

УДК 681.5

Построение эмпирических моделей для управления сложными технологическими процессами

А. В. Деревянко

Харьковский национальный университет им. В.Н. Каразина, Украина

Гибридные аналитические и имитационные модели используются для решения сложных задач в различных областях, при этом их применение в системах управления технологическими процессами не так широко. В работе предлагается подход к созданию гибридной модели и его применение для управления наукоёмкими процессами с использованием в качестве хранилища экспериментальных данных многомерного хранилища.

Ключевые слова: аналитическая модель, имитационное моделирование, эвристическая модель, гибридная модель, управляющая система, многомерное хранилище.

Гібридні аналітичні та імітаційні моделі використовуються для вирішення складних задач в різних областях, при цьому їх використання в системах управління технологічними процесами не так поширене. В роботі пропонується підхід до створення гібридної моделі та його використання для управління складними процесами з використанням у якості сховища даних багатомірної сховища.

Ключові слова: аналітична модель, імітаційне моделювання, евристична модель, гібридна модель, управляюча система, багатомірне сховище.

Hybrid analytic and simulation models are widely used in solving complex problems in a variety of domains, but are less commonly used in technological processes measurement and control systems. This paper reviews hybrid approaches and their applications for science intensive processes control. The MOLAP data warehouse is offered to use as a source of experimental data.

Keywords: analytical model, simulation, heuristic model, hybrid model, control system, multi-dimensional storage.

1. Общая постановка задачи и её актуальность

Развитие возможностей современных вычислительных систем даёт возможность строить развитые компьютеризированные системы для управления технологическими процессами любой сложности. При этом расширение функциональности таких систем за счёт введения в их состав модулей хранения исторических данных, модулей первичной и аналитической обработки превращает систему и информационно-аналитическую (ИАС).

Применимо практически ко всем наукоёмким процессам вне зависимости от их сложности структура ИАС остаётся неизменной. Для её выделения был рассмотрен ряд ИАС, используемых в различных наукоёмких областях: управление нефтегазовыми разработками, электроэнергетика, нанесение ультратонких покрытий в вакууме. На нижнем уровне каждой из перечисленных ИАС находятся датчики параметров технологических процессов и элементы управления, способные воздействовать на протекание процесса. Данные устройства могут быть основаны как на однокристалльных ЭВМ, так и на промышленных контроллерах и ПК. Все полученные данные передаются в центральный управляющий узел, используя различные физические и логические протоколы, где они обрабатываются. В любом случае схематически структура системы может быть представлена как показано на рисунке 1.

При этом выделяют управляющие системы трёх типов:

- сбора данных – без обработки полученных результатов;
- операторные – с первичной обработкой данных и представлением их оператору в удобном виде;
- автоматизированные – с глубоким анализом данных. На основании заложенного алгоритма и накопленных экспериментальных данных возможно принятие решений; данный тип систем может быть замкнутым и работать с минимальным участием оператора или без него.



Рис.1. Структура современной ИАС.

Именно системы 3-го типа представляют наибольший интерес и сложность в разработке. При построении систем для управления плазменно-технологическими процессами для синтеза наноразмерных структур необходимо учитывать многие факторы, выделяющие их из множества остальных. В частности, стоит назвать малые времена, с которыми приходится работать для получения качественного результата. Обеспечение же повторяемости экспериментов возможно только с помощью автоматизированных систем. В связи со спецификой нанотехнологических процессов, их модель описывает диффузный объект (т.е. рабочая модель должна быть статистической).

Несмотря на новизну данного направления, для многих процессов уже отработаны действенные методики, что позволяет проводить эксперимент по заданному алгоритму. Стоит заметить, что зачастую алгоритм имеет сугубо экспериментальную природу и редко выводится из теоретических рассуждений. Например, технология атомного послойного осаждения, описанная в [1]. В настоящее время существует множество компаний, производящих оборудование для создания однородных ультратонких покрытий высокого качества методом АПО. При этом оборудование работает по жёсткому алгоритму.

Стоит заметить, что даже небольшое изменение требований к результату процесса или параметров синтеза структур требует разработки алгоритма с самого начала. В то же время теоретические исследования приводят к результатам, мало пригодным для практических применений.

