УДК 004.89, 004.048

А.Н. Клименко, Н.Ю. Любченко, А.А. Подорожняк

Национальный технический университет "ХПИ", Харьков

ИСПОЛЬЗОВАНИЕ ИНТЕЛЛЕКТУАЛЬНЫХ МЕТОДОВ АНАЛИЗА ПРИ ОБОСНОВАНИИ РАЗРАБОТКИ СУПЕРКОМПЬЮТЕРА

В статье проанализирована возможность применения интеллектуальных методов обработки данных при обосновании разработки суперкомпьютера. Определены требования к структуре, компонентам и их характеристикам, определен круг возможных производителей. Приведен пример обработки данных о лучших суперкомпьютерах за последние десять лет, прогнозирования направлений развития на краткосрочный период и обоснования выбора составляющих, архитектуры и производителя суперкомпьютера.

Ключевые слова: *интеллектуальные методы анализа данных, суперкомпьютер, прогнозирование.*

Введение

Постановка проблемы. В последние годы очевиден взрывной характер усложнения всех сфер жизни общества, в особенности экономической, социальной и, более всего, техносферы. Из-за огромного количества вырабатываемой информации очень малая её часть будет когда-либо увидена и понята человеком. Необходимость интеллектуального анализа данных возникла в конце XX века в результате повсеместного распространения информационных технологий, позволяющих детально протоколировать процессы бизнеса и производства [1].

Термин интеллектуальный анализ данных (ИАД) обозначает не столько конкретную технологию, сколько сам процесс поиска корреляций, тенденций, взаимосвязей и закономерностей посредством различных математических и статистических алгоритмов: кластеризации, создания субвыборок, регрессионного и корреляционного анализа. Цель этого поиска – представить данные в виде, четко отражающем бизнес-процессы, а также построить модель, при помощи которой можно прогнозировать процессы, критичные для планирования бизнеса [2].

В тоже время Украина являясь де-факто высокотехнологической страной, на сегодня не имеет на своей территории современных суперкомпьютеров, позволяющих решать задачи фундаментальной науки, построения долговременных прогнозов погоды, а также различных приложений в аэрокосмической и авто индустрии, ядерной энергетике, при предсказании и разработке месторождений полезных ископаемых, в нефтедобывающей и газовой индустрии и т.д. Поэтому представляет интерес обоснование разработки суперкомпьютера для нужд промышленности, сельского хозяйства и фундаментальной науки с применением современных информационных технологий.

Анализ последних исследований и публикаций. К интеллектуальным средствам анализа данных относятся нейронные сети, деревья решений, индуктивные выводы, методы рассуждения по аналогии, нечеткие логические выводы, генетические алгоритмы, алгоритмы определения ассоциаций и последовательностей, анализ с избирательным действием, логическая регрессия, эволюционное программирование, визуализация данных [3]. Иногда перечисленные методы применяются в различных комбинациях [4]. Нейронные сети относятся к классу нелинейных адаптивных систем с архитектурой, условно имитирующей нервную ткань, состоящую из нейронов. К настоящему времени разработано много программных пакетов, реализующих нейронные сети: Nestor, Cascade Correlation, NeuDisk, Mimenice, Nu Web, Brain Dana, BrainMaker, Neural Professional, HNet, Explorenet 3000, Neuro Solutions, NeuroShell, NeuroWorks, Prapagator, Matlab Toolbox, PathFinder, Neural Analyzer, NeuroPro, НейроОфис.

Деревья решений — метод структурирования задачи в виде древовидного графа, вершины которого соответствуют продукционным правилам, позволяющим классифицировать данные или осуществлять анализ последствий решений. Методы деревьев решений реализованы во многих программных средствах, а именно: C5.0 (RuleQuest, Австралия), Clementine (Integral Solutions, Великобритания), SIPINA (University of Lyon, Франция), IDIS (Information Discovery, США). Индуктивные выводы позволяют получить обобщения фактов, хранящихся в базах данных (БД). Примером системы с применением индуктивных выводов является XpertRule Miner, разработанная фирмой Attar Software Ltd. (Великобритания). Рассуждения на основе аналогичных случаев (Case-based reasoning — CBR) основаны на поиске в БД ситуаций, описания которых сходны по ряду признаков с заданной ситуацией. Примерами систем, использующих CBR, являются: KATE Tools (Acknosoft, Франция), Pattern Recognition Workbench (Unica, США).

