ёМинистерство образования Республики Беларусь

Учреждение образования

Брестский государственный технический университет

Кафедра ИИТ

Лабораторная работа №6

За 7 семестр

По дисциплине «Компьютерное моделирование и анализ данных»

Выполнил:

студент 4 курса

Группы ПО-4(1)

Ковальчук В.В.

Проверил:

Чичурин А.В.

Брест 2022

**Лабораторная работа №6**

**Визуальный анализ данных.**

Визуальный анализ данных – это процесс, в котором исследование данных происходит с помощью наглядного отображения информации. Визуальный анализ необходим для поддержки аналитической работы с данными. Цель визуального анализа – ответ на важные вопросы с использованием данных и фактов.

Визуальный анализ данных позволяет легко работать с неоднородными и зашумленными данными, в то время как не все автоматические методы могут работать с такими данными и давать удовлетворительные результаты. Визуальный анализ данных интуитивно понятен и не требует сложных математических или статистических алгоритмов.

Основной идеей визуального анализа данных является представление данных в некоторой визуальной форме, позволяющей человеку погрузиться в данные, работать с их визуальным представлением, понять их суть, сделать выводы и напрямую взаимодействовать с данными.

Визуальный анализ данных особенно полезен, когда о самих данных мало известно и цели исследования до конца непонятны. За счет того, что пользователь напрямую работает с данными, представленными в виде визуальных образов, которые он может рассматривать с разных сторон под любыми углами зрения, он может получить дополнительную информацию, которая поможет ему более четко сформулировать цели исследования.

Он обычно выполняется в три этапа:

* Беглый анализ – позволяет идентифицировать интересные шаблоны и сфокусироваться на одном или нескольких из них
* Увеличение и фильтрация – идентифицированные шаблоны отфильтровываются и рассматриваются в большем масштабе
* Детализация (по необходимости) – если пользователю нужно получить дополнительную информацию, он может визуализировать более детальные данные

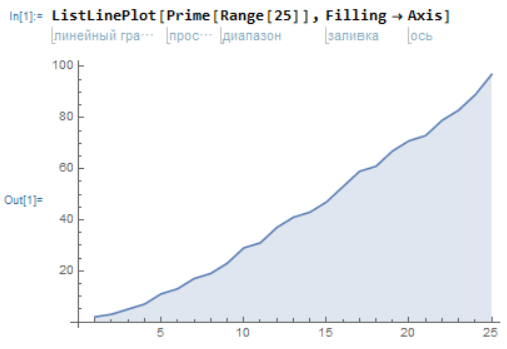
Не все данные могут быть описаны в терминах измерений. К таким данным относятся тексты, гипертексты и т. п. Они характеризуются тем, что не могут быть описаны количественными характеристиками, и, как следствие, не все методы визуализации могут быть для них использованы. Поэтому перед визуализацией требуются дополнительные преобразования, подготавливающие тексты к виду, пригодному для использования методов визуализации.

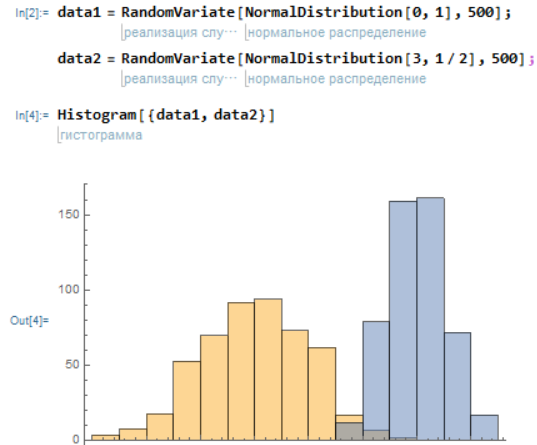
Данные могут состоять в некоторых отношениях с другими частями информации. Для представления таких взаимосвязей широко используются графы. Граф состоит из набора узлов и соединяющих их линий, называемых дугами. Примерами таких данных могут являться: электронная почта, пересылаемая между людьми, гиперссылки между Web-документами, файловая структура диска и т. п. Для отображения таких связей используются специальные методы визуализации.

