Научная визуализация, 2019, том 11, номер 1, страницы 139 - 176, DOI: 10.26583/sv.11.1.11

Модели абстракции данных: выборка (параллельные координаты), фильтрация, кластеризация

Д.В. Манаков

Институт математики и механики им. Н.Н. Красовского Уральского отделения РАН, Екатеринбург, Россия

ORCID: 0000-0001-6852-8096 , manakov@imm.uran.ru

Аннотация

При рассмотрении компьютерной визуализации как самостоятельной дисциплины, необходимо построение ее ментального пространства со своей семантикой, прагматикой и базисом. Тогда любые два специалиста по визуализации смогут говорить на одном языке. Этот базис выбирается из достаточно широкой междисциплинарной области знаний. Верификацию визуализации в духе нечетких множеств определим через отношение двух базисных функций точности и полноты визуализации, она должна гарантировать, что конечному пользователю предложена формально правильная модель визуализации, или, другими словами, что разработчики систем визуализации решили поставленную задачу.

На современном этапе развития компьютерной визуализации критерий полноты является более важным. Сначала необходимо сформировать ментальное пространство, а затем, уточняя семантику, прагматику и базис, заменяя ментальное пространство логическим пространством, перейти к верификации визуализации. Построение монотонно возрастающих базисных функций, например, точность визуализации: постановка задачи, прототип, приложение, сервис, позволяет рассматривать классификацию как непрерывный процесс. Возможные постановки задач рассматриваются как вызовы и определяют не только перспективные направления развития визуализации, но и их множество, что продуцирует функцию полноты.

В секторе компьютерной визуализации ИММ УрО РАН рассматривается возможность разработки он-лайн сервисов параллельных вычислений. На базе конструктора веб-визуализации можно реализовать автономную поддержку стандартных моделей абстракции данных, в частности, фильтрацию, кластеризацию, выборку. Основная часть данной работы содержит обзор этих моделей. С целью выделения общих подходов мы разрабатываем нечеткую верифицированную классификацию, которая учитывает как частоту встречаемости моделей, структурных единиц, информативных признаков, так и математический уровень абстракции данных.

Поскольку визуализация становится средой автоматизированного аналитического процесса, для визуальной аналитики представляют интерес направления, связанные с самоорганизацией, например, диссипативные системы. С этих позиций можно уточнить понятие структурной единицы визуального анализа, включая модели абстракции данных. К структурным единицам визуального процесса относятся визуальная парадигма, анализ чувствительности, рефакторинг, калибровка, предельная неопределенность, веб-визуализация. Построение логического пространства обеспечивает автоматическую верификацию. Мы предлагаем рассматривать структурную единицу как непрерывное отображение класса подмножеств данных на логическое пространство.

Ключевые слова: верификация визуализации, логическое пространство, диссипативные системы, предельная неопределенность, фильтрация, кластеризация, выборка, параллельные координаты.

1. Введение

В секторе компьютерной визуализации ИММ УрО РАН рассматривается возможность разработки он-лайн сервисов параллельных вычислений. На базе конструктора веб-визуализации можно реализовать автономную поддержку стандартных, общепринятых моделей абстракции данных, а на их основе, например, испытательный стенд для оценки и моделирования зрительного восприятия человека в очках виртуальной реальности. Он-лайн сервисы нужны для того, чтобы любой исследователь мог ввести свои данные на вычислитель, выбрать математическую модель и модель визуализации, указать критерий оптимальности, оценить полученные результаты и при необходимости провести рефакторинг моделей. При желании исследователь может предложить свой сервис.

Хотя идея интеграции не является новой, например, можно отметить проект АлгоВики, данная задача по-прежнему остается актуальной. Модели абстракции данных в отличие от визуализации инженерных пакетов зачастую не имеют естественной образности, поэтому структурирование и верификация этих моделей является более трудоемкой задачей. В контексте данной работы под интеграцией понимается ситуация, когда для верификации визуализации недостаточно одного вида отображения или одной модели или одной группы исследователей.

Цель данной работы – не только предложить интегрированную классификацию абстракций данных, выделив стандартные модели, но и оценить дальнейшие перспективы развития с формальных позиций. Формализованная (верифицированная) классификация должна учитывать общетеоретические сведения из области информационной визуализации.

Стоит отметить некоторые проблемы классификации моделей абстракции данных, которые могут иметь формальное решение:

1. Визуальное представление моделей абстракции данных не имеют естественной образности. Можно использовать комбинированные модели, сочетающие модели абстракции данных и квазиестественную образность, например, метафору карты.
2. Высокая степень синонимичности терминов, следовательно, необходимо учитывать их частоту встречаемости (нечеткая модель классификации).
3. Отображение данных на вид отображения не изоморфно, следовательно, необходимо учитывать математическую модель.

Фактически во введении обосновывается структура статьи по степени важности: на первом месте стоит теоретическая часть, затем идет модель классификации и, наконец, обзор моделей абстракции данных, ограниченный рассмотрением и верификацией параллельных координат, фильтрацией данных и кластеризацией.

2. К теории компьютерной визуализации и визуальной аналитики

Визуальная аналитика должна предоставить технологии, которые сочетают сильные стороны человеческой и электронной обработки данных. Визуализация становится средой полуавтоматического (автоматизированного) аналитического процесса, где люди и машины сотрудничают для достижения наиболее эффективных результатов [1]. И все же данная работа ориентирована на компьютерные модели, а не на модели восприятия-понимания человека. Цель визуальной аналитики – сделать способы обработки данных и информации прозрачными для аналитического дискурса (обсуждения).

Итак, визуальная аналитика – наука об аналитических рассуждениях, упрощаемых интерактивными визуальными интерфейсами. Частный случай визуальной аналитики – бизнес-аналитика.

Модели абстракции данных – обобщения, позволяющие абстрагироваться от источника и происхождения (онтологии) данных во время визуального анализа.

Стандартный конвейер (цикл) компьютерного моделирования [2], включающий физическую, математическую, алгоритмическую, программную, визуальную модели и модели визуальной аналитики, можно рассматривать как последовательность отображения данных. Очевидно, что любая формализация повышает уровень абстракции данных, в то время как интерпретация (связанная с визуальной аналитикой) нацелена на понижение этого уровня, реализуемого через обратную связь к любой более низкой модели. Следовательно, интерпретация невозможна без учета контекста –некоторого ментального представления о формальных моделях цикла компьютерного моделирования.

Рассматриваемые в этой статье модели абстракции данных ограничены диапазоном от математической модели до визуальной модели данных. Основные структурные единицы математического уровня абстракции данных, в том числе и по частоте встречаемости – фильтрация, кластеризация, выборка (выбор) – не имеют естественной образности.

Для сравнения приведем пример модели более высокого уровня абстракции. Датацентрическая теория [3] ориентирована на шаблоны (типы, структуры) данных, используемые в программировании. Такие структуры данных, как *k*-деревья, *R*-деревья, ассоциативные массивы, стеки (итераторы) – не только могут быть связаны с оптимизацией вычислительного алгоритма, но и отвечать за способ распараллеливания. Например, *k*-деревья (в частности окта-деревья) применяются для решения задач с различной математической постановкой: параллельные вычисления (алгоритмы с внешней памятью), визуализация (объемный рендеринг с уровнем детализации), дискретная оптимизация (геометрическая аппроксимация транспортной задачи). Таким образом, *k*-деревья являются структурной единицей моделей алгоритмического уровня абстракции данных.

Для того, чтобы компьютерная визуализация считалась самостоятельной дисциплиной, и тем более визуальная аналитика, необходимо сформировать (например, с точки зрения когнитивной психологии) ее ментальное пространство со своей семантикой, прагматикой и базисом. Тогда любые два специалиста по визуализации смогут говорить на одном языке. Этот базис выбирается из достаточно широкой области знаний [4], и должен определять размерность пространства (визуализации). Термин когнитивная размерность, используемый, например, в “usability”, можно считать удачным в качестве дополнительной размерности, например, к декартовым координатам. Когнитивная размерность определяется количеством эвристик или термов, которые изначально считаются ортогональными или независимыми [5]. В то же время подобные определения базиса, на наш взгляд, некорректны в математическом смысле.

Альтернативным решением является понятие структурной (семантической) единицы или просто юнита, подходящие для моделей абстракции данных. С точки зрения теории множеств, структурную единицу можно рассматривать как непрерывное отображение класса подмножеств данных на логическое пространство. В рамках датацентрической теории подобное отображение можно реализовать с помощью ассоциативных массивов с множественным ключом и с множественным значением. Обычное логическое пространство [6] определяется как пара $LS = (V, 2^V)$, где фиксированное непустое множество значений истинности – V рассматривается вместе с множеством своих подмножеств – 2^V. Иными словами, множество истинностных значений называется логическим пространством, если на нем выделены определенные подмножества. Поскольку визуализация по своей природе многозначная и нечеткая, для визуальной аналитики представляют интерес истинностные значения k-логики и нечеткой логики. В области программирования частным случаем непрерывного отображения данных на логическое пространство можно считать домен всех доменов (powerdomain).

Построение логического пространства для системы визуализации обеспечивает автоматическую верификацию деятельности пользователя или агента программирования. Распространенными построениями логического пространства являются: отобра-

жение симметрической группы, *n*-линейное и ограниченное отображение, к которому относятся тестирование и алгоритм MapReduce.

В случае криволинейных координат (трилинейное ограниченное отображение) значение определителя (якобиана) формирует логическое пространство (если определитель больше нуля, то отображение является взаимно-однозначным, следовательно, решение существует и единственное; если определитель равен нулю, то решение неединственное; если определитель меньше нуля образуются восемь видов вырожденных ячеек).

Семантические единицы (различимые по смыслу) – непересекающиеся (частично пересекающиеся) визуальные тексты [7] с истинностным значением, близким к единице, участвуют в формирование базиса. В случае *n*-линейного и ограниченного отображения размерность пространства равна *n*. Для *n*-линейного ограниченного отображения существует взаимно-однозначное отображение на числовую прямую (сюръекция, действие). В этом случае применение теории вероятности или нечетких множеств является оправданным (логическое пространство как вероятностное пространство).

Возможные структурные единицы визуализации: визуальные переменные, знаки, образы, тексты, виды отображения, метафоры визуализации и взаимодействия. Например, визуальные переменные используются в визуальной семиотике.

Аналогично лингвистическому направлению, визуализацию можно рассматривать как представление и как процесс. Выше перечислены структурные единицы представления. Понятие визуальной парадигмы определяется как цель визуализации (целенаправленность, интенциональность). Достижение цели можно рассматривать как процесс.

Цикличность процедуры визуального анализа приводит к появлению понятия об элементарном структурном объекте, повторение которого формирует ход решения. Структурной единицей визуального анализа (визуального процесса) предлагается считать [8] состояние визуальной модели. Интерпретация этого состояния предоставляет наблюдателю объем информации, необходимый для последующего получения общего результата проводимого анализа. Структурная единица может быть самостоятельной визуальной моделью, либо входит в состав более сложного объединения, являющегося средством визуального анализа применительно к поставленной задаче.

Структурная единица визуального анализа является визуально воспринимаемым образом (или другой структурной единицей визуального представления), интерпретируемым как истинностный ответ (в первоисточнике [8] однозначный ответ) на один из промежуточных вопросов. Сложность и структура вопроса определяется уровнем абстракции модели с ограничениями по времени, по вычислительным ресурсам и на модель конечного пользователя.

В терминах системного анализа структурная единица визуального анализа представляет собой управляемую систему *S* с обратной связью *R(t)*. Объем исследуемых данных поступает на неуправляемый вход *V(t)* структурной единицы, см. рис. 1. Результатом является информация, изменяющая состояние выхода *Y(t)*. Особенностью такой системы является возможность изменения состояния управляемого входа системы *U(t)* в зависимости от полученных результатов и благодаря наличию обратной связи *R(t)*. Пользователь *P* является, с точки зрения комплексного подхода к визуальному анализу, обязательным участником структурной единицы. Влияние пользователя проявляется в изменении состояний управляемых входов и выходов, а также в регулировании работы системы *S* и принятии решения о завершении ее функционирования.

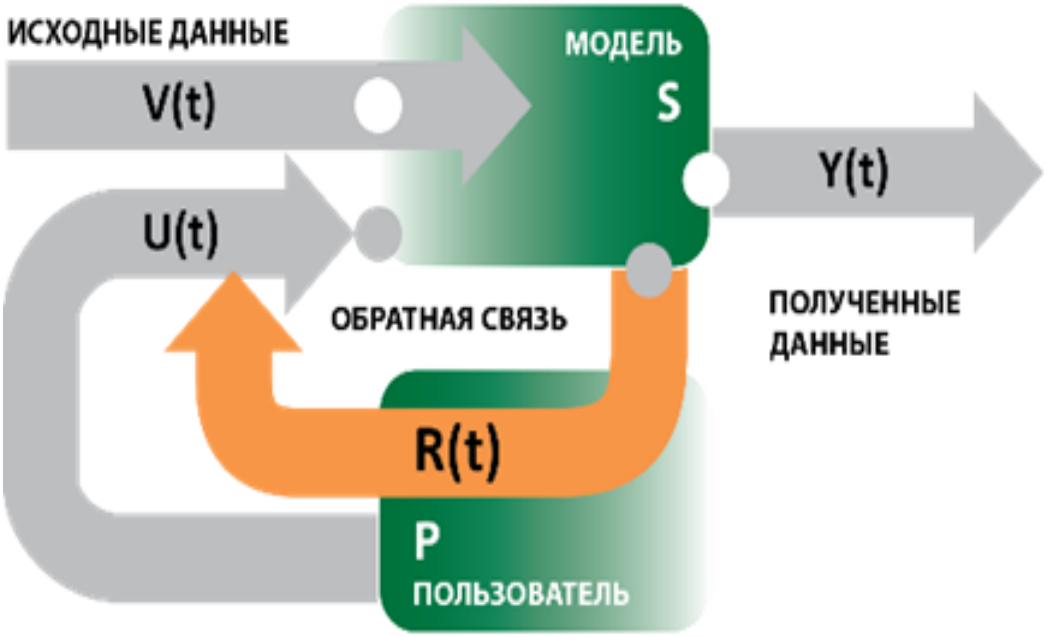

Рис. 1. Структурная единица визуального анализа [2]**.**

Поскольку визуализация становится средой **автоматизированного** аналитического процесса, для визуальной аналитики представляют интерес направления, связанные с **самоорганизацией**: диссипативные системы, автономные вычисления и синергетика. Например, можно уточнить понятие структурной единицы визуального анализа рис. 1 с позиции диссипативных систем.

Диссипативная система - квазистационарная открытая система, характерной особенностью которой является процесс самоорганизации, происходящий в результате действия отрицательного вектора, например, силы трения.

Поясним отличие диссипативной системы от задачи оптимального управления в случае линейных систем:

$$\dot{y} = Ay - Bu,$$

где u – оптимальное управление, которое надо найти.

Поскольку диссипативная система является открытой системой, управление $V(t)$ приходит извне, оно изначально задано. Введение в рассмотрение отрицательного вектора или отрицательной обратной связи $R(t)$ сужает круг задач до моделей с насыщением. Возникновение обратной связи стоит рассматривать не только как результат целенаправленных, но и хаотических, случайных действий пользователя (например, аналогично броуновскому движению), поскольку процесс самоорганизации происходит гораздо быстрее при наличии в системе внешних и внутренних шумов. Управление пользователя $U(t)$ желательно, чтобы было оптимальным $u=u(U(t),V(t))$, по крайней мере оно должно существовать. Например, можно предложить пользователю y_i – монотонно сходящуюся последовательность решений [4].

В принципе, рассмотрение восприятия-понимания человека [9] может базироваться на частном случае диссипативных систем – теории функциональных систем П. К. Анохина. Человек находится в состоянии гомеостаза (почти статического равновесия). Выделяют два типа функциональных систем. Системы первого типа обеспечивают гомеостаз за счёт внутренних (уже имеющихся) ресурсов организма, не выходя за его пределы – замкнутые системы. Системы второго типа поддерживают гомеостаз за счёт изменения поведения, которое формируется в результате отрицательной обратной связи – открытые (диссипативные) системы.

Возможно слияние компьютерных моделей и моделей восприятия-понимания через ввод понятия интеллектуального агента. Под агентом понимают открытую, активную, целенаправленную систему, которая способна сама формировать собственное поведе-

ние в не полностью определенной среде [10]. Интеллектуальный агент – агент, способный к самопознанию.

Ключевым понятием диссипативных систем является энтропия. Например, термодинамическая энтропия - мера необратимой диссипации энергии. Поскольку понятия термодинамической энтропии и информационной энтропии (информативности) эквивалентны, то применение теории диссипативных систем в области информационных технологий и визуализации – обосновано.

Так как данная работа ориентирована на модели абстракции данных, рассмотрим иллюстративный пример динамики диссипативных систем в области информационных технологий (задача кластеризации). Рис. 2 представляет визуальный анализ деятельности кредитных организаций. Он позволяет строить суждения об образовании кластеров (диссипативных структур) и находить объекты (выделено красным цветом), стремящиеся оказаться в кластере [11] (аналогично, модели с насыщением - движения амеб в сторону вещества с наибольшей концентрацией). С точки зрения диссипативных систем на следующем шаге кластер должен разрушиться (обычно по превышению некоторого порогового значения) и на его периферии должен сформироваться новый кластер.

