Министерство 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(2)

Коташевич С.Н.

Проверил: Чичурин А.В.

Брест 2022

**Лабораторная работа №6**

**Визуальный анализ данных.**

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.

В настоящее время существует достаточно большое количество различных видов графических образов, позволяющих представлять результаты анализа в виде, удобном для понимания человеком.

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.

Преимущества визуального анализа:

* Позволяет легко работать с неоднородными и зашумленными данными
* Интуитивно понятен и не требует сложных математических или статистических алгоритмов
* Выполняется быстрее, чем автоматические методы анализа
* Высокая степень конфиденциальности полученных сведений

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

**Характеристики средств визуализации данных.**

Основные характеристики средств визуализации данных:

* Характер данных, которые нужно визуализировать с помощью данного средства
* Методы визуализации и образу, в виде которых могут быть представлены данные
* Возможности взаимодействия с визуальными образами и методами для лучшего анализа данных

Наборы визуальных данных представляют собой матрицы, в которых ряды являются данными, а колонки – атрибутами данных. При этом данные могут характеризоваться одним или несколькими атрибутами. Кроме того, сами данные могут иметь более сложную структуру: иерархическую, текстовую, графическую и т.п. Таким образом, выделяют следующие виды данных, с которыми могут работать средства 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

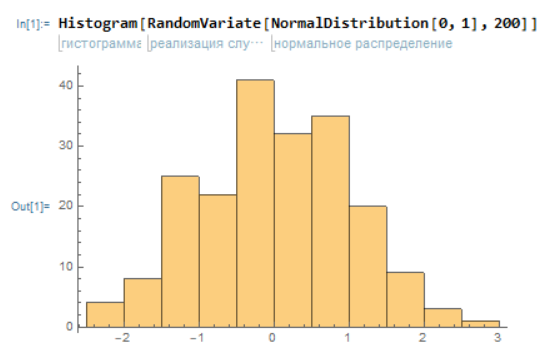
Для визуализации перечисленных типов данных используются различные визуальные образы и методы их создания. Основное требование к образам – это наглядность и удобство анализа данных, которые они представляют. Методы визуализации могут быть как самые простые, так и более сложные, основанные на сложном математическом аппарате. Кроме того, при визуализации могут использоваться комбинации различных методов.

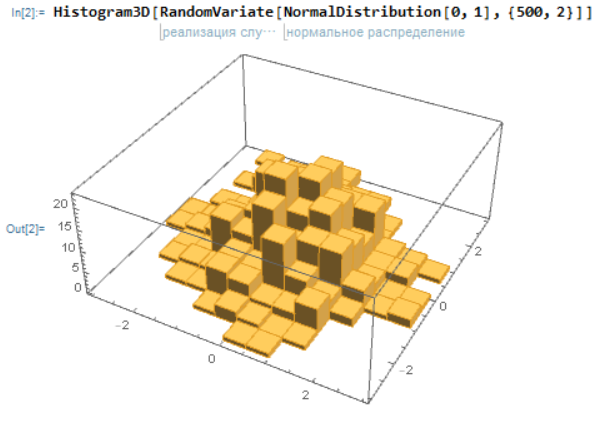
**Методы визуализации**

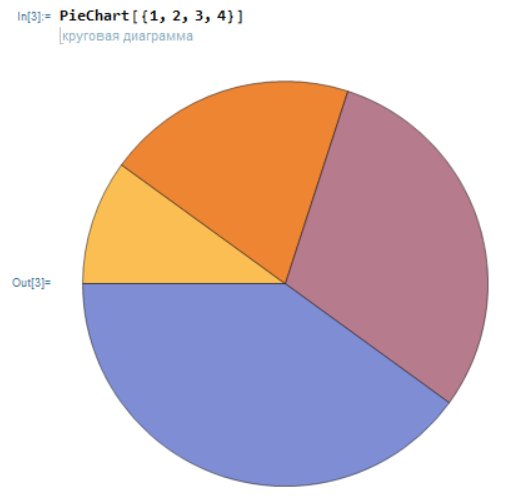
Выделяют следующие типы методов визуализации:

* Стандартные 2D/3D-образы – гистограммы, линейные графики
* Геометрические преобразования – диаграмма разброса данных, параллельные координаты
* Отображение иконок – линейчатые фигуры и звезды
* Методы, ориентированные на пикселы – рекурсивные шаблоны, циклические сегменты
* Иерархические образы – древовидные карты и наложение изменений

