УДК 381.3

А.Е. СКУКИС

ПРОГРАММНАЯ СИСТЕМА ПОДДЕРЖКИ ПРОЦЕССОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА БОЛЬШИХ МАССИВОВ ДАННЫХ

Описаны структура и функциональные возможности программной системы поддержки процесса интеллектуального анализа больших массивов данных. Уделено внимание этапам подготовки данных, организации вычислительного процесса, визуализации результатов. Приведен пример решения с помощью системы задачи булевого квадратичного программирования без ограничений.

Введение. Для математического и программного обеспечения ряда задач, которые возникают при интеллектуальном анализе больших массивов данных, на базе новых эффективных алгоритмов дискретного программирования [1] разработан прототип программной системы (ППС) его поддержки. На примере важных четырех классов задач (задача A задача булевого квадратичного программирования без ограничений, задача B поиска максимальной клики, задача C о максимальном разрезе ориентированного графа, задача D раскраски графа) предложен единый общий подход к их решению, основанный на использовании метода глобального равновесного поиска [2]. Рассматриваемые в рамках интеллектуальном анализе больших массивов данных задачи дискретной оптимизации, как правило, имеют сложную природу, что требует применения эффективных методов их решения. С учетом этого обстоятельства и результатов проведенного анализа существующих программных средств предложены архитектура и функциональное наполнение [3] ППС для поддержки процесса решения названных задач, выполнена ее программная реализация.

Программное обеспечение для комплексного решения задач классов $A-D$ объединено единым информационным подходом. Этот подход означает наличие развитых средств поддержки данных и широкое использование возможностей объектно-ориентированного программирования [4]. При реализации функций ППС учтены специфика задач дискретной

оптимизации и использованы современные технологии программирования. ППС предоставляет пользователям удобный механизм описания моделей решаемых задач, имеет развитый инструмент подготовки данных и организации вычислительного процесса. Функции администрирования ресурсов системы позволяют пользователям проводить вычислительные эксперименты независимо друг от друга. ППС реализован средствами быстрой разработки программ Microsoft Visual C++ 2005 Express Edition с использованием языка программирования C++ [5].

Структура ППС. ППС состоит из файловой базы данных, совокупности прикладных оптимизационных модулей и интерфейса пользователя. Его структура показана на рис. 1.

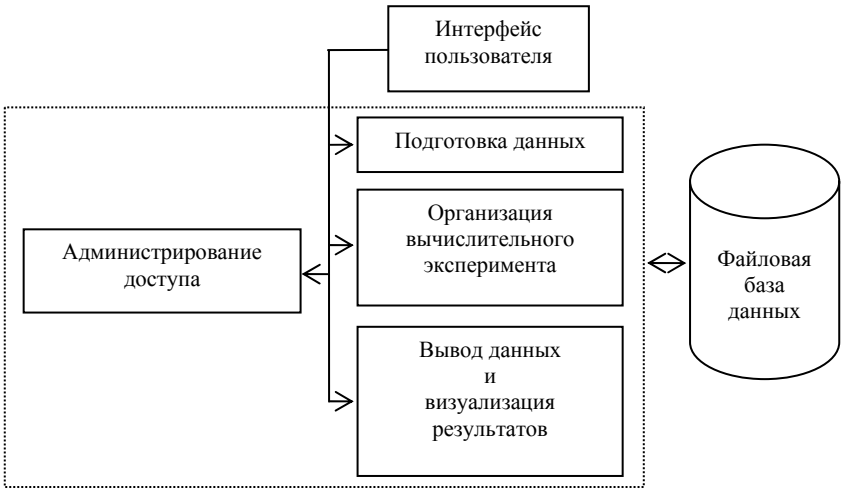

РИС. 1. Структура ППС

Файловая база данных представляет собой структурированный набор папок, которые содержат файлы входных, выходных и рабочих данных прикладных модулей, а также исполняемые файлы алгоритмов решения задач. На рис. 2 показана общая схема структуры файловой базы данных задач. Назначение папок будет описано далее. Интерфейс пользователя содержит средства, позволяющие осуществлять подготовку данных, формировать задания для решения задач, просматривать, редактировать и выводить результаты.

