Теорія оптимальних рішень

УДК ВФ 160.14

В.Б. БИГДАН

МОДЕЛИ УПРАВЛЕНИЯ ПОТОКАМИ ДАННЫХ В ПРОЦЕССАХ РАСПРЕДЕЛЕННОГО ПОИСКА ОПТИМАЛЬНЫХ РЕШЕНИЙ

Рассматриваются особенности реализации некоторых основных аспектов технологии Data Mining в рамках разработанной в Институте кибернетики имени В.М. Глушкова НАН Украины системы оптимизационно-имитационного моделирования NEDISOPT_D. Основное внимание уделяется моделям управления потоками данных в процессах распределенного поиска оптимальных решений с последующим накоплением результатов поиска в хранилищах опыта моделирования.

Введение. В связи с широким использованием методов и средств имитационного моделирования в практике исследования и проектирования сложных стохастических систем актуальными являются вопросы повышения эффективности указанных методов и созданных на их основе подходов.

Проблемы разработки систем управления в области космических исследований, транспорта, финансов, маркетинга, для различных отраслей экономики, особенно в условиях глобализации национальных экономик, потребовали расширения функциональных возможностей современных методов и средств имитационного моделирования. Значительное влияние на формирование такого рода требований также оказали проблемы, возникающие при разработке современных военных стратегий, тактик обучения личного состава групп быстрого реагирования, планов боевых действий применительно к условиям урбанизированных территорий, характеризую-щихся высоким уровнем риска и неопределенности.

Выполненные за последнее десятилетие исследования применительно к проблемам повышения эффективности методов и средств имитационного моделирования способствовали развитию новой методологии Data Farming. В основе этой методологии лежит интеграция возможностей методов имитационного моделирования, методов оптимизации (в первую очередь методов на основе основе эволюционных вычислений), мето-

дов и технологий распределенных вычислений, реализуемых на высокопроизводительных платформах вычислительной техники (кластерных или сетевых архитектурах) [1]. Как правило, процессы исследования и проектирования сложных систем на базе методологии Data Farming сопровождаются генерацией и накоплением больших объемов разнородной информации, составляющей основу практического опыта в соответствующей прикладной области и требующей использования методов интеллектуального анализа. Поэтому в методологии Data Farming существенным образом используются возможности новой информационной технологии Data Mining (discovery-driven data mining). Data Mining обеспечивает широкий спектр методов интеллектуального анализа исторической информации и многомерную визуализацию последней [2].

Концепция Data Farming впервые была предложена в 1998 г. Гари Хорном (Gary E.Horne) при разработке по заказу военно-морского флота США проекта "Альберт". Методология Data Farming обеспечивает высокопроизводительную генерацию и обработку огромных пространств параметров решений, делает возможным оценивание непредвиденных ситуаций (как положительных, так и отрицательных) и осуществление наиболее приемлемого выбора. Вокруг идеи Data Farming в 1999 г. было сформировано международное сообщество, которое регулярно проводит конференции International Data Farming Workshop. Основные результаты, полученные в области применения Data Farming, публикуются в трудах ежегодной конференции Winter Simulation Conference. В настоящее время сформировано одиннадцать интернациональных групп, которые занимаются вопросами практического применения методологии Data Farming в различных прикладных областях. Наиболее активными участниками здесь являются специалисты из США, Германии, Канады, Сингапура, Швеции, Голландии, Египта. Кроме практического применения усилия специалистов направлены на разработку методологических и технологических стандартов для Data Farming [3–5].

В условиях отсутствия в отечественной практике систем моделирования такого рода актуальной становится проблема расширения функциональных возможностей последних.

Постановка задачи и определение целей исследования. В Институте кибернетики имени В.М. Глушкова НАН Украины разработана система оптимизационно-имитационного моделирования NEDISOPT_D, которая соответствует основным концепциям Data Farming, поскольку интегрирует возможности методов имитационного моделирования, методов оптимизации и технологий распределенных вычислений, реализуемых на сетевых архитектурах [6]. Необходимость практического использования системы NEDISOPT_D потребовала расширения ее функциональных возможностей, в первую очередь, на основе методов и средств технологии Data Mining [7].