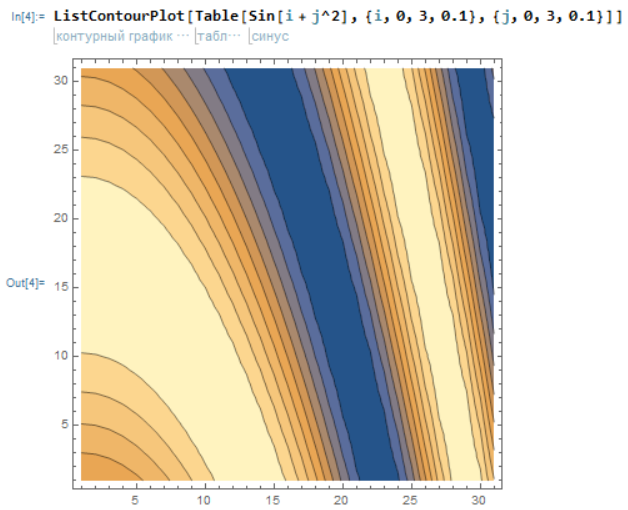
Простейшие методы визуализации, к которым относятся 2D/3D-образы, широко используются в существующих системах. К этим методам относятся: графики, диаграммы, гистограммы. Основным их недостатком является невозможность приемлемой визуализации сложных данных и большого количества данных.

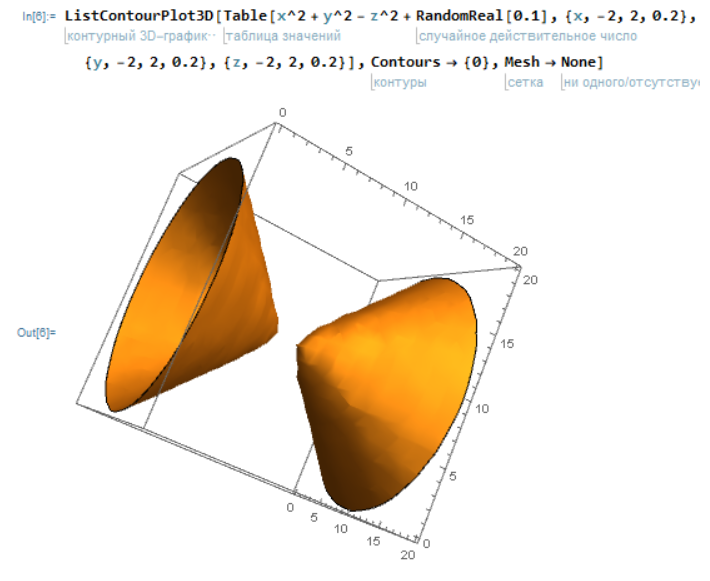




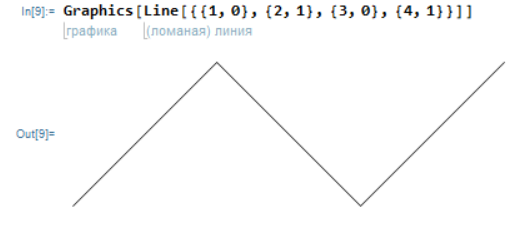


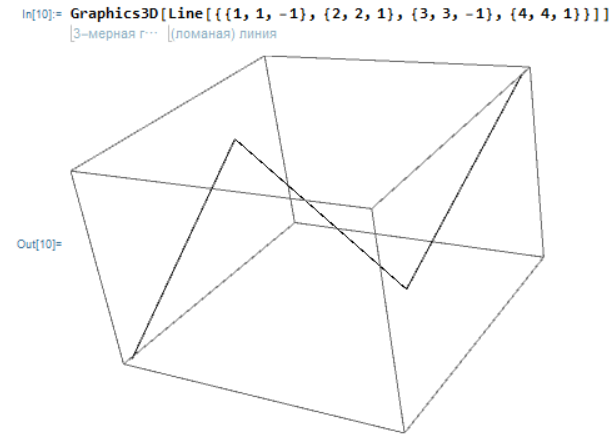
Методы геометрических преобразований визуальных образов направлены на трансформацию многомерных наборов данных с целью отображения их в декартовом и в недекартовом геометрических пространствах. Данный класс методов включает в себя математический аппарат статистики. К нему относятся такие популярные методы,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





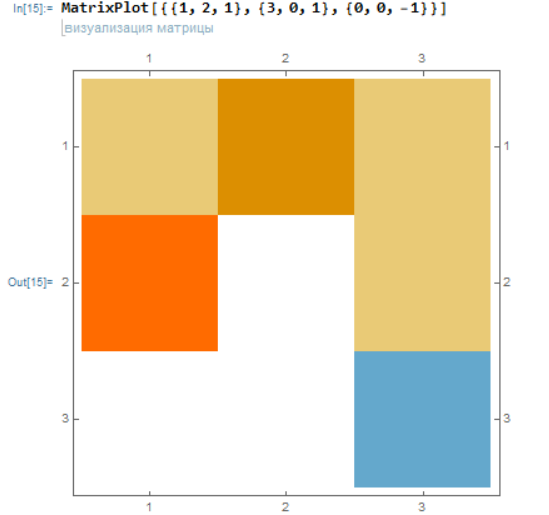
Другим классом методов визуализации данных являются методы отображения иконок. Их основной идеей является отображение значений элементов многомерных данных в свойствах образов. Такие образы могут представлять собой: человеческие лица, стрелки, звезды.



