**Визуальный анализ данных**

1. Выполнение визуального анализа 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.

Визуальный анализ данных обычно выполняется в три этапа:

беглый анализ — позволяет идентифицировать интересные шаблоны и сфокусироваться на одном или нескольких из них;

увеличение и фильтрация — идентифицированные на предыдущем этапе шаблоны отфильтровываются и рассматриваются в большем масштабе;

детализация по необходимости — если пользователю нужно получить дополнительную информацию, он может визуализировать более детальные данные.

В контексте визуализации данных Wolfram – достаточно мощный инструмент, который предполагает:

* **Графическое представление структурированных и неструктурированных данных.**Символьная архитектура и динамический интерфейс языка позволяют применять гибкие подходы к визуализации информации.
* **Редактирование и контроль результатов визуализации в интерактивном режиме.** Wolfram предусматривает множество возможностей контроля над каждым этапом визуализации – от разработки и построения и до окончательного внешнего вида.
* **Импорт и экспорт данных.**Wolfram поддерживает и автоматически обрабатывает данные практически любого известного формата; все они интегрированы в среду программирования в виде символьного обозначения.

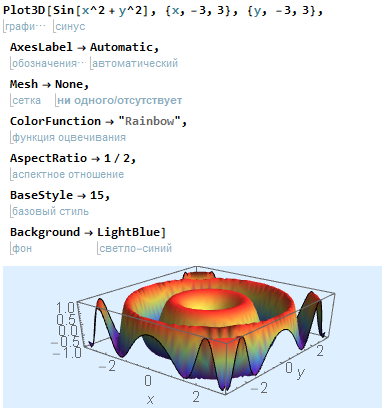
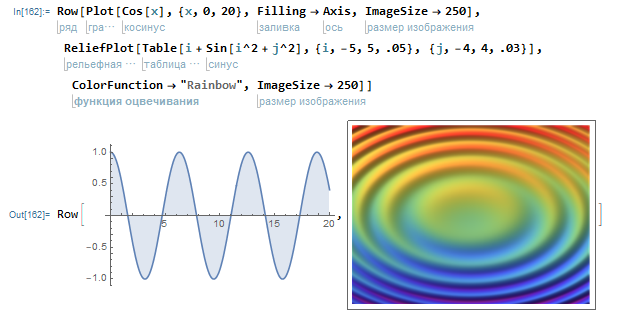
1. Характеристики средств визуализации данных.

Существует достаточно большое количество средств визуализации данных. Рассмотрим более подробно три основные характеристики, которые позволяют определить наилучшее средство для визуализации данных:

* характер 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.

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;
* отображение иконок — линейчатые фигуры (needle icons) и звезды (star icons);
* методы, ориентированные на пикселы — рекурсивные шаблоны, циклические сегменты;
* иерархические образы — древовидные карты и наложение измерений.



В результате применения методов визуализации будут построены визуальные образы, отражающие данные. Mathematica содержит полный набор функций для отображения структурированных и неструктурированных данных в 2d и 3d.

**Графическая визуализация в Wolfram**

Wolfram предоставляет огромное количество возможностей. Среди этих возможностей можно выделить:

* построение кривых, иллюстрирующих поведение одной или нескольких функций одной переменной, заданных аналитическими выражениями или по точкам в декартовых, полярных координатах; функций, заданных в параметрической форме;
* построение кривых, иллюстрирующих поведение функций с логарифмическими и полулогарифмическими масштабами;   
  разные способы оформления изображений трёхмерных поверхностей;
* построение графиков векторных полей, линий тока, линий тока с фоном плотности функции; графиков вихря, градиента, Лапласиана векторного поля;
* построение проекций графиков поверхностей и др.

Любое средство визуализации может быть классифицировано по всем трем параметрам, т. е. по виду данных, с которым оно работает, по визуальным образам, которые оно может предоставлять, и по возможностям взаимодействия с этими визуальными образами. Очевидно, что одно средство визуализации может поддерживать разные виды данных, разные визуальные образы и разные способы взаимодействия с образами.