Нечеткая логика применяется для обработки данных с размытыми значениями истинности, которые могут быть представлены разнообразными лингвистическими переменными. Нечеткое представление знаний широко применяется в системах с логическими выводами (дедуктивными, индуктивными, абдуктивными) для решения задач классификации и прогнози-

рования, например в системе XpertRule Miner (Attar Software Ltd., Великобритания), а также в AIS и NeuFuz и др. Генетические алгоритмы входят в инструментарий ИАД как мощное средство решения комбинаторных и оптимизационных задач. Представителем пакетов из этой категории является GeneHunter фирмы Ward Systems Group. Генетические алгоритмы используются также в пакете XpertRule Miner и др.Эволюционное программирование — суть метода заключается в том, что гипотезы о форме зависимости целевой переменной от других переменных формулируются компьютерной системой в виде программ, процесс построения которых организован как эволюция в мире программ. Методы эволюционного программирования реализованы в системе PolyAnalyst (Unica, США). В современных интеллектуальных средствах анализа данных часто используются комбинированные методы.

Целью статьи является анализ возможности применения базовых существующих алгоритмов интеллектуальной обработки данных при обосновании разработки суперкомпьютера для нужд промышленности, сельского хозяйства и фундаментальной науки с использованием современных информационных технологий.

Основная часть

Проведенный анализ показал, что существует большое количество программных реализаций современных методов и алгоритмов интеллектуального анализа данных. Прежде чем остановиться на одном из пакетов, хотелось бы обобщить возможности интеллектуального анализа данных, особенности его применения и ограничения, налагаемые спецификой применения. ИАД (Data Mining) - это процесс поддержки принятия решений, основанный на поиске в данных скрытых закономерностей (шаблонов информации). При этом накопленные сведения автоматически обобщаются до информации, которая может быть охарактеризована как знания. В общем случае процесс ИАД состоит из трёх стадий: 1) выявление закономерностей (свободный поиск); 2) использование выявленных закономерностей для предсказания неизвестных значений (прогностическое моделирование); 3) анализ исключений, предназначенный для выявления и толкования аномалий в найденных закономерностях. Иногда в явном виде выделяют промежуточную стадию проверки достоверности найденных закономерностей между их нахождением и использованием (стадия валидации).

Все методы ИАД подразделяются на две большие группы по принципу работы с исходными обучающими данными.

В первом случае исходные данные могут храниться в явном детализированном виде и непосредственно использоваться для прогностического моделирования и/или анализа исключений; это так называемые методы рассуждений на основе анализа прецедентов. Главной проблемой этой группы методов является затрудненность их использования на больших объемах данных, хотя именно при анализе больших хранилищ данных методы ИАД приносят наибольшую пользу. Во втором случае информация вначале извлекается из первичных данных и преобразуется в некоторые формальные конструкции (их вид зависит от конкретного метода). Согласно предыдущей классификации, этот этап выполняется на стадии свободного поиска, которая у методов первой группы в принципе отсутствует. Таким образом, для прогностического моделирования и анализа исключений используются результаты этой стадии, которые гораздо более компактны, чем сами массивы исходных данных. При этом полученные конструкции могут быть либо "прозрачными" (интерпретируемыми), либо "черными ящиками" (нетрактуемыми). Прежде чем использовать технологию Data Mining, необходимо тщательно проанализировать ее проблемы, ограничения и критические вопросы, с ней связанные, а также понять, чего эта технология не может.

Data Mining не может заменить аналитика. Технология не может дать ответы на те вопросы, которые не были заданы. Она не может заменить аналитика, а всего лишь дает ему мощный инструмент для облегчения и улучшения его работы.

Сложность разработки и эксплуатации приложений Data Mining. Поскольку данная технология является мультидисциплинарной областью, для разработки приложения, включающего Data Mining, необходимо задействовать специалистов из разных областей, а также обеспечить их качественное взаимодействие.

Квалификация пользователя. Различные инструменты Data Mining имеют различную степень «дружелюбности» интерфейса и требуют определенной квалификации пользователя. Поэтому программное обеспечение должно соответствовать уровню подготовки пользователя. Использование Data Mining должно быть неразрывно связано с повышением квалификации пользователя.

Извлечение полезных сведений невозможно без хорошего понимания сути данных. Необходим тщательный выбор модели и интерпретация зависимостей или шаблонов, которые обнаружены. Поэтому работа с такими средствами требует тесного сотрудничества между экспертом в предметной области и специалистом по инструментам Data Mining. Построенные модели должны быть грамотно интегрированы в бизнес-процессы для возможности оценки и обновления моделей. В последнее время системы Data Mining поставляются как часть технологии хранилищ данных.

Сложность подготовки данных. Успешный анализ требует качественной предобработки данных. По утверждению аналитиков и пользователей баз данных, процесс предобработки может занять до 80% процентов всего Data Mining-процесса. Таким образом, чтобы технология работала на себя, потребуется много усилий и времени, которые уходят на предварительный

анализ данных, выбор модели и ее корректировку.