2. Нерешенные проблемы и цели работы

Таким образом, возникает необходимость в создании управляющих систем, учитывающих как уже известные фундаментальные зависимости, так и результаты проведённых экспериментов. Основой таких систем могут быть феноменологические модели, т.е. предложенные как возможное описание

изучаемого процесса. Примером такой модели может быть феноменологическая модель создания пьезодатчиков на основе многослойных покрытий в вакууме [2]. В этом случае основа модели была получена основываясь на фундаментальные принципы, а затем в ходе экспериментов она была существенно уточнена. Однако данная модель была построена вручную, т.е. анализ полученных результатов и расширение модели проводились оператором. Целью же данной работы было построение системы, способной самостоятельно анализировать исторические данные, обнаруживать зависимости между ними, давать рекомендации по проведению дальнейших экспериментов и находить оптимальные параметры.

3. Выбор класса модели.

Как было сказано ранее, наибольшую эффективность имеют гибридные модели, т.е. объединяющие в себе как практические данные в виде отражения поведения системы, так и аналитические модели, существующие в виде набора уравнений, характеризующих систему. Аналитические данные предоставляют точную статическую информацию, в то время, как данные, полученные практическим путём – приближённую и динамическую (изменяющуюся при поступлении новых данных, опровергающих или уточняющих предыдущие).

Использование сугубо наработанных данных для получения имитационных моделей может потребовать значительных усилий, в то же время подобное моделирование значительно дешевле использования аналитических моделей. С развитием компьютерных технологий и моделирующего программного обеспечения, использование аналитических моделей сократилось. Использование эмпирические модели для моделирования сложных систем стало возможным. Уже многие годы совместное использование эмпирических и аналитических моделей применяется в различных направлениях: разработка структур процессов [3], планирование экспериментов [4], поиск оптимального протекания производственных процессов [5], полупроводниковая промышленность [6]. Число гибридных систем моделирования используемых в информационных технологиях также растёт. Большое количество систем разработано для оценки сетевой производительности [7], разработок многопроцессорных станций [8], оценки производительности систем с распределённой памятью [9] и т.д.

Все используемые модели можно разделить на несколько классов, используя классификацию Шантикумара. В [10] выделяются следующие классы гибридных моделей.

- модели, чья работа определяется либо использованием эмпирической модели, либо аналитической, но не одновременно;
- модели, в которых эмпирическое имитационное и аналитическое моделирование выполняются параллельно, используя результаты работы друг друга;
- модель, в которой имитационная модель используется в качестве подчинённой для получения некоторых значений;
- модель, в которой имитационная модель используется как основная, а аналитическое решение используется для получения входных

параметров (например, заранее известных зависимостей между факторами).

На Рис. 2 показано визуальное представление этих классов.

