УДК 303.732.4:001.5:004.652.5

Модели данных для отдельных проблемно-ориентированных баз данных

Аверин Г.В., Звягинцева А.В.
Донецкий национальный технический университет
averin.gennadiy@gmail.com, anna_zv@ukr.net

Аверин Г.В., Звягинцева А.В. «Модели данных для отдельных проблемно-ориентированных баз данных». *Изучаются методы и средства феноменологического анализа данных для массивов эмпирической или статистической информации, отражающей в виде временных рядов процессы изменения и развития систем различной природы. Основная гипотеза исследования связана с возможностью создания на основе таблично-временных данных множества моделей в виде феноменологических описаний процессов и явлений, отличающихся многомерным полевым представлением массивов количественной информации, а также существованием скалярных полей эмпирических мер для комплексной оценки состояний систем. Данная гипотеза может быть принята или отвергнута на основе обработки имеющихся данных, наиболее адекватные модели могут быть выбраны из множества моделей с использованием полуавтоматических алгоритмов моделирования. Даны предложения по созданию программных продуктов для феноменологического анализа данных. Создание моделей данных позволит повысить эффективность деятельности ученых, аналитиков и экспертов при исследовании природных процессов и явлений, при анализе процессов мирового и регионального развития, при решении актуальных задач охраны окружающей среды, биоразнообразия и промышленной безопасности.*

Ключевые слова: сложные системы, анализ данных, моделирование, феноменологические модели, математическое и программное обеспечение.

Введение

Сегодня феноменологические методы получили широкое распространение в физике сплошных сред и термодинамике. Данное направление в науке и технике имеет большое значение, так как позволяет предложить методы описания объектов и систем многомерной размерности, к которым относятся все природные, биологические и социальные системы, а также сложные технические объекты. Решение данной проблемы возможно в случае наличия данных наблюдений или опыта, которые могут быть представлены в виде обширных структурированных массивов количественной информации. Однако, кроме этого, требуются также новые методы описания такой информации и вычислительные средства для ее обработки и анализа.

Последнее время в целом ряде областей знаний много внимания уделяется созданию универсальных методов моделирования. На повестке дня стоит вопрос создания новой методологии прикладного моделирования, которая позволяла бы использовать общую структурно-логическую схему анализа и построения моделей данных по отношению к различным классам систем и явлений, исходя из применения фундаментальных принципов, общесистемных гипотез и закономерностей в развитии природы и общества. Несмотря на множество исследований в данной области, существенного прогресса в решении данной проблемы пока не наблюдается.

Развитие методологии моделирования, охватывающей разные классы объектов и явлений, позволило бы разработать универсальные алгоритмы анализа данных и моделирования процессов и создать программные продукты, использование которых давало бы возможность исследователю создавать модели изучаемых систем.

Известно, что такие модели отличаются высоким уровнем формализации и универсальностью представления, они могут быть ориентированы на описание самых различных проблемно-ориентированных массивов количественной информации. На основе феноменологических моделей могут быть созданы средства прикладного описания систем, использующие большие массивы данных, а также разработано математическое обеспечение и вычислительные среды для анализа данных и моделирования различных классов систем. Все это в перспективе позволит привлечь множество ученых, экспертов и аналитиков в самых разных областях научной деятельности к анализу опытных и статистических данных и моделированию систем различной природы.

Объектом исследования в данной статье являются массивы данных в виде временных рядов количественных показателей, характеризующих процессы изменения и развития систем различной природы. В свою очередь, целью исследования является теория

 ISSN 2411-720X (Print) ISSN 2413-0834 (Online)

прикладного феноменологического анализа таких данных и методы моделирования состояний систем, для которых имеется обширная совокупность результатов наблюдений или опыта.

Некоторые направления в области создания моделей данных

В климатологии, глобалистике, оценке социально-экономического развития стран и регионов, охране окружающей природной среды и техногенной безопасности сегодня накоплены большие базы данных, позволяющие вести речь об установлении закономерностей процессов изменения и развития систем. Например, анализ данных социально-экономического развития стран и регионов мира не мыслим без информационных систем обработки данных. В этой области исследователь оперирует массивами данных, которые содержат сотни статистических показателей. Современная карта мира включает около 200 стран, многие из которых имеют административное деление на десятки регионов, республик, областей, округов, штатов, провинций, земель и т.д. В свою очередь, ретроспективная глубина данных может составлять десятки лет по каждому объекту с разбивкой на кварталы и даже месяцы. Аналогичным образом, в области изучения данных о климате планеты исследователь вынужден работать с файлами информации объемом от 5 до 20 терабайт, в которых хранятся данные о десятках величин в виде временных рядов с лагами в один час и ретроспективой в десятки лет.

