ПРОГРАММНО-ТЕХНИЧЕСКИЕ КОМПЛЕКСЫ

С.И. ЛАВРЕНЮК, О.Л. ПЕРЕВОЗЧИКОВА

УДК 004

ОПРЕДЕЛЕНИЕ ОПТИМАЛЬНОГО МЕТОДА ПРОГНОЗА ЗАГРУЗКИ КЛАСТЕРНЫХ РЕСУРСОВ И ГРИД-УЗЛОВ

Ключевые слова: *кластер, грид, прогнозирование, числовой ряд, очередь.*

ВВЕДЕНИЕ

При работе вычислительного кластера на полную мощность, когда все узлы постоянно загружены (например, как в кластере СКИТ-3 Института кибернетики НАН Украины), актуальна задача прогнозирования доступности необходимых ресурсов или их объема в определенные временные периоды. На кластере СКИТ-3 [1, 2] работают пользователи из многих институтов НАН Украины. Данный кластер является также одним из мощных узлов Украинского Академического Грида (УАГ) [3]. В связи с этим в нем обслуживаются несколько очередей задач — локальная (для пользователей кластера) и грид-очередь (для пользователей УАГ).

Время выполнения задачи на кластере определим формулой

$$T_{all} = T_q + T_r, \tag{1}$$

где T_q — время нахождения задания в локальной очереди, T_r — время выполнения задания.

Время выполнения задачи в грид-сети определим формулой

$$T'_{all} = T_s + T_q + T_r + T_u, \tag{2}$$

где T_s — время поступления задачи и данных на грид-узел, T_q — время нахождения задания в локальной очереди грид-узла, T_r — время выполнения задания, T_u — время получения результата выполнения задания на указанный ресурс.

В общем случае T_q — время ожидания освобождения необходимых ресурсов для выполнения задания. Чтобы оценивать и минимизировать T_{all} и T'_{all}, нужно прежде всего оценить значения T_q. По результатам проведенного оценивания можно выбрать грид-узлы с оптимальным T_q и T_r для запуска задач и минимизации T_{all} и T'_{all}. Тогда задача сводится к нахождению и выбору оптимальных методов прогнозирования состояния ресурсов локальной очереди и грид-очереди, а также определению оптимального числа периодов приемлемого прогноза.

РЕШЕНИЕ ЗАДАЧИ

Для решения поставленной задачи использован пакет автоматического построения числовых прогнозов PREDICTOR. Этот пакет интерактивного прогнозирова-

ния встраивается как дополнение в MS Excel и имеет для пользователя четыре режима сложности обработки с разными интеллектуальными средствами [4].

1. Для начинающих пользователей режим **Мастер** позволяет быстро получить прогноз путем пошаговых операций установления сезонности, выбора метода (модели), визуального контроля качества работы модели и записи результата в виде ряда чисел. При этом оптимальные величины всех необходимых параметров PREDICTOR подбирает автоматически.

2. Для квалифицированных пользователей **диалоговая среда итеративного прогнозирования** иногда намного более длительный процесс, чем экспресс-прогнозирование. Однако преимущество этого режима состоит в том, что пользователь может «конструировать» прогноз путем коррекции параметров после их автоматического подбора пакетом, сравнения альтернативных вариантов прогноза одного и того же ряда (визуального и количественного, используя статистические показатели) и сохранения полученной таким образом эмпирической модели прогнозирования одних данных для использования ее с другими данными, описывающими подобные процессы.

3. В режиме **функций электронной таблицы** предполагается, что параметры прогнозной модели уже определены, а результаты прогнозирования должны стать исходными данными для последующих вычислений в электронной таблице. При этом необходимо, чтобы при изменении исходных данных, которые не являются результатом каких-либо вычислений, происходил автоматический пересчет всех зависимых от них звеньев, в частности результатов прогноза, — поэтому модели на этом уровне организованы в виде функций MS Excel. PREDICTOR включает 25 методов прогнозирования, среди которых: простые и линейные скользящие средние; сглаживание — простое, адаптивное, линейное по Холту, линейное по Брауну, квадратичное по Брауну, аддитивное сезонное по Винтерзу, сезонное по Холту–Винтерзу, сезонное по Брауну–Харриссону; регрессия — авторегрессия, S-кривые, кривая Гомпертца, логистическая кривая, популярные и определяемые пользователем тренды; методология Бокса–Дженкинса, ARARMA, ARIMA-модели с сезонностью в AR и MA, обобщенная адаптивная фильтрация (GAF); множественная регрессия; интерполяционные аппроксимирующие нейронные сети [5–7].