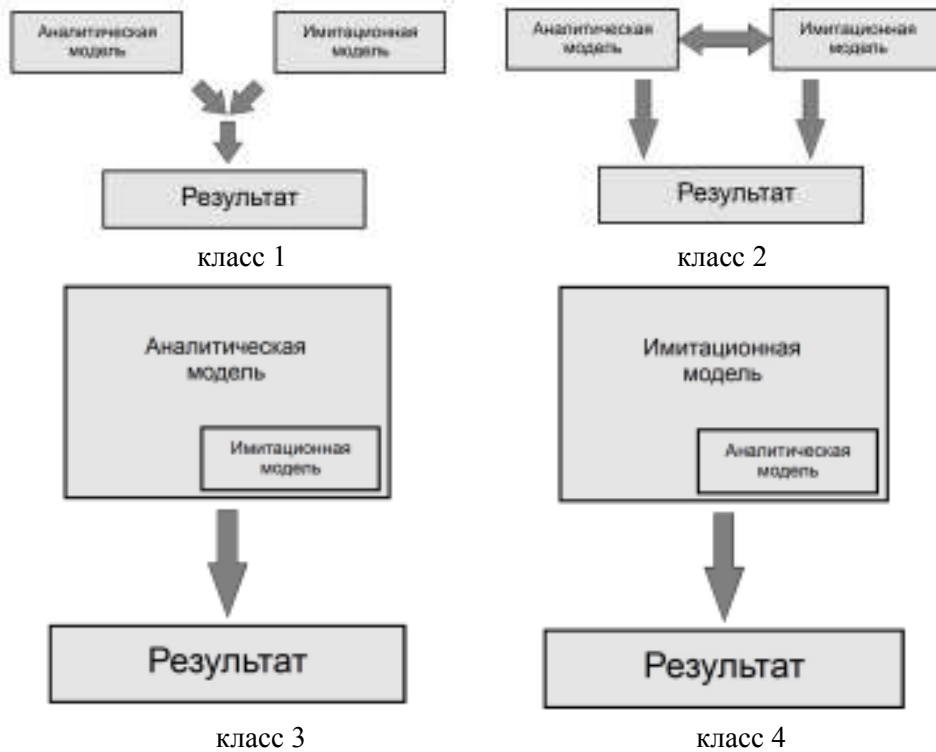


Рис. 2 Классы гибридных моделей

В работе предлагается использовать модель 3 класса. При этом имитационное моделирование играет значительную роль в исследовании поведения системы. Результаты используются для получения аналитических моделей для прогнозирования работы системы. При этом может быть, что аналитическую модель без упрощений создать нельзя, а приближённая модель не обладает достаточной точностью. Таким образом, модели данного класса применимы в случае:

- когда связь между целевым фактором и параметрами системы неизвестна или нет её чёткого понимания, что делает аналитические модели очень сложными для разработки; в этом случае, когда аналитические модели очень дороги, недостоверны или неосуществимы, имитационное моделирование может помочь в понимании связи между всеми факторами, что сделает возможным разработку аналитической модели;
- во многих прикладных задачах, в частности в исследуемой, на полезный сигнал часто накладывается множество шумов, которые практически невозможно учесть; имитационное моделирование позволяет исследовать поведение динамической системы и выделить

ключевые параметры для оценки с тем, чтобы затем эти параметры были включены в аналитическую модель. Модель подобного класса можно найти в [11].

4. Технология хранения экспериментальных данных.

Для сохранения данных и их последующей обработки было предложено использовать ставшие уже популярными хранилища данных. Хранилищами данных (ХД) называют предметно-ориентированные интегрированные базы данных, преимущественно используемые в системах поддержки принятия решений. В ХД часто накапливаются гетерогенные данные от многих распределённых источников, которые хранятся как исторические. Для выделения нужных данных и представления в виде реляционной таблицы используется механизм витрин данных. При этом зачастую, ХД представляет собой объединённые, созданные ранее витрины. При работе с ХД наибольшую трудность вызывают проблемы, связанные с её размерами: помимо предоставления системы хранения больших объёмов информации необходимо обеспечить механизм получения нужных данных с использованием простых запросов, требуемое быстродействие (для выполнения аналитической обработки в режиме реального времени OLAP) и простоту в использовании.

Данные, хранимые в ХД, поступают автоматически из указанных источников (не требуя интерактивного обновления) и находятся в режиме «только для чтения». Для этого в составе ХД присутствуют специализированные программы – экстракторы, способные обнаруживать появление новых данных в определённом источнике, извлекать и добавлять их в хранилище.

Как было описано в [12] для использования концепции ХД для управления технологическими процессами в роли одного из источников данных возможно использовать реализованную аппаратно систему сбора данных. Тогда при использовании метода ХД «получить данные от указанного источника» может быть использовано либо извлечение данных из указанного документа, либо произведения запроса и получение результатов измерений от какого-либо датчика. И в том, и в другом случае данные заносятся в ХД как исторические.

Для выбора иерархии ХД был произведён обзор наиболее активно используемых. Наибольшую популярность в последнее время приобрели технологии построения ХД вида «звезда» и её расширение – «снежинка» [13]. Для исследования таких систем используют пространственную модель, в которой база данных состоит из центральной таблицы фактов и набора окружающих её таблиц измерений, каждая из которых соответствует одному из полей таблицы фактов. В терминах реляционных БД таблица фактов содержит вторичные ключи, связанные с первичными ключами в каждой из таблиц измерений. В связи со схожестью такой структуры в её графическом представлении со звездой, схема называется звездообразной. Дальнейшее развитие пространственно модели [14] привело к появлению модели сущность-взаимосвязь, в которой помимо описанных выше принципов используется теория нормализации. Т.о. звездообразная схема может быть преобразована в схему типа «снежинка», в которой обеспечивается иерархичность каждого из полей таблиц измерений благодаря использованию подтаблиц измерений.