Сегодня данные наблюдений или опыта обрабатывают в одном из известных программных продуктов для анализа информации: среда R, Statistica, SPSS, EpiInfo или в других программах для обработки специализированных данных. Подобный подход используется большинством исследователей, аналитиков и экспертов при анализе данных.

Ряд зарубежных IT-компаний разрабатывают методы описания и моделирования больших данных [1]. Например, в США в таких программах участвует ряд государственных организаций и частных компаний, тесно связанных с деятельностью Разведывательного сообщества. Такие разработки, как правило, относятся к продукции двойного назначения. В этой области наиболее известны продукты компаний Cloudera и Palantir Technologies, которые ведут работы по сбору и анализу данных о сетевой активности, изучению данных методами data mining, исследованию метаданных, выполнению аналитических работ в различных областях деятельности и осуществлению прогностических исследований. Технологическая деятельность компании Palantir (*http://www.palantir.com*) связана с интеграцией содержимого различных баз данных для социальных и финансовых систем, выделением связанных фрагментов и получением аналитических выводов. Заказчиками на выполнение таких работ, чаще всего, являются армейские, военно-морские, и финансовые службы, а также полиция.

Технологически подобные продукты объединяют методы интеграции и представления разнородных данных по общей форме в одной базе данных, используют поисковые механизмы и разные способы составления запросов, применяют различные аналитические алгоритмы (генетические алгоритмы, эвристические алгоритмы поиска, нейронные алгоритмы, статистические методы и т.д.). Это позволяет предложить эксперту простой в использовании инструмент для анализа данных. Окончательные выводы по решению задачи принимаются экспертом или группой экспертов. Программное обеспечение компании Palantir рассчитано на высококвалифицированных экспертов, имеет открытый интерфейс, расширяемую архитектуру и возможности адаптации к прикладной области.

Вторым важным примером является проект Recorded Future, который в 2010 году начал выполняться за счет средств Google и инвестиционного фонда Разведывательного сообщества In-Q-Tel. Данный проект предполагает использование поисковика третьего поколения, который может вести поиск объектов и отражать связи между объектами и их характеристиками, осуществляет структурирование всего информационного поля поиска по интересующим событиям, мнениям и реакциям людей на эти события и дает возможность исследовать явные и латентные связи, анализировать тренды и вести оценку отношений и связей на информационном поле. В настоящее время Recorded Future используется в разведке и госбезопасности, а также в сфере бизнеса и финансов.

Третьим примером аналитических систем нового поколения является платформа Quid, которая позволяет вести научно-техническое прогнозирование и поиск исполнителей для решения задач в сфере развития технологий, используя патентные данные ведущих стран мира и информацию открытых научно-технических баз данных. Данная система уже построена с учетом качественной модели развития технических и технологических решений, которая получила название «техноценоза». Клиентами данной системы являются ведущие корпорации, разведывательные и военные структуры США.

Сегодня можно уже говорить о новых возможностях создания моделей данных применительно к некоторым видам баз данных. Например, компания Palantir Technologies в своих продуктах применяет онтологические модели данных, под которыми понимается логический подход к формализации знаний в определенной предметной области в виде схемы, отражающей структуру данных, состоящую из объектов, поделенных на классы, связей между ними, а также из правил и ограничений, принятых в этой области. Компания Quid Inc. применяет в известном продукте анализа данных Quid модель изучаемых систем, которая получила название «техноценоза». Компания Google формирует модели данных путем структурирования всего информационного поля изучаемых данных.

Авторы данной статьи предлагают в предметных областях использовать феноменологические модели данных в виде сред моделирования, для которых на основе изучаемых данных в каждом конкретном случае устанавливаются модельные зависимости и параметры. Все это позволяет предлагать новые методы анализа данных и создавать модели данных в прикладных областях. Подобные подходы широко используются в физике сплошных сред и термодинамике.

Данную идею образно можно пояснить на примере. Предположим, что мы имеем большую базу данных термодинамических свойств различных веществ, но не имеет теории их описания, то есть количественные модели данных отсутствуют. Применение существующих методов обработки данных не позволило бы разработать теорию термодинамики, которая интегрирует знания, полученные поколениями ученых. Статистические методы обработки данных, методы интеллектуального анализа данных и т.п. являются итерационным средством для поиска различных закономерностей, а модели количественных данных уже требуют значительного интеллектуального труда в течение длительного времени, благодаря чему определяются основные закономерности. Однако, так как теория термодинамики разработана, имеются эффективные алгоритмы, которые построены с учетом существующих термодинамических зависимостей и которые позволяют с высокой степенью достоверности определять термодинамические свойства веществ. Данные алгоритмы реализуются сегодня в различных современных программных продуктах для определения свойств веществ.