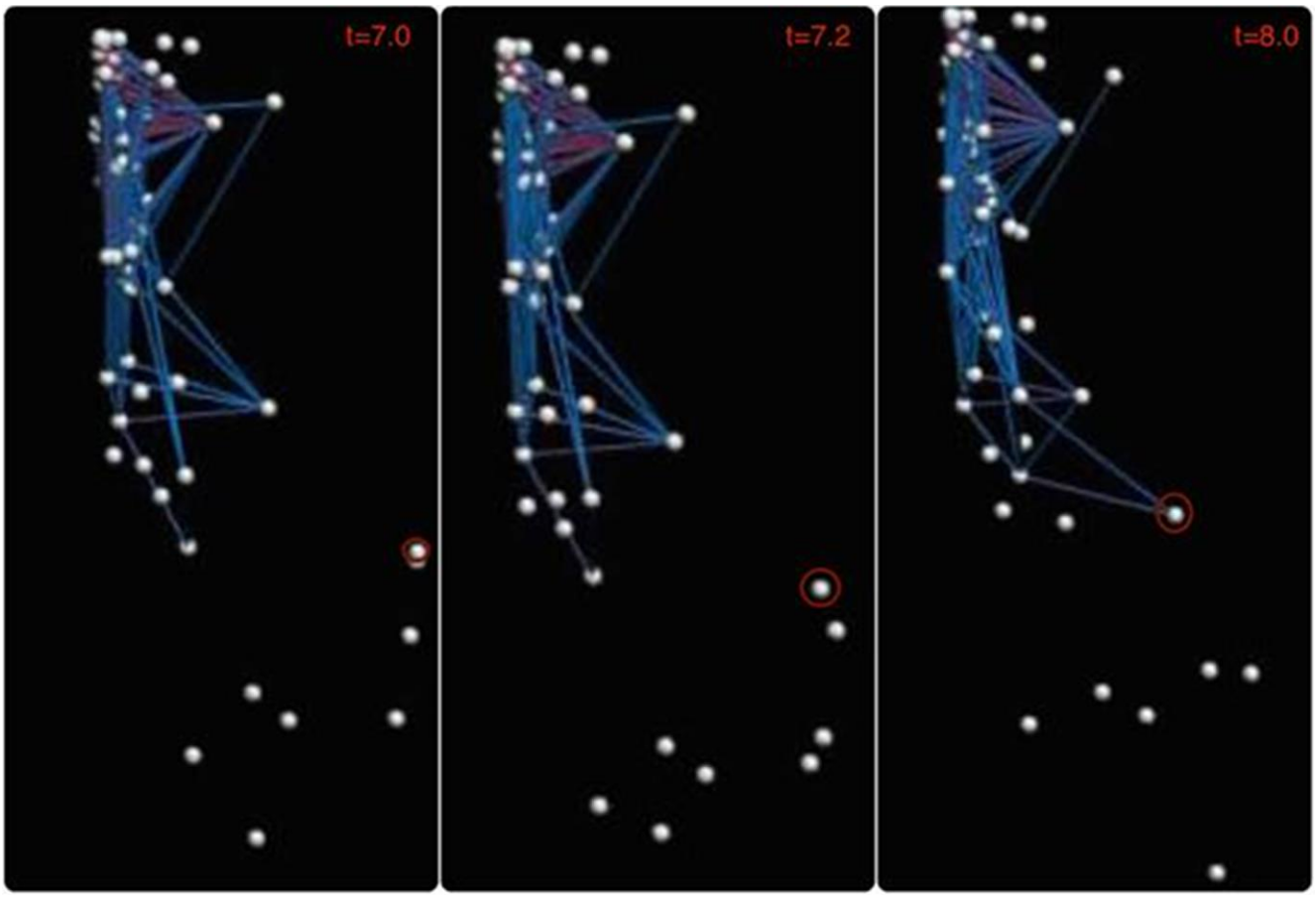

Рис. 2. Кластер и объект, стремящейся оказаться в кластере [11].

Негосударственные интернет-деньги, включая криптовалюту, формируют новый кластер. Перспективным направлением в области параллельных и распределенных вычислений является модель потока данных. Если в области финансов примером модели потока данных является Blockchain (способ последовательно записать такой поток), то в области визуализации – Vistrail [12]. Vistrail это классический конструктор модели потока данных, например, параметры модели можно не только менять, но и связывать (используя графы, таблицы). Vistrail использует графическую библиотеку VTK. При переносе объектно-ориентированных библиотек в область параллельных

вычислений существуют определенные ограничения на реализацию непосредственного взаимодействия, в частности, параллельный VTK ориентирован на параллельный рендеринг, поэтому ввод вида отображения multiple-view (рис. 3 - множеству фиксированных параметров ставится в соответствие множество изображений) является вынужденным.

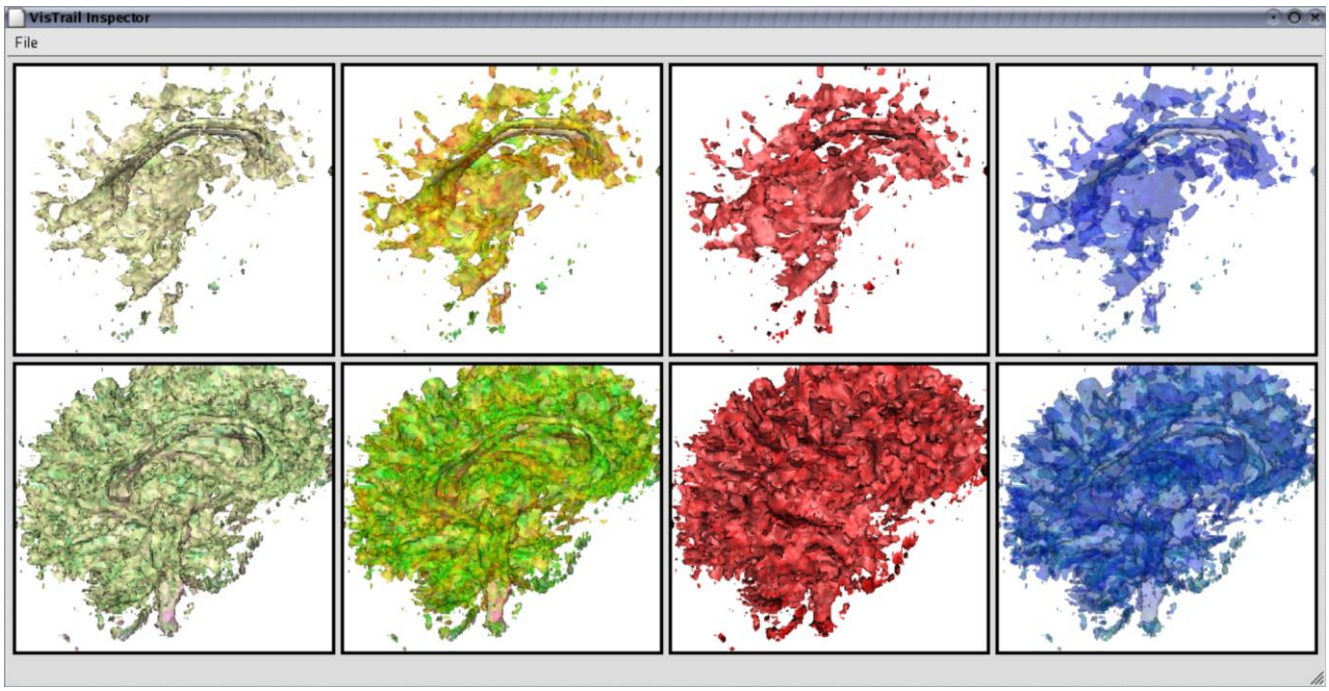

Рис. 3. Вид отображения multiple-view [12]

Формализация модели потока данных может основываться на денотационной семантике Скотта для λ-исчисления [13]. Для построения непрерывного отображения топологический подход, основанный на построении замыкания с такими определяющими свойствами, как частичный порядок и существование супремума, является конструктивным.

В качестве примера в области визуализации можно привести семиотическое определение метафоры визуализации, рассматриваемой как непрерывное отображение исходного множества на целевое множество [4] (или любой другой структурной единицы визуализации). В стандартное определение метафоры по Лакоффу [14] добавлено только свойство непрерывности. Хотя монотонность является следствием частичного порядка и ограниченности, в [4] предлагается рассматривать монотонное отображение с целью расширения класса решаемых задач.

Визуализация по своей природе многозначная и нечеткая; можно вести понятие k-монотонности и α-монотонности соответственно для k-логики и нечеткой логики так, чтобы отображение по-прежнему оставалось непрерывным. Для непрерывной модели разделение визуализации на представление и процесс не существенно. Наравне с топологическим подходом (денотационная семантика) непрерывную модель можно определить через малое изменение параметров модели (операционная семантика). Например, модель процесса исследования визуализации (A Model for the Visualization Exploration Process) [15] основана на исчислении параметров. Если существует визуальная парадигма, то процесс визуализации можно рассматривать как направленное действие. Для любого направленного действия существует конечный предел.

Непрерывность модели – непрерывность цикла прохождения задания и непрерывность познания, включая визуальный анализ результатов вычислений. Таким образом, для систем визуализации возможна постановка таких стандартных в математическом плане задач:

1. Определение информативных признаков, то есть, какими свойствами должна обладать система визуализации, чтобы работа с ней была эффективной, и построение по ним логического пространства.
2. Задача визуального анализа чувствительности решения в зависимости не только от параметров прикладной задачи (модели), но и от параметров параллельной программы, и от параметров визуализации [7].

Анализируя визуальные модели, скорее стоит говорить о ментальной модели, чем о математической модели. Хотя формальное понимание может помочь в разработке новых систем визуализации и в уточнении текущих [15].

В результате сформировалось направление по верификации визуализации, которое должно предложить пользователям формально правильные системы визуализации. Начиная с 2012 года, проходят Европейские семинары по проблеме воспроизводимости, верификации и валидации в визуализации (EuroRV3) [16].

Можно ввести две базисные функции или меры: полнота верификации и точность верификации, связанная с распространением неопределенности. В работе [17] вводится понятие верифицируемой визуализации, которая отслеживает, как распространяется погрешность (неопределенность) на всем этапе вычислительного конвейера, включая визуализацию.

Одновременно с верификацией визуализации необходимо рассматривать валидацию визуализации. Валидация – мера адекватности. В математическом моделировании адекватность определяется как соответствие теории (формальной модели) практике. Так формально правильная модель может быть неадекватной. Математические направления, используемые для формализации визуализации, можно рассматривать как возможный способ структурирования.

При исследовании данных большого объема и сложной структуры возможностей одной модели недостаточно для принятия решения. Комбинированная (нелинейная) модель – построение из набора моделей, каждая из которых соответствует некоторому фрагменту исходного объема данных, либо отвечает на вопрос анализа лишь частично [8] (ответ как истинностное значение). Комбинированная модель имеет когнитивное значение, превосходящее сумму значений частных моделей.

После постановки проблемы “больших данных” и задач их визуализации по этим вопросам опубликовано большое количество работ. Есть понимание того, что формируется новая предметная область. Сформировалась структура новой дисциплины, хотя и требующая дальнейшей детализации [18].

Сравнивая структуру обработки “больших данных” (рис. 4) и структурную единицу визуального анализа (рис. 1) стоит подчеркнуть наличие дополнительного элемента – моделирования и еще одной обратной связи – рефакторинг модели. Визуальный анализ больших данных не только невозможен без некоторого ментального представления о формальной модели, но и должен уточнять эту модель или создавать новую. Рефакторинг модели (программы, базы данных, визуализации) – это уточнение параметров модели в результате обратной связи и постановки задачи минимума, или верификации и валидации модели [4]. Переопределение понятия рефакторинга связано с необходимостью рассмотрения программ визуализации, не только как замкнутых, но и как открытых детерминированных систем, например, визуальный агент может иметь адаптивное поведение. Частный случай рефакторинга – калибровка параметров модели. Рефакторинг как метод визуальной аналитики сочетает сильные стороны человеческой (обратная связь) и электронной (критерий оптимальности) обработки данных.

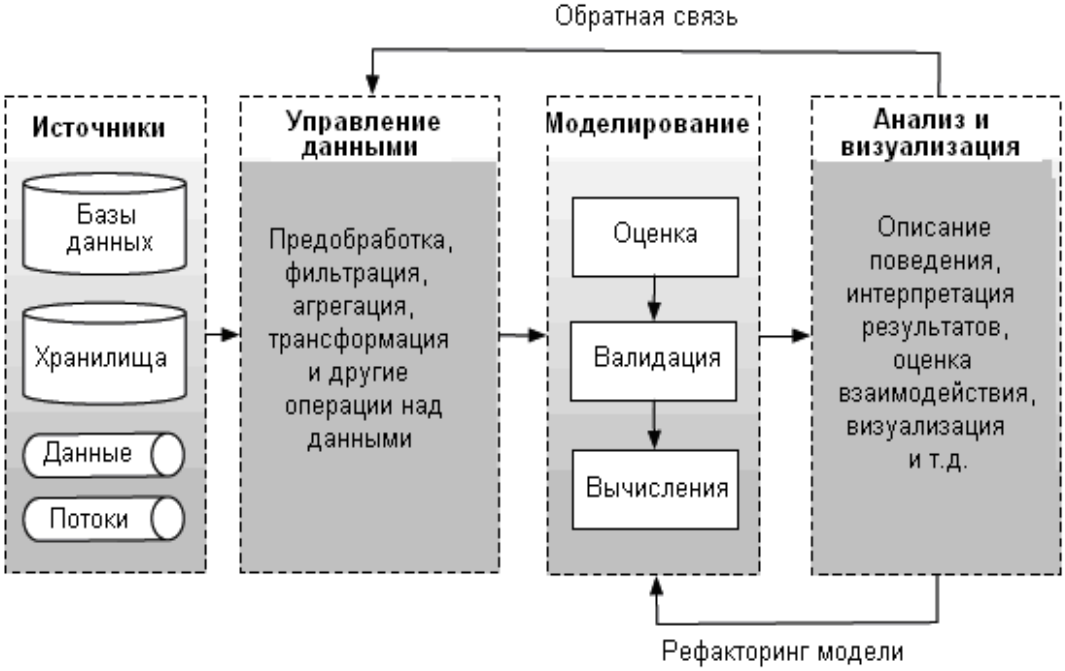

Рис. 4. Структура обработки “больших данных”.

Большие данные - предельный (на данный момент) случай обработки данных, при котором универсальные подходы к анализу и визуализации не работают или неэффективны. Тогда в качестве больших данных могут рассматриваться многомерные и многокатегориальные данные, данные большого объема, данные с неполной информацией (модель с неопределенностью).

Предельный случай формирует вызовы, на которые необходимо ответить, чтобы двигаться дальше. Решение возникающих проблем приводит к тому, что сегодняшние “большие данные” завтра становятся нормой [19]. При анализе и визуализации больших данных рассмотрение предельной неопределенности – неопределенности, которая имеет конечный предел в конкретной метризуемой топологии, является вынужденным.

Модели абстракции данных: фильтрация, кластеризация, выборка должны быть ориентированы на обработку больших данных. В данной работе они верифицируются в рамках экономической модели, которая допускает стандартную оценку эффективности. Используемое в экономике определение эффективности (продуктивности) через отношение продукта к источнику продукта (ресурсам) достаточно адекватно (логическое значение близко к единице). Формально, это скорость или полный дифференциал:

$$V = \frac{df}{dt} \quad df = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} dx_i \qquad (1,2)$$

где f – продукт, x_i – i-й источник продукта.

Также возможно рассмотрение данного отношения и с позиций нечетких множеств (нечеткая экономическая модель), так как экономическая эффективность всегда меньше или равна единице, например, вследствие закона сохранения массы. Поскольку значение эффективности должно быть оптимальным, ищется решение оптимизационной задачи, например, на нахождение минимума функционала:

$$\sum_{i=1}^{n} \frac{f_i}{x_i} \to \min \qquad (3)$$

С целью валидации моделей стандартно используют измерение потерь – L. Очевидно, что мера потерь является двойственной мерой эффективности $|V|=1-|L|$, а задачи $V \to max$ и $L \to min$ эквивалентными. В качестве потерь рассматривают относительную эффективность в области параллельных вычислений, например, закон Амдала, а в области визуализации, например, через отношение расстояний, площадей, объемов.

В работе [20] рассматривается мера разницы гистограмм (агрегированная дисперсия), частным случаем которой является мера «ближайший соседний», используемая для поиска похожих изображений. Там же приведены и другие примеры отношений и соответствующие ссылки. Стандартно относительную эффективность определяют через отношение двух базисных функций, таких как точность и полнота визуализации, что соотносится с определением верификации визуализации. Аналогично мера разницы гистограмм определяется через «уровень абстракции - соотношение между размерами абстрагированного и исходного набора данных (точность визуализации), и качество абстракции данных (полнота визуализации) - степень, в которой абстрагированный набор данных представляет собой исходный набор данных».

Относительную эффективность можно рассматривать и как условную вероятность. Если в критерий оптимальности добавить коэффициенты, то возможна калибровка параметров модели. Например, в случае трилинейного вероятностного пространства: $k_1 p_{1,2} + k_2 p_{1,3} + k_3 p_{2,3} = 1$, где $p_{i,j}$ - условные вероятности. Стратегия визуальной калибровки достаточно простая: $k_1 = 1$, k_2 непрерывно изменяется, например, используя слайдер, k_3 вычисляется, например, через дисперсию. Исследователь, выбирая наилучшее визуальное представление, калибрует модель.

Нечеткие модели достаточно распространены (логическое значение близко к единице), в том числе, и в области компьютерной визуализации [21]. Например, семинар EuroRV3 [16] имеет подзаголовок на 2018 год: «визуализация с неопределенностью». Частный случай нечетких моделей – частота встречаемости слов, информативных признаков.

Если нечеткие множества упрощенно можно рассматривать как теорию вероятности на множестве (топология поточечной сходимости), то теория возможности описывается в компактно-открытой топологии, изначально на вложенных интервалах и является наиболее распространенным случаем модели с неопределенностью. Существует достаточно большое количество математических направлений, свидетельствующих о востребованности моделей с неопределенностью, ориентированных на перенос теории возможности с интервала в топологию больше, чем числовая прямая: грубые множества, информационный разрыв теории принятия решения, теория монотонной меры, в которой используется пространственное расширение интервала – функция девиации (отклонение).

В работе [21] для верификации визуализации предлагается комбинированная модель с неопределенностью, включающую теорию монотонной меры и расширение понятия треугольного нечеткого числа в рамках аффинной арифметики (первой степенью многочлена, заданной через отношение, что соответствует нечеткой экономической модели нашей классификации).