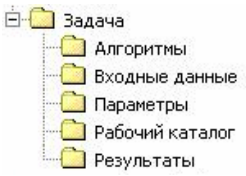

РИС. 2. Схема структуры файловой базы данных

Доступ к ППС авторизован, что позволяет распределять ресурсы системы между пользователями, список которых формируется и сопровождается администратором системы. Реквизитами пользователя являются индивидуальный идентификационный код, пароль и путь к части базы данных.

Функциональные возможности. Администрирование доступа и настройка параметров ППС предусматривают определение индивидуальных параметров проведения вычислительного эксперимента. Последний характеризуется датой проведения, вариантом входных данных, версией прикладных модулей, которые реализуют алгоритмы решаемых задач, местом сохранения выходных данных.

Организация вычислений предусматривает определение варианта входных данных задачи, алгоритма ее решения, способа проведения вычислений. Система предоставляет возможность решения задач как в автономном, так и в пакетном режимах. Процесс вычислений может быть остановлен для уточнения значений параметров задачи и алгоритма. По его завершению предусмотрено просмотр, редактирование и печать выходных данных.

Контекстная подсказка и детальная помощь дают возможность правильно применять ППС для анализа больших массивов данных на основе включенных в ее состав алгоритмов дискретной оптимизации.

Основной компонент интерфейса пользователя – главная экранная форма. Она содержит главное меню системы, функциональные кнопки быстрого доступа к пунктам главного меню, кнопки максимизации, минимизации, нормализации и закрытия формы. Общение пользователя с ППС осуществляется с помощью главного меню, которое содержит перечень функций, предназначенных для проведения вычислительного эксперимента.

Главное меню состоит из пунктов “Задачи”, “Помощь”, “Параметры” и “Выход”. Пункт меню “Задачи” содержит подпункты с названиями решаемых с помощью ППС задач, пункт “Помощь” – подпункты детального описания математических моделей задач классов $A - D$ и этапов их решения. Пункт меню “Параметры” включает настройки параметров системы и функции ее администрирования. В ППС также реализован доступ к отдельным пунктам меню с помощью кнопок быстрого доступа, которые располагаются на панели под строчкой главного меню.

Применение ППС. Рассмотрим принцип применения ППС для интеллектуального анализа больших массивов данных на примере решения задачи A булевого квадратичного программирования без ограничений. Математическая модель задачи может быть представлена в виде:

найти

$$\max\left\{ f(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij} x_i x_j \,\middle|\, x \in B^n \right\},$$

где q_{ij} – элементы симметричной действительной матрицы Q порядка n, B^n – множество n-мерных векторов с координатами 0 или 1.

Для решения задачи необходимо подготовить входные данные. Они могут быть сформированы программно – датчики случайных чисел, результаты вычислений с помощью других программ и систем; подготовлены с использованием произвольного редактора текстовых файлов. Данные размещаются в базе данных, структура которой показана на рис. 3.

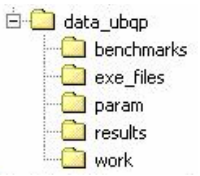

РИС. 3. Структура базы данных задачи *A*

Папка ”benchmarks” содержит файлы входных данных задачи. Формат файлов – текстовый. Их названия (название варианта входных данных) формируются произвольным образом. В первой строчке файла указывается размерность матрицы Q (см. описание математической модели) и количество ее ненулевых элементов. Следующие строчки – это позиции и значения ненулевых элементов.

В папке ”exe_files” размещены exe-файлы программ, которые реализуют алгоритмы решения задачи. Для задачи *A* в системе представлен один алгоритм глобального равновесного поиска (ГРП), которому соответствует exe-файл с именем ubqp_ges_pr1.

Папка ”param” содержит файлы с параметрами алгоритма. Для каждого алгоритма существует файл начальных значений параметров и файлы со значениями параметров, которые не совпадают с начальными. Название файла начальных значений параметров формируется следующим образом: param_<название варианта входных данных>_<название алгоритма>_initial. Названия других файлов формируются системой автоматически и трактуются следующим образом: param_<название варианта входных данных>_<название алгоритма>_N, где N – порядковый номер варианта параметров.