Исходя из имеющихся количественных данных в прикладных областях, необходимо развивать практику использования в исследованиях биологических и социальных процессов естественнонаучных и информационных методов, основанных на феноменологических подходах анализа и описания опытных данных. Это будет способствовать повышению научного и технического уровня обработки опытной и статистической информации.

Примеры данных и источники информации

В сети Интернет имеются структурированные данные для систем различной природы. Такие данные можно получить там, где есть стандартизованные методики сбора и обработки информации и применяются различные модели для оценки их точности и достоверности, используются методы сглаживания, описания и ассимиляции данных и т.д. Например, так формируются архивы климатических данных Всемирного климатического центра. Данные проходят сложные процедуры ассимиляции исходных данных, в результате чего создаются архивы повторного анализа данных (AMIP/DOE Reanalysis aka NCEP/NCAR R2, ERA-Interim, MERRA, 20CR – 20 Century Reanalysis).

В науках об обществе можно использовать данные международных организаций, которые приведены в таблице 1. Информация должна быть структурирована по общей форме и интегрирована в массив данных. Среди региональных баз данных социально-экономических величин в качестве примера можно выделить базу данных Государственной автоматизированной системы ГАС «Управление» Российской Федерации, которая включает в себя более 20 компонентов и аспектов развития, в которые сведены свыше 6000 социально-экономических показателей развития для всех регионов в территориальном разрезе и в динамике с 1995 года.

При изучении биоразнообразия, могут быть использованы следующие базы данных:

- глобальная база данных по биоразнообразию – GBIF. – Электр. ресурс. URL: AnAge: http://data.gbif.org/welcome.htm/;
- The Animal Ageing and Longevity Database. – Электр. ресурс. URL: http://genomics.senescence.info/species/.

При изучении данных о характеристиках звезд могут быть использованы базы данных космических звездных каталогов, например, каталоги Hipparcos, Tycho, Tycho-2 и др. Существует также обширная база данных (https://www.quandl.com), где представлено более 12 миллионов временных рядов различных процессов и явлений.

Для указанных выше баз данных возможно создание феноменологических моделей данных.

ISSN 2411-720X (Print)
ISSN 2413-0834 (Online)

Таблица 1. – Некоторые статистические базы данных о развитии стран мира

Название	Краткая характеристика	Адрес доступа
Статистика ООН	Статистика глобального и национального уровня, собранная ООН по различным аспектам развития стран	http://data.un.org
Статистика Конференции ООН по торговле и развитию	Статистика в области международной торговли, инвестиций и развития экономики	http://unctad.org/en/Pages/Statistics.aspx
Статистика Международного валютного фонда	Статистические данные по всевозможным финансовым и экономическим показателям	http://www.imf.org/external/data.htm
База данных Программы развития ООН	Статистические данные Программы развития ООН	http://hdr.undp.org/en/data
Статистика ВТО	База данных статистики для стран мира по торговой политике, доступам на рынки, экспорту и импорту	https://www.wto.org/english/res_e/statis_e/statis_e.htm
Статистика Всемирного банка	Более 2000 показателей развития стран мира ретроспективой до 50 лет	http://data.worldbank.org/
Справочник ЦРУ по странам мира	Подробная статистика и фактическая информация по всем странам мира	https://www.cia.gov/library/publications/the-world-factbook/
Статистика Международного энергетического агентства (МЭА)	База данных стран мира по производству и потреблению основных источников энергии	http://www.iea.org/statistics
Статистика ООН по вопросам образования, науки и культуры (ЮНЕСКО)	Более 1000 индикаторов и данных по вопросам образования, грамотности, науки и технологий для более 200 стран	http://www.uis.unesco.org/Pages/default.aspx
Статистика статистической службы ЕС	Базы статистических данных по странам Евросоюза	http://ec.europa.eu/eurostat/help/new-eurostat-website
Базы данных Trading economics	Статистические данные по 196 странам для 300 000 показателей	http://Tradingeconomics.com

Предлагаемые методы анализа и моделирования данных

Предлагаемые методы ориентированы на данные, представленные в виде таблично-временных массивов информации. Для целого ряда систем самой разной природы возможно формирование таких массивов информации. Обычно такие данные имеют структуру таблиц в виде матриц «объекты – параметры», причем множество таблиц (t) упорядочено по времени, например, годам, месяцам, часам, секундам и т.д. В качестве объектов для систем определенной природы выступают однотипные объекты, например, вещества, организмы, биологические виды или особи, изделия, устройства, установки, здания, природные объекты одного класса, технические системы, близкие по технологии производства, профильные предприятия, города, районы, страны, граждане государств и т.д. В качестве параметров (показателей), отражающих свойства определенных видов систем, могут быть различные физические, химические, биологические, социально-экономические, природно-ресурсные, технологические или индификационные величины, имеющие количественное измерение.