Число диапазонов (интервальное число, range number) – гибридное число, которое явно фиксирует все соответствующие типы неопределенности в одной величине и сопоставляет с каждой величиной степень достоверности. Определим число диапазонов $\check{r}$ как аффинную форму (т.е. первой степенью многочлена):

$$\frac{\tilde{r}}{\eta(\widetilde{r})} = r_0 \oplus (\oplus_{s=1}^{N} \frac{r_s \widetilde{\delta_s}}{\eta(s)}) \qquad (4)$$

где '~' указывает на нечеткое число, $\oplus$ является нечетким оператором сложения (например, для параллельных вычислений важен порядок суммирования), символ деление "-" переопределен, как оператор отношения, связывающий числитель со знаменателем, в духе обозначений нечетких множеств, введенных Заде . Это дает число в виде $\frac{w}{\eta(w)}$, где w – значение, $\eta: W \to [0,1]$ - мера доверия (надежности), связанная с количеством $w \in W$. Из $\eta(w)=1$ следует, что число w является абсолютно надежным, в то время как из $\eta(w)=0$ следует, что оно полностью ненадежно. Это определение в обобщенной форме: значение = приблизительное значение ± отклонение, где r_0 приблизительное значение, отклонение задается суммированием членов, дающих положитель-

ные и отрицательные вклады. Отклонение – это сумма неспецифической, нечеткой и случайной неопределенности. Отдельные члены многочлена, вида $r_s\delta_s$, называются членами отклонения и представляют форму неопределенности из определенного источника s, которых N.

Мы предлагаем [4] другое расширение и аннотацию нечеткого числа – параметрическое (темпоральное) или локально компактное нечеткое число. Продемонстрируем это понятие на примере параллельных вычислений. Сложность алгоритма задается иногда, как дробь, такая как, например, 2/3, где 3 - сложность последовательного алгоритма, а 2 - сложность передачи данных. Вообще говоря, это могут быть не числа, а функции, зависящие от количества процессоров:

$$[2,2]/[3,3](p)=L(p) \quad (5)$$

Темпоральное нечеткое число допускает рассмотрение предельной неопределенности (нечетко-темпоральная модель), например, критерий оптимальности в данном случае будет иметь вид:

$$\lim_{p\to\infty} L(p) \to min \quad (6)$$

С точки зрения математического моделирования формула (6) описывает модель с насыщением, следовательно, возможна интерпретация формулы (5) как “нечеткого” сходящегося ряда *o(2)/o(3)(p)*.

Для диссипативных систем в качестве параметра предпочтительнее рассматривать информационную энтропию, в данном случае, p/N, где N – количество данных. В рамках визуальной аналитики вышеприведённый критерий оптимальности эквивалентен нахождению минимума таких абстрактных метрик, как когнитивное расстояние, информационный разрыв, инсайт. Как уже отмечалось, относительную эффективность определяют через отношение двух базисных функций, таких как точность и полнота визуализации. Если в качестве параметра определить количество данных, то с помощью нечетко-темпоральной модели можно описать визуализацию с уровнем детализации.

С целью выделения общих подходов авторы разрабатывают нечеткую верифицированную классификацию, которая учитывает как частоту встречаемости моделей, структурных единиц, информативных признаков, так и математический уровень абстракции данных. К структурным единицам визуального процесса относятся визуальная парадигма, анализ чувствительности, рефакторинг, калибровка, предельная неопределенность, веб-визуализация. Построение логического пространства обеспечивает автоматическую верификацию. Авторы предлагают рассматривать структурную единицу как **непрерывное** отображение класса подмножеств данных на **логическое пространство**. В частности n-линейное, логическое пространство может формировать n мер оценки визуализации или информативных признаков.

3. Модель классификации

Верифицированная классификация должна обеспечить переход от декларативных определений терминов к формальным определениям, для того чтобы их можно было рассматривать как структурные единицы визуализации.

Одним из параметров эффективности теории принятия решения является уровень доверия. Результаты тестирования показывают, что текстовое представление информации вызывает большее доверие, чем визуальное. Это можно объяснить тем, что в тексте меньше многозначности и неопределенности по сравнению с визуализацией. Уровень доверия можно увеличить, если текстовая и визуальная информация будут согласованы.

Следовательно, целью классификации в рамках визуальной аналитики можно считать построение непрерывного отображения семантических единиц (подмножеств визуальных текстов) из X на пространство терминов (определений) Y. Ориентируясь на интегрированную классификацию, представим модель классификации как топологи-

ческое произведение нечетких множеств, где X, Y – нечеткие множества, на которых заданы меры потерь. Произведение этих мер будет падать и, следовательно, уровень доверия будет увеличиваться.

Данная модель достаточно валидна. Так, в сфере визуализации хорошо зарекомендовало себя применение бинарного вида отображения, который можно рассматривать как произведение бикомпактов или в алгебраическом смысле, когда заданы два окна визуализации (два множества), между которыми определены некоторые операции [4]. Частным случаем бинарного вида отображения является рисунок с пояснением – сторинг («рассказывать историю»).

Модель классификации можно интерпретировать и как нечеткую экономическую модель, например, модель спроса и предложения. Для модели с насыщением известно, что начальные данные (например, определения терминов) необходимо находить так, чтобы они были близки к равновесному состоянию, которое в визуальной форме может быть представлено логистической кривой. В то же время для модели с неопределенностью можно предположить, что равновесное состояние существует, но оно не известно. В данном случае конструктивным подходом является рассмотрение относительной эффективности или скорости сближения двух решений.

Основная задача модели классификации – посчитать уровень доверия (степень достоверности информации) для семантических единиц и соответствующих терминов. Основным критерием является их частота встречаемости. Эта задача достаточно стандартная (алгоритм MapReduce) и распространенная (в некотором корпусе текстов накапливают статистику по информативным признакам, например, валидация, эффективность). С точки зрения визуальной аналитики представляют интерес выводы об изменении условной вероятности, например, вывод о том, что перед этапом массового производства резко возрастает интерес к вопросу об эффективности.

Идея создания онтологии знаний востребована для оценки достоверности информации. Хотя существует достаточно много формальных определений метаонтологий, в модели классификации ставится ударение на применение количественных (нечетких) мер. Можно ввести, например, три вероятностные меры: онтология (происхождение определения в смысле неполной информации), гносеология (смысл), целеполагание (соответствие цели). Считая их независимыми, меры можно перемножить. Если полученное значение, например, больше одной второй, то информацию можно считать достоверной. Данный подход аналогичен методу голосования, но в котором один эксперт голосует за разные меры.

Необходимо отметить, что любое декларативное определение является незамкнутым по сравнению с формальным. При наличие формальной модели уровень доверия можно считать равным единице. Можно ввести бесконечно много мер оценки декларативного определения. С одной стороны, чем больше мер оценки, тем лучше, но практически их количество должно быть ограничено [4].

Мера происхождения определения должна учитывать принадлежность термина к циклу компьютерного моделирования. Например, фильтрация данных – часть графического конвейера: фильтрация, геометрическая обработка, растеризация. А в работе [20] определяется абстракция данных визуального уровня моделирования. Абстракция данных – процесс сокрытия деталей данных, сохраняющий основные характеристики данных.

Смысловая мера должна учитывать применимость определения к параллельным вычислениям, большим данным, непрерывному отображению или визуальной аналитике и может быть основана на сравнении семантических единиц.

Смысловое определение фильтра можно вывести из сравнения фильтрации и хеширования. Фильтр – это любая поэлементная операция над данными, изменяющая их количество [4]. Из этого определения следует, что объединение объектов или слияние баз данных является фильтром. Текстовое сжатие изменяет не количество данных, а их длину, следовательно, не является фильтром. Аналогично, изменение цвета, вращение,

перемещение, увеличения (без изменения уровня детализации) объекта меняют некоторые значения, а не количество данных, то есть эти операции не являются фильтрами. Для обеспечения векторного (конвейерного) параллелизма нужна возможность применения фильтра к любому элементу множества. Параметрический фильтр задает частичный порядок, следовательно, фильтрация данных может рассматриваться как непрерывный процесс, в том числе и в области параллельных вычислений.

Слайсинг – построение срезов. Частный случай фильтрации данных, когда функция от данных равна константе. Кроме сечения плоскостью, это такие стандартные виды отображения, как изолинии и изоповерхности. Условие равенства константе фактически сокращает размерность данных на единицу. Изменение константы приводит к построению фазового пространства. В качестве константы может выступать идентификатор функции или максимальная длина графа [19].

Достаточно часто слайсинг онтологически связывают с компьютерной томографией. Цель, которая вносит определенные нюансы в визуальную парадигму. Так, стандартными задачами являются: восстановление трёхмерного изображения по набору срезов, построение изоповерхностей и их скролинг (например, по плотности ткани), выделение особенностей (задача распознавания). Для аддитивных технологий визуальная парадигма слайсера заключается в обеспечении качества 3d печати – в частности, востребованы задачи оптимизации формы.

Выборка или выбор (самплинг) – простой фильтр, в котором условие выбора данных ограничено проверкой на равенство. На уровне моделирования выборка однозначно связана с математической статистикой.

Кластеризация относится к процессу разбиения набора данных на группы объектов, основанных на сходстве объектов или их близости на некоторую дистанционную меру. Кластеризация является методом агрегирования, поскольку кластер рассматривается как объект более высокого уровня, который представляет все объекты, которые он содержит [20]. Рассматривая эти определения, напрашивается смысловое определение кластеризации, основанное на сравнение с агрегацией. Методы агрегации описывают соотношение часть – целое. Не всегда это соотношение является мерой или мерой расстояния, например, в случае агрегирования (наследования) классов.

Мера целеполагания, определяется критерием оптимальности. Возможно наличие несколько противоречивых целей, так же как и критериев эффективности.

Если рассматривать непрерывную модель классификации, то наиболее значимой мерой является новизна, которую можно ввести через количество изменений в определении термина. Например, аналогично метрике Левенштейна – количество изменений при переходе от одного кортежа к другому, в частности, используемой для отладки эффективности параллельных программ [22].

Еще раз подчеркнем, что цель данной работы – предложить верифицированную классификацию моделей абстракции данных. Байесовский вывод и нейронные сети также можно рассматривать как стандартные модели абстракции данных, поскольку они достаточно однозначно определяются в первом случаем математическим уровнем моделирования, а во втором – алгоритмическим. Остановимся только на таких моделях, как фильтрация, кластеризация и параллельные координаты.

4. Аксиоматика параллельных координат

Термин параллельные координаты (PrC) активно используется в литературе. На диаграмме PrC множество осей размещаются параллельно друг другу, чтобы можно было выявить зависимости между переменными. С точки зрения экономической модели PrC является диаграммой аддитивной эффективности, где эффективность отображается последовательно по каждой переменной и непрерывно. В данном случае *непрерывность представления* является аксиомой или обоснованным допущением: если бы эффективность изображалась с разрывами на осях координат (действительно как сум-

ма), то идентификация функций была бы затруднена. Для построения PrC можно использовать формулу (4), определив нечеткий оператор сложения. В модели классификации аксиома рассматривается как информативный признак семантической единицы PrC (аксиоме ставится в соответствие вид отображения).

Аксиома – непрерывность процесса. Результат – переупорядочение PrC.

PrC можно рассматривать как слайсинг, где в качестве константы выступает идентификатор функции эффективности - r_0 в формуле (4). Изменение константы приводит к построению фазового пространства. Константы задаются кластерным (естественным) или случайным образом. Как правило, эффективность возрастает по всем координатам, в этом случае предпочитают рассматривать мажорируемую эффективность. Процесс визуализации можно рассматривать как направленное действие, имеющее конечный предел (крайне правое значение эффективности). Если для мажорируемых PrC решить обратную задачу, то можно уточнить значение констант, изначально заданных случайным образом, см. рис. 5. Обычно в литературе PrC рассматривают не как слайсинг, и как выборку (что подтверждает необходимость использования нечеткой модели классификации), поскольку PrC эмпирически связаны с математической статистикой, но очевидно, что математическая модель должна быть вариационной, например, вероятностной.

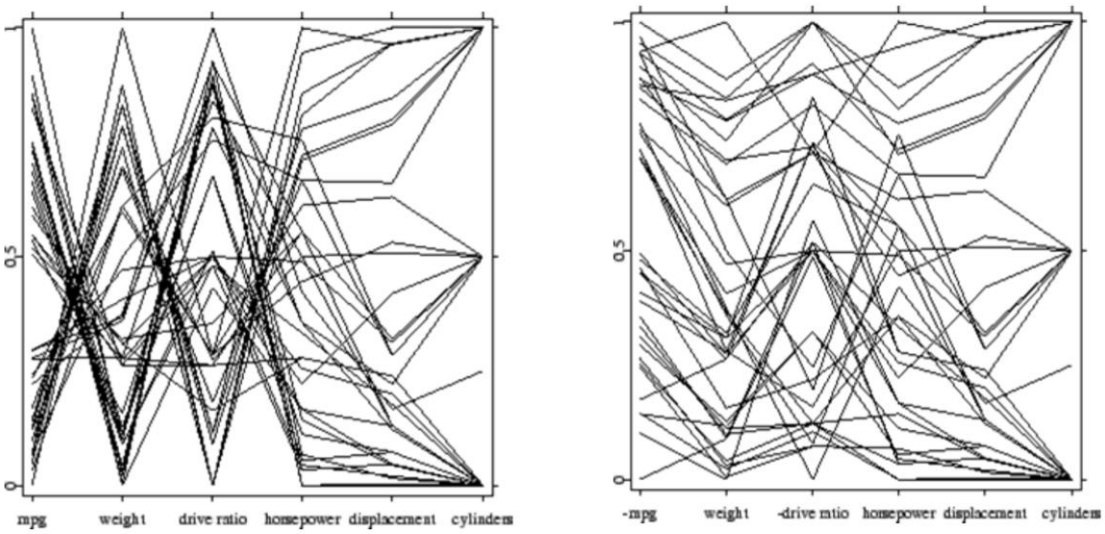

Рис. 5. Оригинальные и переупорядоченные PrC [23].

Гладкие PrC

Обоснование: аксиома гладкости гарантирует, что существует взаимно однозначное отображение на числовую прямую (действие) и как следствие, что вычислительный метод сходится. Например, кривизна линий вычисляется через их взаимное притяжение, рис. 6.

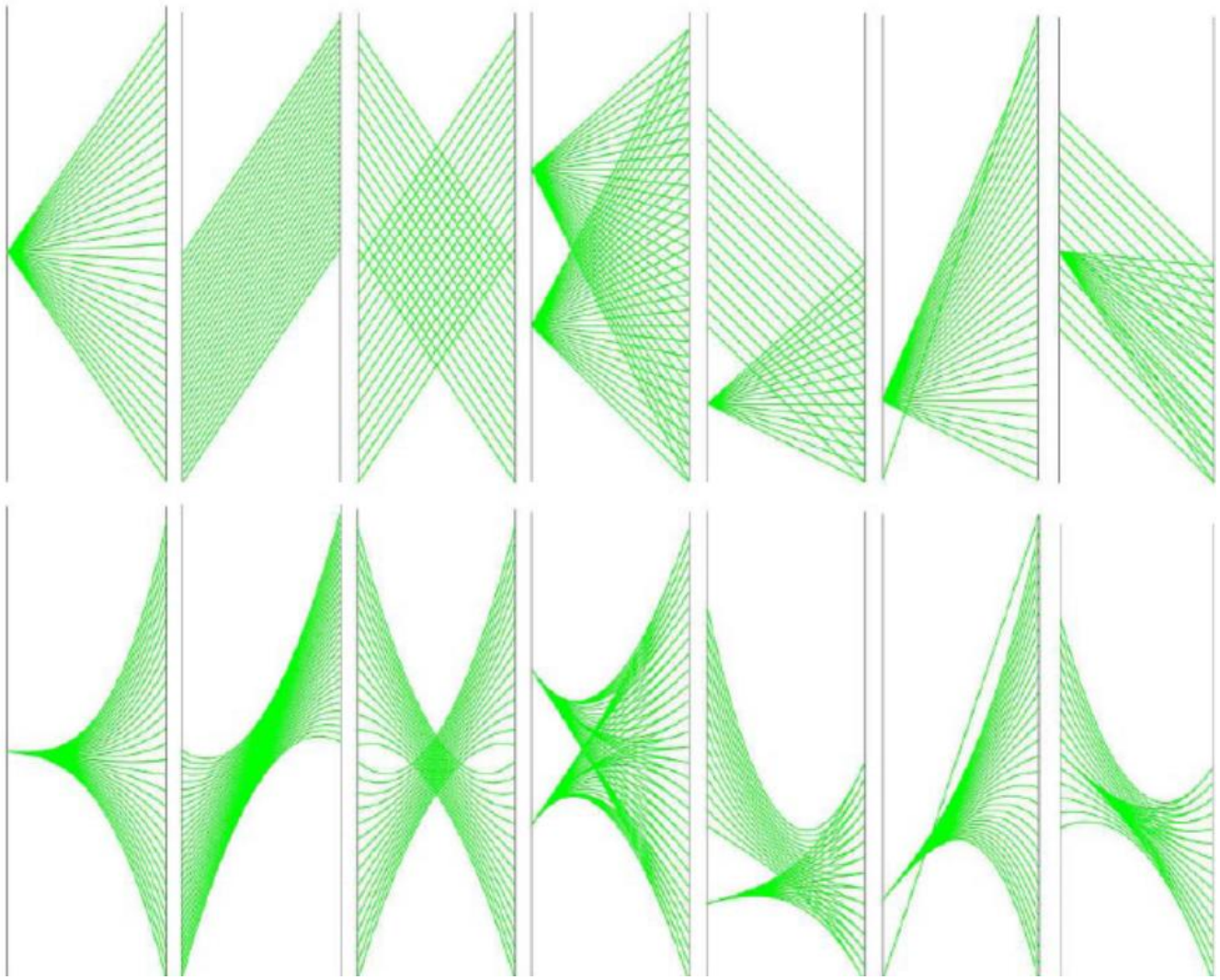

Рис. 6. Оригинальные и гладкие PrC [23].

Оптимальные (дважды дифференцируемые) PrC

На рис. 7 приведен пример экстраполяции сплайнами. Экстремальные точки в комбинированной модели могут быть использованы для задачи кластеризации. В [23] используют термин – scattering (рассеивание точек). Иногда этот термин семантически связывают с визуализацией **хаотических** данных. Интересно то, что диссипация также переводится, как рассеивание.