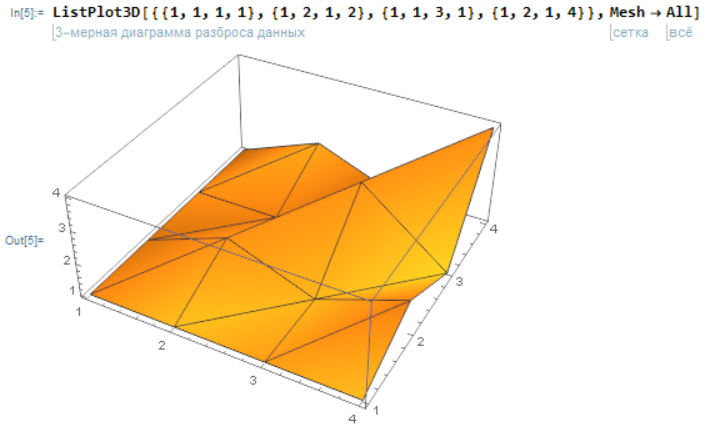
* Одномерные данные – одномерные массивы, временные ряды
* Двумерные данные – точки двумерных графиков, географические координаты
* Многомерные данные – финансовые показатели, результаты экспериментов
* Тексты и гипертексты – газетные статьи, Web-документы
* Иерархические и связанные – структура подчиненности в организации, электронная переписка людей, гиперссылки документов
* Алгоритмы и программы – информационные потоки, отладочные операции

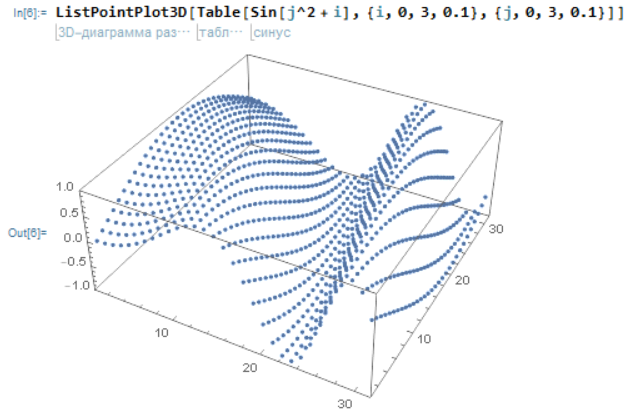
Очевидно, что количество визуальных образов, которыми могут представляться данные, ограничиваются только человеческой фантазией. Основное требование к ним — это наглядность и удобство анализа данных, которые они представляют. Методы визуализации могут быть как самые простые (линейные графики, диаграммы, гистограммы и т. п.), так и более сложные, основанные на сложном математическом аппарате. Кроме того, при визуализации могут использоваться комбинации различных методов. Выделяют следующие типы методов визуализации:

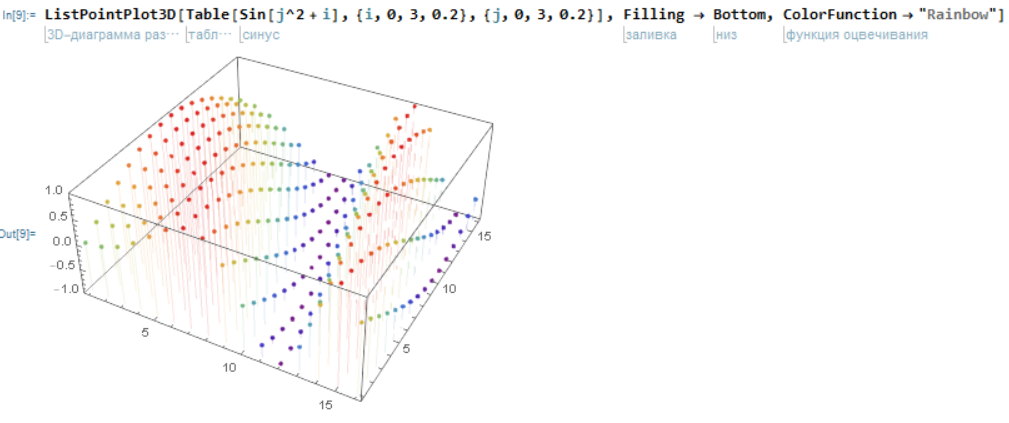
* Стандартные 2D/3D-образы – гистограммы, линейные графики











* Геометрические преобразования – диаграмма разброса данных, параллельные координаты

Методы геометрических преобразований визуальных образов направлены на трансформацию многомерных наборов данных с целью отображения их в декартовом и в недекартовом геометрических пространствах. Данный класс методов включает в себя математический аппарат статистики. К нему относятся такие популярные методы, как диаграмма разброса данных, параллельные координаты, гипердоли.

На диаграмме с параллельными координатами каждой переменной присваивается собственная ось, и все оси располагаются параллельно друг другу.