Исходя из сказанного выше, для определенного объекта каждый параметр в таблично-временном массиве данных будет представлен временным рядом из опытных точек в количестве (t), которые задаются с определенным лагом. Структура таблично-временных данных показана на рисунке 1.

Таким образом, каждый объект в определенный момент наблюдения находится в некотором состоянии и характеризуется совокупностью параметров. Состояния объектов изменяются с течением времени. Подобный подход позволяет определить состояние объекта как совокупность его наблюдаемых свойств, параметры которых формируются под действием условий окружающей среды в конкретный момент времени.

На рисунке 2 показано место моделей данных в общем процессе построения модели объекта или процесса. Если процесс построения теории прикладной науки разделить условно на этапы: получение опытных данных, установление закономерностей, разработка теории, то модели данных являются составной частью феноменологии данной науки.

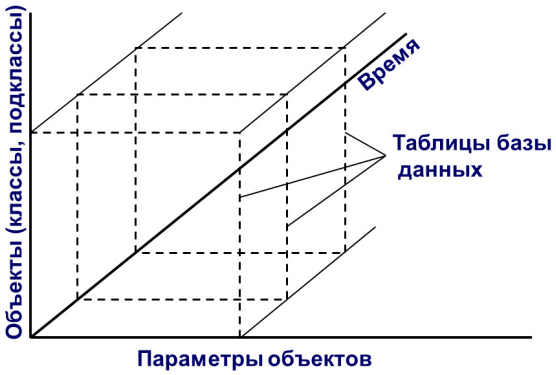

Рисунок 1. – Структура таблично-временных массивов данных

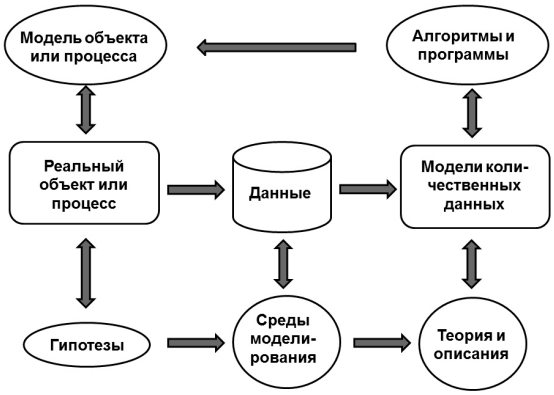

Рисунок 2. – Модели данных и модели объектов и процессов

Предположим, что для k однотипных объектов, формирующих систему определенной природы, в таблично-временных массивах данных содержится количественная информация о m параметрах, характеризующих множество самых различных свойств данной системы. Выберем из общего числа m всех параметров n атрибутивных показателей $(n < m)$, которые характеризуют базовые свойства изучаемых объектов, тогда имеющиеся массивы данных опыта или наблюдений могут быть отражены в n-мерном пространстве атрибутивных переменных p точками, где $p = k \cdot t$.

Перечень атрибутивных показателей определяется сложившимися в научном сообществе представлениями о поведении изучаемой системы, корреляционным анализом данных или другими методами установления наиболее значимых переменных. Количество объектов, параметров и таблиц не ограничивается и определяется только ресурсами вычислительной системы.

Любое множество n переменных для параметров свойств задает n-мерное пространство $\{z_1, z_2, ..., z_n\}$, являющееся декартовым произведением областей значений всех переменных z_k данного множества. Точки этого пространства соответствуют n-мерным наборам значений всех переменных $z_1, z_2, ..., z_n$. Таким образом, состояние любого объекта в n-мерном пространстве в каждый момент времени будет отображаться многомерной точкой $M = M(z_1, z_2, ..., z_n)$, процесс изменения состояния объекта во времени – многомерной кривой, которая описывается точкой M в этом пространстве.

Рассмотрим сложное совместное событие A_i одновременного наблюдения n параметров и определим, что состояние определенного объекта в заданный момент времени будет характеризоваться не только совокупностью параметров свойств для этого объекта, которые отображаются точкой M_i, но и данным наблюдаемым событием. Будем считать, что существует вероятность данного события. Назовем данную статистическую вероятность вероятностью состояния изучаемой системы. Статистические вероятности для сложного события A_i одновременного наблюдения n заданных параметров могут быть найдены с использованием различных алгоритмов перебора, группировки и подсчета частот благоприятных событий в общей выборке наблюдений. Существуют вычислительные алгоритмы, позволяющие для выборки из p наблюдений определить статистическую вероятность события A_i для каждой опытной точки M_i в наблюдаемом n-мерном пространстве переменных [2]. Основное условие для определения статистической вероятности связано с тем, что количество данных наблюдений должно быть достаточно большим, соизмеримым с числом $N = d \cdot f^n$, где f – число интервалов группирования данных для одной переменной, которое обычно принимается равным от 10 до 15, а d – число опытных данных на одном интервале группирования $(d = 5 - 7)$.