Цель работы – создание моделей и методов, поддерживающих управление потоками данных в процессе реализации оптимизационно-имитационных экс-

периментов и определяющих структурную организацию информационного хранилища опыта моделирования в соответствующих прикладных областях (исторических данных). При этом должна быть обеспечена возможность переиспользования такого рода данных, их эффективный поиск и удобный пользовательский доступ к ним.

Разработка моделей управления потоками данных для системы NEDISOPT_D должна базироваться на таких основных концепциях технологии Data Mining как "шаблоны" (pattern) и "хранилища данных" (data warehouse), а также следующих принципах реализации системы NEDISOPT_D: использование концепций "имитационное приложение", "оптимизационно-имитационная интеграция", "популяция решений"; представление программной среды системы NEDISOPT_D в виде многослойного сценария; представление сценариев оптимизационных стратегий и используемых ими данных в стандартизованных форматах; распределенный поиск оптимальных решений; реализация оптимизационно-имитационных экспериментов в формате сессий моделирования [8].

Согласно концепций, принятых в современной технологии Data Mining. шаблоны выступают в роли информационных моделей, отражающих структуру многоаспектных взаимоотношений исторических данных. Отличительной особенностью таких моделей является наличие разнородной информации (количественные, качественные и текстовые данные). Основное требование к таким моделям – представление информационных взаимосвязей в компактной форме, удобной для понимания человеком. Поскольку исторические данные в хранилищах размещаются в формате шаблонов, то для эффективного поиска нужной информации шаблоны должны иметь соответствующие ключи.

Модели-шаблоны для управления потоками входной информации сессии моделирования. На рисунке представлены потоки данных и основные сценарии системы NEDISOPT_D, формирующие иерархически структурированную программную среду поддержки сессий моделирования.

В соответствии с используемыми по ходу сессии моделирования классами данных (входные, выходные) хранилище данных системы NEDISOPT_D делится на две секции: для хранения входных данных и для результатов оптимизационно-имитационных экспериментов, объединяемых в хранилище опыта моделирования.

Входные данные для главного сценария MSCN, управляющего ходом выполнения сессии моделирования, объединены в группу конфигурационных параметров, шаблон для которых представлен в табл. 1. Ключом для данного шаблона является фраза "конфигурационные параметры сессии моделирования". Параметры nComp и maxNumAppl определяют конфигурацию сети, формируемую для сессии моделирования, и количество параллельно исполняемых приложений на компьютерах сети соответственно.