Большой процент ложных, недостоверных или бессмысленных результатов. С помощью Data Mining можно отыскивать действительно очень ценную информацию, которая вскоре даст большие дивиденды, в частности в виде финансовой и конкурентной выгоды. Однако Data Mining достаточно часто делает множество ложных и не имеющих смысла открытий. Многие специалисты утверждают, что Data Mining-средства могут выдавать огромное количество статистически недостоверных результатов. Чтобы этого избежать, необходима проверка адекватности полученных моделей на тестовых данных.

Наличие достаточного количества репрезентативных данных. Средства Data Mining, в отличие от статистических, теоретически не требуют наличия строго определенного количества ретроспективных данных. Эта особенность может стать причиной обнаружения недостоверных, ложных моделей и, как результат, принятия на их основе неверных решений. Необходимо осуществлять контроль статистической значимости обнаруженных знаний.

Исследования отмечают, что существуют как успешные решения, использующие Data Mining, так и неудачный опыт применения этой технологии. Области, где применения технологии Data Mining, скорее всего, будут успешными, имеют такие особенности: требуют решений, основанных на знаниях; имеют изменяющуюся окружающую среду; имеют доступные, достаточные и значимые данные; обеспечивают высокие дивиденды от правильных решений. В качестве исходной базы данных использовался проект по составлению рейтинга и описания 500 самых мощных компьютерных систем мира, который ведется начиная с июня 1993 года по два раза в год [5]. В этой базе данных хранится место, занимаемое данным компьютеромв рейтинге; Rmax — наивысший результат, полученный при использовании системы тестов Linpack, измеряется в TFLOPS; Rpeak — теоретическая пиковая производительность системы, измеряемая в TFLOPS; Число процессорных ядер, задействованных во время прохождении теста Linpack; название процессоров и вид соединения между процессорными узлами; производитель платформы или оборудования; название организации, использующей суперкомпьютер; страна нахождения суперкомпьютера; год, в котором суперкомпьютер введён в строй и другие данные. В качестве программного пакета использовалась аналитическая платформа Deductor 5.2 [6], особенностью которой является способность решать типовые задачи анализа бизнес-данных и практически полная функциональность бесплатной Academic версии предоставляемой компанией *BaseGroup Labs* для образовательных целей. Результаты, полученные при помощи векторного анализатора при определении параметров распределенных по частоте сигналов, показаны на рис. 1 – 4 .