Существование статистических вероятностей сложных событий является основной вероятностной закономерностью реальной действительности и связано со свойством устойчивости относительных частот событий. Данное свойство справедливо для систем различной природы и является универсальной особенностью в поведении всех систем. Так как мы можем по обширным совокупностям результатов наблюдений алгоритмически определить статистическую вероятность состояния изучаемой системы, то в таблично-временных массивах данных можно искать вероятностные связи.

ISSN 2411-720X (Print)
ISSN 2413-0834 (Online)

Следует отметить, что статистические вероятности наиболее характерных событий, отражающих особенности в изменении и развитии конкретных систем, могут выступать как некоторые комплексные характеристики систем. Поэтому каждой *n*-мерной точке в пространстве $\{z_1, z_2, ..., z_n\}$ могут быть поставлены в соответствие вероятности различных характерных событий.

Можно пойти дальше и считать, как и в статье [3], что имеются самые разные комплексные характеристики, определяющие состояния систем наряду с параметрами свойств. Аналогичным образом каждой *n*-мерной точке могут быть поставлены в соответствие некоторые эмпирические величины, которые комплексно отражают состояния объектов и связаны с параметрами свойств. В общем случае назовем подобные величины эмпирической мерой состояний системы. Будем считать, что эмпирическая мера W может определяться в опыте на основе некоторых процедур измерений, оценок или расчетов. Величина W не может являться параметром одного из свойств системы $z_1, z_2, ..., z_n$. В качестве эмпирической меры могут выступать различные комплексные величины, например, вероятность различных характерных событий, количество теплоты, температура, стоимость объектов, различные индексы, определяемые экспертным путем, опытные величины, имеющие тесную связь со многими параметрами свойств системы, и т.д.

Таким образом, на основе переменных $z_1, z_2, ..., z_n$ можно сформировать *n*-мерное пространство координат $\{z_1, z_2, ..., z_n\}$, в котором возможные состояния системы образуют некоторую область Ω_n, охватывающую все наблюдаемые в опыте точки. Каждой точке M_i можно поставить в соответствие некоторую эмпирическую меру состояния W_i.

Первая фундаментальная гипотеза, которая принимается при построении количественных моделей данных, состоит в том, что мы предполагаем непрерывность области Ω_n. Это означает, что в пространстве состояний Ω_n существует бесконечное множество состояний для некоторой генеральной совокупности объектов системы определенной природы и точки состояний $M(z_1, z_2, ..., z_n)$ непрерывно заполняют это пространство. Будем также считать, что опытные точки $M(z_1, z_2, ..., z_n)$ являются ограниченной выборкой наблюдений из данной генеральной совокупности.

Вторая фундаментальная гипотеза предполагает существование некоторой эмпирической меры для комплексной оценки состояний изучаемой системы. Для построения моделей количественных данных принимаем гипотезу о непрерывности эмпирической меры в области Ω_n. Другими словами мы предполагаем существование скалярного поля эмпирической меры в многомерном пространстве Ω_n вида $W = W(M)$.

Все сказанное выше позволят сформулировать следующие аксиомы.

1. Пусть в пространстве состояний Ω_n некоторой системы каждой точке M поставлено в соответствие действительное число W, которое будем называть эмпирической мерой состояния системы.

2. Величина $W = W(M)$ является функцией точки и образует скалярное поле, которое является непрерывным в области Ω_n.

Для построения модели описания процессов системы можно использовать гипотезу, что скалярное поле эмпирической меры W может быть аналитически описано в окрестности произвольной точки M. Будем считать, что вблизи точки M осуществляется процесс изменения состояния системы. Для задания скалярного поля эмпирической меры $W = W(M)$ как функции независимых переменных $z_1, z_2, ..., z_n$ необходимо определить функцию точки. Предположим, что в области Ω_n можно задать аналитическую непрерывную функцию $\theta(z_1, z_2, ..., z_n)$, на основе которой будет формироваться математическая модель. При известном виде функции $\theta(z_1, z_2, ..., z_n)$ и значениях переменных $z_1, z_2, ..., z_n$ в области Ω_n можно построить еще одно скалярное поле, которое будем называть средой моделирования.