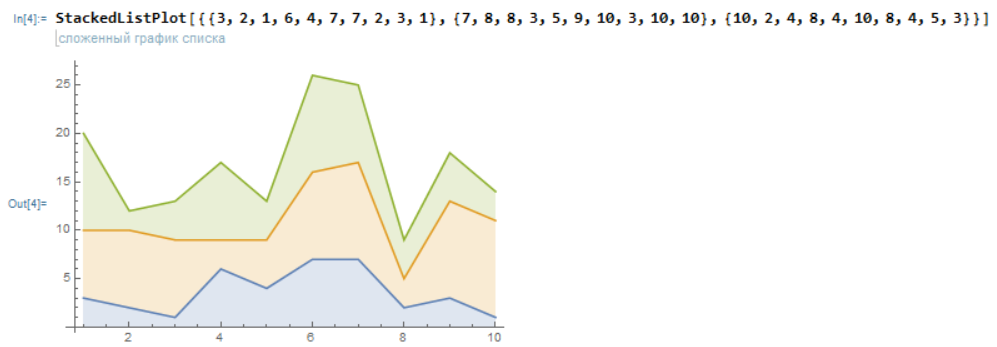
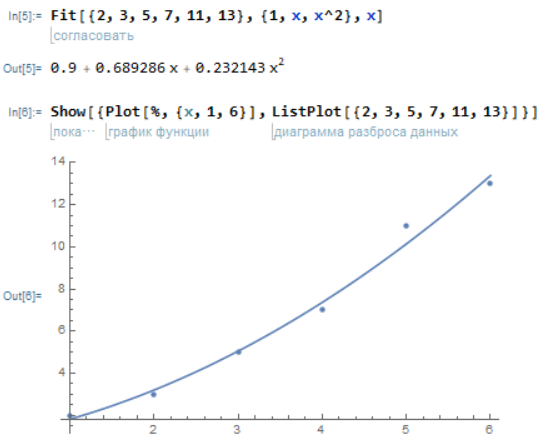
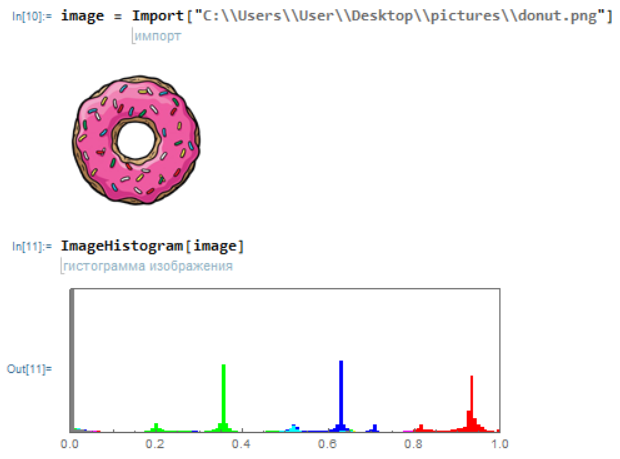


диаграмма разброса данных



* Методы, ориентированные на пикселы – рекурсивные шаблоны, циклические сегменты

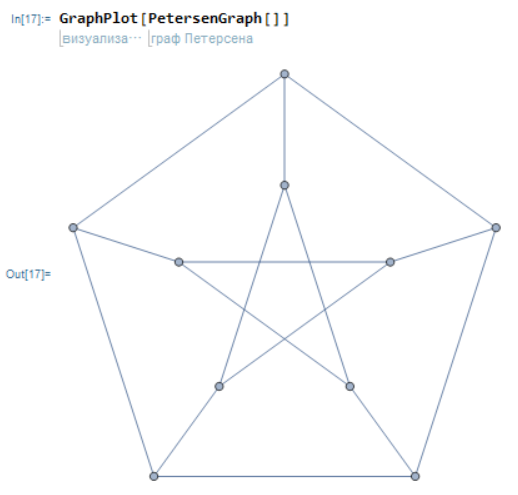
Основной идеей методов, ориентированных на пикселы, является отображение каждого измерения значения в цветной пиксел и их группировка по принадлежности к измерению. Так как один пиксел используется для отображения одного значения, то данный метод позволяет визуализировать большое количество данных.

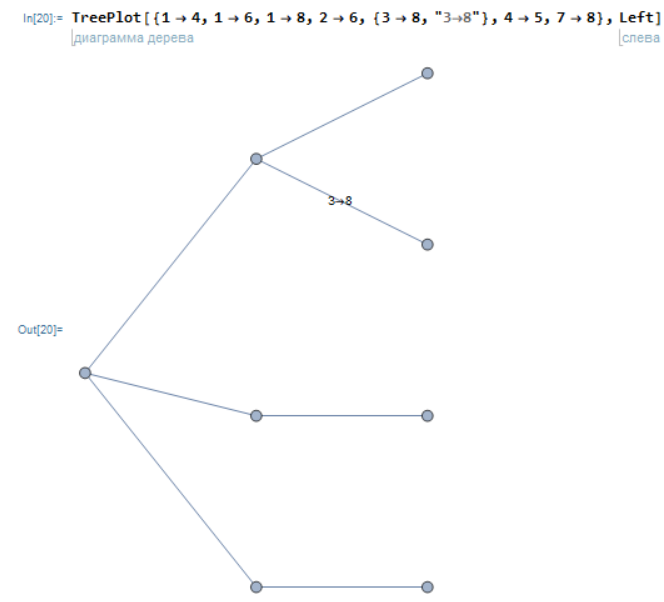




* Иерархические образы – древовидные карты и наложение изменений

Методы иерархических образов предназначены для представления данных, имеющих иерархическую структуру. В случае многомерных данных должные быть правильно выбраны измерения, которые используются для построения иерархии.





