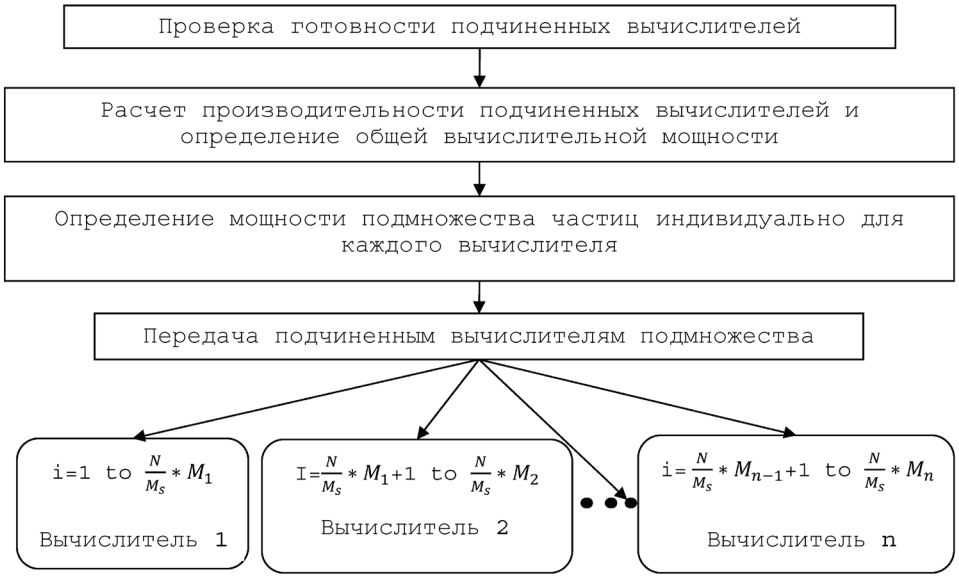

Алгоритм распределения нагрузки между вычислителями в соответствии с производительностью каждого вычислителя.

Для оценки эффективности разработанной модели распределенных вычислителей проведено тестовое моделирование системы оксид кремния SiO2 в программном комплексе ИИС «MD-SLAG-MELT»[5]. Тестирование проходило на 16 вычислителях AMD и Intel разной производительной мощности. В моделирование сравнивалось время, затраченное на проведение эксперимента с числом частиц в системе 105 и числом шагов 5000. Результаты компьютерных экспериментов показаны в таблице.

Результаты тестового моделирования системы SiO_2 (в часах)

N Метод загрузки	1000	4200	12000	100000
Равномерная загрузка	0,9	1,75	3,7	11,5
Загрузка по вычислительным мощностям	0,42	0,81	2,9	9,4

Таким образом, использования модели балансировки нагрузки позволит в равной степени задействовать все вычислители на максимальную мощность в течении всего времени выполнения моделирования системы N-частиц. Данный подход обеспечивает наиболее равномерное распределение нагрузки между вычислителями, т.к. позволяет гибко регулировать объемы обрабатываемых подмножеств и обеспечивает эффективность работы всей системы приближенную к оптимальной, что не позволяют обеспечить ранее описанные методы для гетерогенной вычислительной среды. Кроме того, данная модель позволяет обеспечить оптимальную эффективность использования вычислителей, как для гетерогенной, так и для однородной среды.

Список литературы

1. Воронова Л.И., Григорьева М.А., Воронов В.И., Трунов А.С. Программный комплекс «MD-SLAG-MELT» для моделирования наноструктуры и свойств многокомпонентных расплавов. – Расплавы. 2013. № 4. С. 36-49.

2. Трунов А.С., Воронова Л.И., Шалабай Т.С. Метод параллельного расчета коррелированной системы n-частиц на графическом процессоре. – Современные наукоемкие технологии. 2013. № 6. С. 117-119.

3. Воронова Л.И., Трунов А.С., Воронов В.И. Разработка методов параллельного расчета коррелированной многочастичной системы на графическом процессоре. – Вестник Российского государственного гуманитарного университета. 2013. № 14. С. 236-247.