Алгоритм ubqp_ges_pr1 использует шесть параметров: номер решения задачи; максимальное количество попыток решения задачи; количество повторов основного цикла ГРП; время (в секундах), отведенное для решения задачи; значение целевой функции, при достижении которого необходимо завершить решение задачи; значение параметра tabu алгоритма табу, используемого алгоритмом ГРП. Значения этих параметров указываются в отдельной строчке файла параметров.

В папку ”result” записываются файлы решения задачи. Формат файлов текстовый. Имена файлов формируются системой автоматически и трактуются следующим образом:

<название варианта входных данных>_record_<название алгоритма>_ <номер варианта параметров>;

<название варианта входных данных>_record_solution_<название алгоритма>_ <номер варианта параметров>.

Файлы, в именах которых есть ключевое слово record, содержат номер решения задачи; время, за которое оно получено; найденное значение целевой функции; наилучшее известное значение целевой функции.

Файлы, в именах которых есть словосочетание record_solution, содержат решение задачи, найденное значение целевой функции.

Папка ”work” содержит рабочие файлы, которые формируются и используются алгоритмом решения задачи.

Для начала решения задачи необходимо выбрать пункт меню “Задачи” –> –>“Задача *A* булевого квадратичного программирования без ограничений”. При выборе этого пункта или при нажатии кнопки быстрого доступа к пункту меню на экране появится форма, которая состоит из закладок: “Математическая модель”, “Входные данные”, “Параметры”, “Выполнение”, отображенных далее на соответствующих рисунках.

Закладка “Математическая модель” содержит краткое описание математической модели задачи.

Закладка “Входные данные” (рис. 4) предназначена для определения даты проведения вычислительного эксперимента, каталога задачи; выбора (автономного или пакетного) режима ее решения, алгоритма решения задачи и варианта входных данных.

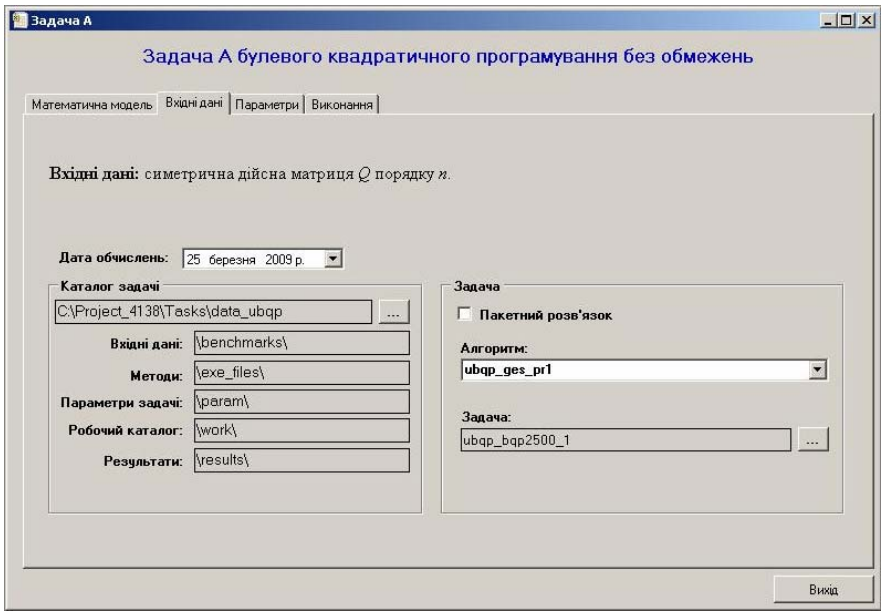

РИС. 4. Форма для задачи *A*. Закладка “Входные данные”

Группа полей “Каталог задачи” определяет путь к базе данных задачи. Корневой каталог выбирается в диалоговом режиме, а структура вложенных папок определяется системой автоматически.

Поле “Пакетное решение” предназначено для определения режима решения задачи. Если это поле выбрано, то вычисления будут проводиться для нескольких вариантов входных данных. В противном случае – для одного.

