Министерство образования Республики Беларусь

Учреждение образования

«Брестский государственный технический университет»

Кафедра ИИТ

Лабораторная работа №6

«Компьютерное моделирование и Анализ данных»

Выполнила:

студентка 4-го курса

группы ПО-4(2)

Юрьев В. А.

Проверил:

Чичурин А. В.

Брест 2022

**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.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 и под любыми углами зрения, в прямом смысле этого слова, он может получить дополнительную информацию, которая поможет ему более четко сформулировать цели исследования.

Таким образом, визуальный анализ данных можно представить как **процесс генерации гипотез.** При этом сгенерированные гипотезы можно проверить или автоматическими средствами (методами статистического анализа или методами Data Mining), или средствами визуального анализа. Кроме того, прямое **вовлечение пользователя** в визуальный анализ имеет **два основных преимущества** перед автоматическими методами:

- 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.

**Три основные характеристики визуализации:**

- характер отображаемых данных, которые нужно визуализировать с помощью данного средства;

- методы визуализации и образы, в виде которых могут быть представлены данные;

- возможности взаимодействия с визуальными образами и методами для лучшего анализа данных.

Наборы визуализируемых данных, как и в Data Mining, представляют собой матрицы, в которых [ряды](https://intellect.icu/category/ryady#term-ryady)являются данными (например, записями об экспериментах, покупки в магазине и т. п.), 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.

Для визуализации перечисленных типов данных используются различные визуальные образы и методы их создания. 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.

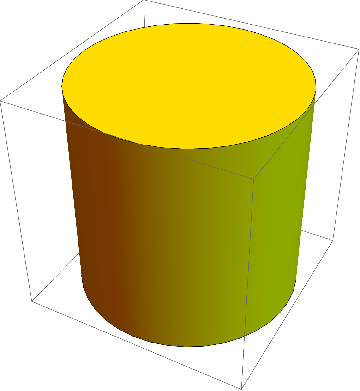
Простейшие методы визуализации, к которым относятся 2D/3D-образы, широко используются в существующих системах (например, в Microsoft Excel). К этим методам относятся: графики, диаграммы, гистограммы и т. п. Основным их недостатком является невозможность приемлемой визуализации сложных данных и большого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

При исследовании большого количества данных важно иметь возможность разделять наборы данных и выделять интересующие поднаборы — фильтровать образы. При этом важно, чтобы данная возможность предоставлялась в режиме реального времени работы с визуальными образами (т. е. интерактивно). Выбор поднабора может осуществляться или напрямую из списка, или с помощью определения свойств интересующего поднабора. Выбор из списка неудобен при большом количестве поднаборов, в то же время запросы не всегда позволяют получить желаемый результат.

**Graphics3D**

Mathematica включает в себя графические примитивы, при помощи которых Вы можете создавать двумерную и трехмерную графику любой сложности.

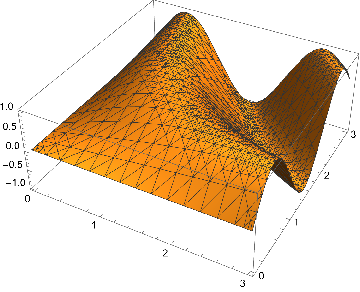
Функция Graphics3D преобразует трёхмерные примитивы в трехмерные графические объекты, добавляя в структуру примитива стили (цвет, толщину и т.д.).



**Plot3D**

Mathematica позволяет создавать не только статичную, но и анимированную графику.

Первый аргумент – выражение f, график которого должен быть построен. Второй и третий аргументы определяют границы изменения переменных x и y задаваемые в виде списков. Обычно функция вычисляется на сетке 15 на 15 и каждый кусок закрашивается согласно светоотражающей модели.



**Импорт и экспорт**

**Импорт и экспорт** может обрабатывать не только табличные данные, но и данные, соответствующие графике, звукам, выражениям и даже целым документам. Импорт и экспорт часто могут вывести соответствующий формат для данных, просто посмотрев на расширение имени файла для файла, в котором хранятся данные.

**Import**

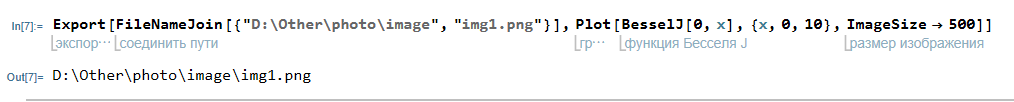
Импортирует данные из источника, возвращая их представление Wolfram Language.