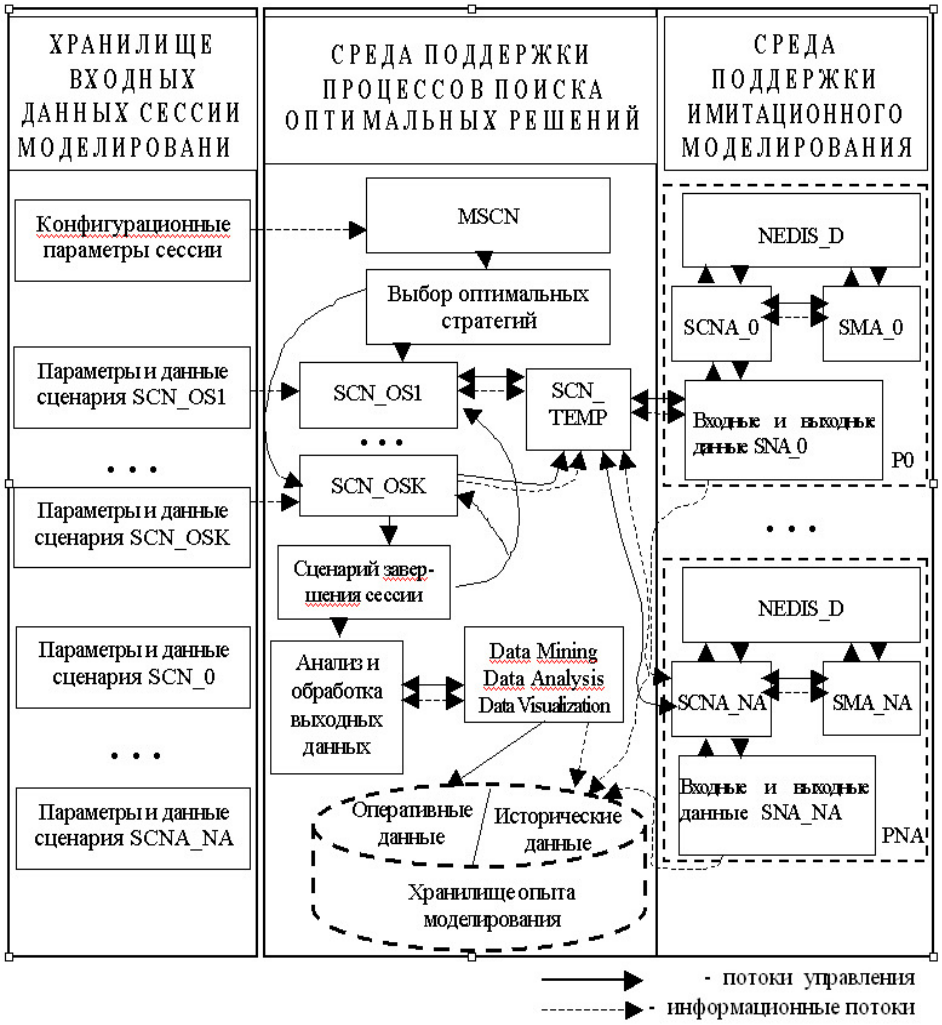

РИСУНОК. Схема потоков данных в системе NEDISOPT_D

Параметр typeOptimize, задающий композицию оптимизационных стратегий, определяет конфигурацию программной среды сессии моделирования как динамически формируемую цепочку сценариев, реализуемых в процессе поиска соответствующих оптимальных решений на распределенной архитектуре.

Следует заметить, что параметры nOpts и maxNumAppl являются также управляющими для сценария SCN_TEMP, который обеспечивает интерфейс (по данным и потокам) между средой поддержки процессов поиска оптимальных решений и средой поддержки имитационного моделирования.

Входная информация для сценариев SCN_OS1,…, SCN_OSK содержит два фрагмента: управляющие параметры и начальная популяция решений (множество характеристик оцениваемых альтернатив при различных уровнях факторов).

ТАБЛИЦА 1. Шаблон для конфигурационных параметров

№ п/п	Конфигурационные параметры сессии моделирования	
	Имя	Назначение
1	nOpts	Число оптимизационных стратегий в текущей сессии моделирования
2	nComp	Максимальное число компьютеров в сети, сконфигурированной применительно к задачам текущей сессии моделирования
3	maxNumAppl	Максимально допустимое число параллельно исполняемых процессов (приложений)
4	typeOptimize	Композиция сценариев оптимизационных стратегий применительно к текущей сессии моделирования
5	signPrintRes	Композиция типовых фрагментов результатов моделирования выдаваемых на печать
6	populSize	Размер начальной популяции
7	genSize	Количество генов в хромосомах
8	sortDirFit	Направление сортировки fitness-значений
9	signarch	Признаки архивации результатов сессии

Типовой шаблон управляющей информации для оптимизационных стратегий на примере генетического алгоритма (ГА) представлен в табл. 2. Ключевой здесь является фраза "управляющие параметры ГА". Естественно, что для схемы последовательного перебора вариантов (шаблон с ключевой фразой "управляющие параметры последовательного перебора вариантов") множество параметров будет иметь другой перечень.

Для всех типов оптимизационных стратегий разработан шаблон с ключевой фразой "популяция решений", содержащий характеристики множества альтернатив, которые будут оцениваться в процессе реализации оптимизационно-имитационных экспериментов.