4. Трунов А.С., Воронова Л.И., Воронов В.И. Разработка методов распределения для высокопроизводительных вычислений в многочастичных системах. – Международный журнал прикладных и фундаментальных исследований. 2013. № 10-2. С. 192-194.

АВТОМАТИЗАЦИЯ РАНЖИРОВАНИЯ ТИПОВ ВИЗУАЛИЗАЦИИ В DATA MINING ДЛЯ НЕРЕЛЯЦИОННЫХ БАЗ ДАННЫХ

Хомутова Е.В., Воронова Л.И.

Московский технический университет связи и информатики, Москва, e-mail: evenie@mail.ru

Основным инструментом практически любого бизнеса в современном мире является анализ информации. Без анализа данных невозможно провести оценку выбора пользователя («пользовательской корзины»), произвести сегментацию рынка и прогнозирование, оценить риски и разработать стратегию дальнейшего развития. Наиболее полно решить все перечисленные выше задачи способны программные продукты, реализующие методы технологии Data Mining. В статье рассмотрены варианты автоматизации одного из модулей Data Mining – визуализации результатов анализа данных для нереляционных баз данных.

Data Mining – (рус. добыча данных, интеллектуальный анализ данных, глубинный анализ данных) – собирательное название, используемое для обозначения совокупности методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности [1]. В более узком смысле, Data Mining – это интеллектуальный анализ данных (ИАД), направленный на выявление скрытых, неявных, ранее неизвестных закономерно-

стей. Технология Data Mining включает в себя такие модули, как классификация, кластеризация, прогнозирование, поиск данных и визуализация.

Являясь достаточно молодой методикой (с 1989 г. [2]), Data Mining представляет собой обширное поле для исследований и модернизации существующих методов анализа данных. Наименее изученным и разработанным модулем является модуль визуализации, который выбран в данной работе как тема научного исследования.

На данный момент сложно найти область производства, в которой в том или ином виде не применяется ИАД. Зачастую наработки Data Mining используются в сферах, далеких от информационных технологий и потребители программных продуктов анализа данных могут не иметь профессионального образования в данных областях, что требует от программных приложений для ИАД удобства и простоты использования.

Современные программные продукты, обеспечивающие анализ данных, уже решили проблемы автоматизации предварительной обработки данных для анализа. Так, программный комплекс PolyAnalyst компании Megaputer позволяет решить проблемы предобработки для таких модулей, как прогнозирование, классификация, кластеризация, группирование по родству, анализ связей и т.д. [3]. А флагман предсказательного моделирования – KXEN – и вовсе проводит первичную оценку данных на основе существующей базы знаний (то есть практически без вмешательства пользователя), разрешая возникающие конфликты в режиме диалога [4].

Однако и PolyAnalyst, и KXEN не производят предварительной оценки результатов анализа данных для выбора метода их визуализации: таблицы, графики, диаграммы, гистограммы и многое другое. Подобно программному пакету Excel, они предоставляют пользователю право выбрать нужный ему тип из каталога имеющихся видов визуализации. Данный подход является не совсем корректным, поскольку пользователь может не знать, какой тип визуализации лучше подходит для его задачи, либо какие вообще типы визуализации существуют в данном пакете. Поэтому возникает задача автоматизации ранжирования видов визуализации, которая рассматривается в данной статье.

Рассмотрим концепцию предварительной обработки данных для выбора типа визуализации. Основной алгоритм состоит из следующих шагов:

Выбор данных для анализа;

Анализ типа данных и количества их групп;

Выявление математических закономерностей исследуемого набора данных;

Сравнение полученных на предыдущем шаге закономерностей с параметрами каждого типа визуализации из имеющейся базы знаний типов визуализации;

Предоставление пользователю наиболее подходящего типа визуализации.