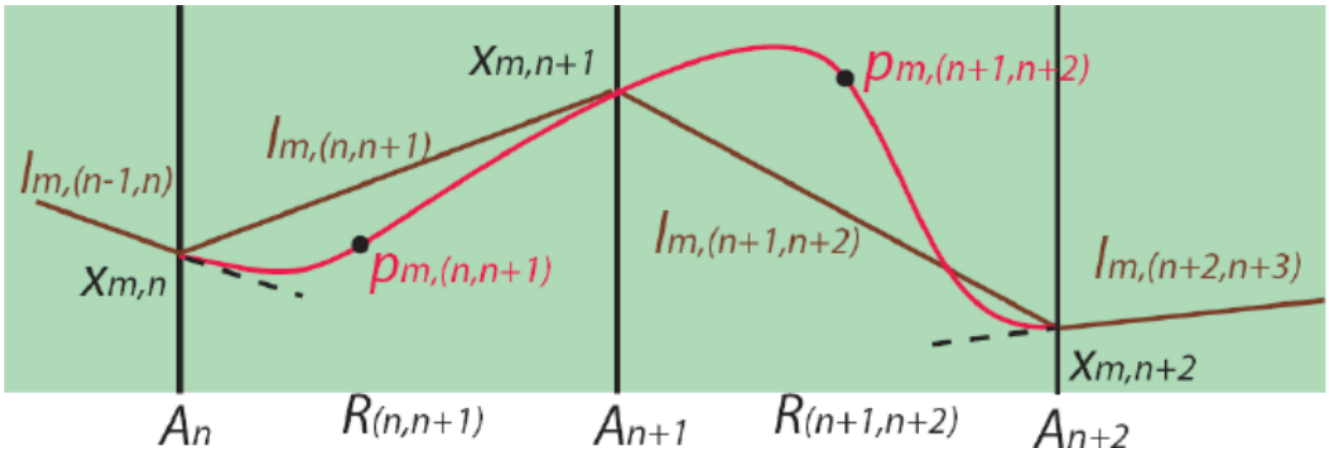

Рис. 7. Оптимальные PrC.

Раскрашенные PrC

Цвет, вероятно, является единственной линейной визуальной переменной из всех возможных, поэтому его наиболее часто используют как дополнительную когнитивную размерность в области визуализации. (Форму объекта формально можно рассматривать как линейный параметр, что с точки зрения зрительного восприятия человека, скорее всего, не верно. Так, в компьютерной графике окружность аппроксимируется многоугольником. Даже с очень плохой точностью многоугольник воспринимается как окружность, семиугольник уже с трудом отличим от восьмиугольника.) В частности, для PrC цвет используется для идентификации функций эффективности. На рис. 8 представлены PrC, раскрашенные естественным (кластерным) способом.

Градиентное раскрашивание является стандартным (всего три варианта, например, линейно по значению). На рис. 9 представлены иерархические PrC [24], раскрашенные на компакте (линейно от среднего значения).

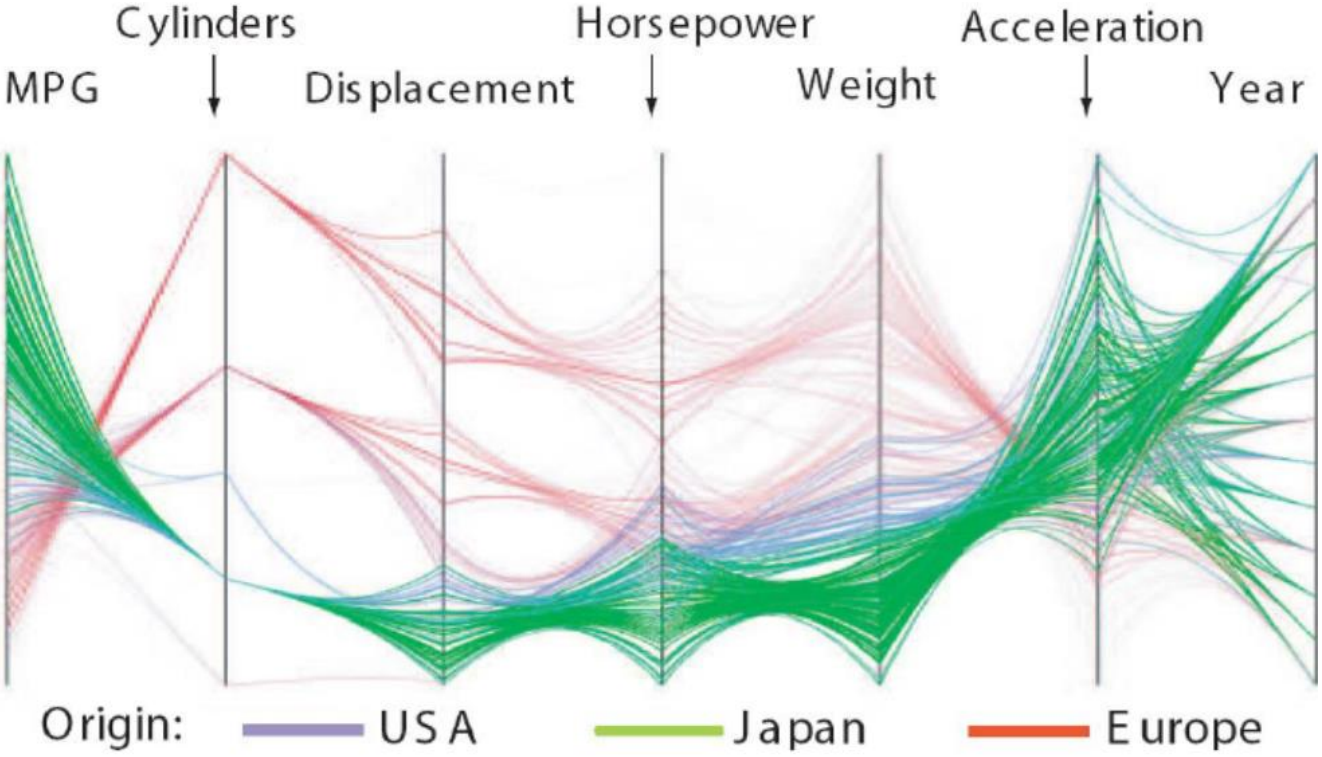

Рис. 8. Естественные PrC [23].

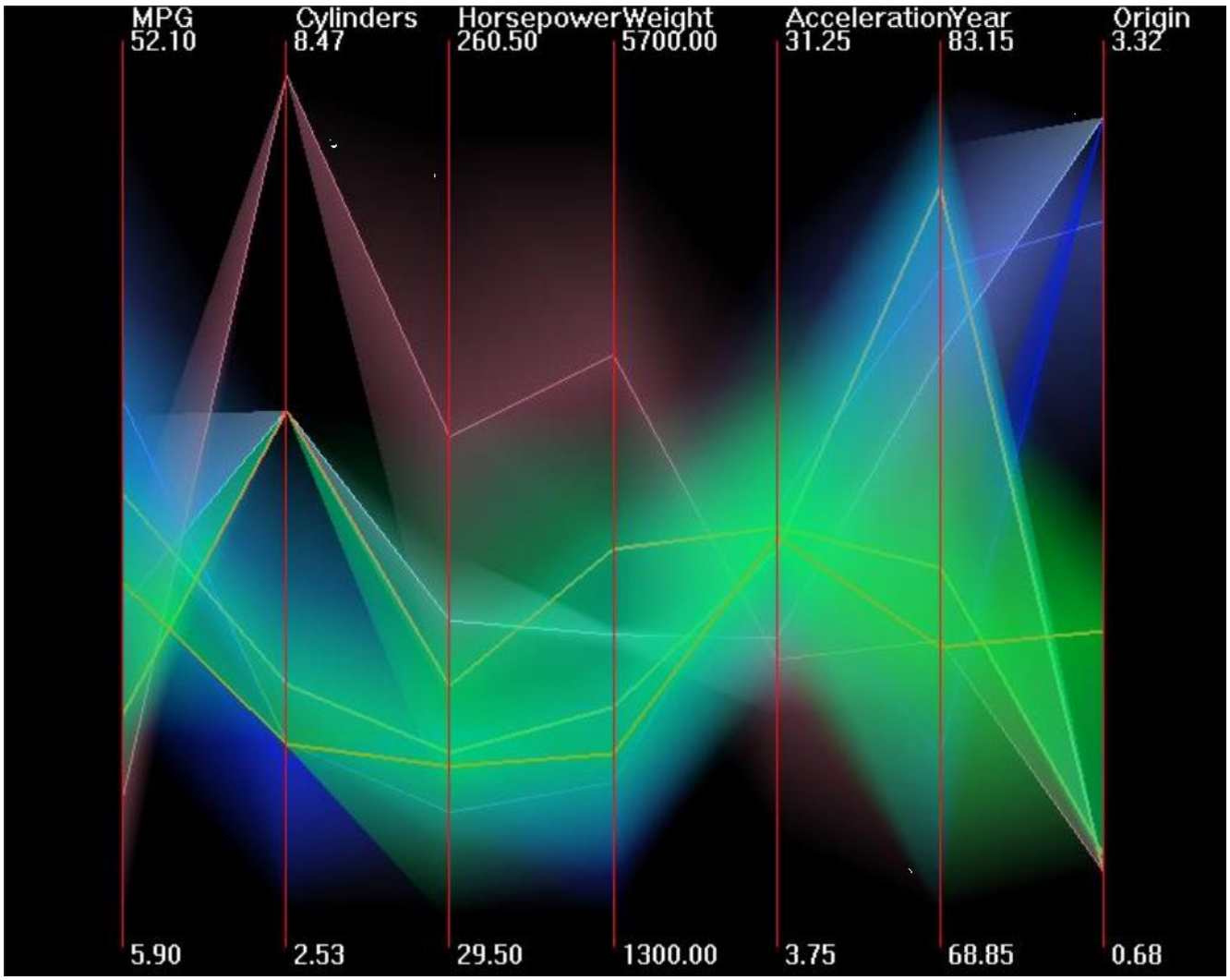

Рис. 9. Иерархические PrC [24].

Взвешенные PrC

В работе [25] не используют термин PrC, но применяют тот же самый вид отображений в рамках теории принятия решений, где переменные (ресурсы) называются критериями, а функции эффективности – альтернативами. Среди априорных методов решения многокритериальных задач выбора самой распространенной является построение обобщенного критерия с использованием весов относительной важности критериев. Рис. 10 демонстрирует визуальный анализ чувствительности (устойчивости, при малом изменении параметра) результата выбора от совокупности значений весов всех критериев. Веса критериев представлены в виде столбиковых диаграмм.

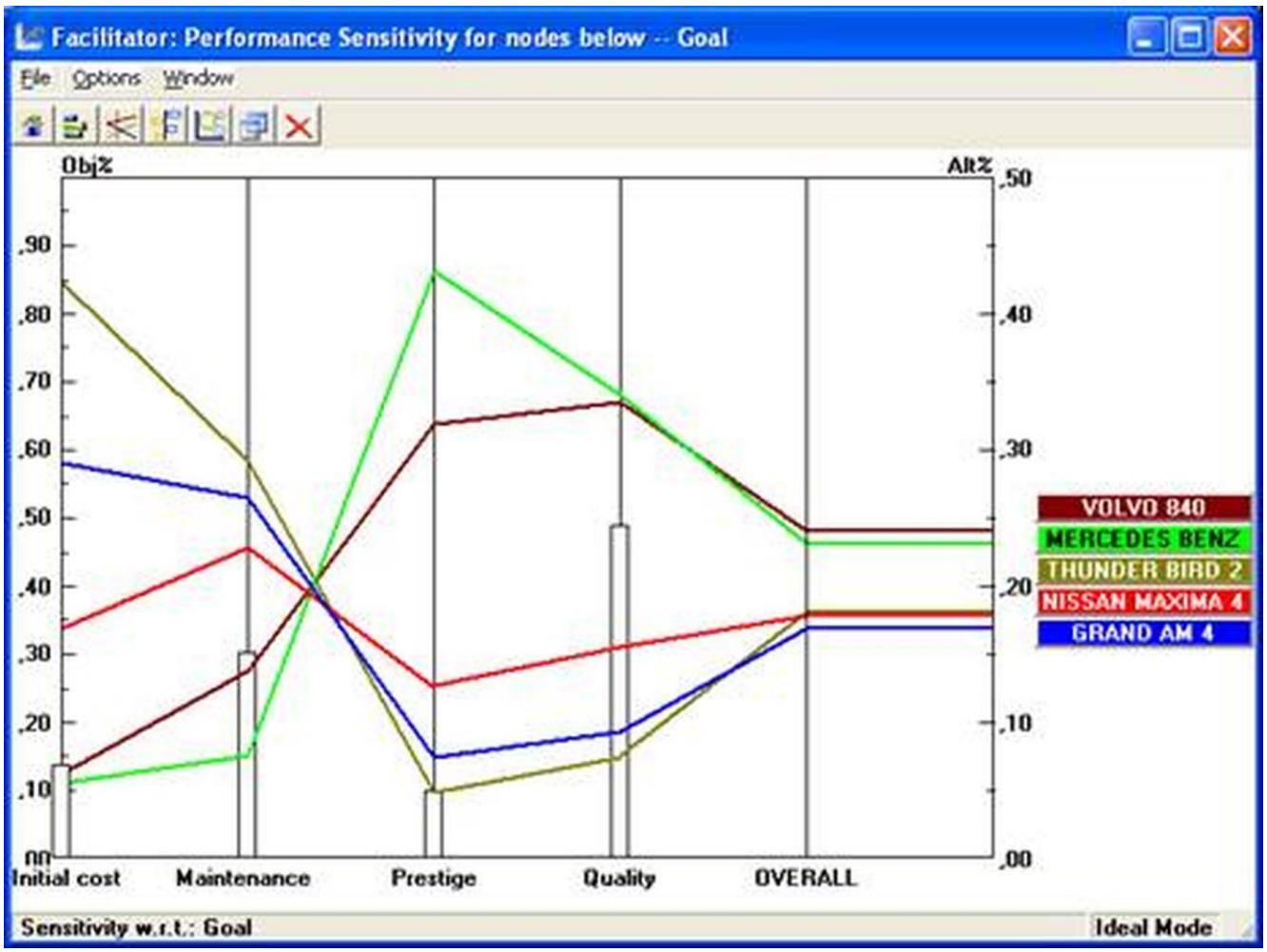

Рис. 10. Взвешенные PrC [25].

В данном примере визуальный анализ чувствительности является NP-полной задачей, и подобные интерфейсы при большом количестве критериев не эффективны. Вариант с пропорциональным изменением весов является не достаточно гибким. С точки зрения визуальной аналитики предпочтительнее вариант с калибровкой параметров (весов) модели. Можно предложить комбинированную модель визуализации, сочетающую PrC и визуализацию с неопределенностью, см. рис. 11.

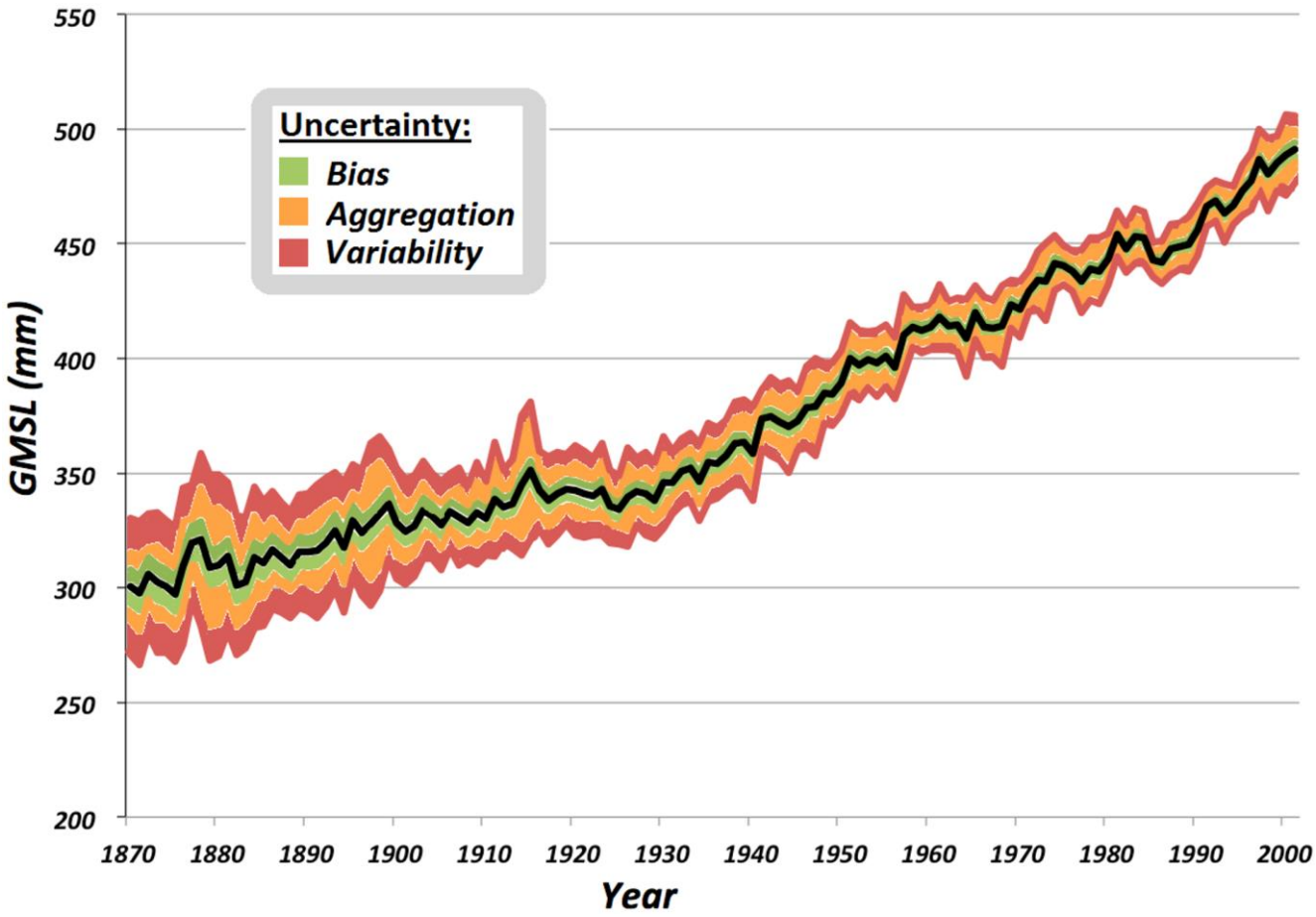

Рис. 11. Визуализации с неопределенностью [21].