Исходя из этого, в общем случае для построения феноменологической модели изучаемой системы сформулируем аксиому.

3. Пусть в пространстве состояний Ω_n некоторой системы скалярные поля величин W и θ однозначно связаны между собой. Если в окрестности любой точки M объект системы осуществляет некоторый процесс l, то для линии процесса l справедливо соотношение $dW = c_l \cdot d\theta$, где c_l – эмпирические величины, которые являются функциями процесса.

В ряде работ авторами показано, что аксиом (1) – (3) достаточно для построения феноменологических описаний данных, представленных таблично-временными массивами информации. Данные описания связаны с уравнениями Пфаффа вида [2, 3]:

$$dW = c_1 \cdot \left(\frac{\partial \theta}{\partial z_1}\right) dz_1 + ... + c_n \cdot \left(\frac{\partial \theta}{\partial z_n}\right) dz_n , \qquad (1)$$

где феноменологические величины c_l определяются по имеющимся данным наблюдений. Для многих классов функций $\theta(z_1, z_2, ..., z_n)$ решения уравнений Пфаффа позволяют получить общие интегралы, которые по своему виду близки к функциям состояния и широко используются в термодинамике – это энтропия и термодинамические потенциалы. Энтропия является характеристической функцией пространства состояний системы. Также в пространстве Ω_n для целого ряда функций существует общий интеграл $U(z_1, z_2, ..., z_n) = C$, который может выступать потенциалом пространства состояний системы.

Энтропия s и потенциал U являются естественными криволинейными координатами пространства состояний пространстве Ω_n и могут быть приняты в качестве обобщенных характеристик для комплексной оценки состояния систем различной природы. Их наиболее важной особенностью является то, что данные величины являются функциями состояния системы при справедливости условия существования скалярного поля эмпирической меры W. Изменение данных функций зависит только от начального и конечного состояния системы и не зависит от пути перехода системы между этими состояниями.

В данном варианте построения теории важным является выбор среды моделирования θ. В общем случае среда моделирования в области Ω_n может быть представлена виде различных функциональных зависимостей относительно атрибутивных параметров: мультипликативными, степенными, аддитивными, экспертными или иными зависимостями, входящими в классы однородных или мультипликативных функций. Установлено [2], что при этих условиях среда моделирования θ в пространстве Ω_n для многих переменных позволяет при феноменологических описаниях использовать квазилинейные многомерные уравнения в частных производных первого порядка, которые тесно связаны с уравнениями Пфаффа.

Существует множество способов формирования различных сред моделирования и исследователь должен выбирать эти среды, исходя особенностей предметной области, принятых гипотез или предположений, сложившихся в научном сообществе представлений или интуитивных подходов.

Например, среда моделирования может быть представлена как многомерная геометрическая вероятность, как мультипликативная степенная функция относительно параметров $z_1, z_2, ..., z_n$, из геометрических представлений однородности пространства Ω_n, в виде экспертных зависимостей, т.е. соответственно в виде:

$$\theta = \beta \cdot \frac{z_1}{z_{10}} \cdot \frac{z_2}{z_{20}} \cdot ... \cdot \frac{z_n}{z_{n0}}; \qquad (2)$$

$$\theta = \beta \cdot \left(\frac{z_1}{z_{10}}\right)^{\alpha_1} \cdot \left(\frac{z_2}{z_{20}}\right)^{\alpha_2} \cdot ... \cdot \left(\frac{z_n}{z_{n0}}\right)^{\alpha_n}; \qquad (3)$$

$$\theta = (z_1 - z_{10})^2 + (z_1 - z_{10})^2 + ... + (z_n - z_{n0})^2 ; \qquad (4)$$

$$\theta = \beta_1 \cdot \frac{z_1}{z_{10}} + \beta_2 \cdot \frac{z_2}{z_{20}} + ... + \beta_n \cdot \frac{z_n}{z_{n0}}, \qquad (5)$$

где z_{k0} – некоторые опорные значения переменных; β_k – весовые или стандартизированные коэффициенты. Может быть также предложено множество других видов сред моделирования.

Функции (2) – (5) входят в класс однородных функций, поэтому для них существуют решения уравнения (1). Для каждой такой функции определяется свой вид зависимости для энтропии и потенциала. Например, для среды моделирования (2) энтропия и потенциал имеют следующий вид:

$$s - s_0 = c_1 \cdot \ln\left(\frac{z_1}{z_{10}}\right) + ... + c_n \cdot \ln\left(\frac{z_n}{z_{n0}}\right); \qquad (6)$$

$$U - U_0 = \frac{z_1^2 - z_{10}^2}{c_1} + ... + \frac{z_n^2 - z_{n0}^2}{c_n}. \qquad (7)$$

Естественно, что при разработке программных продуктов необходимо создание библиотеки для выбора сред моделирования.