ТАБЛИЦА 2. Шаблон для управляющих параметров ГА

№ п/п	Управляющие параметры ГА	
	Имя	Назначение
1	cBegPopForm	Признак формирования начальной популяции
2	typeCrossover	Тип кроссовера
3	typeSelect	Тип процедуры отбора
4	thResh	Порог отсечения
5	nRunup	Количество смен поколений на этапе разгона процесса эволюции
6	nGenPhase	Количество смен поколений в рамках каждого этапа
7	nPhase	Общее количество этапов
8	epsilon	Относительная точность определения

Входные данные для сценариев SCNA_0,…, SCNA_NA представляются множеством шаблонов, которые должны создаваться разработчиками приложения применительно к конкретной проблемной области и поставленным задачам исследований: шаблон для информации, управляющей прогоном имитационной модели, шаблон для информации, управляющей регистрацией и накоплением статистики, шаблон для представления числовых характеристик исследуемой системы (множество структурных, потоковых, функциональных характеристик), шаблоны для представления размеров штрафов за простои оборудования и размеров инвестиционных вкладов в изменение инфраструктуры исследуемых систем.

Следует заметить, что разрабатываемые шаблоны делятся на проблемно-ориентированные и проблемно-независимые (системные). Шаблоны для конфигурационных и управляющих параметров относятся к системным, остальные шаблоны – проблемно-ориентированные, а их структура и взаимосвязи между отдельными элементами определяются спецификой проблемной области и задачами конкретных сессий моделирования.

Модели-шаблоны для представления результатов сессии моделирования. Композиция фрагментов-шаблонов выходной информации, выдаваемой на печать или в хранилище опыта моделирования, определяется конфигурационным параметром signPrintRes.

В целом результаты сессии моделирования представляются последовательностью шаблонов, которые объединяются в соответствующие фрагменты. Первый фрагмент всегда представляется двумя шаблонами: заголовочным с ключевой фразой "Результаты сессии моделирования" и сведениями о дате и времени начала и завершения сессии; второй шаблон содержит сведения о конфигурационных параметрах сессии. Этот фрагмент является проблемно-независимым и обязательно включается в поток выходной информации сессии моделирования.

Результаты поиска оптимальной альтернативы в пространстве параметров-решений (факторов) представляются шаблонами в виде причинно-следственной матрицы. Идея такой матрицы была предложена Каору Исикава, японским специалистом в области процессов управления качеством.

Общий вид шаблона для причинно-следственной матрицы, принятой в системе NEDISOPT_D, представлен в табл. 3.

ТАБЛИЦА 3. Причинно-следственная матрица

Номер альтернативы	Результаты поиска оптимальной альтернативы							
	Характеристики альтернативы				Оценки альтернативы			
1	$fact_{11}$	$fact_{12}$	…	$fact_{1n}$	$resp_{11}$	$resp_{12}$	…	$resp_{1k}$
…	…	…	…	…	…	…	…	…
m	$fact_{m1}$	$fact_{m2}$	…	$fact_{mn}$	$resp_{m1}$	$resp_{m2}$	…	$resp_{mk}$

В табл. 3 приняты следующие обозначения: $fact_{ij}$ – характеристики альтернативы (факторы или переменные решений); $resp_{ij}$ – оценки альтернатив, представленные откликами имитационных моделей с обязательным включением значения функции цели. В приведенной матрице альтернативы упорядочиваются в соответствии с полученными значениями функции цели на основе значений конфигурационного параметра sortDirFit.

Число элементов такой матрицы определяется спецификой соответствующей области приложений и поставленных задач исследования или проектирования.

Поскольку в системе NEDISOPT_D реализованы оптимизационные стратегии на базе методов последовательного перебора вариантов и генетического алгоритма и дополнительно реализована схема репликационных прогонов для получения оценок статистической достоверности результатов поиска, то разработаны шаблоны для трех типов причинно-следственных матриц. Каждой причинно-следственной матрице предшествует информация, представляющая управляющие параметры соответствующей стратегии оптимизации (по аналогии с табл. 2).

Таким образом, фрагменты выходной информации для оптимизационных стратегий представляются двумя шаблонами: управляющими параметрами и причинно-следственной матрицей.