Данный метод требует предварительной разработки базы знаний для программного продукта. Параметрами каждого типа визуализации могут служить, например, типы данных в столбце и число столбцов. Так, для числовых данных или данных, имеющих тип «дата», при наличии двух столбцов для анализа приемлемей использовать графики или точечные диаграммы, поскольку они показывают динамику развития или распределение статических данных. Для случая, когда один из столбцов для анализа данных имеет символьный тип данных (CHAR), а второй – числовой или датированный, больше подходит отображение данных в виде линейных столбцов, круговой диаграммы или диаграммы-пирога (“pie chart”). Если же для анализа представлено более двух столбцов, то оптимальным вариантом является визуализация данных в виде гистограммы, позволяющей наглядно отобразить сравнение изменений нескольких групп данных. При этом важно отметить, что столбцы, содержащих символьные типы данных, преимущественно будут интерпретированы как группы данных (т.е. число столбцов в гистограмме), а датированные или числовые типы данных в свою очередь будут располагаться на осях X и Y соответственно.

Похожее решение проблемы автоматизации выбора типа визуализации уже было предложено корпорацией Microsoft, но на сегодняшний момент еще не было реализовано [5]. Однако предложенный метод пригоден только для реляционных баз данных, что заставляет пользовать проводить предварительную обработку данных, приводя их к табличному виду, а это является достаточно сложной, а порой и невозможной задачей.

Поскольку технология ИАД Data Mining направлена на анализ как реляционных, так и нерелляционных данных, ниже приводится принцип ранжирования видов визуализации для XML-файлов. Данный принцип будет базироваться на описанной выше технологии анализа типов данных и количества анализируемых атрибутов (столбцов), однако будет добавлен модуль, осуществляющий предварительное преобразование XML-файла к реляционному виду.

XML – рекомендованный W3C язык разметки. Спецификация XML описывает XML-документы и частично описывает поведение XML-процессоров[6]. XML-файлы имеют программный интерфейс DOM и иерархическую структуру, представленную корневым тэгом со вложенными в него элементами. [7] Таким образом программа-парсер, проводящая предварительную обработку XML-файла должна выполнить следующие шаги:

• Нахождение тэга по заданному пользователем имени;

• Фиксация значения, содержащегося в данном тэге;

• Подсчет данных тэгов и фиксация содержащихся в них значений;

• Проверка на наличие вложенных тэгов;

• Фиксация значения, содержащегося во вложенных тэгах;

• Подсчет данных тэгов и фиксация содержащихся в них значений.

Программа-парсер работает до тех пор, пока не будут разобраны все одноименные тэги в выбранном для анализе XML-файле. По завершении извлеченные из файла данные преобразуются к табличному виду (может происходить конвертация в любой подходящий тип файла: xls, csv, dtb и др.), и далее происходит анализ типов данных и количества атрибутов, описанный в данной статье выше.

Стоит заметить, что все предобработки, а также сам процесс ранжирования типов визуализации происходит автоматически, что значительно упрощает работу пользователей. Им необходимо ввести только имя тэга, по которому нужно построить диаграмму. В то время, как в аналогичном продукте корпорации Microsoft пользователь должен указать все необходимые для анализа столбцы таблицы.

Предложенный метод автоматизации ранжирования типов визуализации для нерелляционных баз данных имеет ряд преимуществ перед предложенным методом компании Microsoft. Среди них можно выделить следующие:

Уникальность: на данный момент отсутствуют реализации аналогичного программного продукта;

Универсальность: метод работает как с реляционными, так и с нетабличными БД;

Простота использования пользователем: при анализе нерелляционных баз данных пользователю необходимо указать только имя тэга;

Дешевизна реализации: описанный в данной статье метод достаточно прост для реализации, поскольку является лишь одним компонентом программного комплекса, осуществляющего интеллектуальный анализ данных.

Выводы

В работе проведен анализ предметной области, связанной с применением средств визуализации технологии Data Mining. Рассмотрены актуальные программные приложения. Обозначена концепция выявления параметров данных, влияющих на тип их визуализации. Описаны алгоритмы работы анализаторов для реляционных данных и предложен алгоритм ранжирования типов визуализации для нерелляционных данных.