3) Методы визуализации.

Одним из методов визуализации является метод геометрических преобразований. Основная идея методов геометрических преобразований — визуализировать преобразования и проекции данных в декартовом и в недекартовом геометрических пространствах. Wolfram позволяет выполнять следующие геометрические преобразования: точки и матрицы, поверхностные и объемные графики, контуры, параллельные координаты, текстуры и растры.

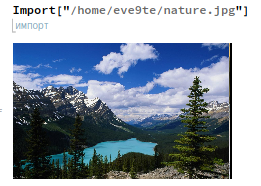
4) Отображение иконок.

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

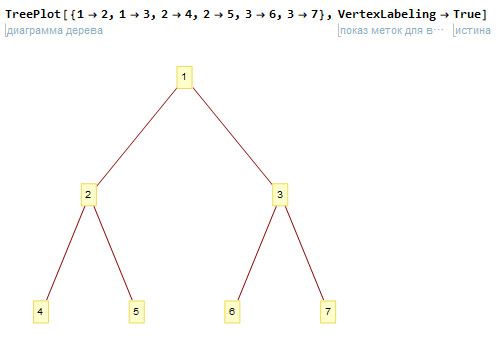
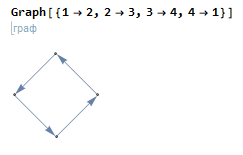


Так и импортировать файлы с изображениями.

А также использовать их при построении трёхмерных графиков.

5) Иерархические образы.  
 Иерархические образы используются для отображения иерархий и отношений в данных. Wolfram позволяет работать с такими типами иерархических данных как графы и деревья.  
 Система*Mathematica* предоставляет современную функциональность для анализа и синтезирования графов и сетей. При работе с графами доступны следующие возможности:

* Поддержка ориентированных, неориентированных и взвешенных графов.
* Непосредственная поддержка случайных графовых распределений.
* Выделение графовых элементов и подграфов с помощью языка шаблонов системы *Mathematica*.
* Эффективное тестирование графов.
* Более 15 метрик и мер центральности для квалифицирования графов и сетей.
* Эффективные функции нахождения кратчайшего пути, циклов и маршрутов и др.

****

**Извлечение знаний из Web — Web Mining**

1. Проблемы анализа информации из Web.

Всемирная сеть сейчас содержит огромное количество информации, знаний. Однако это многообразие данных скрывает в себе проблемы, которые могут возникнуть не только при анализе, но и при поиске необходимой информации в Интернет.

* Проблема поиска нужной информации связана с тем, что пользователь не всегда сразу может найти необходимые ему электронные ресурсы.
* Проблема обнаружения новых знаний. Даже если найдено множество информации, для пользователя извлечение полезных знаний является довольно трудоемкой и непростой задачей.
* Проблема изучения потребителей связана с предоставлением пользователю информации, которая оказалась бы ему интересна.

1. 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.   
 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 информации из найденных Web-ресурсов.   
 3. Обобщение — обнаружение общих шаблонов в отдельных и пересекающихся множествах сайтов.   
 4. Анализ — интерпретация найденных шаблонов.   
 На этапе обобщения к обработанной информации применяются методы Data Mining. На этом этапе важную роль играет человек, учитывая также тот факт, что на последнем этапе он должен будет интерпретировать полученные результаты.

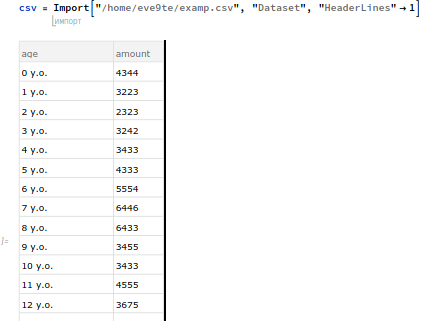
1. Web Mining и другие интернет-технологии.

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.