Из сказанного выше видно, что предлагаемый метод феноменологического анализа данных для массивов количественной информации тесно связан с логикой построения теории термодинамики, так как изначально вводятся феноменологически определяемые величины c_l, характеризующие процессы изменения состояний объектов. Множество данных величин для каждой элементарной области пространства Ω_n определяется из соотношений $dW = c_l \cdot d\theta$ при условии, что задана система определения величин W и θ. Это позволяет с высокой точностью моделировать процессы изменения состояний систем различной природы на основе использования проблемно-ориентированных массивов информации и феноменологических описаний данных, особенно в случаях, когда имеется обширная совокупность результатов наблюдений или опыта.

ISSN 2411-720X (Print)
ISSN 2413-0834 (Online)

Особенность предложенного подхода заключается в том, что исходные гипотезы могут быть приняты или отвергнуты на основе обработки данных опыта или наблюдений, характеризующих поведение той или иной системы. Подобная проверка носит итерационный характер и связана с перебором различных сред моделирования, определением феноменологических констант и оценкой критериев, определяющих точность моделей.

Данный подход, принятый при построении моделей данных, широко используется в термодинамике и физике сплошных сред. Естественно, что он может быть реализован и в других предметных областях. Например, для социально-экономических данных Программы развития ООН и Всемирного банка авторами в качестве примера была разработана методика и показана возможность построения феноменологических моделей развития стран в пространстве многих переменных. Решение отдельных тестовых задач при изучении развития стран мира показало, что данный метод позволяет построить феноменологическую теорию развития стран мира [4].

Аналогичным образом, были построены феноменологические модели при анализе и описании токсикологических данных, полученных при негативных воздействиях вредных веществ на животных и человека [2]. Все это указывает на перспективность данного направления исследований, ориентированного на построение моделей данных.

При справедливости принятых гипотез, которые могут быть проверены на исходной информации, изучаемые массивы данных могут служить основой для создания базы знаний в виде соотношений, характеризующих системы определенной природы.

Предложения по созданию программных продуктов для анализа данных

Предложенные методы могут быть реализованы в программных продуктах, связанных с анализом данных. Работая с такой вычислительной средой эксперты и аналитики могут не только искать закономерности в данных, но и строить модельные описания данных, тем самым, создавая феноменологическую теорию для изучаемого класса систем. В процессе выполнения работ по построению феноменологических моделей данных необходимо провести большой объем вычислительных работ. Это связано с необходимостью перебора различных сред моделирования, алгоритмическим расчетом значений эмпирических мер, изучением различных связей в массивах информации и определением феноменологических констант, поиском оптимальных моделей для описания данных, выполнением определенного набора однотипных расчетов для каждой таблицы базы данных. Поэтому видна необходимость автоматизации таких расчетов. Подобный программный продукт должен интегрировать несколько приложений и сред для хранения, обработки, анализа и описания данных:

- хранилище массивов данных для систем различной природы;
- приложение, которое обеспечивает ввод и импорт данных, характеризующих системы различной природы, а также преобразование входных количественных данных в формат для работы со средой анализа данных *R* и в собственный формат данных для работы с моделирующей средой продукта;
- среду анализа данных R, которая позволяет осуществить визуализацию данных, их предварительный анализ, провести исследование данных с целью установления особенностей и закономерностей, использовать более 5000 различных функций визуализации, обработки и анализа данных и т.д.;
- приложение для выделения атрибутивных параметров и формирования перечней таких параметров на основе методов установления значимых переменных;
- библиотеку для выбора сред моделирования при решении прикладных задач;
- вычислительную среду для анализа данных, поиска феноменологических закономерностей и моделирования процессов изменения систем различной природы;
- интерфейс для взаимодействия сервисов и работы вычислительной среды.

Исходя из сказанного выше, можно разработать математическое обеспечение и специальный программный продукт для феноменологического анализа данных и моделирования различных классов систем.

Выводы

Таким образом, как видно из приведенного материала, для целого ряда предметных областей можно разработать феноменологические модели изучаемых систем. Предложенные методы позволяют разработать математическое обеспечение и вычислительные средства для описания количественных данных наблюдений или опыта. Это будет способствовать повышению научного и технического уровня обработки больших объемов количественной информации. Создание моделей данных позволит повысить эффективность деятельности ученых, аналитиков и экспертов при исследовании различных процессов и явлений.