Для упрощення структур баз даних, способних забезпечити простоту таких запитів використовуються індекси: індекси об'єднання, битові, з'явисті [15]. Однак, ці методи застосовні тільки, якщо показники застосовують мале число варіантів значень, і абсолютно непридатні, якщо цих значень нескінченне множинство.

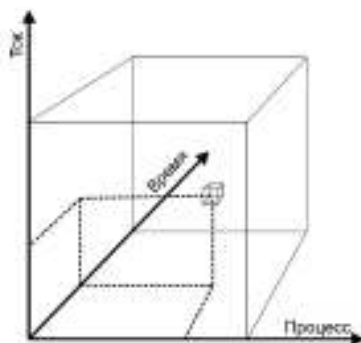


Рис. 3. Структура бази даних ХД нанесення покриттів

Поэтому в нашем случае была использована многомерная база данных (технология MOLAP), в которой в терминологии ХД существует факт – ячейка базы данных, содержащая одно, либо несколько значений – и измерения, которые в данном случае представлены осями куба, который, как правило, является гиперкубом. Преимуществами такого построения является, прежде всего, возможность архивирования накопленных данных, так как большая часть ячеек обычно не заполнена. Кроме того, поиск нужных элементов становится более быстрым, т.к., во-первых, есть возможность использования многомерной индексации массива гиперкуба, а, во-вторых, работа с ячейками по большей части становится арифметической операцией, а не работой с индексами.

Несмотря на большую популярность и широкое использование хранилищ данных, на сегодняшний день существует только один соответствующий программный продукт, выпускаемый как независимый – Hyperion фирмы Oracle, в то время как остальные системы управления ХД являются сопутствующими продуктами фирм-производителей СУБД. Поэтому для поставленной задачи хранилище было реализовано самостоятельно как многомерный массив с возможностью его частичной архивации и обладающий некоторыми методами:

- извлечения нужного элемента;
- извлечения нужной последовательности;
- извлечения нужной поверхности;
- принудительная экстракция данных из указанного источника (как правило, из системы сбора данных).

Описанное хранилище позволяет вести достаточно эффективную обработку наработанных данных и создавать эмпирические модели проводимых процессов.

При этом первым этапом является пассивный эксперимент, когда в ходе обычного проведения эксперимента информация о его протекании поступает в хранилище. Т.е. оператор (или управляющая система) выступает в роли

пассивного наблюдателя, задачей которого является правильное планирование эксперимента: оптимальная организация сбора информации и решение вопросов выбора количества и частоты измерений, выбор методов обработки полученных данных. На втором этапе (активного эксперимента) все факторы рассматриваются как независимые и управляемые.

5. Получение функциональных зависимостей.

Для быстрого извлечения ключевых данных из хранилища удобно использовать ряд методик, которые позволят проанализировать большие объёмы информации и выделить нужную его часть. Они включают в себя такие средства как кластеризацию и статистический анализ (включающий в себя выбор данных и многомерный анализ).

Кластеризация – это широко используемый метод добычи данных, в ходе которого вводится определённая численная метрика, и данные из хранилища обрабатываются и разбиваются на группы (кластеры), при этом между элементами одной группы наблюдается слабое различие, а разных групп – сильное. В качестве очень мощного средства добычи данных кластеризация хорошо изучена и применяется для многих целей [16]. Уже было разработано, усовершенствовано и внедрено множество алгоритмов и методик [17, 18, 19]. Наиболее популярными из них являются кластеризация и выборка.

Для создания модели протекающего процесса требуется найти функциональные зависимости между параметрами, образующими хранилище. Теме поиска таких зависимостей уделено большое внимание в [20, 21], множество соответствующих алгоритмов приведено также в [22]. Один из видов таких методов – многопараметрический, который включает в себя регрессионный анализ, анализ основных компонент, ковариационный анализ, канонический ковариационный анализ и т.д.

В работе требуется использовать метод для определения аппроксимирующей функции так, чтобы коэффициент корреляции данных и функции был максимален, в то время как ошибка – минимальна. Очевидно, что для этого необходимо в качестве кандидатов рассматривать большой набор функций. Основными преимуществами такого метода являются: отсутствие необходимости предполагать, какой вид имеет функция и небольшие вычислительные требования. Последнее преимущество особенно важно, так как в рассматриваемом случае размер хранилища может быть очень большим.