4. Быстрое получение прогноза дает **пакетное прогнозирование**, в процессе которого пользователю необходимо указать только источник данных, длину прогноза, общие параметры, характеризующие временной ряд, и выбрать методы для тестирования из предложенного списка. Более квалифицированный пользователь может настроить весовые коэффициенты статистических оценок, по которым ведется отбор оптимального метода. Оптимальные параметры, метод десезонирования пакет подбирает автоматически, тестируя при этом полученную модель, а в конце работы предлагает список из четырех оптимальных методов и прогнозы, построенные с их применением, т.е. дает возможность автоматически получить результат с минимальным числом шагов.

В пакете PREDICTOR поддерживается динамическое прогнозирование по мере поступления новых данных, управление сценариями и повторное их использование. PREDICTOR задуман как средство, объединяющее простоту и наглядность электронных таблиц MS Excel и мощные возможности методов численного прогнозирования.

Этот пакет автоматически строит тестовые прогнозы по ряду или нескольким рядам данных и выбирает четыре лучших метода, которые эффективнее аппроксимируют имеющиеся данные. Далее можно в ручном режиме менять параметры выбранных методов и сравнивать полученные результаты прогноза с имеющимися данными.

В качестве исходных данных о состоянии ресурсов вычислительных кластеров использованы данные базы мониторинга состояния грид-узлов УАГ и

грид-узлов коллаборации Nordugrid [8]. В базе мониторинга хранятся данные о загруженности кластеров локальными задачами (локальная очередь) и грид-задачами (грид-очередь). Взяты данные за февраль-март 2010 года и выбраны 12 узлов, на которые регулярно поступают задания: пять узлов УАГ, семь узлов Nordugrid (один из Швеции, два из Дании, четыре из Финляндии). Данные поступают в базу каждые 15 минут. Для удобства исследования длинного ряда проведена свертка данных и вычислена средняя загрузка узла за день [9].

Для каждого узла отдельно для локальной и грид-очереди автоматически выбраны четыре лучшие модели для прогнозирования.

На рис. 1 показаны предложенные пакетом PREDICTOR четыре лучшие модели для грид-узла nordug.bitp.kiev.ua по прогнозированию состояния локальной очереди задач.

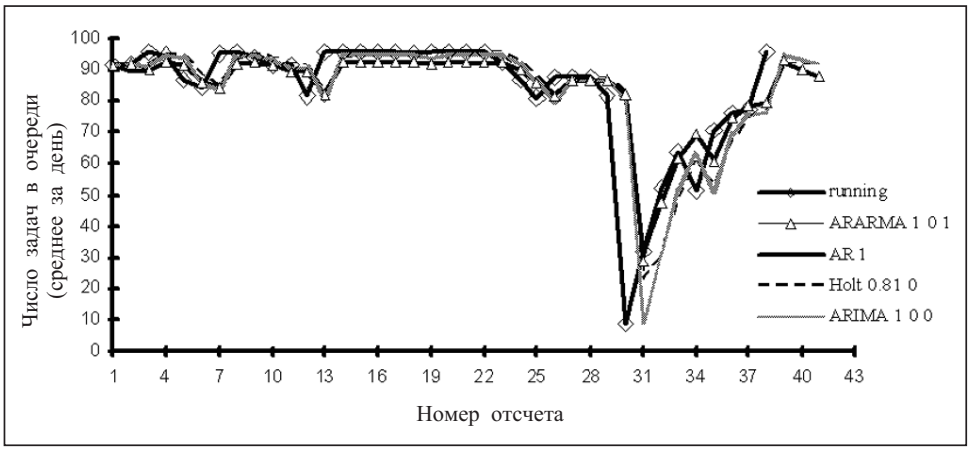

Рис. 1

На рис. 2 показаны предложенные пакетом PREDICTOR четыре лучшие модели для грид-узла nordu.hpcc.ntu-kpi.kiev.ua по прогнозированию состояния грид-очереди задач.