Литература

1. Аналитика двойного назначения // Открытые системы, № 10, 2013.
2. Аверин Г.В. Системодинамика. – Донецк: Донбасс, 2014. – 405 с. – Электр. ресурс. URL: http://www.chronos.msu.ru/ru/rrules/item/sistemodinamika-2 (05.07.14).
3. Аверин Г.В., Звягинцева А.В. Взаимосвязь термодинамической и информационной энтропии при описании состояний идеального газа // Системный анализ и информационные технологии в науках о природе и обществе. Донецк: Друк-инфо, 2013. №1 (4) – 2 (5). – С. 46 – 55. – Электр. ресурс. URL: http://sait.csm.donntu.org (11.07.14).
4. Применение методов интеллектуального анализа данных при оценке развития Украины // Геотехническая механика. Днепропетровск, 2013. – Выпуск 112. – С. 257 – 270. – Электр. ресурс. URL: http://geotm.dp.ua/index.php/ru/2013-god/vypusk-112 (06.06.14).

References (transliteration)

1. Analitika dvojnogo naznachenija [Analysis of dual purpose] // Otkrytye sistemy, no 10, 2013.
2. Averin G.V. Sistemodinamika [Systemdynamics]. – Doneck: Donbass, 2014. – 405 p. – Elektr. resurs. URL: http://www.Chronos.msu.ru/ru/rrules/item/ sistemodinamika-2 (05.07.14).
3. Averin G.V. and Zvjaginceva A.V. Vzaimosvjaz' termodinamicheskoj i informacionnoj jentropii pri opisanii sostojanij ideal'nogo gaza [The relationship of the thermodynamic entropy and information in the description of the ideal gas] // Sistemnyj analiz i informacionnye tehnologii v naukah o prirode i obshhestve. Doneck: Druk-info, 2013. – no 1 (4) – 2 (5). – pp. 46 – 55. – Elektr. resurs. URL: http://sait.csm.donntu.org (11.07.14).
4. Primenenie metodov intellektual'nogo analiza dannyh pri ocenke razvitija Ukrainy [The use of data mining techniques in the evaluation of the development of Ukraine] // Geotehnicheskaja mehanika. Dnepropetrovsk, 2013. – Issue 112. – pp. 257 – 270. – Elektr. resurs. URL: http://geotm.dp.ua/index.php/ru/2013-god/vypusk-112 (06.06.14).

Аверін Г.В., Звягінцева Г.В. «Моделі даних для окремих проблемно-орієнтованих баз даних». *Вивчаються методи та засоби феноменологічного аналізу даних для масивів емпіричної або статистичної інформації, яка відображає у вигляді часових рядів процеси зміни й розвитку систем різної природи. Основна гіпотеза дослідження пов'язана з можливістю створення на основі таблично-тимчасових даних безлічі моделей у вигляді феноменологічних описів процесів і явищ, що відрізняються багатовимірним польовим поданням масивів кількісної інформації, а також існуванням скалярних полів емпіричних заходів для комплексної оцінки станів систем. Дана гіпотеза може бути прийнята або відкинута на основі обробки наявних даних, найбільш адекватні моделі можуть бути обрані з безлічі моделей з використанням напівавтоматичних алгоритмів моделювання. Надано пропозиції щодо створення програмних продуктів для феноменологічного аналізу даних. Створення моделей даних дозволить підвищити ефективність діяльності вчених, аналітиків та експертів при дослідженні природних процесів і явищ, при аналізі процесів світового та регіонального розвитку, при вирішенні актуальних завдань охорони навколишнього середовища, біорізноманіття та промислової безпеки.*

Ключові слова: складні системи, аналіз даних, моделювання, феноменологічні моделі, математичне та програмне забезпечення.

Averin G.V., Zviagintseva A.V. "Data models for certain problem-oriented databases". *The paper studies phenomenological analysis methods and tools applicable to empirical or statistical information that in the form of time series reflects the change and development processes of different nature systems. The main research hypothesis is related to the possibility of using tabular-temporal data to create the set of models in the form of phenomenological descriptions of processes and events. The models differ from others by representing quantitative information with multidimensional field as well as existence of scalar fields of empirical measures for complex assessment of the system states. This hypothesis may be accepted or rejected based on the processing of available data. The most adequate models may be selected among the set of models using supervised modeling algorithms. The paper gives the proposals for creating phenomenological analysis software. The creation of data models may help to increase the effectiveness of the work conducted by researchers, analysts and experts during the exploration of natural processes and events, analyses of global and regional development processes, solving tasks of environmental protection, biodiversity and industrial safety.*

Keywords: complex systems, data analysis, modeling, phenomenological models, mathematical and software support.

Статья поступила в редакцию 07.08.2014
Рекомендована к публикации д-ром техн. наук Ф.В. Недопекиным