Рис. 10 можно определить как компактные (иерархические [24]) PrC, отличающиеся от мажорируемых и монотонно возрастающих PrC только топологией сходимости (PrC на компакте).

Трехмерные диаграммы эффективности

С точки зрения зрительного восприятия 3d визуализация предпочтительнее 2d, но это утверждение справедливо только для видов отображений, имеющих естественную (неабстрактную) образность. В качестве обоснованности применения 3d визуализация или, например, расширенной реальности может выступать следующий критерий оптимальности: $dimX \times Y \to min$, где X – декартово пространство, $dimY$ – когнитивная размерность (размерность логического пространства). Если идентификацию функций рассматривать как дополнительную когнитивную размерность к размерности декартовых координат, то при переходе от 2d визуализация к 3d вместе с масштабируемостью (информативностью) растет и избыточность изображения (размерность топологического произведения пространств). Например, звездный глиф как перенос радарной диаграммы в 3d нужно оценивать через отношение объемов, а не площадей, что с позиции зрительного восприятия, возможно, является негативным фактором.

На рис. 12 [26] схематично представлен нечеткий оператор сложения (4), в котором веса распределены пропорционально, и соответствующие ему 3d компактные (иерархические [20]) PrC. Доверительный интервал (круг) может рассматриваться как дополнительное условие фильтрации данных и может применяться для анализа чувствительности. В данном случае возможно применение квазиестественного вида отображения. Например, если $r_o = r_o(x, y)$, где x,y – географические координаты организации, а 3d диаграмма эффективности отражает информативные признаки деятельности некоторого класса организаций, то трехмерная визуализация позволяет учесть горизонтальные (причинно-следственные) связи между организациями.

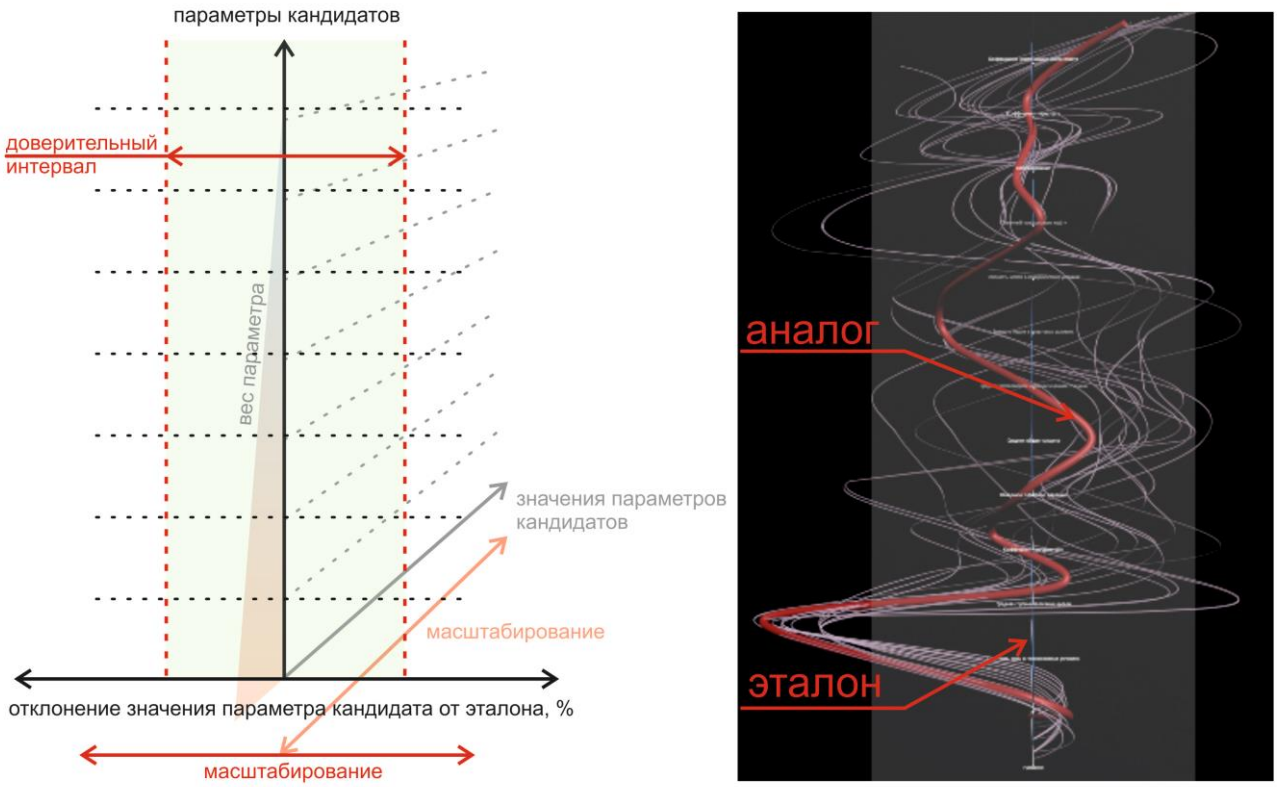

Рис. 12. 3d компактные PrC [26].

Трехмерная диаграмма эффективности (параллельных вычислений) в виде градиента представлена на рис. 13. Эффективность как функция многих переменных соответствует формуле (2) и зависит от количества процессоров и от количества данных. Данный подход позволяет использовать методики представления трехмерных объектов, характерные для научной визуализации.

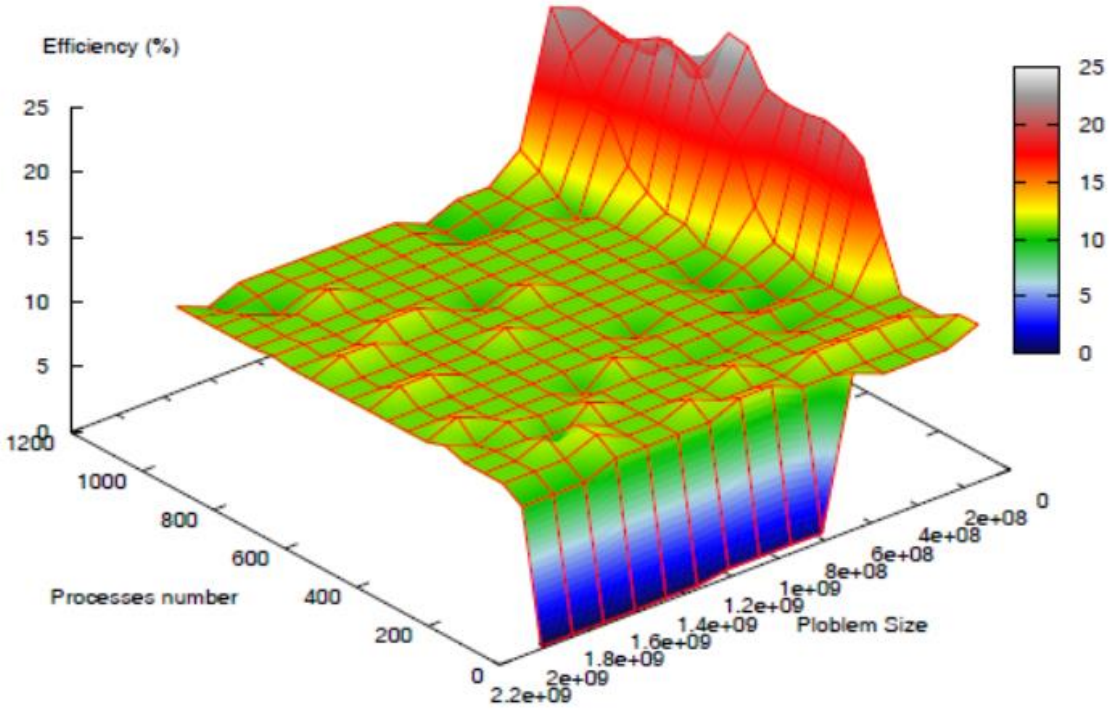

Рис. 13. Эффективность – функция многих переменных, которая зависит от количества процессоров и от количества данных [27].

Комбинированные модели

Как уже отмечалось, эти модели представляют особый интерес для визуальной аналитики. На рис. 14 представлена комбинированная модель в виде конвейера: PrC, кластеризация, граф принятия решения [23].

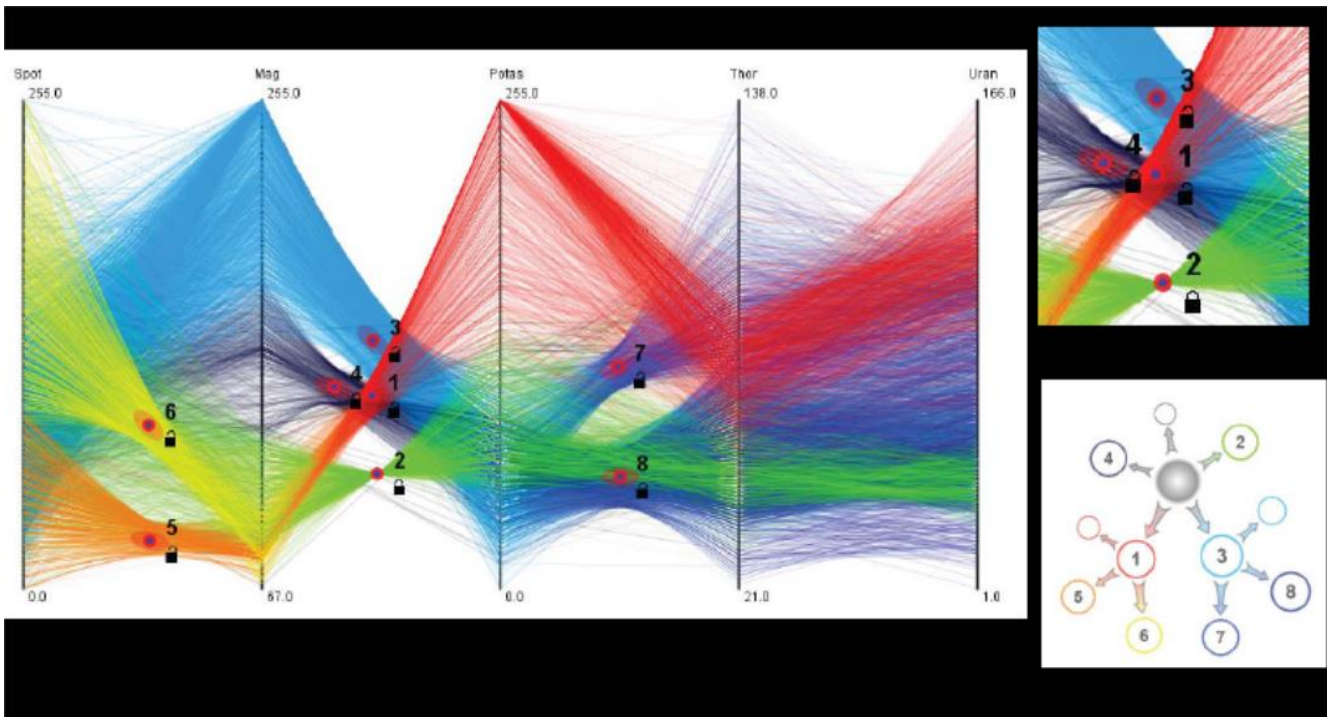

Рис. 14. Конвейер моделей: PrC, кластеризация, граф принятия решения [23].

Матрица “scatterplots” (рассеивания) – алгебраическая метафора множества всех проекций на все (двумерные) подпространства, показанная рис. 15 [20] слева, также может рассматриваться как комбинированная модель. Могут проектироваться не только сырые данные, но и PrC, и их рассеянные точки, например, экстремальные. Если рассмотреть непрерывное отображение симметрической группы на логическое пространство (логическую симметричную матрицу со значениями от нуля до единицы, единица на главной диагонали), то для визуальной аналитики представляет интерес решение задач экстраполяции и распознавания. В работе [24] вместо матрицы “scatterplots” предлагается “Hierarchical Dimensional Stacking” – рекурсивное отображение, см рис. 14 справа; в частности возможно вейвлет-преобразование. В данном случае для визуального анализа ментальной модели недостаточно, слишком много вариантов отображений. Но разнообразные визуальные матричные представления могут быть востребованы для визуализации алгоритмов, в частности матрица “scatterplots”, показанная на рис. 15 слева, может рассматриваться как иллюстрация к методу многомерного шкалирования (метод кластеризации).

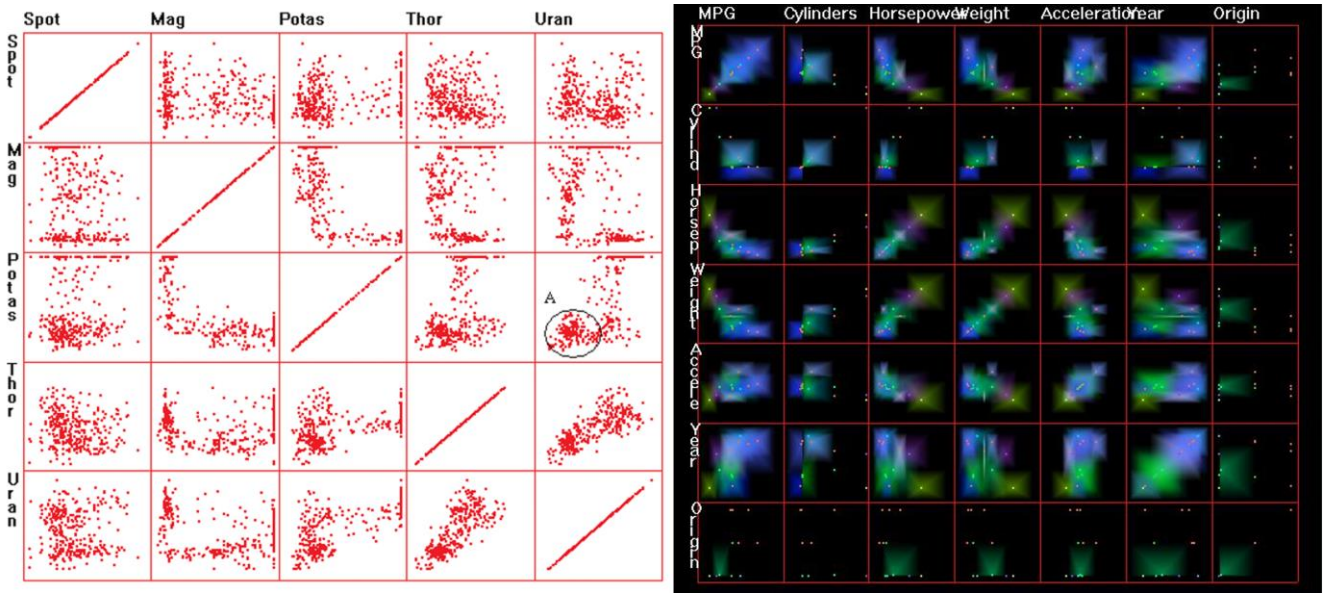

Рис. 15. Матрица “scatterplots” [20] и “Hierarchical Dimensional Stacking” [24].

Параллельные вычисления и *визуализация программного обеспечения*.

При построении диаграмм эффективности параллельных вычислений дифференцирование ведется не по времени, а, например, по количеству процессоров, что для темпоральной (однопараметрической) модели не существенно. Следовательно, вся аксиоматика PrC применима и в области параллельных вычислений. Визуальная парадигма, в данном случае отладка эффективности, вносит определенные особенности в визуальный анализ.

PrC легко переносятся в область визуализации программного обеспечения, где в качестве ресурсов для параллельных вычислений обычно рассматривают время чистого счета процессора или количество попаданий в кэш, а в общем случае – любая информация, которую можно собрать о программе. В случае параллельной программы цветом можно отображать номер процессора. Если отслеживать только минимальное и максимальное значения ресурсов, очевидна формализация и интерпретация в рамках теории возможности. При достаточно мелком разбиении, например, по времени счета или длине программы интервал значений ресурсов должен сойтись в точку, конечно, если выполняется свойство монотонной сходимости, и лучше к наиболее эффективному значению. Это и является целью отладки эффективности параллельных программ или правилом интерпретации. В результате получается вид отображения, известный в литературе, как информационная фреска или фреска Джердинга [28], см. рис. 16.