A picture containing icon

Description automatically generated

**Export**

Экспортирует данные в файл, преобразуя их в формат, соответствующий расширению файла.



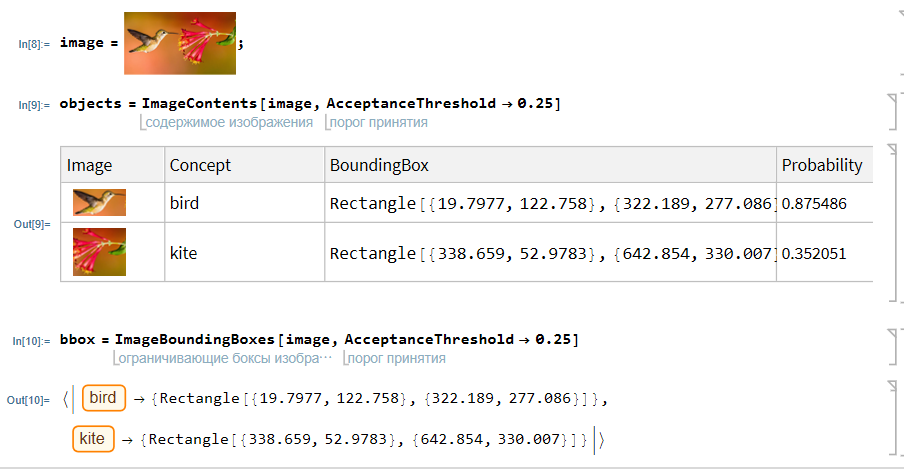
Chart, line chart

Description automatically generated with medium confidence

Используя различные современные методы, Wolfram Language обеспечивает немедленные функции для идентификации изображений и обнаружения и распознавания объектов, а также извлечения признаков. Wolfram Language поддерживает определенные геометрические объекты, такие как края и углы, а также общие ключевые моменты, которые можно использовать для регистрации и сравнения изображений.

**ImageContents, ImageBoundingBoxes**

Дает набор данных идентифицированных сущностей на изображении.



**HighlightImage**

Дает сопоставление списков ограничительных рамок для каждой идентифицированной категории объектов на изображении.

Graphical user interface, text, application

Description automatically generated

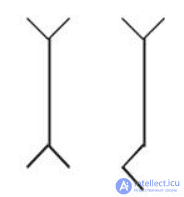
## ****Отображение иконок****

Подход, основанный на **отображении иконок**, предполагает каждому объекту данных ставить в соответствие некоторую иконку. При этом атрибуты объекта должны отображаться различными визуальными свойствами иконок. Иконки могут комбинироваться в матрицы или графики и, таким образом, предоставляют возможность анализировать все объекты в целом.

Использование иконок предполагает следующие методы визуализации:

* линейчатые фигуры;
* "лица Чернова";
* цветные иконки;
* глифы 1 и др.

**Линейчатая фигура** представляет собой иконку с некоторым количеством ветвей (линий). Например, на рисунке представлены две фигуры, имеющие тело (длинная линия) и ветви (четыре коротких линии).



Каждый объект представляется отдельной фигурой. Атрибуты объекта отображаются с разными наклонами и местоположением линий.

**Глиф** — визуальное представление символа шрифта, образ символа шрифта, а также печатное изображение символа шрифта.

Другим хорошо известным методом отображения иконок **является метод "лиц Чернова".** Этот метод предполагает использовать для представления объектов образы человеческих лиц. При этом каждый атрибут отображается определенной характеристикой человеческого лица: длиной, формой и т. п.

Цветные иконки представляют атрибуты объектов цветом, формой, размером, границами, ориентацией

Существует два подхода к раскрашиванию иконок:

1.Закрашивается линия, которая соответствует отдельному атрибуту.

2.Закрашивается часть иконки, соответствующая атрибуту.

**Глифы** представляют собой пикселы с более чем одним измерением. Они размещаются на 2D-площадке и их позиции описываются двумя атрибутами, тогда как другие атрибуты представляются цветом и формой. Некоторые модификации метода применяют глифы в виде цветов, звезд и др

Wolfram позволяет очень гибко манипулировать самыми различными видами иконок. И позволяет как рисовать их из пикселей прямо в коде:

Icon

Description automatically generated

**Извлечение знаний из Web – Web Mining**

1. Проблемы анализа информации из Web.