Поле со списком “Алгоритм” определяет алгоритм, которым будет решаться задача.

Поле “Задача” предназначено для определения входных данных задачи. Информация в нем может быть представлена в двух вариантах. Первый – поле “Пакетное решение” не выбрано. Тогда, при нажатии на кнопку [...], которая находится справа от поля “Задача”, в диалоговом режиме выбирается вариант входных данных. Название этого варианта переносится в поле “Задача”.

Второй вариант – поле “Пакетное решение” выбрано. В этом случае при нажатии на кнопку [...], которая находится справа от поля “Задача”, в диалоговом режиме выбирается несколько вариантов входных данных. Диалог проходит в два этапа. На первом этапе предлагается определить, будет ли проводиться добавление варианта входных данных в пакет или нет. Если да, то на втором этапе выбирается вариант входных данных. По завершению диалога в поле “Задача” будет помещен текст “Пакетное решение”.

Закладка “Параметры” (рис. 5) предназначена для определения набора входных параметров алгоритма, которым будет решаться задача. Значения параметров можно просмотреть, отредактировать и сохранить. Выбор варианта входных параметров производится в диалоговом режиме. Значения параметров отображаются в соответствующих полях групп “Параметры задачи” и “Параметры алгоритма”.

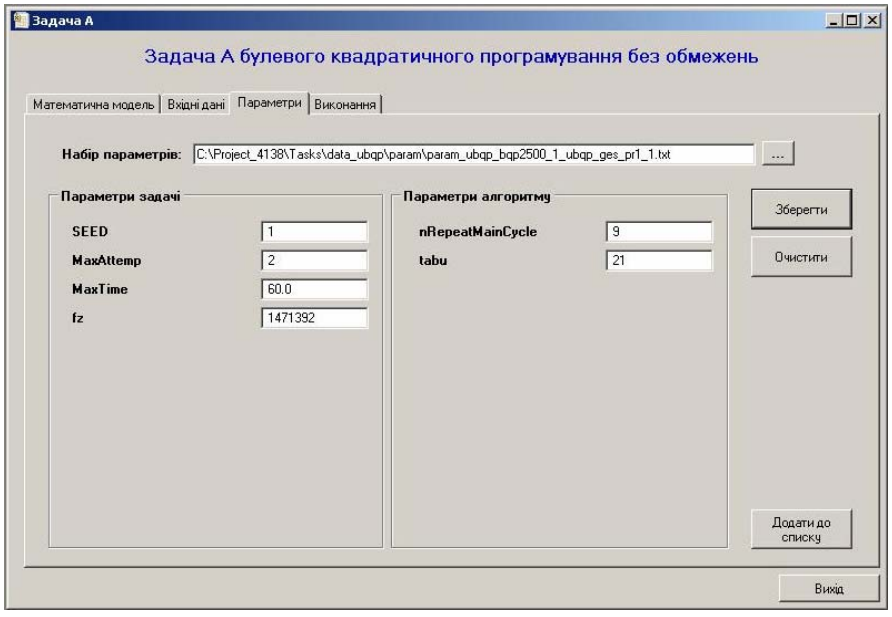

РИС. 5. Форма для задачи *A*. Закладка “Параметры”

Закладка “Выполнение” (рис. 6) предназначена для управления процессом вычислений и просмотра выходных данных задачи.