Список литературы

1. http://ru.wikipedia.org/wiki/Data_mining
2. Коваленко О.С. Обзор проблем и перспектив анализа данных
3. http://megaputer.ru/data_mining.php
4. http://www.kxen.com/
5. http://www.findpatent.ru/patent/248/2488159.html
6. http://ru.wikipedia.org/wiki/XML
7. http://ru.wikipedia.org/wiki/Document_Object_Model

Секция «Математическое моделирование химико-технологических процессов», научный руководитель – Антипина С.Г.

ИЗУЧЕНИЕ СКОРОСТИ МОНОМОЛЕКУЛЯРНОЙ ХИМИЧЕСКОЙ РЕАКЦИИ

Базова А.В., Самохвалова И.О., Антипина С. Г.

Волжский политехнический институт, филиал Волгоградского государственного технического университета, Волжский, e-mail: gelj@live.ru

В истории химической кинетики мономолекулярные реакции занимают особое место, так как статистическая обработка экспериментальных данных при изучении скорости мономолекулярной химической реакции используется повседневно в химической промышленности. Мономолекулярные реакции – химические реакции, в которых одна молекула подвергается превращению. К мономолекулярным реакциям относят многочисленные реакции распада сложных молекул (диссоциация) и изомеризации. Скорость химической реакции – количество вещества, которое вступает в реакцию или образуется при реакции за единицу времени в единице объема системы.

На практике при изучении скорости мономолекулярной химической реакции получили зависимость количества вещества в реакционной смеси к моменту времени, прошедшего от начала опыта:

x	3	6	9	12	15	18	21	24
y	57,6	41,9	31,0	22,7	16,6	12,2	8,9	6,5

Значительное число нелинейных зависимостей, встречающихся в химической практике, может быть описано следующими уравнениями:

$$y = a \cdot b^x;\ y = a \cdot x^b;\ y = \frac{x}{a + bx}.$$

Первое и второе уравнения легко привести к линейному виду, прологарифмировав их:

$$(1)\ \ln y = \ln a + x \cdot \ln b \Rightarrow Y = A + Bx,$$

где $Y = \ln y,\ A = \ln a,\ B = \ln b$;

$$(2)\ \ln y = \ln a + b \cdot \ln x \Rightarrow Y = A + bX,$$

где $Y = \ln y,\ A = \ln a,\ X = \ln x$.

Для приведения третьего уравнения к линейному виду выполним преобразование:

$$y = \frac{x}{a + bx} \Rightarrow \frac{x}{y} = a + bx \Rightarrow Y = a + bx,$$

где $Y = \frac{x}{y}$.

На рис. 1. отображены диаграммы значений новых переменных (X_i; Y_i) для каждой из рассмотренных моделей. Для первой модели точки располагаются вдоль некоторой прямой, поэтому зависимость y от x выражается формулой $y = a \cdot b^x$.

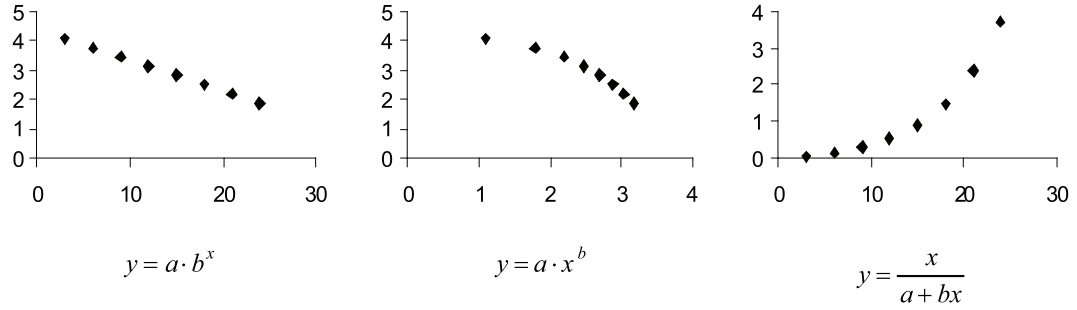

Рис. 1. Диаграммы рассеяния