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. Основное отличие этой технологии от IR заключается в том, что она работает с самим документом и ищет в нем релевантную информацию, в то время как IR работает с множеством документов, извлекая из него релевантные документы.

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

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.   
   Кроме большого размера модели, векторное представление документов имеет еще один существенный недостаток: оно не обрабатывает синонимы — документы считаются семантически далекими друг от друга, если в них нет одинаковых слов. Данный недостаток устраняется методом скрытой семантической индексации . Согласно этому методу, пространство термов сингулярным разложением (отбрасываются наименее значимые сингулярные значения) приводится к пространству ортогональных факторов (некоррелируемых "индексных термов"). В результате документы одной тематики, но которые не используют одинаковые термины, размещаются в одном разделе, и их стемминг сокращает слова с общими морфологическими корнями.   
    Необходимо обратить внимание, что эффективность вариантов препроцессинга, направленных на сокращение размера набора свойств, может быть различна для разных областей.  
   Отдельно стоит остановиться на задаче определения событий и трекинге. Данная задача является частной задачей более широкого направления автоматизированной обработки новостных данных — Topic Detection and Tracking (TDT) . В TDT выделяют следующие направления исследований:  
     
   - разбиение потока на сюжеты;  
   - идентификация новых событий;  
   - определение связей между новостными историями;  
   - отслеживание интересующей пользователя информации.   
     
   5) Слабоструктурированные документы.   
    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.  
   Как и в случае с неструктурированными документами, к полученным представлениям применяются общие методы Data Mining.

Область применения методов довольно широка:

- гипертекстовая классификация;

- классификации и кластеризации;

- изучение отношений между Web-документами;

- извлечение шаблонов или правила;

- поиск шаблонов и слабоструктурированных данных.

6) 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 формирование запросов к Web;

извлечение информации и интеграция;

создание и реструктуризация Web-сайта.

Хотя первые две задачи относятся к категории извлечения Web-контента, не все методы, применяемые при их решении, относятся к этой категории. Это связано с отсутствием машинного обучения или использованием методов Data Mining в процессе их решения. Обычно методы извлечения Web-контента пытаются выявить структуру Web-документа или преобразовать его для сохранения в базе данных таким образом, чтобы улучшить информационное управление и cделать возможным запрос к нему.

С точки зрения размещения Web-контента в базе данных целью, большей частью, 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. Этого 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.   
 DataGuide — сжатый вид схемы слабоструктурированных данных. Для практического применения и из-за сложности вычислений DataGuide часто аппроксимируется. Некоторые приложения не решают задачу поиска глобальной схемы, они концентрируются на задачах поиска подсхем в слабоструктурированных данных.  
 Другое применение методов данной категории Web Mining — это формирование многослойной базы данных (MLDB), в которой каждый уровень создается обобщением низших уровней и использует специальный язык запросов для Web Mining, чтобы извлекать некоторые знания из MLDB.

Из-за различий представления Web-документов, используемых в извлечении Web-контента для их сохранения в БД, большинство из используемых методов отличаются от классических методов Data Mining, которые действуют на плоских данных и требуют дополнительной модификации.

Необходимо отметить, что существует также большая область, связанная с анализом мультимедийной информации, которая в большом объеме присутствует в Web-документах.