Для применения методик нахождения функциональной зависимости между факторами необходимо подготовить хранилище, чтобы уменьшить количество содержащихся в нём данных. Для этого необходимо выполнить следующие шаги:

- предварительная обработка данных;
- сокращение пространства поиска решений;
- кластеризация.

После этого выполняется поиск функции, для которой наилучшим образом выполняется функциональная зависимость. В качестве критерия на лучшую пригодность использовались корреляционные коэффициенты Пирсона, которые

вычислялись для каждой из 34 рассматриваемых функций. Перечень из 34 функций был взят, опираясь на исследования [23].

6. Результаты функционального тестирования системы.

Для тестирования системы было создано многомерное хранилище, имеющее 415 осей и содержащее 166 млн. элементов. Затем был проведён ряд запросов по нахождению зависимостей между парами осей. В качестве стоимости работы алгоритма было взято время его выполнения в мкс. На рисунке 4 представлена зависимость времени поиска функциональных зависимостей в первоначальной совокупности элементов, в образце и в сегментированном образце в зависимости от размера первоначального массива.

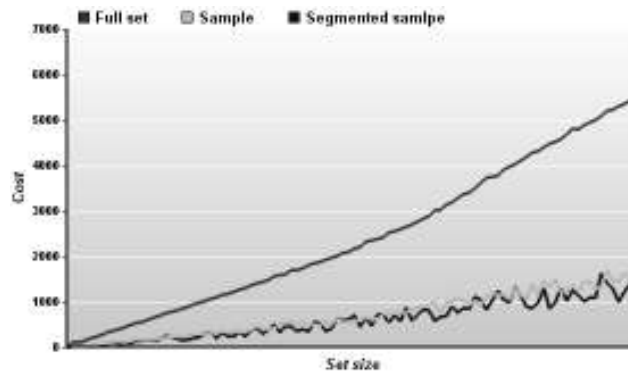


Рис. 4. Зависимость времени поиска функциональных зависимостей для полного набора, выборки и сегментированной выборки

На рисунке 5 показана зависимость времени сокращения числа элементов с сегментированием и без него в зависимости от размера первоначального массива. Как видно, при практически сходных дальнейших затратах на построение аналитической модели время на сегментирование требуется значительное. Поэтому в дальнейшем было принято решение отказаться от использования кластеров как от неэффективного средства.

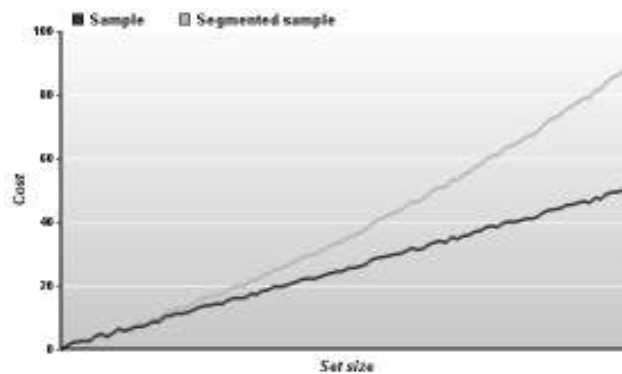


Рис.5. Зависимость стоимости получения выборки и сегментированной выборки из полного набора.

Таким образом, используя приведенную методику из первоначальных 166 млн. элементов (415 осей хранилища), участвовавших в построении модели, после сокращения их числа осталось лишь 80000 элементов (129 осей), при этом оставшиеся данные обладают всеми свойствами полного набора и позволяют строить эффективные модели на них. При этом видно, что время поиска зависимостей приемлемо для управления реальными процессами, что показывает эффективность использования хранилищ экспериментальных данных такого вида.