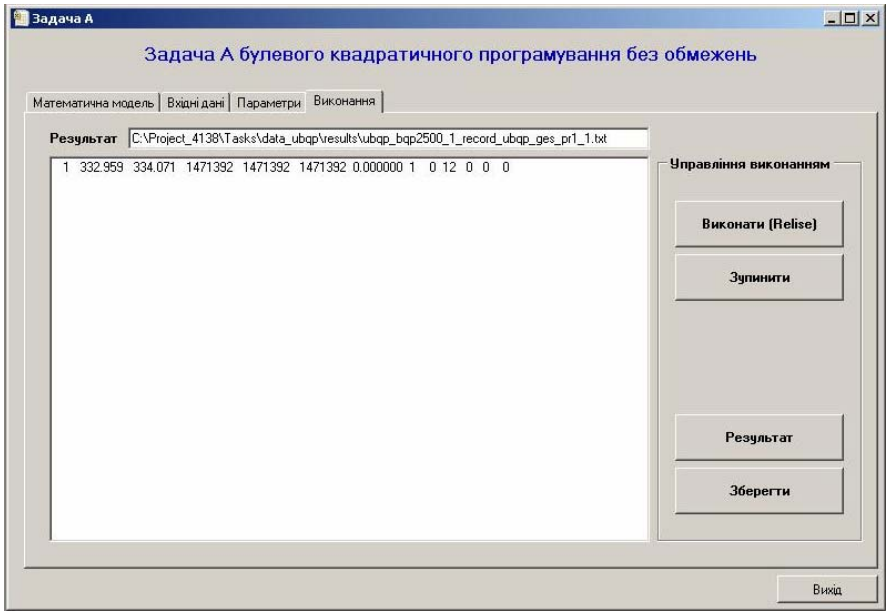

РИС. 6. Форма для задачи *A*. Закладка “Выполнение”

Для запуска выбранного алгоритма на выполнение предназначена кнопка “Выполнить (Relise)”. Процесс вычислений отображается в отдельном окне. Последний можно остановить. Для этого предназначена кнопка “Остановить”.

По окончанию процесса вычислений результаты решения задачи можно просмотреть, отредактировать и напечатать. Для этого предназначены соответственно кнопки “Результат”, “Редактировать / Сохранить”.

Получить детальную информацию о задаче, работе с ППС можно воспользовавшись пунктом главного меню “Помощь”.

Заключение. Разработан ППС, предназначенный для поддержки процесса решения задач, которые возникают при интеллектуальном анализе больших массивов данных, с помощью новых эффективных алгоритмов дискретного программирования. При его программной реализации была учтена специфика задач дискретного программирования, применена объектно-ориентированная технология программирования.

Созданный ППС предоставляет пользователям удобный механизм описания моделей решаемых задач, обладает развитым инструментом подготовки данных и организации вычислительного процесса. Средства визуализации результатов решения задач дают возможность интерпретировать данные в терминах предметной области. Функции администрирования ресурсов системы позволяют пользователям системы проводить вычислительные эксперименты независимо друг от друга.

О.Є. Скукіс

ПРОГРАМНА СИСТЕМА ПІДТРИМКИ ПРОЦЕСІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ВЕЛИКИХ МАСИВІВ ДАНИХ

Описані структура та функціональні можливості програмної системи підтримки процесу інтелектуального аналізу великих масивів даних. Приділено увагу етапам підготовки даних, організації обчислювального експерименту, візуалізації результатів. Наведено приклад розв'язання з допомогою системи задачі булевого квадратичного програмування без обмежень.

O.E. Skukis

THE PROGRAM SYSTEM FOR SUPPORT LARGE SCALE DATA MINING PROCESSES

The structure and functional capabilities of program system for support large scale Data Mining processes are described. The stages of data preparation, computational process organization and results visualization are considered. An example of unconstraint binary quadratic problem solving by the system is provided.

1. *Сергиенко И.В.*, *Шило В.П.* Задачи дискретной оптимизации: проблемы, методы решения, исследования. – Киев: Наук. думка, 2003. – 264 с.
2. *Шило В.П.* Метод глобального равновесного поиска // Кибернетика и системный анализ. – 1999. – № 1. – С. 74–81.
3. *Скукис А.Е.* Объектно-ориентированный подход к построению программных систем для решения задач дискретной оптимизации // Компьютерная математика. – 2007. – Вып. 2. – С. 80–85.
4. *Буч Г.* Объектно-ориентированный анализ и проектирование с примерами приложений на C++, 2-е изд. / Пер. с англ. – М.: Изд-во Бином, 1998. – 560 с.
5. Visual C++ 2005: базовый курс. : Пер. с англ. – М.: ООО "Изд. дом «Вильямс»", 2007. – 1152 с.

Получено 09.04.2009

Об авторах:

Скукис Алексей Евгеньевич,
аспирант, научный сотрудник
Института кибернетики имени В.М. Глушкова НАН Украины.
e-mail askukis@yahoo.com