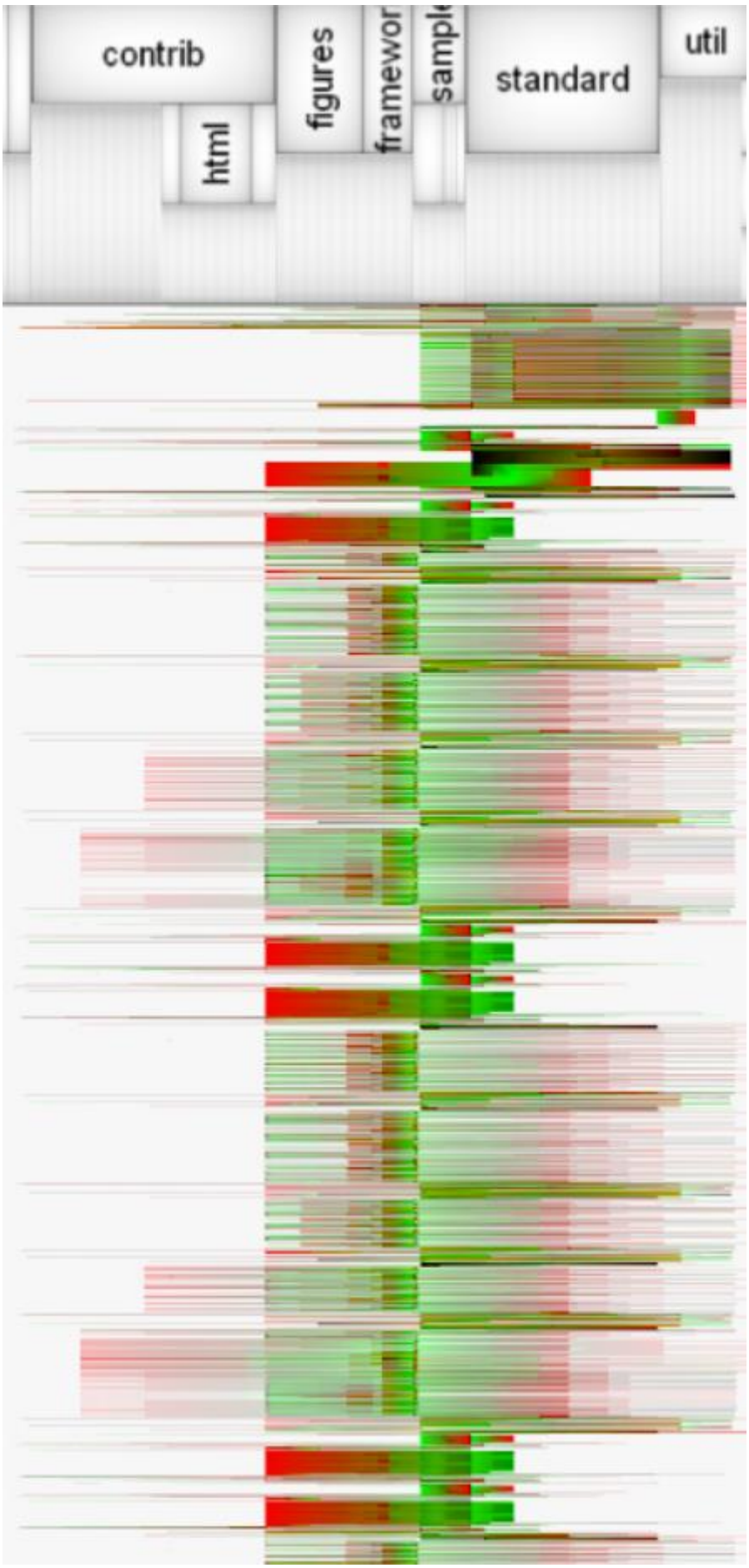

Рис. 16. EXTRAVIS информационная фреска [28].

EXTRAVIS – инструмент для визуализации больших трасс. Предлагаемые методы визуализации должны быть количественно оценены с целью понимания программы. Общие подходы в литературе разделяются на статические и динамические Важное преимущество динамического анализа (отладки) – его точность, поскольку он рассматривает фактическое поведение системы. Среди недостатков его неполнота, так как собранные данные принадлежат сценарию, который был выполнен, и также существуют известные проблемы масштабируемости [28].

В зависимости от уровня сложности PrC можно отнести как к фильтрации данных, так и к слайсингу. Обычно PrC относят к выборке, поскольку эта семантическая единица связана с обработкой статистической информации и информационной визуализацией. В отличие от PrC технология фильтрации данных востребована и в области научной визуализации.

5. Фильтрация данных

Фильтрация данных как модель абстракции данных достаточно распространена, существуют работы по информационной визуализации, связанные с фильтрацией для распределенных баз данных, а понятие фильтр используется в ряде графических библиотек (например, VTK). К сожалению, не так много работ, связанных с большими данными и супервычислениями в этом направлении.

Уже упоминался Vistrail [12] – конструктор модели потока данных, использующий графическую библиотеку и фильтры VTK, ориентированный на параллельный рендеринг.

Стоит отметить диссертацию [29], в одной из частей которой рассматриваются фильтры построения изолиний и изоповерхностей как задача дискретной оптимизации нахождения оптимального подграфа в графе. В принципе, построение изолиний можно свести к задаче коммивояжера и использовать разные приближенные методы дискретной оптимизации. В частности, метод эластичной сети как вариант нейронной сети должен хорошо распараллеливаться. В нейронную сеть можно ввести неопределенность и рассматривать город не как точку, а как окружность. В результате возникает комбинированная модель – слияние метода дискретной оптимизации и линейного программирования. А возможность изменения радиуса окружности позволяет провести визуальный анализ чувствительности.

Применение методов дискретной оптимизации перспективно, но не элементарно для фильтрации данных, в том числе для объемного рендеринга и воксельной графики. Для воксельной графики существенным является ограничение по памяти, следовательно, для больших данных необходимо использовать визуализацию с уровнем детализации. Стандартной реализацией является применение окта-деревьев. Существует достаточно много работ по этому направлению, приведем лишь ссылку на публикацию [30], использующую и определяющую термин «фильтрация данных» и связанную с компьютерной томографией (восстановление объемного изображения по набору проекций). Далее остановимся на наших работах по фильтрации данных: *фильтрация данных как процесс (3d сетка) и фильтрация результатов поиска (контекстное облако тегов).*

Фильтрация данных как процесс (3d сетка)

В рамках концепции параллельной фильтрации данных можно рассматривать фильтры как преобразования, обеспечивающие целостное восприятие или детализацию [4]. Правда, при этом возможна значительная потеря качества. Поэтому целесообразно использовать комплексные (множественные) виды отображения. Очевидно, что одновременный вывод на экран всего слишком большого объема данных влечет за собой снижение уровня детализации.

В качестве упрощенного примера нечетко-темпоральной модели рассмотрим применение фильтра – сечение плоскостью. Дана трехмерная сетка со скалярным значе-

нием в узлах сетки, которое стандартно отображается цветом. В этом смысле, метрика цветовой дифференциации совместима с топологией числовой прямой. Количество узлов сетки большого объема пусть будет N^3. Видеокарта способна обработать только количества данных порядка N^2 (то есть плоскость). Рассмотрим параметрическое нечеткое число как отношение функции точности к функции полноты (эффективность процесса фильтрации данных, использующего фильтр сечение плоскостью):

$$V(N) = \frac{[1,2]}{[2,3]}(N)$$

Базисная функция полнота – темпоральное нечеткое число ***[2,3](N)***. Проще говоря, если вывести *N* сечений плоскостью, то получится полное (равное единице) представление о трехмерной сетке. Точность определим через количество узлов сетки в примитиве визуализации – темпоральное нечеткое число ***[1,2](N)***.

Для данного примера максимальное значение эффективности равное единице достигается, когда точно отображается одна плоскость или когда выводятся N плоскостей, каждая из которых содержит N узлов сетки. Эффективность в данном примере – это эффективность без учета задачи взаимодействия (управления). Выбор оптимального решения остается за пользователем, но в пределах двух интервалов.

Для облегчения задачи управления часто говорят об автоматическом выделении особенностей, например области сгущения сетки. С точки зрения модели с насыщением можно говорить о выборе оптимальных начальных данных. Существование решения задачи управления могла бы обеспечить монотонная последовательность решений, в данном примере можно было бы вывести сначала одну плоскость, затем две и так далее *N*. Но в постановке задачи предполагалось, что у нас данные большого объема, то есть за раз точно можно отобразить только одну плоскость.

Напрашивается решение: выводить по одной плоскости, но достаточно быстро, например, со скоростью 25 кадров в секунду, тем самым обеспечив иллюзию непрерывности. Такой режим управления можно назвать режим “радара”. В результате возникает необходимость в рассмотрении еще одной базисной функции – скорости дискретизации, зависящей от функций полноты и точности визуализации.

Все эти рассуждения говорят о том, что можно определить и верифицировать эффективность. В рамках той же модели можно применить и другие фильтры, а затем сравнить их эффективность. В специализированной системе визуализации данных [33] нами был предложен конвейер фильтров, включающий пространственную фильтрацию и фильтрацию «по значению». Функция пространственной фильтрации (полноты) реализована с помощью метафоры “альфа-сферы”, см. рис. 17.

Для фильтрации по значению (точность) введен дополнительный компонент пользовательского интерфейса - поле диапазонов. Диапазон представляет собой отрезок [***a,b***], заданный на оси соответствующей характеристики, и только из этого интервала происходит отображение и интерполяция данных. Метафора “альфа-сферы” - это тот же диапазон, но в сферических координатах, где функциональное значение полноты - прозрачность объектов, находящихся внутри области, прямо пропорционально удаленности объектов от центра области.

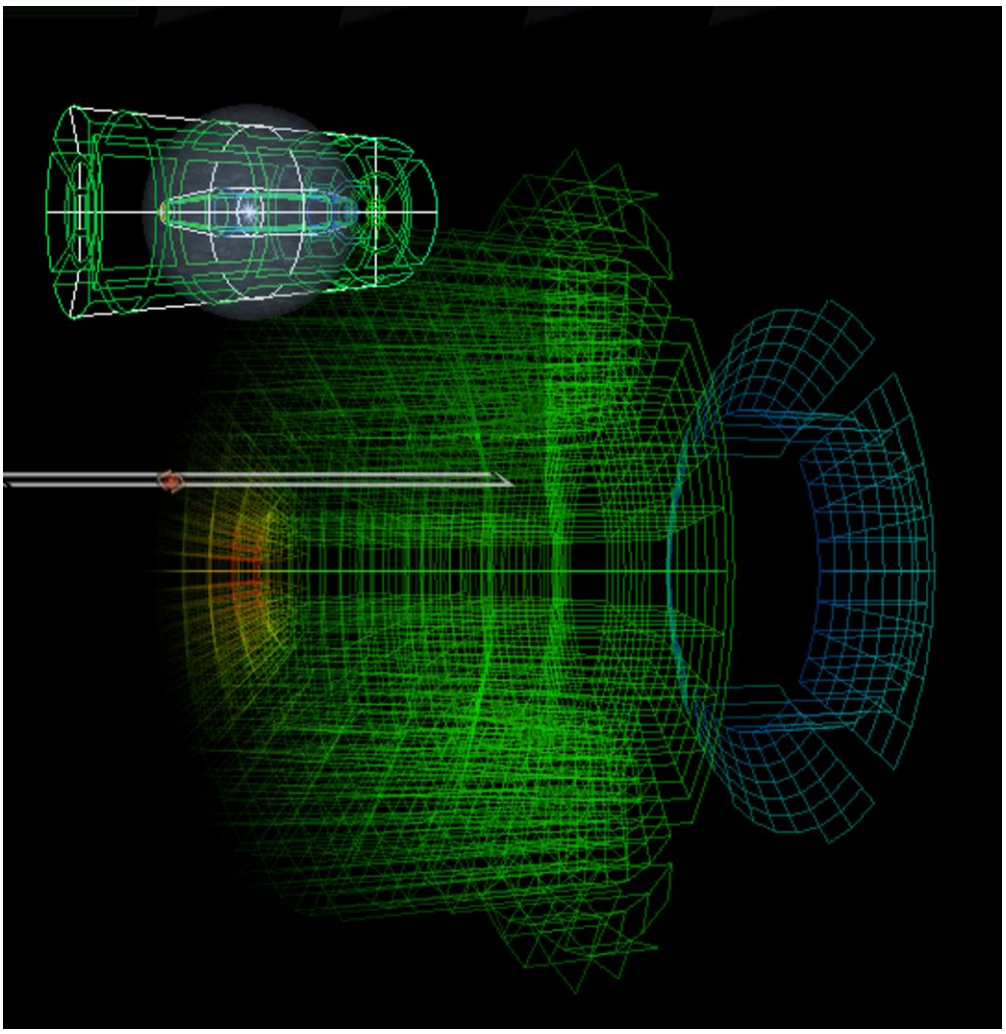

Рис. 17. Пространственная фильтрация с помощью метафоры ”альфа-сферы” [31].

Рис. 17 является бинарным видом отображения, включающим миникарту (слева вверху) и основной вид отображения.

Подчеркнем, что вышеприведенные примеры визуализации верифицированы при помощи нечеткой экономической модели через отношение базисных функций точности и полноты визуализации. Если “альфа-сферу” или фильтр сечения плоскостью рассматривать как эпсилон-покрытия, то возможно применять модели дискретной оптимизации.

Фильтрацию данных как процесс и нахождение оптимального фильтра можно рассматривать и как задачу оптимального управления [4], возможен рефакторинг данной модели с позиции диссипативных систем. Видов отображения, используемых в научной визуализации для визуализации сеток, достаточно много, таким образом, множественных видов отображения еще больше. Сравнение и оценку которых целесообразно построить на основе формальных моделей, чтобы предложить оптимальный вариант.

В [32] описан множественный вид отображения, предлагаемый в системе просмотра медицинских данных. Этот вид отображения включает два ортогональных сечения плоскостью с прокруткой и воксельный вид отображения. Плюсом данной системы является поддержка веб-технологий. Не заостряя внимание на удаленной и он-лайн визуализации, можно указать на то, что наличие данных свойств также можно рассматривать, как меры оценки эффективности и адекватности систем визуализации.

Фильтрация результатов поиска (контекстное облако тегов).

Параллельная фильтрация данных, наряду с параллельным рендерингом, активно применяется для сокращения объема визуализируемых (отображаемых) данных. В этом разделе рассмотрим задачу информационной фильтрации, а именно отображение

результатов поиска. Вероятно, эта задача формально более простая по сравнению с отображением сеточных данных. Результаты поиска в Интернете плохо масштабируются, известным решением данной проблемы является кластеризация. Применяются частные случаи комбинированной модели фильтрация на кластере, например, поиск фильмов [33].

Мы предлагаем более общий подход, основанный на фильтрации данных [34] и обеспечивающий достаточно высокий уровень формализации или верификации. В этом случае для нас важен выбор задачи (информационной фильтрации данных). На этой задаче можно продемонстрировать применение модели с неопределенностью. Она удобна для проверки адекватности модели, если основываться, например, на эвристическом подходе (тестирование). Решение данной задачи требует описание трех основных частей: *архитектурного решения, вида отображения и формальной модели*.

Архитектурное решение соответствует облачным вычислениям. Модуль-посредник перехватывает результаты поиска через Google API и реструктуризирует данные в виде ассоциативных массивов. В результате взаимодействие в клиентской части строится на основе хеширования, то есть практически без пересчета.

Вид отображения выбран бинарный, включающий миникарту и основной вид отображения – вывод непосредственно текста документа, реализация которого очевидно возможна через гиперссылку или в виде всплывающей подсказки. Поэтому основное внимание уделено реализации миникарты, по сути – это расширение облака тегов.

Облако тегов – популярная в Интернете визуальная метафора, генерирующая вид отображения, где упорядоченному списку (например, слов по частоте встречаемости) ставится в соответствие двухмерное изображение в декартовых координатах, в котором можно варьировать позицию объекта (слова), его размер, цвет. Нетрудно видеть, что это определение можно применить и для других случаев (квазитекстовых и квазиестественных), например, посещаемость сайта (статистической информации ставится в соответствие карта мира, в данном случае возможно применение трехмерного изображения или сферических координат). В общем случае, можно рассмотреть нечеткое множество X, где для каждого элемента множества (объекта) задана функция принадлежности $\mu_x(X) \in [0,1]$.

Формальная модель. Результаты поиска (нечеткое множество X) представлено в виде массива ассоциативных массивов, где центральный элемент массива – строка поиска, см. рис. 18. Определяются две функции принадлежности:

1. Ширина контекста или функция отклонения (количество слов слева и справа от строки поиска), реализующая фильтрацию данных в рамках компактно-открытой топологии;
2. Частота встречаемости слова, зависящая от типа документа с расширением pdf, doc, html соответственно отображается градацией цвета R,G,B, реализующая хеширование данных и брашинг (выделение цветом).

В данном случае фильтр можно рассматривать, как реализацию семантического запроса с неопределенностью. Поскольку определены две функции принадлежности, то визуальный анализ чувствительности можно проводить по двум типам параметров. Результаты поиска, размещенные внизу, также включают название документа и гиперссылку, которые идентифицируются и детализируют визуальное представление через взаимодействие с конкретным словом (ключом) и выделением черным цветом данной ветви.

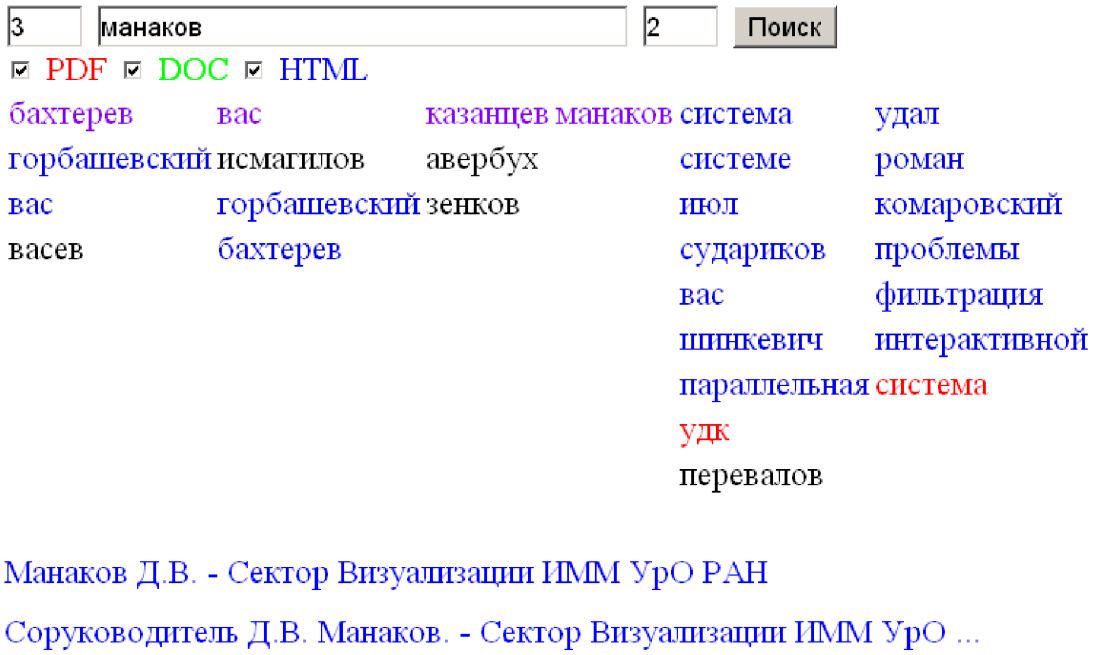

Рис. 18. Облако тегов для информационной фильтрации данных.

