структуры базы данных для хранения информации, алгоритма построения имитационной модели.

В настоящее время известно множество визуальных средств моделирования. Среди них универсальным средством имитационного моделирования для рассматриваемого класса задач является система SIMUL8 [5].

### Разработка математической модели функционирования транспортной сети

Наиболее распространенный и наглядный подход к описанию транспортной сети — это представление ее в виде графа. Вершины графа представляют собой остановочные пункты (остановки) — специально оборудованные места для посадки и высадки пассажиров. Ребра графа отождествляются с перегонами между остановочными пунктами. Их основными характеристиками являются расстояние и максимальная разрешенная скорость движения. Транспортные средства в сети движутся по заранее определенным маршрутам. Маршрут — это перечень остановочных пунктов, которые в предопределенном порядке объезжает транспортное средство. Каждый маршрут характеризуется узловыми точками. Это остановочные пункты, на которых фиксируется соответствие текущего времени и времени подъезда транспортного средства по расписанию движения к данной остановке.

Представление транспортной сети в виде графа дает информационной системе возможность использовать большое количество уже существующих алгоритмов для программной оптимизации транспортных перевозок по результатам выполнения модели.

Режим работы маршрутной транспортной сети подчиняется, в первую очередь, расписанию движения подвижного состава. В нем содержится информация о количестве подвижного состава на каждом маршруте, числе рейсов каждого транспортного средства и времени прохождения контрольных точек маршрута. Таким образом, под оптимизацией работы ГПТ, в первую очередь, подразумевается определение оптимального расписания движения транспорта, а также характеристик маршрутных транспортных средств, к которым относятся скорость перемещения, вместимость, стоимость обслуживания, стоимость проезда и др. [6—8].

Составление расписания является обычной задачей линейного программирования, однако на практике выполнение ее не так уж просто. В расписание движения вносит свои коррективы окружающая обстановка на дорогах — пробки, аварии, светофоры. В связи с этим расписание, составленное обычными методами, целесообразно проверять на имитационной модели (которая учитывает вышеупомянутые нюансы) и по результатам ее выполнения можно вносить коррективы в режим работы транспортной сети.

Так как сеть ГПТ является сложной динамической системой и характеризуется большим количеством стохастической информации, то и сама математическая модель основана на различного рода вероятностях.

Исходная информация для математической модели функционирования транспортной сети: $n_{ik}^{l}(t)$ — число пассажиров, севших на $i$-й маршрут на $k$-й остановке в момент времени $t$; $n_{ik}^{m}(t)$ — число пассажиров, вышедших на $k$-й остановке с $i$-го маршрута; $L_j^i$ — расстояние между $i$-м и $j$-м остановочным пунктом; $N_i$ — число транспортных средств на $i$-м маршруте.

$R_{ic} = \{t_{ik1}, t_{ik2}, ..., t_{ikn}\}$ — расписание движения ГПТ, где $i$ — номер маршрута; $c$ — номер транспортного средства на маршруте; $n$ — число узловых точек маршрута; $kn$ — номер узловой точки маршрута; $t$ — время прибытия транспортного средства.

Весь период времени моделирования работы транспортной сети разбивается на конечное число интервалов $T = \{t_1, t_2, ..., t_s\}$, где $T$ — период времени функционирования транспортной сети; $s$ — число интервалов, на которое был разбит весь период времени.

Вероятность, с которой с $i$-й остановки в $k$-й маршрут в момент времени $t$ сядет пассажир, равна

$$p_{ij}^{l} = \frac{\sum\limits_{t} n_{ik}^{l}(t)}{\sum\limits_{t}\sum\limits_{r=1}^{n} n_{rk}^{l}(t)}. \qquad (1)$$

Вероятность, с которой на $i$-й остановке с $k$-го маршрута выйдет пассажир, равна

$$p_{ij}^{m} = \frac{\sum\limits_{t} n_{ik}^{m}(t)}{\sum\limits_{t}\sum\limits_{r=1}^{n} n_{rk}^{m}(t)}. \qquad (2)$$

Интервал появления пассажиров на $i$-й остановке, которые поедут на $k$-м маршруте, будет

$$\Delta t^{l}(t_s) = \frac{t_s}{\sum\limits_{t}\sum\limits_{r=1}^{n} n_{rk}^{l}(t_s)}. \qquad (3)$$

Интервал времени, с которым пассажиры выходят на $i$-й остановке с $k$-го маршрута, равен

$$\Delta t^{m}(t_s) = \frac{t_s}{\sum\limits_{t}\sum\limits_{r=1}^{n} n_{rk}^{m}(t_s)}. \qquad (4)$$

Скорость движения транспорта между остановочными пунктами

$$V = \frac{t_{k_{i-1}} - t_{k_i}}{L_{t_{k_{i-1}}}^{t_{k_i}}}. \qquad (5)$$

Время прибытия транспортного средства на $b$-й остановочный пункт

$$T_b^{prib} = t_{k_i} + VL_b^{k_i} + \alpha_1(t), \qquad (6)$$

где $t_{k_i}$ — время подъезда к ближайшей предшествующей контрольной точке; $L_b^{k_i}$ — расстояние от ближайшей узловой точки до текущей остановки; $\alpha_1(t)$ — случайная величина, отражающая неточность выполнения расписания, возникающую под влиянием пробок и аварийных ситуаций на дорогах, а также работы светофоров.