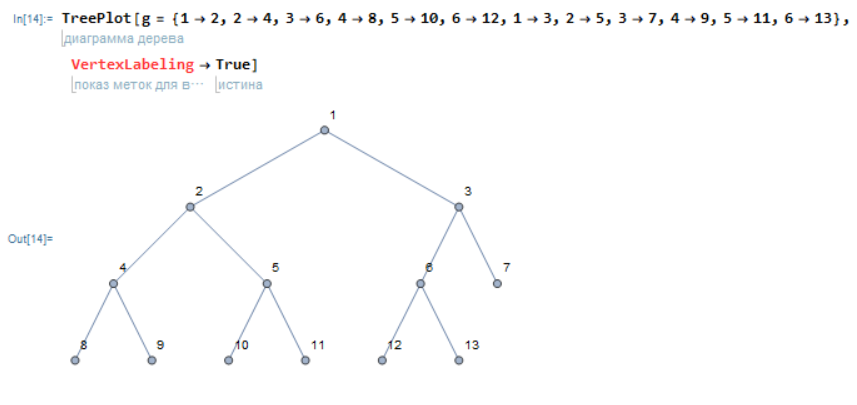


Основной идеей методов, ориентированных на пикселы, является отображение каждого измерения значения в цветной пиксел и их группировка по принадлежности к измерению. Так как один пиксел используется для отображения одного значения, то данный метод позволяет визуализировать большое количество данных.





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**

**Проблемы анализа информации из Web**

Выделяют следующие проблемы работы с информацией Web:

* Поиск значимой информации. Пользователи в поиске информации могут самостоятельно перемещаться 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 символам. Однако использование поисковых систем имеет следующие 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 возникает 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. Одной из них является Web Mining. 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

**Web Mining и другие интернет-технологии.**

Web Mining, являясь инструментов для обработки и анализа Web-ресурсов, рассматривается в одном ряду с такими интернет-технологиями, как получение информации (IR) и извлечение информации (IE).

Технология IR заключается в получении документов из Web-среды, релевантных запросу пользователей. При этом очень часто полученные документы включают в себя как релевантные, так и нерелевантные документы. Целью IE является извлечение необходимых фактов из Web-документов.

Как следует из анализа, различные методы и технологии могут использоваться совместно, взаимно улучая друг друга.

**Извлечение Web-контента в процессе информационного поиска**

Извлечение Web-контента включает в себя методы извлечения полезной информации из Web-ресурсов, таких как содержание, данные, документы. 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, но также к данным, хранящимся в БД компаний, и другим ресурсам.

Различают два основных типа документов: неструктурированные и почти структурированные. К неструктурированному типу относятся все текстовые документы, не имеющие определенной структуры. К почти структурированным относятся документы, имеющие структуру в целом, но позволяющую вхождение в структурные элемент неструктурированного текста. К таким документам относятся HTML, XML.

Большинство методов анализа неструктурированного текста использует представление текстового документа в виде множества или вектора слов. Данный подход также широко применяется в методах Text Mining. При этом в такие представления помещаются отдельные слова без учета их расположения, связи с другими словами, контекста и других лингвистических особенностей.

Извлечение Web-контента из **слабоструктурированных документов** использует более развитые средства представления текста. Это в первую очередь связано с тем, что в документах уже выделены некоторые структурные элементы. Практически все методы в этой области для представления документа используют HTML-структуры внутри документов. Некоторые методы используют также для представления гиперссылки между документами

**Извлечение Web-контента для формирования баз данных**

Задача извлечения Web-контента для его размещения в базе данных относится к проблеме управления информацией и обработки запросов к ней. Существуют три класса задач, относящихся к этой проблеме:

* Моделирование и форматирование запросов к Web
* Извлечение информации и интеграция
* Создание и реструктуризация Web-сайта

Обычно методы извлечение Web-контента пытаются выявить структуру Web-документа или преобразовать его для сохранения в базе данных таким образом, чтобы улучшить информационное управление и сделать возможным запрос к нему. С точки зрения размещения Web-контента в базе данных целью является построение модели данных и объединение их таким образом, чтобы поиск мог выполняться не только по ключевым словам, но и по запросам, более приближенным к естественному языку. Это можно достичь построением схемы Web-документов, формированием хранилища, базы знаний или виртуальной базы данных. Исследования в этой области большей частью имеют дело со слабоструктурированными данными. Слабоструктурированные данные из представления базы данных часто ссылаются на данные, которые имеют некоторую структуру, но не жесткую схему.