Поскольку отклики имитационной модели часто наблюдаются как объекты типа "гистограмма", "очередь", "устройство", то требуемые для представления таких объектов шаблоны определяются семантикой объектов.

При этом фрагменты выходной информации структурированной с учетом специфики таких объектов содержат заголовочную информацию (с ключевыми словами "гистограмма", "устройство" "очередь") и имя соответствующего объекта. Экспорт таких информационных шаблонов в выходной поток сессии регламентируется значением конфигурационного параметра signPrintRes.

Заключение. В результате исследований на основе таких концепций информационной технологии Data Mining как "шаблон" и "хранилище данных" разработаны стандартизованные форматы для представления входных и выходных данных системы NEDISOPT_D.

Архивация в хранилищах опыта моделирования множества трасс поиска оптимальных решений в стандартизованных форматах обеспечит эффективные процедуры поиска соответствующей информации.

К перспективным направлениям исследований следует отнести разработку методов и средств интеллектуального анализа данных, форматированных и структурированных согласно разработанным шаблонам.

В.Б. Бігдан

МОДЕЛІ КЕРУВАННЯ ПОТОКАМИ ДАНИХ У ПРОЦЕСАХ РОЗПОДІЛЕНОГО ПОШУКУ ОПТИМАЛЬНИХ РІШЕНЬ

Розглядаються особливості реалізації деяких основних аспектів технології Data Mining у рамках розробленої в Інституті кібернетики імені В.М. Глушкова НАН України системи оптимізаційно-імітаційного моделювання NEDISOPT_D. Основна увага приділяється моделям керування потоками даних у процесах розподіленого пошуку оптимальних рішень із наступним накопиченням результатів пошуку у сховищах досвіду моделювання.

V.B. Bigdan

MODELS OF DATA FLOWS CONTROL IN THE PROCESSES OF THE DISTRIBUTED SEARCH OF OPTIMAL DECISIONS

The features of realisation are considered in regard to some basic aspects of Data Mining technology in the frame of optimisation-simulation system NEDISOPT_D developed at V.M.Glushkov Institute of cybernetics of NAS of Ukraine. The emphasis is made upon the models of data flows control in the processes of the distributed search of optimal decisions with the subsequent search results accumulation in warehouses of experience of modelling.

1. *Davis Dan M., Baer Garth D., Gottschalk Thomas D.* 1st Century Simulation: Exploiting High Performance Computing and Data Analysis // Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC). – 2004. – N 1517. – P. 1 – 14.
2. *Brady Thomas F. Yellig Edward.* Simulation data mining: a new form of computer simulation output // Proc. of the 2005 Winter Simulation Conf. – 2005. – N 05 – 030. – P. 285 – 289.
3. *Horne Gary E., Meyer Theodore E.* Data Farming: Discovering Surprise // Proc. of the 2005 Winter Simulation Conf. – 2005. – P. 1082 – 1087.
4. *Horne Gary E., Schwierz Klaus-Peter* AFarming Around The World Overview // Proc. of the 2008 Winter Simulation Conf. – 2008. – P. 1442 – 1447.
5. *Chua C.L., Sim CPT W.C.* Automated red teaming: an objective-based data farming approach for red teaming // Proc. of the 2008 Winter Simulation Conf. – 2008. – P. 1456 – 1462.
6. *Галаган Т.Н., Пепеляев В.А., Сахнюк М.А., Черный Ю.М., Шваб Н.Д.* О моделях сценариев распределенного поиска оптимальных решений // Компьютерная математика. – 2007. – № 2. – С. 144 – 156.
7. *Дюк В.* Data Mining – интеллектуальный анализ данных. // http://www.iteam.ru/publications/it/section_92/article_1448/. – 2003. – С. 1 – 13.
8. *Галаган Т.Н., Пепеляев В.А., Сахнюк М.А.* Особенности реализации многослойного сценария распределенного поиска оптимальных решений. // Проблеми програмування. – 2008. – № 2 – 3. – С. 636 – 640.

Получено 20.03.2009