Число пассажиров в $c$-м автобусе $k$-го рейса на $i$-й остановке будет

$$Q_{ic}^{k} = Q_{ic}^{k-1}(t) + N_{ic}^{\text{вошед}} - N_{ic}^{\text{вышед}}, \qquad (7)$$

где

$$N_{ic}^{\text{вошед}} = \frac{t_s}{\Delta t^{l}(t_s)} p_{ic}^{l},\ N_{ic}^{\text{вышед}} = \frac{t_s}{\Delta t^{m}(t_s)} p_{ic}^{m}. \qquad (8)$$

Большинство исходных данных, используемых в модели (1)—(8), получены в результате проведения натурных исследований в течение продолжительного промежутка времени пассажиропотоков на примере г. Тамбова.

## Реализация имитационной модели и системы ее автоматизированного построения в среде SIMUL8

Для реализации имитационной модели была выбрана система имитационного моделирования динамических процессов SIMUL8 [9]. Данная система имеет мощный набор инструментов для проведения моделирования и дальнейшей обработки результатов. В основе разработки имитационной модели в среде SIMUL8 лежит объектно-ориентированный подход. Каждому реальному объекту или процессу ставится в соответствие объект (либо сочетание объектов) моделирования SIMUL8.

При создании имитационной модели ГПТ можно выделить следующие соответствия:

- остановочный пункт — имитируется объектом Work Center;
- транспортное средство — Work Item;
- процесс появления пассажиров на остановке — Work Entry Point;
- очередь пассажиров на остановке, а также очередь транспортных средств перед ней — моделируются с помощью объектов типа storage bin;
- маршруты и расписание движения хранит в себе объект Job Matrix.

Для хранения информации о текущем числе пассажиров, общем числе перевезенных людей и т. д. используются объекты label, которые присоединяются к объектам типа Work Item.

Для хранения всей необходимой информации о транспортной сети была разработана структура базы данных в СУБД MS Access (рис. 1).

Для удобства наполнения базы данных информацией, а также автоматизированного построения модели в среде SIMUL8 по информации из БД на языке Visual Basic была разработана программа, рабочее окно которой представлено на рис. 2 (см. четвертую сторону обложки).

Результатом работы данной программы является готовая имитационная модель сети ГПТ в среде SIMUL8 (рис. 3). На рис. 4 (см. четвертую сторону