1. Исследование использования Web-ресурсов.  
    Процесс исследования использования Web-ресурсов обычно включает в себя только три фазы:  
    1. Препроцессиг  
    2. Извлечение шаблонов  
    3. Анализ шаблонов.  
   В отличие от предыдущих двух задач Web Mining, 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-адресса, ссылки на страницы, а также дату и время доступа к ним;  
    пользовательские профили — данные, которые обеспечивают демографическую информацию (пол, возраст, социальное положение и др.) о пользователе, а также регистрационную информацию.  
   Данные об использовании Web собираются в различных источниках, которые можно разделить на следующие основные группы: серверы, клиенты, прокси.  
    На стороне сервера информация для анализа может извлекаться из логов, трафика сервера, куки-файлов, запросов пользователей и др.  
    Логи Web-серверов являются важным источником информации, т. к. они в явном виде содержат описание действий посетителей сайтов. Однако до сервера доходит не вся информация о действиях пользователей, так просмотр пользователем страниц, сохраненных у него в кэше, никак не отражается на сервере. Кроме того, информация, заключенная в HTTP Post-запрос, не будет сохраняться в логе сервера.  
    Для извлечения информации из входящего в сервер трафика может использоваться технология анализатора пакетов (packet sniffer). Данная технология позволяет извлекать информацию напрямую из TCP/IP-пакетов, поступающих на Web-сервер. Куки-файлы автоматически генерируются для каждого пользовательского браузера, чтобы поддержать сессионное взаимодействие. Они хранят информацию, идентифицирующую пользовательскую сессию. Пользовательские запросы содержат информацию о потребностях пользователей, искомых документах и др.  
    На стороне клиента потенциально может находиться достаточно много информации о поведении пользователей, начиная от просматриваемых страниц и заканчивая щелчками мыши. Однако основная трудность заключается в извлечении этой информации, т. к. пользователей сайтов бесконечно много и заранее невозможно предсказать, какие пользователи будут обращаться к сайтам.  
    Ни один из видов источников не хранит в себе полной информации, а следовательно, не может рассматриваться как единственный поставщик данных для анализа. Для повышения качества анализа поведения пользователей в нем должна участвовать информация от всех видов источников данных.

Вся информация, поставляемая от перечисленных ранее источников, может быть описана в следующих терминах, определенных группой W3C Web Characterization Activity (WCA).

**Пользователь** (user) — индивидуум, получающий доступ к файлу одного или нескольких Web-серверов через браузер.

**Просмотр страницы** (page view) — включает в себя одновременное отображение нескольких файлов в браузере пользователя.

**Потоки кликов** (click streams) — последовательность открытия Web-стра- ниц (вызываемых кликами). Не всегда можно восстановить полную последовательность, основываясь только на информации со стороны сервера, т. к. запросы, обрабатываемые клиентом или прокси-сервером, "не известны".

**Пользовательская сессия** — последовательность просматриваемых страниц (поток кликов) одним пользователем на разных Web-сайтах.

**Сессии сервера** (server sessions) — набор просматриваемых страниц в рамках одной пользовательской сессии с одного Web-сервера.

**Эпизоды** (episodes) — любой семантически значимый поднабор пользовательской или серверной сессии.

1. 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. Например, для идентификации пользователя недостаточно только IP-адреса, т. к. возможны следующие ситуации.

- Один IP-адрес — несколько серверных сессий. Интернет-провайдеры обычно имеют пул прокси-серверов, доступных пользователям. В одно и то же время пользователь может просматривать страницы одного сайта с различных прокси-серверов.

- Несколько IP-адресов — одна серверная сессия. Некоторые интернетпровайдеры или инструменты безопасности случайным образом каждый пользовательский запрос ассоциируют с разными IP-адресами. В результате одна серверная сессия ассоциируется с несколькими IP-адресами.

- Несколько IP-адресов — один пользователь. Пользователь может получать доступ к Web-ресурсам с разных машин, а следовательно, будет ассоциирован с разными IP-адресами. Это усложняет отслеживание повторяющихся запросов от одного пользователя.

- Множество агентов — один пользователь. Пользователь, работающий с более чем одним браузером или пользующийся агентами, даже с одной машины будет распознаваться как несколько пользователей.

При условии если удалось идентифицировать пользователя (с помощью кукифайлов, логина или анализа его IP/агента/пути), необходимо поток кликов от каждого пользователя разделить на сессии. Эта задача усложняется тем, что тяжело определить, когда пользователь покинул сайт, т. к. обычно при анализе запросы данного пользователя к другим серверам не доступны. Кроме того, часто сессия прерывается по тайм-ауту, если пользователь не возобновляет

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.