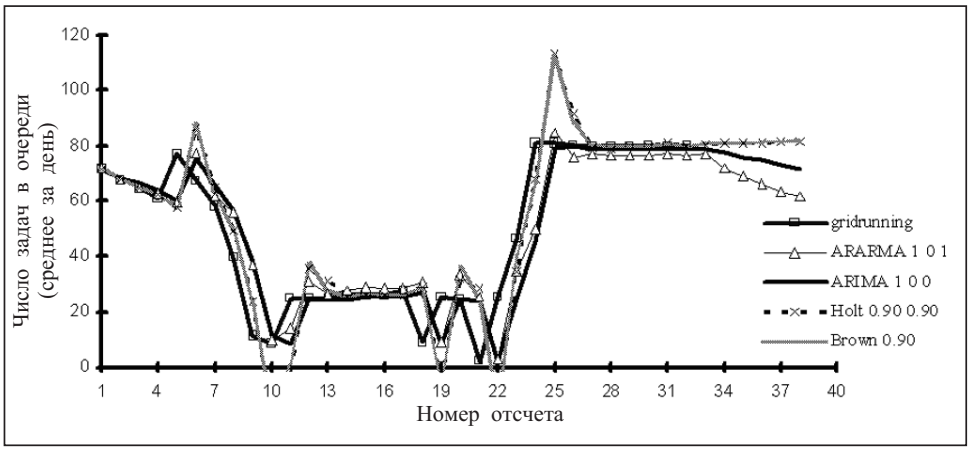

Рис. 2

На рис. 3 показаны предложенные пакетом PREDICTOR четыре лучшие модели для грид-узла svea.c3se.chalmers.se по прогнозированию состояния грид-очереди задач.

Результаты выбора моделей прогнозирования приведены в табл. 1.

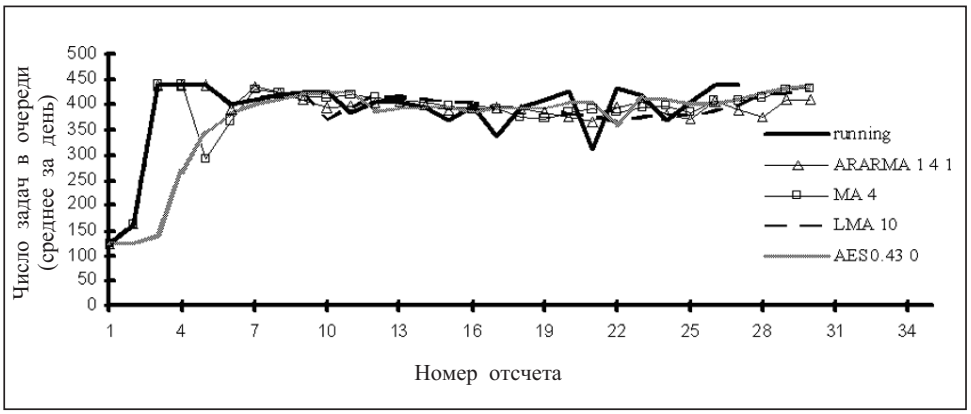

Рис. 3

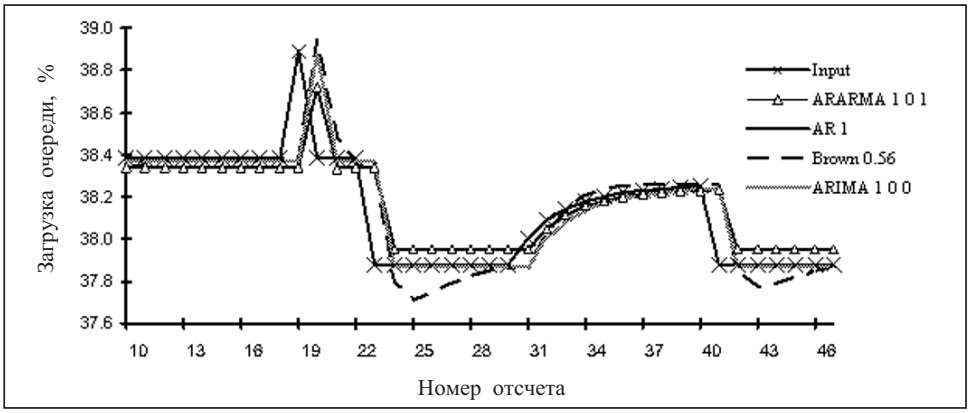

Рис. 4

Рис. 5

На основании анализа табл. 1 определены по ранжиру четыре метода, используемых чаще других при выборе четырех лучших моделей для прогнозирования: ARARMA, AR, ARIMA, LMA.