Облако тегов может рассматриваться как граф, но двунаправленный (в компактно-открытой топологии) и отображаемый в виде таблицы. Именно представление данных в виде графа обеспечивает семантический зуминг.

Как уже отмечалось, формализация нужна для оценки эффективности, которую определим в соответствие с моделью классификации как произведение двух мер масштабируемости (информативности) и избыточности. Масштабируемость – это отношение количества строк в облаке тегов к количеству в стандартной выдаче на поисковый запрос (нечеткая экономическая модель). Избыточность – это остаток сходящегося ряда $o(n) = 1 - \sum_{i=1}^{n}(1/2)^n$, где n – ширина контекста. В данном случае условной вероятностью можно пренебречь, так как зависимость этих мер достаточно слабая, что подтверждается опытным путем, например, при изменении ширины контекста на единицу. Очевидно, что контекстное облако тегов хорошо масштабируется, но облегчит ли оно интерпретацию результатов поиска? В рамках предлагаемой модели интерпретация должна упроститься, если пользователь, конечно, имеет представление о работе алгоритма MapReduce. Адекватность модели предполагалось проверить на основе тестирования, по единственному критерию – помогает ли контекстное облако тегов поиску. Желающие могут самостоятельно оценить перспективу проекта и качество взаимодействия по ссылке: http://www.cv.imm.uran.ru/oblako. Также в рамках исследования планировалось рассмотреть применение конвейера фильтров, например, для поиска соавторов.

Визуальные представления в виде графа широко распространены, и поэтому обзоры этого направления должны быть более формализованы. Очевидно, что такие визуальные отображения как граф принятия решения, “flow graph” должны рассматриваться не только как представление, но и как процесс.

В этой главе была верифицирована фильтрация данных как процесс для сеточных данных и информации структурированной по частоте встречаемости в рамках нечеткой экономической модели. С достаточно простой логикой: параметризованный фильтр → частичный порядок + монотонность → непрерывное отображение → построение логического пространства (например, определение функции принадлежности) → верификация → рефакторинг модели.

Далее рассмотрим такую абстракцию данных как кластеризация.

6. Кластеризация

Среди многообразия определений кластеризации можно привести в качестве примера работу [20], в которой кластеризация рассматривается как метод агрегирования, и работу [35], в которой кластеризация рассматривается как частный случай диаграммы рассеивания и которая ориентирована на дизайнерское направление или на визуальную семиотику, см. рис. 19.

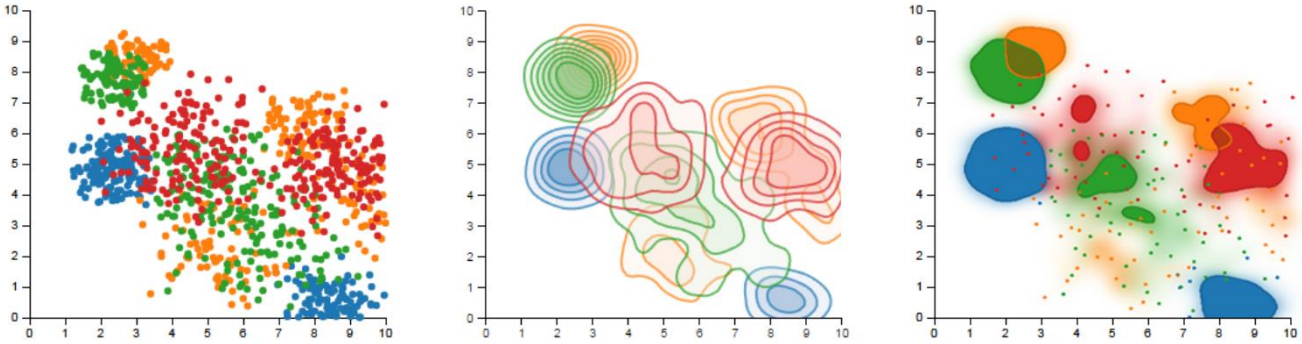

Рис. 19. Различный дизайн диаграммы рассеивания [35].

Мы ориентируемся на достаточно однозначную классификацию кластеризации по математическим методам: регрессионный анализ, метод главных компонент, эластичная карта. Методы кластеризации многократно описаны в литературе, в то же время диаграмма рассеивания не является единственной семантической единицей кластеризации. Уже упоминалась матрица рассеивания; также распространенным случаем является отображение данных в трехмерное пространство, например, с выбором трех главных компонент. Поэтому остановимся только на методе «эластичная (упругая) карта», имеющем определенные особенности отображения данных.

Эластичная карта.

Основой для построения упругой карты является двумерная прямоугольная сетка G, вложенная в многомерное пространство, которая аппроксимирует данные и обладает регулируемыми свойствами упругости по отношению к растяжению и изгибу [36]. Расположение узлов сетки ищется в результате решения оптимизационной задачи на нахождение минимума функционала аналогично формуле (3):

$$D = \frac{D_1}{|X|} + \lambda\frac{D_2}{m} + \mu\frac{D_3}{m} \to min$$

где $|X|$ - число точек в многомерном объеме данных X; m - число узлов сетки, λ, μ - коэффициенты упругости, отвечающие за растяжение и изогнутость сетки соответственно;D_1, D_2, D_3,, - слагаемые, отвечающие за свойства сетки.

D_1 - мера близости расположения узлов сетки к данным:

$$D_1 = \sum_{i,j}\sum_{x\in K_{i,j}} |x - r^{i,j}|^2$$

где $K_{i,j}$ - подмножества точек из X, для которых узел сетки $r^{i,j}$ является ближайшим:

$$x \to r^{i,j}$$

$$|x - r^{i,j}|^2 \to min$$

$$K_{i,j} = \{x \in X, x \to r^{i,j}\}$$

D_2 - мера растянутости сетки:

$$D_2 = \sum_{i,j} |r^{i,j} - r^{i,j+1}|^2 + \sum_{i,j} |r^{i,j} - r^{i+1,j}|^2$$

D_3 - мера изогнутости (кривизны) сетки:

$$D_3 = \sum_{i,j} \left|2r^{i,j} - r^{i,j-1} - r^{i,j+1}\right|^2 + \sum_{i,j} |2r^{i,j} - r^{i-1,j} - r^{i+1,j}|^2$$

Варьирование параметров упругости заключается в построении упругих карт с последовательным уменьшением коэффициентов упругости, в силу чего карта становится более мягкой и гибкой, наиболее оптимальным образом подстраиваясь к точкам исходного многомерного объема данных. После построения упругую карту можно развернуть в плоскость для наблюдения кластерной структуры в изучаемом объеме данных. Упругая карта плохо масштабируется. Для решения этой проблемы можно использовать подход, называемый “квази-зум” (рис. 20), заключающийся в вырезании области сгущения из рассматриваемого облака многомерных данных и построения для вырезанной области упругой карты заново.

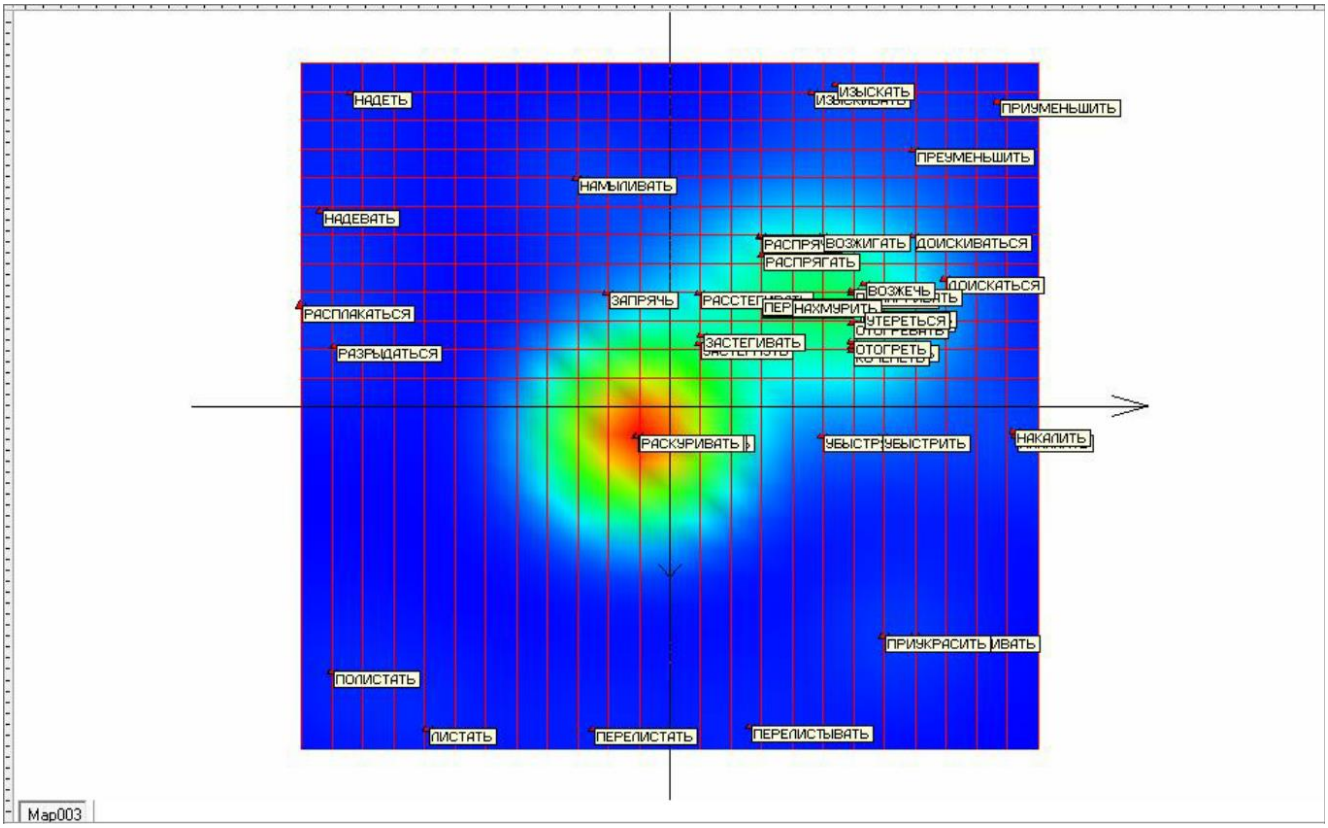

Рис. 20. Результаты применения подхода “квази-зум” для разделения слипшихся точек (развертка упругой карты с раскраской по плотности данных) [36].

Меры качества кластеризации в идейном плане ничем не отличаются от измерения эффективности визуализации, и придумывание им собственных названий зачастую только засоряет учебные курсы. Как уже отмечалось, мера эффективности является двойственной мерой потерь, как и меры гомогенности (h) и полноты (c):

$$h = 1 - \frac{H(C|K)}{H(C)} \qquad c = 1 - \frac{H(K|C)}{H(K)}$$

где H – энтропия, K - результат кластеризации, C - истинное разбиение выборки на классы.

Эластичная карта и проективные методы кластеризации могут рассматриваться как методы трансформации данных. Все же основными информативными признаками кластеризации являются следующие: визуальная парадигма - разбиение данных на классы и наличие меры расстояния.

Как уже отмечалось, модели абстракции данных не имеют естественной образности, поэтому с позиции зрительного восприятия и визуальной аналитики желательно по возможности использовать комбинированные модели, сочетающие модели абстракции данных и квазиестественную образность, например, карту местности (см. рис. 21) или метафору карты (города, ландшафта).

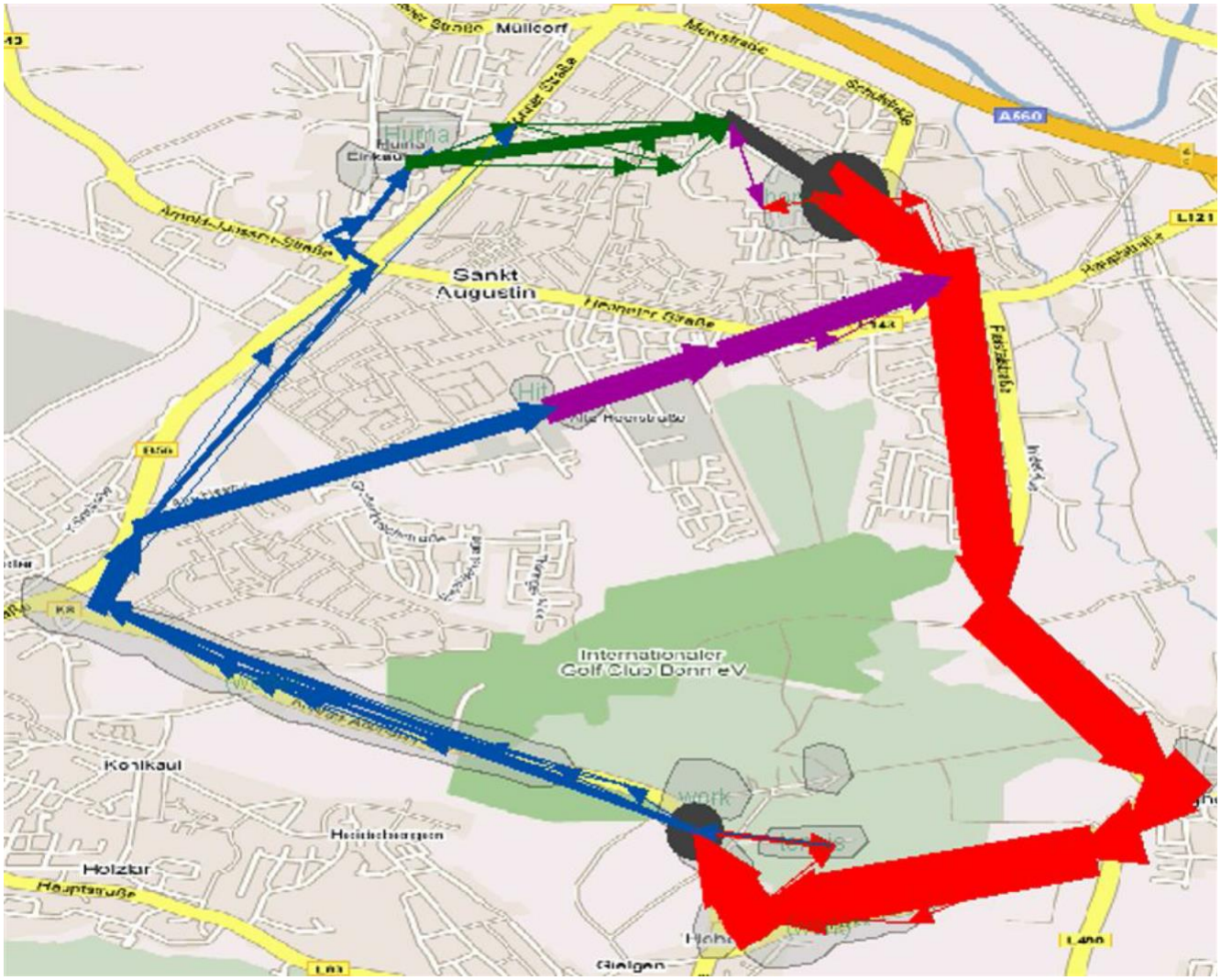

Рис. 21. Результат кластеризации и суммирования данных движения абонентов сотовой связи (маршруты между значимыми местами) [1].

7. Заключение

В работе предложена нечеткая верифицированная классификация моделей абстракции данных (параллельные координаты, фильтрация, кластеризация), учитывающая как частоту встречаемости подходов, так и математический уровень абстракции данных. В основе классификации лежит понятие структурной единицы визуального анализа, определяемое авторами как непрерывное отображение класса подмножеств данных на логическое пространство. Построение логического пространства в каждом отдельном случае обеспечивает автоматическую верификацию как систем визуализации, так и действий пользователя.

Верификация визуализации определяется через отношение двух базисных функций точности и полноты визуализации. На современном этапе развития компьютерной визуализации критерий полноты является более важным. Авторы считают, что обзор по параллельным координатам является наиболее полным, по фильтрации данных пол-

ным только в отношении параллельных вычислений, по кластеризации только в методологическом плане.

Необходимо подчеркнуть, что все модели абстракции данных, изложенные в работе, верифицированы единым образом в рамках нечеткой экономической модели. Математические направления (теория принятия решений, дискретная оптимизация, оптимальное управление), применяемые для верификации визуализации, можно рассматривать как альтернативный способ структурирования обзоров, приведены примеры.

Модели абстракции данных широко распространены и постоянно развиваются, поэтому классификации этих моделей должны не только отражать стандартные подходы, но и предлагать перспективные направления исследования, хотя бы в виде постановок задач. Модели абстракции данных не требуют априорной информации о данных и не зависят от характера данных, происхождения данных [36] и физической модели. В то же время они востребованы в разных математических дисциплинах. Например, методы кластеризации являются составной частью как машинного обучения без учителя, так и математического распознавания образов. Так, в ИММ УрО РАН в1990 годы был разработан пакет Квазар [37], а позднее прототип "Квазар-Офлайн" – распределенный вычислительный комплекс для решения задач распознавания образов.

Он-лайн сервисы параллельных вычислений имеют определенные интеграционные преимущества. Например, они позволяют проводить визуальный анализ чувствительности решения в зависимости не только от параметров прикладной задачи (модели), но и от параметров параллельной программы, и от параметров визуализации. Хотя сектор компьютерной визуализации ориентирован на разработку специализированных систем и системное программирование, ничто не мешает выложить их составные части (сервисы) в общий доступ через интернет, используя конструктор веб-визуализации. Сервисы являются семантическими единицами, формирующими базис, а он-лайн сервисы позволяют рассматривать цикл прохождения задания для параллельных вычислений как непрерывный процесс.