**Извлечение знаний из Web**

Всемирная паутина WWW в настоящее время является наиболее богатым источником информации и знаний. Она содержит огромное количество документов, данных, аудио- и видеофайлов. Однако пользователи Интернета сталкиваются с большими проблемами не только при анализе, но и при поиске нужной им информации. Выделяют следующие проблемы работы с информацией из Web:

* **Поиск значимой информации**

Пользователи в поиске информации могут самостоятельно перемешаться от сайта к сайту или пользоваться популярными в настоящее время поисковыми системами. Последние по введенным ключевым словам предоставляют списки ссылок на страницы, на которых представлена информация, соответствующая введенным ключевым словам. Однако использование поисковых систем имеет проблемы. небольшой процент действительно нужной информации среди множества ссылок, которые предоставляют поисковые системы. низкая повторяемость вызовов, связанная с невозможностью индексировать все Web-ресурсы. В результате возникают трудности поиска неиндексированной информации, которая может быть необходима для пользователя.

* **Создание новых знаний вне информации, доступной на Web**

Эта проблема может рассматриваться как часть проблемы поиска значимой информации. Она возникает уже после выполнения поиска информации и связана с извлечением полезных знаний из того множества информации, которое было найдено поисковой системой по запросу пользователя.

* **Персонализация информации.**

Данная проблема связана с типом и представлением информации в зависимости от смысла, вкладываемого пользователем.

* **Изучение потребителя или индивидуального пользователя**

Эта проблема связана с предоставлением пользователю именно той информации, которую он хочет получить. Для этого требуется настройка и персонализация поисковой системы для конкретного потребителя или пользователя.

**Этапы Web Mining**

Для решения перечисленных проблем используются различные технологии, напрямую или косвенно разрешающие их. К таким технологиям относятся: базы данных, информационный поиск, обработчики естественных языков и др. Технология Web Mining также направлена как на прямое, так и на косвенное решение перечисленных проблем.

Выделяют следующие этапы применения Web Mining:

* Поиск ресурсов — локализация неизвестных документов и сервисов в Web.
* Извлечение информации — автоматическое извлечение определенной информации из найденных Web-ресурсов.
* Обобщение — обнаружение общих шаблонов в отдельных и пересекающихся множествах сайтов.
* Анализ — интерпретация найденных шаблонов.

**Категории Web Mining**

В области Web Mining выделяют следующие направления анализа:

* извлечение Web-контента (Web Content Mining);
* извлечение Web-структур (Web Structure Mining);
* исследование использования Web-ресурсов (Web Usage Mining).

Извлечение Web-контента включает в себя методы извлечения полезной информации из Web-ресурсов, таких как содержание, данные, документы и др. Актуальность данного направления возрастает в связи с тем, что в настоящее время прослеживается тенденция предоставления компаниями доступа к своим ресурсам. Это относится не только к статической информации, представленной в виде HTML-страниц, но также к данным, хранящимся в БД компаний, и другим ресурсам (таким, например, как электронные магазины). Безусловно, остается часть данных, к которым доступ невозможен. К этой категории относятся и закрытая (конфиденциальная) информация, а также информация, которая не может анализироваться в виду своей динамичности (например, динамические страницы, формируемые по запросам пользователей). Особенностью Web-ресурсов является разнородность представленной информации: текстовые файлы, изображение, звук, видео, метаданные, а также гиперссылки.

При извлечении Web-структур строятся модели, отображающие взаимосвязи между Web-страницами. Модель основывается на топологии гиперссылок с или без описания этих ссылок. Такая модель может использовать категоризацию Web-страниц и быть полезна для генерации информации об отношении и подобности между Web-сайтами. Данная категория Web Mining может быть использована для распознавания авторских сайтов и обзорных сайтов по темам.

Исследование использования Web-ресурсов анализирует информацию, генерируемую в процессе пользовательских сессий и поведений пользователей. В отличие от первых двух категорий Web Mining, которые работают с первичной информацией, исследование использования Web работает со вторичной информацией, порождаемой как результат взаимодействия пользователей с Web-ресурсами. К таким источникам информации относятся протоколы доступа Web-серверов, протоколы прокси-серверов, протоколы браузеров, пользовательские профили, регистрационные данные, пользовательские запросы, куки, клики мышками, прокручивание и многое другое, что делает пользователь в процессе взаимодействия с Web-ресурсами.