Качество прогнозирования состояния ресурсов кластера с помощью выбранных моделей проверено на данных о состоянии локальной очереди кластера СКИТ. На рис. 4 показан график выбора модели для прогнозирования загрузки локальной очереди.

На рис. 5 показан график выбора модели для прогнозирования количества задач на кластере СКИТ.

Таблица 1

Метод прогнозирования	Результаты выбора методов	
	локальная очередь	грид-очередь
ARARMA (Autoregressive moving-average model)	10	8
AR (Autoregression)	6	4
ARIMA (Box–Jenkins)	6	6
LMA (Linear Moving Averages)	5	3
AES (Adaptive Exponential Smoothing)	4	4
MA (Single Moving Averages)	4	3
Brown (Linear ES by Brown)	3	3
QBrown (Quadratic ES by Brown)	3	4
Sc (S-Curve Fitting)	3	6
Holt (Linear ES by Holt)	1	1
Sg (Gompertz' S-Curve Fitting)	1	1
Sl (Logistic S-Curve Fitting)	1	1

Программный пакет PREDICTOR автоматизировал многие процессы проведения моделирования и эксперимента по построению прогнозов данных. Это позволяет избежать разработки дополнительных программных модулей и потери времени на решение второстепенных задач моделирования.

По выбранным моделям построены и проанализированы прогнозы с горизонтом на десять точек. Эти модели дают приемлемый прогноз с небольшим расхождением длины горизонта до шести точек.

ЗАКЛЮЧЕНИЕ

Для прогнозирования локальной и грид-очереди кластеров с допустимым качеством построения прогноза эффективны авторегрессионные модели ARARMA и ARIMA, строящие приемлемые прогнозы состояния локальной и грид-очереди. Рационально строить прогнозы с горизонтом до шести точек. В рамках одного конкретного грид-узла, при регулярном поступлении задач, изменение очереди можно аппроксимировать теми же моделями, что и локальной очереди кластера, который является грид-узлом. Таким образом можно оценивать T_q и оптимизировать T_{all} и T'_{all}.

Задачу выбора оптимальных моделей прогноза необходимо усложнить с использованием произвольного множества связанных эконометрических рядов данных, например зависимости количества задач в локальной и грид-очереди от времени (суток, рабочих и выходных дней и т.д.). При этом проводится дополнительное аналитическое разделение указанных рядов на подряды. Целесообразно выделить такие индикаторы задач, как количество запрашиваемых процессоров, тип задачи, идентификация пользователя и др. Необходимо также проверять влияние сезонности данных (неделя, месяц, квартал, год) на качество прогноза.

Использование связанных рядов данных повысит точность прогнозирования для планирования загруженности ресурсов как всей грид-системы, так и отдельных вычислительных кластеров.

СПИСОК ЛИТЕРАТУРЫ

1. Коваль В., Сергієнко І. СКІТ — український суперкомп'ютерний проект // Вісн. НАН України. — 2005. — № 8. — С. 3–13.
2. СКИТ-3 (http://icybcluster.org.ua/).
3. Украинский Академический Грид (http://grid.bitp.kiev.ua/).
4. Интеллектуальные пакеты статистического прогнозирования / О.Л. Перевозчикова, И.Н. Пшонковская, Т.К. Терзян, В.Г. Тульчинский и др. // Упр. системы и машины. — 1997. — № 6. — С. 56–67.
5. Box G.E.P., Jenkins G.M. Time series analysis: Forecasting and control. — San Francisco: Holden-Day, Inc., 1976. — 423 p.
6. Cohn D.A., Ghahramani Z., Jordan M.I. Active learning with statistical models // Artif. Intel. Res. — 1996. — N 4. — P. 129–145.
7. Geman S., Bienenstoak E., Doursat R. Neural networks and the bias/variance dilemma // Neural Computation. — 1992. — N 4. — P. 1–58.
8. ARC Grid Monitor (http://www.nordugrid.org/monitor/).
9. Лавренюк А.Н., Лавренюк С.И., Грипич Ю.А Построение базы данных состояния grid-узлов на основе использования активных экспериментов // Розподілені комп'ютерні системи: Зб. праць ювіл. міжнар. наук.-практ. конф. РКС-2010 (6–8 квітня 2010). — Київ: НТУУ «КПІ», 2010. — С. 22–25.

Поступила 17.05.2010