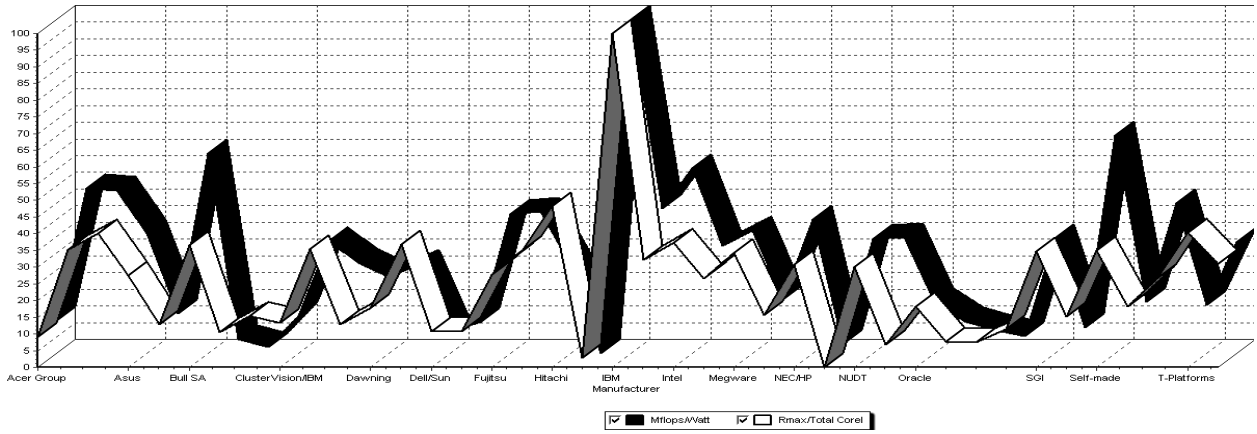

Рис. 1. График зависимости Mflops/Watt и Rmax/Total cores от производителей

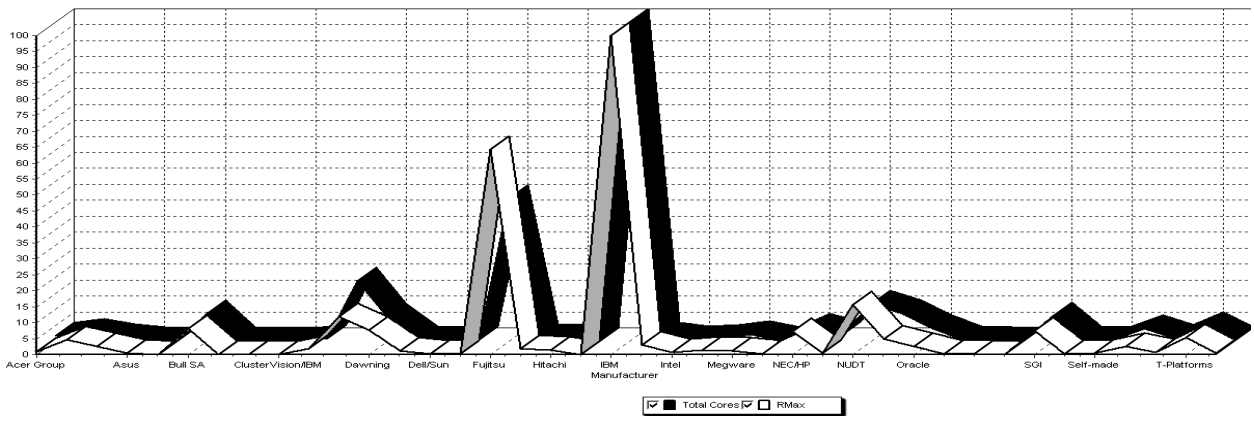

Рис. 2. График зависимости Total cores и Rmax от производителей

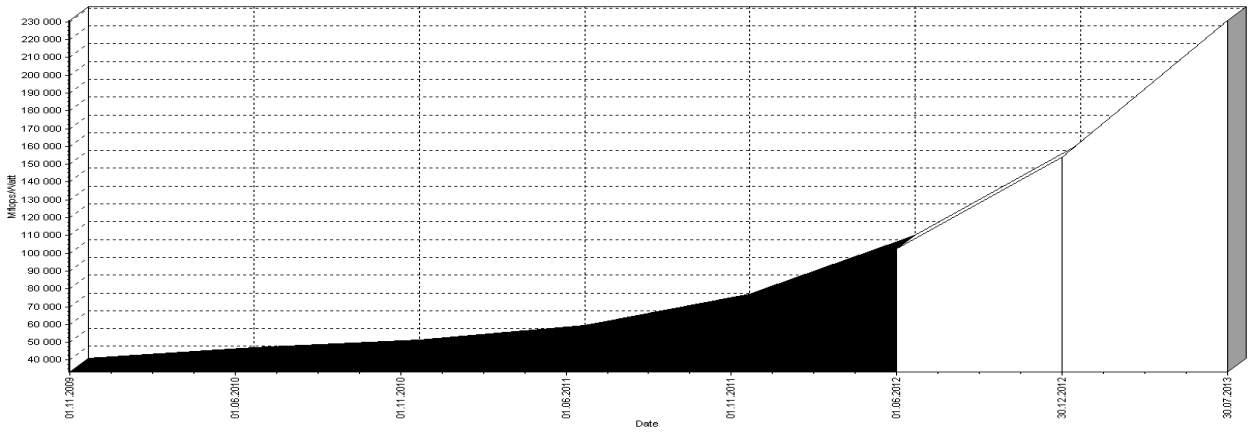

Рис. 3. График зависимости Mflops/Watt от времени, включая прогнозируемые значения

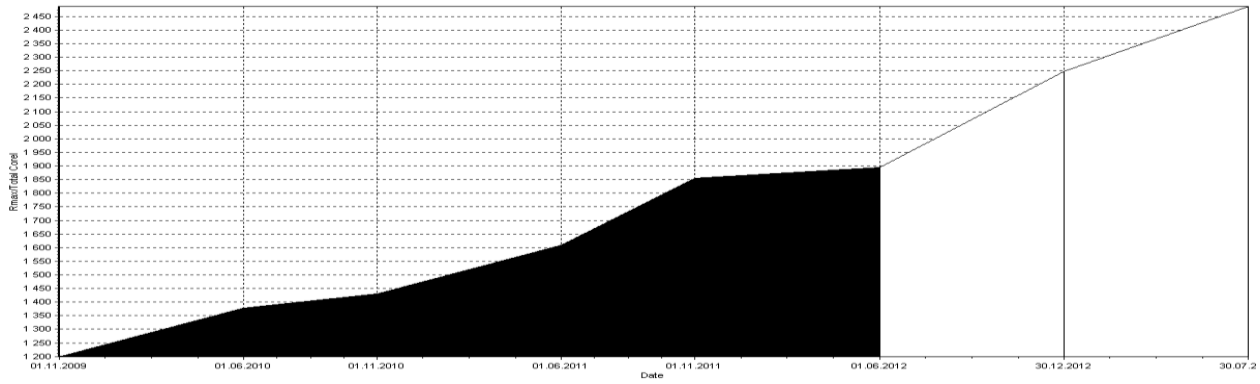

Рис. 4. График зависимости Rmax/Total cores от времени, включая прогнозируемые значения