Литература

1. Keim, Daniel A., Gennady Andrienko, Jean-Daniel Fekete, Carsten GÖRG, Jörn Kohlhammer, Guy Melançon, 2008. Visual Analytics : Definition, Process, and Challenges. In: Kerren, A., Stasko, JT, Fekete, J-D, and North, C (Eds). Information Visualization - Human-Centered Issues and Perspectives, Vol. 4950, LNCS State-of-the-Art Survey. Springer: Berlin, 2008. Pp. 154-175
2. Самарский А.А. Математическое моделирование и вычислительный эксперимент // Вестник АН СССР 1979, N 5. Стр. 38--49.
3. Purchase, HC, Andrienko, N, Jankun-Kelly, TJ, Ward, M. Theoretical foundations of information visualization. In: Kerren, A, Stasko, JT, Fekete, J-D, and North, C (Eds). Information Visualization - Human-Centered Issues and Perspectives, Vol. 4950, LNCS State-of-the-Art Survey. Springer: Berlin, 2008. Pp. 49–64.
4. Манаков Д., Авербух В. Верификация визуализации // Научная визуализация 2016. Кв.1. Том 8. N: 1. Стр. 58 - 94.
5. Green T.R.G., Petre M. Usability analysis of visual programming environments: a “cognitive dimensions” framework // J. Visual Languages and Computing, 7, 1996. Pp. 131-174.
6. Тарасов В.Б. Универсальная логика, грануляция информации и искусственный интеллект. http://www.raai.org/news/pii/ppt/2015/tarasov2015.ppt
7. Д. В. Манаков, В. Л. Авербух, П. А. Васёв. Визуальный текст как истинностное подмножество универсального пространства // Научная визуализация 2016. Кв.4. Том 8. N: 4. Стр. 38 - 49.
8. А.А. Захарова, Е.В. Вехтер, А.В. Шкляр. Методика решения задач анализа данных при использовании аналитических визуальных моделей // Научная визуализация 2017. Кв.4. Том 9. N: 4. Стр. 78 - 88. А.А.

9. Манаков Д.В. Модели восприятия визуальной информации // GraphiCon 2017. 27-я Международная конференция по компьютерной графике и машинному зрению. Труды конференции. ПГНИУ. Пермь. 2017. Стр. 129-132.
10. Тарасов В.Б., Калуцкая А.П., Святкина М.Н. Гранулярные, нечеткие и лингвистические онтологии для Обеспечения взаимопонимания между когнитивными агентами // Материалы II Междунар. науч.-техн. конф. «Открытые семантические технологии проектирования интеллектуальных систем» (OSTIS–2012). Минск: БГУИР. 2012. Стр. 267–278.
11. Д.Д. Попов, И.Е. Мильман, В.В. Пилюгин, А.А. Пасько. Решение задачи анализа многомерных динамических данных методом визуализации // Научная визуализация 2016. Кв.1. Том 8. N: 1. Стр. 55 – 57.
12. Louis Bavoil, Steven P. Callahan, Patricia I. Crossno, Iuliana Freire, Carlos E. Scheidegger, Claudio T. Silva, and Huy T. Vo. VisTrails: Enabling interactive multiple-view visualizations. IEEE Visualization, 2005.
13. Scott D. S. Data types as lattices // Proceedings of the International Summer Institute and Logic Colloquium, Kiel, in Lecture Notes in Mathematics. Springer-Verlag. 499. Pp. 579-651.
14. Lakoff G. The contemporary theory of metaphor // Metaphor and Thought. (2nd ed.). Cambridge: Cambridge University Press, 1993, Pp. 202-251.
15. T. Jankun-Kelly, K. Ma, and M. Gertz. A model for the visualization exploration process. In IEEE Visualization, 2002.
16. The EuroRV3: EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization. http://www.eurorvvv.org/
17. Kirby R., Silva C. The need for verifiable visualization // IEEE Computer Graphics and Applications, 28(5) Sep 2008. Pp. 78–83.
18. Assuncao M.D., Calheiros R.N., Bianchi S., Netto M.A.S., Buyya R. Big Data Computing and Clouds: Challenges, Solutions, and Future Directions // arXiv:1312.4722v2, 22 Aug 2014. http://arxiv.org/pdf/1312.4722.pdf
19. Авербух В.Л., Манаков Д.В. Анализ и визуализация “больших данных” // Труды международной научной конференции “Параллельные Вычислительные Технологии” (ПаВТ'2015). Екатеринбург, 31 марта - 2 апреля 2015. Челябинск, Издательский центр ЮУрГУ. 2015. Стр.332-340.
20. Q. Cui, M. Ward, E. Rundensteiner, and J. Yang. Measuring data abstraction quality in multiresolution visualizations. IEEE TVCG, 12(5):709–716, 2006.
21. Fout N., Ma K.-l. Reliable Visualization: Verification of Visualization based on Uncertainty Analysis/ Tech. rep., University of California, Davis, 2012.
22. Choudhury A.N.M.I., Bei Wang, Rosen P., Pascucci, V. Topological analysis and visualization of cyclical behavior in memory reference traces // IEEE Pacific Visualization Symposium, PacificVis 2012, Korea, February 28 - March 2, 2012. IEEE. 2012. Pp. 9-16.
23. 23.Multi-Dimensional Data Visualization (Parallel Coordinates) .http://www.math.pku.edu.cn/teachers/yaoy/math112230/Lecture20_YuanXR_visualization.pdf
24. J. Yang, M.Ward, and E. Rundensteiner. Hierarchical exploration of large multivariate data sets. Data Visualization: The State of the Art 2003, Pp. 201–212.
25. А.П. Нелюбин, Т.П. Галкин, А.А. Галаев, Д.Д. Попов, С.Ю. Мисюрин, В.В. Пилюгин. Использование визуализации при решении задач многокритериального выбора // Научная визуализация 2017. Кв.4. Том 9. N: 5. Стр. 59 - 70.
26. Д.А. Завьялов, А.А. Захарова, А.В. Шкляр. Визуальные методы оценки и модели данных при проектировании разработки месторождений углеводородов // GraphiCon 2017. 27-я Международная конференция по компьютерной графике и машинному зрению. Труды конференции. ПГНИУ. Пермь. 2017. Стр. 112-115.

27. Теплов А.М. Об одном подходе к сравнению масштабируемости параллельных программ // Вычислительные методы и программирование. 2014. Т. 15. Выпуск 4. Стр. 697-711
28. Cornelissen B., Zaidman A., Van Rompaey B., van Deursen A. Trace Visualization for Program Comprehension: A Controlled Experiment // Proc. 17th IEEE Int'l Conf. Program Comprehension, 2009. Pp. 100-109.
29. М.В.Якобовский. Вычислительная среда для моделирования задач механики сплошной среды на высокопроизводительных системах (Автореферат диссертации на соискание ученой степени доктора физико-математических наук), Москва, 2006
30. Zheng Z, Xu W, Mueller K. VDVR: Verifiable visualization of projection-based data. IEEE Transactions On Visualization and Computer Graphics. 16: 2010. Pp. 1515-1524.
31. Горбашевский Д.Ю., Казанцев А.Ю., Манаков Д.В. Параллельная фильтрация в системе визуализации параллельных вычислений // ГрафиКон'2006, 1-5 июля 2006. Россия. Новосибирск, Академгородок. Труды Конференции. Новосибирск. ИВМиМФ СО РАН. 2006. Стр. 333-336.
32. http://slicedrop.com/
33. C. Ahlberg and B. Shneiderman. Visual information seeking using the filmfinder. Proc. ACM SIGCHI Conference on Human Factors in Computing Systems, 2:433, 1994.
34. Манаков Д.В., Судариков Р.О. Облако тегов для информационной фильтрации данных // XIV Международная конференция "Супервычисления и Математическое Моделирование". Тезисы. ФГУП РФЯЦ ВНИИЭФ. Саров. 2012, Стр. 121-123.
35. Alper Sarikaya and Michael Gleicher. "Scatterplots: Tasks, Data, and Designs." IEEE Transactions on Visualization and Computer Graphics 28 (2018).
36. А.Е. Бондарев, В.А. Галактионов, Л.З. Шапиро. Визуальный анализ и обработка многомерных данных // GraphiCon 2017. 27-я Международная конференция по компьютерной графике и машинному зрению. Труды конференции. ПГНИУ. Пермь. 2017. Стр. 103-107.
37. Казанцев В.С. Задачи классификации и их программное обеспечение (пакет КВАЗАР). М.: Наука, 1990. - 136 с.

Scientific Visualization, 2019, volume 11, number 1, pages 139 - 176, DOI: 10.26583/sv.11.1.11

Data abstraction models: sampling (parallel coordinates), filtering, clustering

D.V. Manakov

N.N. Krasovskii Institute of Mathematics and Mechanics of the Ural Branch of the Russian Academy of Sciences, Ekaterinburg, Russia

ORCID: 0000-0001-6852-8096 , manakov@imm.uran.ru

Abstract

When considering computer visualization as an independent discipline, it is necessary to build its mental space with its semantics, pragmatics and basis. Thus any two visualization specialists will be able to speak the same language. This basis is chosen from a sufficiently wide interdisciplinary field of knowledge. Verification of visualization in the spirit of fuzzy sets is defined in terms of the ratio of two basis functions of accuracy and completeness of visualization, it must ensure that the end user is offered a formally correct model of visualization or in other words that the developers of visualization systems have solved the task.

At the present stage of the development of computer visualization, the criterion of completeness is more important. First, it is necessary to form a mental space, and then, by clarifying the semantics, the pragmatics and the basis, replacing the mental space with a logical space, go to the verification of visualization. The construction of monotonically increasing basic functions, for example, the accuracy of visualization: statement of the problem, prototype, application, service, allows to view the classification as a continuous process. Possible statements of problems are considered as challenges and determine not only prospective directions of visualization development, but also their set produces a completeness function.

In the computer visualization sector of IMM UrB RAS, the possibility of developing online parallel computing services is considered. Based on the web-visualization constructor, one can implement stand-alone support for standard data abstraction models, in particular, filtering, clustering, and sampling. The main part of this paper contains an overview of these models. In order to identify common approaches, we develop a fuzzy verified classification that takes into account both the frequency of occurrence of models, structural units, informative features, and the mathematical level of data abstraction.

Since visualization becomes the environment of an automated analytical process, the directions related to self-organization, for example, dissipative systems, are of interest for visual analytics. From these positions, it is possible to clarify the notion of a structural unit of visual analysis, including data abstraction models. The structural units of the visual process include the visual paradigm, sensitivity analysis, refactoring, calibration, limited uncertainty, web-visualization. Building a logical space provides automatic verification. We propose considering the structural unit as a continuous mapping of the class of subsets of data to a logical space.

Keywords: verification of visualization, logical space, dissipative systems, limited uncertainty, filtering, clustering, sampling, parallel coordinates.

References

1. Keim, Daniel A., Gennady Andrienko, Jean-Daniel Fekete, Carsten GÖRG, Jörn Kohlhammer, Guy Melançon, 2008. Visual Analytics : Definition, Process, and Challenges. In: Kerren, A., Stasko, JT, Fekete, J-D, and North, C (Eds). Information Visualization - Human-Centered Issues and Perspectives, Vol. 4950, LNCS State-of-the-Art Survey. Springer: Berlin, 2008. Pp. 154-175
2. Samarskiy A.A. Mathematical modeling and computational experiment // Bulletin of the Academy of Sciences of the USSR 1979, N 5. Pp. 38--49.
3. Purchase, HC, Andrienko, N, Jankun-Kelly, TJ, Ward, M. Theoretical foundations of information visualization. In: Kerren, A, Stasko, JT, Fekete, J-D, and North, C (Eds). Information Visualization - Human-Centered Issues and Perspectives, Vol. 4950, LNCS State-of-the-Art Survey. Springer: Berlin, 2008. Pp. 49–64.
4. D. Manakov, V. Averbukh. Verification of visualization. Scientific Visualization. 2016. Quarter 1. Volume 8. Number 1. Pp. 58-94.
5. Green T.R.G., Petre M. Usability analysis of visual programming environments: a “cognitive dimensions” framework // J. Visual Languages and Computing, 7, 1996. Pp. 131-174.
6. Tarasov V.B. Universal logic, information granularity and artificial intelligence. http://www.raai.org/news/pii/ppt/2015/tarasov2015.ppt
7. D.V. Manakov, V.L. Averbukh, P.A. Vasev. Visual text as truth subset of the universal space. Scientific Visualization. 2016. Quarter 4. Volume 8. Number 4. Pp. 38-49.
8. A.A. Zakharova, E.V. Vekhter, A.V. Shklyar. Methods of solving problems of data analysis using analytical visual models. Scientific Visualization. 2017. Quarter 4. Volume 9. Number 4. Pp. 78-88.
9. D.V. Manakov. The visual information perception models. International Conference Graphicon 2017, Perm, PSU, Russia. Pp. 129-132.
10. Tarasov V. B., Kalutskaya A. P., Svyatkina M. N. Granular, fuzzy and linguistic ontologies to achieve understanding between cognitive agents // Proceedings of the II Intern. Conf. “Open semantic technology of intelligent systems” (OSTIS–2012). Minsk. 2012. Pp. 267-278.
11. D.D. Popov, I.E. Milman, V.V. Pilyugin, A.A. Pasko. Solution to a multidimensional dynamic data analysis problem by the visualization method. Scientific Visualization. 2016. Quarter 1. Volume 8. Number 1. Pp. 55 – 57.
12. Louis Bavoil, Steven P. Callahan, Patricia I. Crossno, Iuliana Freire, Carlos E. Scheidegger, Claudio T. Silva, and Huy T. Vo. VisTrails: Enabling interactive multiple-view visualizations. IEEE Visualization, 2005.
13. Scott D. S. Data types as lattices // Proceedings of the International Summer Institute and Logic Colloquium, Kiel, in Lecture Notes in Mathematics. Springer-Verlag. 499. Pp. 579-651.
14. Lakoff G. The contemporary theory of metaphor // Metaphor and Thought. (2nd ed.). Cambridge: Cambridge University Press, 1993, Pp. 202-251.
15. T. Jankun-Kelly, K. Ma, and M. Gertz. A model for the visualization exploration process. In IEEE Visualization, 2002.
16. The EuroRV3: EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization. http://www.eurorvvv.org/
17. Kirby R., Silva C. The need for verifiable visualization // IEEE Computer Graphics and Applications, 28(5) Sep 2008. Pp. 78–83.
18. Assuncao M.D., Calheiros R.N., Bianchi S., Netto M.A.S., Buyya R. Big Data Computing and Clouds: Challenges, Solutions, and Future Directions // arXiv:1312.4722v2, 22 Aug 2014. http://arxiv.org/pdf/1312.4722.pdf
19. Averbukh V. L., Manakov D. V/ Big Data analysis and visualization. Proceedings of the international scientific conference "Parallel Computing Technologies" (PaVT'2015).

Ekaterinburg, March 31 - April 2, 2015. Chelyabinsk, Publishing Center of SUSU. 2015. Pp.332-340.
20. Q. Cui, M. Ward, E. Rundensteiner, and J. Yang. Measuring data abstraction quality in multiresolution visualizations. IEEE TVCG, 12(5):709–716, 2006.
21. Fout N., Ma K.-l. Reliable Visualization: Verification of Visualization based on Uncertainty Analysis/ Tech. rep., University of California, Davis, 2012.
22. Choudhury A.N.M.I., Bei Wang, Rosen P., Pascucci, V. Topological analysis and visualization of cyclical behavior in memory reference traces // IEEE Pacific Visualization Symposium, PacificVis 2012, Korea, February 28 - March 2, 2012. IEEE. 2012. Pp. 9-16.
23. Multi-Dimensional Data Visualization (Parallel Coordinates). http://www.math.pku.edu.cn/teachers/yaoy/math112230/Lecture20_YuanXR_visualization.pdf
24. J. Yang, M.Ward, and E. Rundensteiner. Hierarchical exploration of large multivariate data sets. Data Visualization: The State of the Art 2003, Pp. 201–212.
25. A.P. Nelyubin, T.P. Galkin, A.A. Galaev, D.D. Popov, S.Yu Misyurin, V.V. Pilyugin. Usage of visualization in the solution of multicriteria choice problems. Scientific Visualization . 2017. Quarter 4. Volume 9. Number 5. Pp. 59-70.
26. D.A. Zavyalov, A.A. Zakharova, A.V. Shklyar .Visual assessment methods and data models in the development of hydrocarbon fields. International Conference Graphicon 2017, Perm, PSU, Russia. Pp. 112-115.
27. Teplov A. An approach to comparing the scalability of parallel programs // Numerical Methods and Programming. 2014. V. 15. Issue 4. Pp. 697-711.
28. Cornelissen B., Zaidman A., Van Rompaey B., van Deursen A. Trace Visualization for Program Comprehension: A Controlled Experiment // Proc. 17th IEEE Int'l Conf. Program Comprehension, 2009. Pp. 100-109.
29. M.V. Yakobovsky. Computational environment for modeling problems of continuum mechanics on high-performance systems (Abstract of the thesis for the degree of Doctor of Physical and Mathematical Sciences), Moscow, 2006
30. Zheng Z, Xu W, Mueller K. VDVR: Verifiable visualization of projection-based data. IEEE Transactions On Visualization and Computer Graphics. 16: 2010. Pp. 1515-1524.
31. Gorbashevskiy D., Kazantsev A. Manakov D. Parallel Filtering in Parallel Computing Visualization System // International Conference Graphicon 2006, Novosibirsk, Akademgorodok, Russia. Pp. 333-336.
32. http://slicedrop.com/
33. C. Ahlberg and B. Shneiderman. Visual information seeking using the filmfinder. Proc. ACM SIGCHI Conference on Human Factors in Computing Systems, 2:433, 1994.
34. Manakov D., Sudarikov R. Tag cloud for the information data filtration. XIV International Seminar "Super-Computations and Computer Simulations". Sarov, 2012. Pp. 121-123.
35. Alper Sarikaya and Michael Gleicher. "Scatterplots: Tasks, Data, and Designs." IEEE Transactions on Visualization and Computer Graphics 28 (2018).
36. A.E. Bondarev, V.A. Galaktionov, L.Z. Shapiro Visual analysis and processing of multidimensional datasets. International Conference Graphicon 2017, Perm, PSU, Russia. Pp. 103-107.
37. Kazantsev V.S. Classification tasks and their software (QUASAR package). M .: Science, 1990.