ЛИТЕРАТУРА

1. Т.В. Семикина, Э.Ф. Венгер, В.Н. Комащенко, Технология атомного послойного осаждения: история развития, принцип действия, основные области применения. Сборник научных работ «Фізико-хімічні основи формування і модифікації мікро- та наноструктур», Харьков, Украина, 2008, т. 1, сс. 76-79.
2. Д.В. Великодный, Т.М. Гричановская, Л.В. Ордодворец, Ч. Панчел, И.Е. Проценко, Тензорезистивные свойства многослойных плёнок на основе хрома и железа или никеля и ванадия. Сборник научных работ «Фізико-хімічні основи формування і модифікації мікро- та наноструктур», Харьков, Украина, 2008, т. 1, сс. 90-92.
3. E. Kozan, Comparison of analytical and simulation planning models of seaport container terminals, *Transportation Planning and Technology* 20 (3) (1997) 235-248.
4. J.F. Claver, B. Coste, Paste Plant Design and Control. A new approach, in: *Light Metals 1993 Proceedings of the 122nd Annual Meeting*, 1993, pp. 641-645.
5. M.V.F. Pereira, M.E.P. Maceira, G.C. Oliveira, L.M.V.G. Pinto, Combining analytical models and Monte Carlo techniques in probabilistic power system analysis, *IEEE Transaction and Power Systems* 7 (1) (1992) 265-272.
6. N.G. Pierce, Golden nuggets of AMHS modeling and design for semiconductor wafer fabrication in: *Proceedings of the IEEE/SEMI 1994 Advanced Semiconductor Manufacturing Conference and Workshop*, 1994, pp. 1-14.
7. K.M. Sage, E.A. Sykes, Evaluation of routing related performance for large scale packet switched networks with distributed, adaptive routing policies, *Information and Decision Technologies* 19 (6) (1994) 543-562.
8. P.W. Dowd, K. Bogineni, K.A. Aly, J.A. Perreault, Hierarchical scalable photonic architectures for high performance processor interconnection, *IEEE Transaction and Computers* 42 (9) (1993) 1105-1120.
9. R. Pozo, S.L. Smith, Limited resource scheduling in multifrontal algorithms for space linear systems in: *Proceedings of the Scalable High-Performance Computing Conference*, 1994, pp. 593-600.
10. J.G. Shanthikumar, R.G. Sargent, Unifying view of hybrid simulation/analytic models and modeling. *Operations Research* 31 (6) (1983) 1030-1052.
- A. Alan, B. Pritsker, Developing analytic models based on simulation results, in: *1989 Winter Simulation Conference Proceedings*, 1989, pp. 653-660.
11. А.В. Деревянко, А.Н. Стервиедов, М.Ю. Силкин. Стабилизация процесса ионно-лучевого осаждения наноразмерных плёнок нитридов и

- оксинитридов металлов. Физическая инженерия поверхности, т.6, № 1-2, 2008, сс. 114-120.
12. M. Levene, G. Loizou, Why is the snowflake schema a good data warehouse design? *Information Systems*, 28 (2003), pp. 225-240.
 13. R. Kimball, L. Reeves, M. Ross, W. Thornthwaite, *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing and Deploying Data Warehouses*, Wiley, Chichester, 1998.
 14. K. Rudin, C. Buss, W. Inmon, *Data Warehouse Performance*, John Wiley & Sons, 1998
 15. K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review. *ACM Computing Surveys* 31 (3) (1999), pp. 264-323.
 16. D. Cheng, R. Kannan, S. Vempala, G. Wang, A divide-and-merge methodology for clustering. *ACM Transactions on Database Systems* 31 (4) (2006), pp. 1499-1525.
 17. S.V. Jagadish, L.V. Lakshmanan, D. Srivastava, Snakes and sandwiches: optimal clustering strategies for a data warehouse, in: *ACM Proceedings: International Conference on Management of Data*, Philadelphia, USA, pp. 37-48.
 18. C.R. Palmer, C. Faloutsos, Density biased sampling: an improved method for data mining and clustering, in: *ACM Proceedings: International Conference on Management of Data*, Texas, USA, 2000, pp. 82-92.
 19. K. Vu, K.F. Hua, H. Cheng, S. Lang, A non-linear dimensionality-reduction technique for fast similarity search in large database, in: *ACM Proceedings: International Conference of Management of Data*, (2006) Chicago, USA, pp. 527-538.
 20. D. Zhang, Z. Zhou, S. Chen, Semi-supervised dimensionality reduction, in: *SIAM Proceedings: International Conference on Data Mining*, (2007).
 21. I.K. Fodor, *A Survey of Dimension Reduction Techniques*. US Department of Energy, Lawrence Livermore National Laboratory.
- A. Kuri-Morales, F. Rodriguez-Eraza, A search space reduction methodology for data mining in large databases, *Engineering Application of Artificial Intelligence* 22 (2009) pp. 57-65.

Надійшла у першій редакції 06.07.2009, в останній - 13.10.2009.

© Деревянко А. В., 2009