Исследуемый сигналоследовательный анализ спектра дает хорошие результаты только при исследовании непрерывных стационарных сигналов. Сигналы систем связи с пакетной передачей, часовым разделением каналов псевдослучайной перестройкой несущей, частоты появляются в полосе обзора лишь на короткое время. В связи с этим неискаженное представление таких сигналов в частотной области можно получить только с помощью параллельного анализа. Более того, поскольку момент появления и длительность импульсного сигнала в общем случае неизвестны, его выявление и оценку параметров придется выполнять одновременно. Для этого векторный анализатор использует специальный режим с постоянной записью следующих друг за другом реализаций комплексной огибающей в буферную память. Выполнив параллельный анализ спектра и вычисление модулирующих процессов для каждой реализации можно обнаружить сигнал и получить представление об эволюции его спектральных и часовых характеристик во времени.

Выводы

В статье проанализирована возможность применения методов интеллектуального анализа данных при формировании реальных решений, а именно при обосновании разработки суперкомпьютера.

Определены требования к структуре, компонентам и их характеристикам, определен круг возможных производителей. Приведен пример обработки данных о современных суперкомпьютерах за последние десять лет, прогнозирования направлений развития на краткосрочный период и обоснования выбора составляющих, архитектуры и производителя суперкомпьютера. Темой дальнейших исследований является применения описанных методов для анализа разнородных классов данных, а также синтез алгоритмов для конкретных приложений.

Список литературы

1. Wang J. Data Mining: Opportunities and Challenges / Wang J. (Ed.). Hershey: Idea Group Publishing, 2003. – 468 p.

2. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям. – СПб.: Питер, 2009. – 624 с.

3. Шумейко А.А., Сотник С.Л. Интеллектуальный анализ данных – Днепропетровск: Белая, 2012. – 212 с. http://www-01.ibm.com/software/analytics/spss/.

4. Компания ZSoft. Библиотека Xelopes [Электронный ресурс]: http://www.zsoft.ru/page.php?14.

5. TOP500 – рейтинг и описание 500 самых мощных компьютерных систем мира [Электронный ресурс]. – Режим доступа к ресурсу: http://www.top500.org/.

6. Компания BaseGroup Labs. Аналитическая платформа Deductor [Электронный ресурс]. – Режим доступа к ресурсу: http://www. basegroup.ru/deductor/

Поступила в редколлегию 10.10.2012

Рецензент: д-р техн. наук, проф. В.Б. Кононов, Харьковский университет Воздушных Сил им. И. Кожедуба, Харьков.

ВИКОРИСТАННЯ ІНТЕЛЕКТУАЛЬНИХ МЕТОДІВ АНАЛІЗУ ПРИ ОБГРУНТУВАННІ РОЗРОБКИ СУПЕРКОМП'ЮТЕРА

А.М. Клименко, Н.Ю.Любченко, А.О. Подорожняк

У статті проаналізовано можливість застосування інтелектуальних методів обробки даних при обгрунтуванні розробки суперкомп'ютера. Визначено вимоги до структури, компонентів та їх характеристиками, визначено коло можливих виробників. Наведено приклад обробки даних про кращі суперкомп'ютери за останні 10 років, прогнозування напрямів розвитку на короткостроковий період і обгрунтування вибору складових, архітектури та виробника суперкомп'ютера.

Ключові слова: *інтелектуальні методи аналізу даних, суперкомп'ютер, прогнозування*.

USE DATA MINING METHODS IN JUSTIFYING DEVELOPMENT THE SUPERCOMPUTER

A.N. Klimenko, N.Y. Lubchenco, A.A. Podorozhnyak

In the article analyzed the possibility of application of data mining methods in justify development the supercomputer. The requirements for the structure, components, and their characteristic, the number of possible vendors. The example of the data about the best supercomputers over the past ten years, predicting areas of development in the short term, and justify the selection of the components, architecture, and supercomputer manufacturer.

Keywords: *data mining, the supercomputer, prediction.*