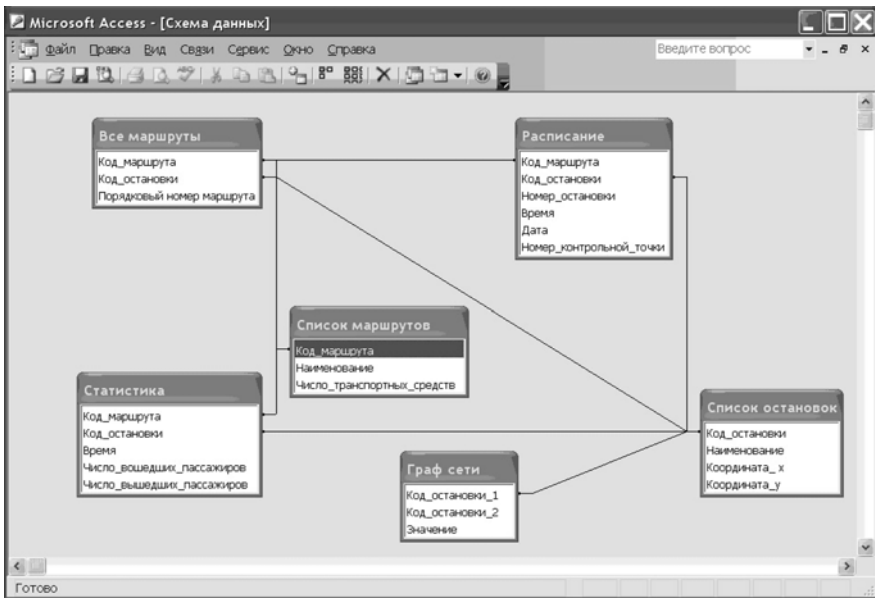



**Рис. 1. Схема базы данных для хранения информации о сети ГПТ**

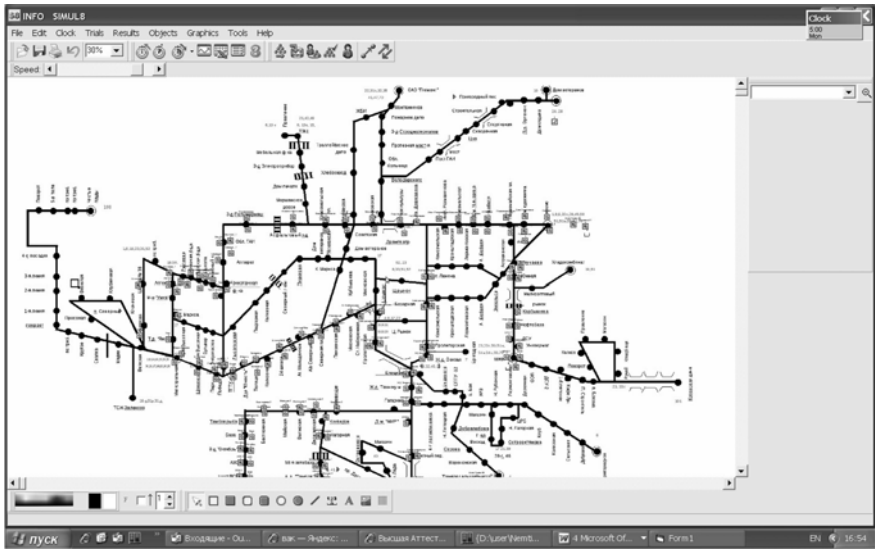


**Рис. 3. Визуализация общего вида модели сети ГПТ г. Тамбова**

обложки) представлены результаты моделирования отдельного остановочного пункта, в том числе диаграмма зависимости числа человек на остановке от времени.

## Заключение

Разработанная авторами математическая модель функционирования маршрутной транспортной сети и информационная система, реализованная в среде SIMUL8, позволяет автоматизировать процесс построения имитационной модели и обработку результатов моделирования. Полученная модель является мощным инструментом анализа, оптимизации и развития маршрутной транспортной сети. С ее помощью можно моделировать различные ситуации и заранее продумывать пути решения проблем, возникающих в системе ГПТ.

### Список литературы


1. **Хемди А., Таха Н.** Имитационное моделирование. Введение в исследование операций. М.: Вильямс, 2007. 737 с.

2. **Строгалев В. П., Толкачева И. О.** Имитационное моделирование. М.: Изд-во МГТУ им. Н. Э. Баумана, 2008. С. 697—737.

3. **Сафронов Э. А.** Транспортные системы городов и регионов: учеб. пособие. М.: Издательство АСВ, 2005. 272 с.

4. **Сафронов К. Э., Киммель Д. С.** Начало реформирования ГПТ — совершенствование маршрутных сетей городов // Автомобильный транспорт. 2004. № 5. С. 57—58.

5. **Немтинов В. А., Немтинова Ю. В., Донских Ю. А.** Оперативное управление транспортными потоками в населенных пунктах с использованием системы моделирования динамических процессов // Вестник компьютерных и информационных технологий. 2009. № 3. С. 12—14.

6. **Донских Ю. А.** Моделирование маршрутных транспортных сетей населенных пунктов // Материалы III Всероссийской студенческой научно-практической конференции. 2010. С. 27—28.

7. **Немтинов В. А., Донских Ю. А.** Имитационная модель транспортных потоков г. Тамбова: Свидетельство о регистрации программ для ЭВМ № 2010610706 в Федеральной службе по интеллектуальной собственности, патентом и товарным знаком. Дата регистрации 20 января 2010 г.

8. **Немтинов В. А., Донских Ю. А., Немтинов К. В.** Имитационное моделирование транспортных систем // XIII научная конференция ТГТУ. Фундаментальные и прикладные исследования, инновационные технологии, профессиональное образование. Тамбов, 2010. С. 33—35.

9. **SIMUL8.** Animate Your Business. Simulation Software. SIMUL8 Corporation. 2002. 362 с.


# CONTENTS

**Адрес редакции:**

107076, Москва, Стромынский пер., 4

Телефон редакции журнала **(499) 269-5510**

E-mail: it@novtex.ru

Дизайнер *Т.Н. Погорелова*. Технический редактор *Е. В. Конова*.

Корректор *М.Г. Джавадян*.

Сдано в набор 07.02.2012. Подписано в печать 22.03.2012. Формат 60×88 1/8. Бумага офсетная.

Усл. печ. л. 9,8. Заказ IT312. Цена договорная.

Журнал зарегистрирован в Министерстве Российской Федерации по делам печати, телерадиовещания и средств массовых коммуникаций.

Свидетельство о регистрации ПИ № 77-15565 от 02 июня 2003 г.

Оригинал-макет ООО "Адвансед солюшнз". Отпечатано в ООО "Адвансед солюшнз".

105120, г. Москва, ул. Нижняя Сыромятническая, д. 5/7, стр. 2, офис 2.