Всемирная сеть сейчас содержит огромное количество информации, знаний. Пользователи на различных условиях могут просматривать всевозможные документы, аудио- и видеофайлы. Однако это многообразие данных скрывает в себе проблемы, которые могут возникнуть не только при анализе, но и при поиске необходимой информации в Интернет.

1. Этапы Web Mining.

Web Mining – технология, использующая методы Data Mining для исследования и извлечения информации из Web-документов и сервисов. Выделяют следующие этапы применения Web Mining:  
1. Поиск ресурсов – локализация неизвестных документов и сервисов в Web.  
2. Извлечение информации – автоматическое извлечение определенной информа-ции из найденных Web-ресурсов.   
3. Обобщение – обнаружение общих шаблонов в отдельных и пересекающихся множествах сайтов.   
4. Анализ – интерпретация найденных шаблонов.

**WolframAlpha** – это смесь алгоритмов и использование искусственного интеллекта для поиска нужной информации. То есть он сам найдёт на основе входных данных нужную мне информацию. Всё, что мне нужно будет сделать – это обработать её.

1. Web Mining и другие интернет-технологии.

Web Mining, являясь инструментом для обработки и анализа Web-ресурсов, рассматривается в одном ряду с такими интернет-технологиями, как получение информации (Information Retrieval – IR) и извлечение информации (Information Extraction – IE). Однако, имея с ними много общего, Web Mining имеет также существенные отличия. Рассмотрим некоторые из них.

1. Извлечение Web-контента в процессе информационного поиска.

Методы извлечения Web-контента в процессе информационного поиска во многом зависят от типа анализируемых документов. Различают два основных типа: 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 структурный элемент неструктурированного текста. К таким документам относятся HTML, XML и др.

1. Слабоструктурированныедокументы.

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.

1. 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:

1. моделирование и формирование запросов к Web;
2. извлечение информации и интеграция;
3. создание и реструктуризация Web-сайта.

7) Исследование использования Web-ресурсов.  
Процесс исследования использования Web-ресурсов обычно включает в себя только три фазы:  
1. Препроцессиг  
2. Извлечение шаблонов  
3. Анализ шаблонов.

8) Этап препроцессинга.

Для решения задачи исследования использования Web на этапе препроцессинга в массиве анализируемых данных должны быть выделены перечисленные сущности. Это значительно усложняет неполнота данных, получаемых с одного источника.

9) 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.

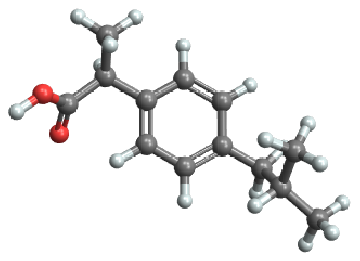
10) Этап анализа шаблонов и их применение

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.

**WolframAlpha**

Oтправляет запрос в адрес Wolfram| Альфа и импортирует выходные данные.

c



Вывод информации об акциях компании IBM с 1 января 2019 по 1 января 2021:

Mean 120.45$ Highest 141.76$ Lowest 101.87$ Volatility 28% Change 15%

Получение информации о количестве сотрудников в Epam при помощи прямого запроса данных:

**Вывод:**

- 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 ними.

- Визуальный анализ данных обычно выполняется в три этапа: беглый анализ, уве-личение и фильтрация, детализация по необходимости.

- Выделяют три основные характеристики средств визуализации: характер отобра-жаемых данных, методы визуализации, возможности взаимодействия с визуальными образами.

- Web Mining включает в себя следующие этапы: поиск ресурсов, извлечение ин-формации, обобщение и анализ.

- Различают следующие категории задач Web Mining: извлечение Web-контента, извлечение Web-структур и исследование использования Web-ресурсов.

- В решении задачи извлечения структуры Web используются подходы из области социальных сетей, библиометрики, ранжирования документов и т. п.

- В задаче исследования использования Web анализу подвергаются вторичные дан -ные о взаимодействии пользователя с Web: протоколы работы, куки, авторизация и т. п.

- Существуют два основных подхода анализа использования Web-ресурсов: преобразование данных использования Web-сервера в реляционные таблицы до выполнения адаптированных методов Data Mining и использование информации из файла протокола непосредственно, применяя специальные методы предварительной обработки.