**Исследование использования Web-ресурсов**

Процесс исследования использования Web-ресурсов обычно включает в себя только три фазы:

* Препроцессинг
* Извлечение шаблонов
* Анализ шаблонов

В этой задаче исходными данными являются не сами страницы и их содержимое, а информация, записываемая в результате взаимодействия пользователей с Web. Выделяют следующие типы данных, применяемые в задаче исследования использования Web-ресурсов:

* Использование – данные, которые описывают использование страниц, такие как IP-адреса, ссылки на страницы, а также дату и время доступа к ним
* Пользовательские профили – данные, которые обеспечивают демографическую информацию (пол, возраст, социальное положение) о пользователе, а также регистрационную информацию

На стороне сервера информация для анализа может извлекаться из логов, трафика сервера, куки-файлов, запросов пользователей.

На стороне клиента потенциально может находиться достаточно много информации о поведении пользователей, однако основная трудность заключается в извлечении этой информации, т.к. пользователей сайтов бесконечно много и заранее невозможно предсказать, какие пользователи будут обращаться к сайтам. Предлагаются подходы, основанные на использовании агентов и на модификации соответствующих браузеров с целью встраивания в них средств мониторинга действия пользователя.

Прокси-серверы являются промежуточным уровнем между клиентов и сервером и используются для кэширования часто запрашиваемых пользователем страниц. Они могу хранить действительные запросы от нескольких клиентов к различным сайтам, которые не дошли до серверов.

Ни один из видов источников не хранит в себе полной информации. Поэтому для повышения качества анализа поведения пользователей в нем должна участвовать информация от всех видов источников данных.

**Этап препроцессинга**

Для решения задачи исследования использования Web на этапе препроцессинга в массиве анализируемых данных должны быть выделены перечисленные (основные) сущности.

На этапе препроцессинга кроме выделения основных элементов, необходимых для дальнейшего анализа, часто выполняется фильтрация по контексту. Такая задача может возникнуть, например, если аналитика интересует использование Web-ресурсов по определенной тематике. Для такой фильтрации применяют методы анализа Web-контента.

**Этап извлечения шаблонов**

Для извлечения шаблонов из информации об использовании Web-ресурсов применяются различные методы как классической статистики, так и относящиеся к области Data Mining.

Методы статистики часто используются для анализа посещений сайтов и трафиков. Такой вид получаемых знаний может быть полезным для улучшения производительности систем, повышения безопасности систем, решения задач модификации сайтов и обеспечения поддержки для решения маркетинговых задач.

Методы генерации ассоциативных правил могут быть использованы для выявления наиболее часто совместно запрашиваемых страниц, объединенных одной серверной сессией. Методы кластеризации применяются для кластеризации пользователей, так и для кластеризации страниц. Она позволяет группировать пользователей со схожим поведением просмотра страниц. Методы классификации могут быть использованы для развития профилей пользователей, относящихся к определенному классу или категории. Методы обнаружения шаблонов в последовательности применяются для выявления межсессионных шаблонов, в которых элементы следуют друг за другом в упорядоченном по времени множестве сессий и эпизодов.

**Этап анализа шаблонов и их применение**

Последним этапом в исследовании использования Web-ресурсов является анализ извлеченных шаблонов. Целью анализа является отфильтровать наиболее интересные шаблоны и отбросить ничего не значащие шаблоны. Методология анализа во многом зависит от области применения, в которой он выполняется. Более общей формой анализа шаблонов является механизм запроса знаний, такой как SQL. Другой метод заключается в загрузке данных в куб данных и применения к нему OLAP-операций. Методы визуализации могут выделять характерные шаблоны или тренды в данных. Контент или информация о структуре может быть использована для фильтрации страниц.

Области применения систем анализа использования Web-ресурсов:

* Персонализация – обеспечивает для каждого пользователя индивидуальный подход.
* Улучшение системы – анализ позволяет рассматривать изменения трафика, обращения к страницам, поведение пользователей. Результаты анализа могу быть применены для разработки политики кэширования, балансировки нагрузки и распределения данных. Кроме того, позволяет выявлять атаки на сайты.
* Модификация сайтов – анализ обеспечивает дизайнера сайта обратной связью от пользователей и информацией, необходимой для принятия решения об изменения структуры и его содержания.
* Бизнес-интеллект – выполняет анализ информации об использовании пользователями данных с Web-сайтов, совместно с маркетинговой информацией из электронной коммерции.