1. 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 посещения сайтов и трафиков. Так, анализ сессионных файлов позволяет выполнить различные виды дескриптивного статистического анализа (вычислить частоту, математическое ожидание, дисперсию и т. п.) для просмотров страниц, времени просмотра и длины навигационного пути. Многие инструменты анализа трафика позволяют получить такие характеристики, как наиболее часто посещаемые страницы, среднее время посещения страниц или средняя длина пути перемещения по страницам. Подобные отчеты могут содержать низкоуровневые ошибки, связанные с невозможностью определить неавторизированные точки входа или недействительные URI.  
    Методы генерации ассоциативных правил могут быть использованы для выявления наиболее часто совместно запрашиваемых страниц, объединенных одной серверной сессией. Эти страницы могут быть связаны не напрямую друг с другом (т. е. не иметь прямых ссылок друг на друга).   
    Методы кластеризации в области исследования использования Web-ресурсов применяются как для кластеризации пользователей, так и для кластеризации страниц. Кластеризация пользователей позволяет группировать пользователей с похожим поведением просмотра страниц. Такие знания полезны для того, чтобы сделать выводы о демографии пользователей и выполнении маркетинговой сегментации рынка в электронной коммерции или обеспечении пользователей персональным Web-контентом.  
    Методы классификации могут быть использованы для развития профилей пользователей, относящихся к определенному классу или категории. Это требует построения и выбора функции, которая бы наилучшим образом описывала свойства данного класса.   
    Методы обнаружения шаблонов в последовательностях применяются для выявления межсессионных шаблонов, в которых элементы следуют друг за другом в упорядоченном по времени множестве сессий и эпизодов. Данный подход может помочь в прогнозировании структуры будущих посещений, что в свою очередь помогает в правильном размещении рекламы, нацеленной на конкретную пользовательскую аудиторию.   
    Моделирование использования Web-ресурсов позволяет обеспечить не только теоретическую основу поведения пользователей, но и помочь в предсказании. Это способствует увеличению продаж продуктов, размещаемых на сайте, а также улучшению навигации.
2. 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. Более общей формой анализа шаблонов является механизм запроса знаний, такой как SQL. Другой метод заключается в загрузке данных в куб данных для применения к нему OLAP-операций. Методы визуализации, такие как раскрашивание или графическое изображение шаблонов, могут выделять характерные шаблоны или тренды в данных.

В литературе приводится классификация существующих систем анализа использования Web-ресурсов. Классификация выполняется по пяти характеристикам:

- источнику данных: сторона сервера, сторона клиента и прокси;

- типу данных: структуры, контент и информация об использовании;

- количеству пользователей: однопользовательские и многопользовательские;

- количеству сайтов: один или множество сайтов;

- области применения.

Выделяют следующие области применения систем анализа использования Web-ресурсов.

- Персонализация (Personalization) — обеспечивает для каждого пользователя индивидуальный подход и является одной из важнейших задач для многих Web-систем (например, систем электронной коммерции).

- Улучшение систем (System Improvement) — анализ использования Web-ресурсов позволяет рассматривать (выявлять закономерности и взаимосвязи) изменения трафика, обращения к страницам, поведение пользователей. Результаты анализа могут быть применены для разработки политики кэширования, балансировки нагрузки и распределения данных.

- Модификация сайтов (Site Modification) — анализ использования Web-ресурсов обеспечивает дизайнера сайта своего рода обратной связью от пользователей и информацией, необходимой для принятия решения об изменении структуры и его содержания.

- Бизнес-интеллект (Business Intelligence) — выполняет анализ информации об использовании пользователями данных с Web-сайтов, совместно с маркетинговой информацией из электронной коммерции.

# **Выводы**

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

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

- Извлечение Web-контента может проводиться в целях информационного поиска и с целью сохранения его в базе данных, также различают неструктурированные и слабоструктурированные Web-документы.

- Для извлечения Web-контента из неструктурированных документов используют модели их представления и методы, заимствованные из Text Mining.

- В задаче извлечения Web-структур для представления Web используют направленные и ненаправленные графы.

- 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 данные 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.