***Визуальный анализ данных***

1. ***Выполнение визуального анализа данных.***

Основной идеей визуального анализа данных является представление данных в некоторой визуальной форме.

Визуальный анализ данных можно представить, как процесс генерации гипотез. При этом сгенерированные гипотезы можно проверить или автоматическими средствами или средствами визуального анализа.

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 предполагает:

* Графическое представление структурированных и неструктурированных данных.
* Редактирование и контроль результатов визуализации в интерактивном режиме.
* Импорт и экспорт данных.

1. ***Характеристики 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.

Наборы визуализируемых данных, как и в Data Mining, представляют собой матрицы, в которых ряды являются данными, а колонки — атрибутами данных. 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:

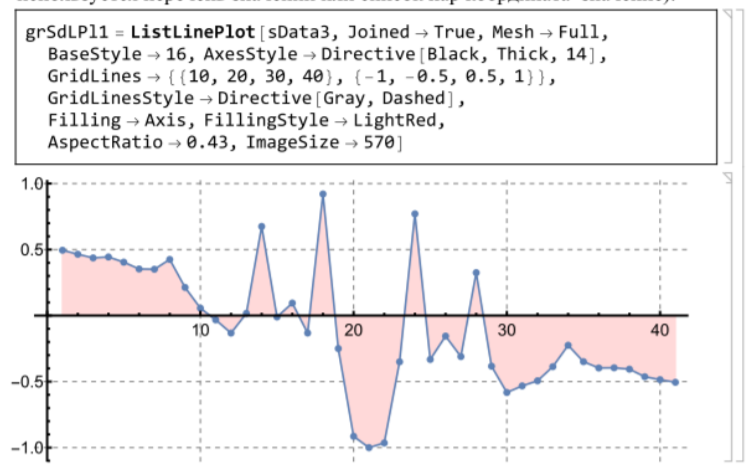
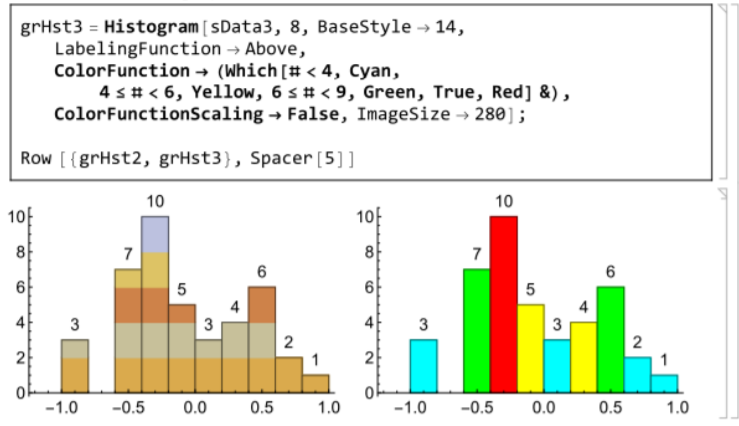
* одномерные данные — одномерные массивы, временные ряды;
* двумерные данные — точки двумерных графиков, географические координаты;
* многомерные данные — финансовые показатели, результаты экспериментов;
* тексты и гипертексты — газетные статьи, Web-документы;
* иерархические и связанные — структура 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. Выделяют следующие типы методов визуализации:

* 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.

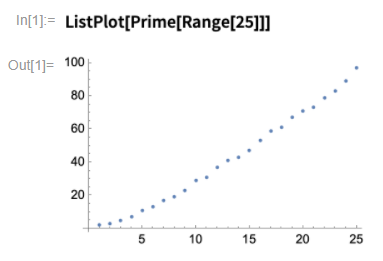
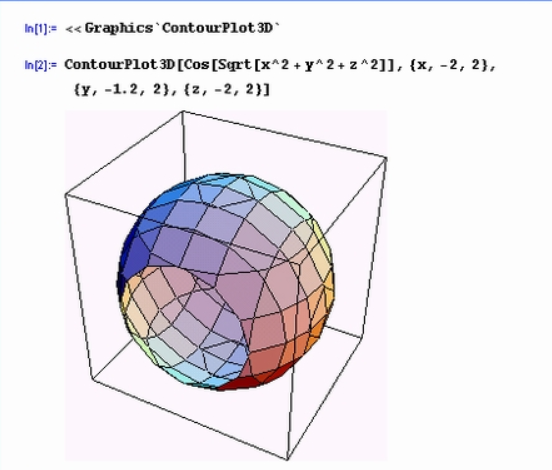
***Графическая визуализация в Wolfram***

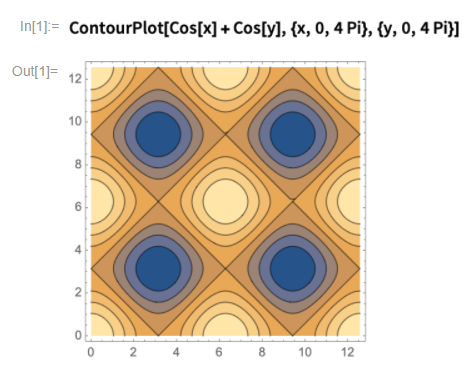
Wolfram предоставляет огромное количество возможностей. Среди этих возможностей можно выделить:

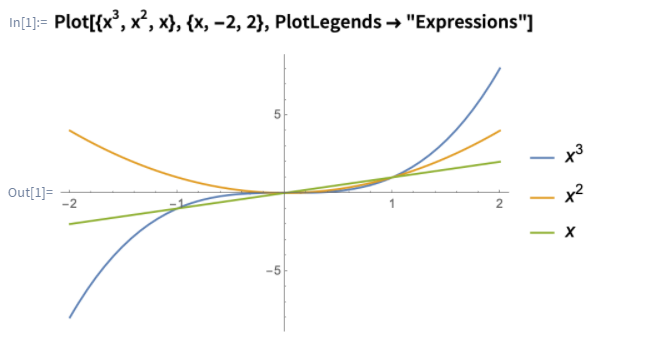
* построение кривых, иллюстрирующих поведение одной или нескольких функций одной переменной, заданных аналитическими выражениями или по точкам в декартовых, полярных координатах; функций, заданных в параметрической форме  
  
* построение столбиковых, секторных, финансовых диаграмм; представление данных, аппроксимация сплайнами, поддержка сплайнов, включая неоднородный рациональный В-сплайн  
  
* разные способы отображения на плоскости трёхмерных поверхностей
* разные способы оформления изображений трёхмерных поверхностей
* построение графиков векторных полей, линий тока, линий тока с фоном плотности функции
* построение проекций графиков поверхностей;
* построение пространственных фигур стереометрии, полиэдров;
* построение реалистических графиков рельефа
* формирование трехмерных графических объектов, получаемых вращением кривых относительно некоторой оси
* графы и сети полностью интегрированы в системупостроение графиков в основных картографических системах
* возможности импорта рисунков и редактирования получаемых изображений инструментами встроенного редактора графики.

***3) Методы визуализации.***

Основная идея методов геометрических преобразований — визуализировать преобразования и проекции данных в декартовом и в недекартовом геометрических пространствах. Wolfram позволяет выполнять следующие геометрические преобразования:

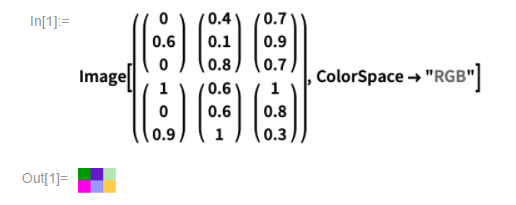
* точки и матрицы;  
  
* поверхностные и объемные графики, контуры;  
  



* параллельные координаты;  
  

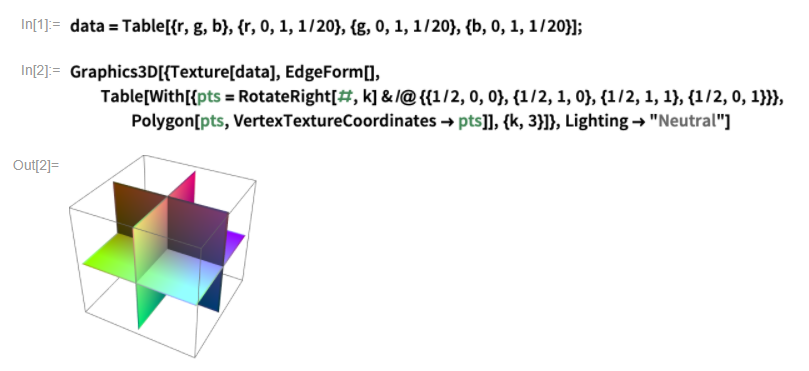
***4) Отображение иконок.***

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.***

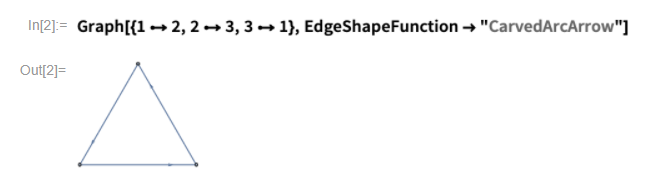
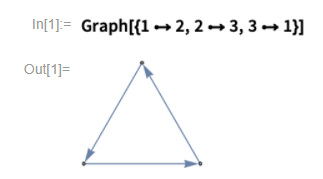
Иерархические образы используются для отображения иерархий и отношений в данных.

Wolfram позволяет работать с такими типами иерархических данных как:

* Графы
* Деревья

Система*Mathematica* предоставляет современную функциональность для анализа и синтезирования графов и сетей.

Примеры построения графов:



Деревья являются частными случаями графов, исходя из этого можно сделать вывод что для деревьев доступны такие же операции, как и для графов.

***Извлечение знаний из Web — Web Mining***

Web Mining — это процесс извлечения данных из веб-ресурсов, который, как правило, имеет больше практическую составляющую нежели теоретическую. Основная цель Web Mining — это сбор данных (парсинг) с последующим сохранением в нужном формате. Фактически, задача сводится к написанию HTML парсеров, и как раз об этом поговорим более детально.

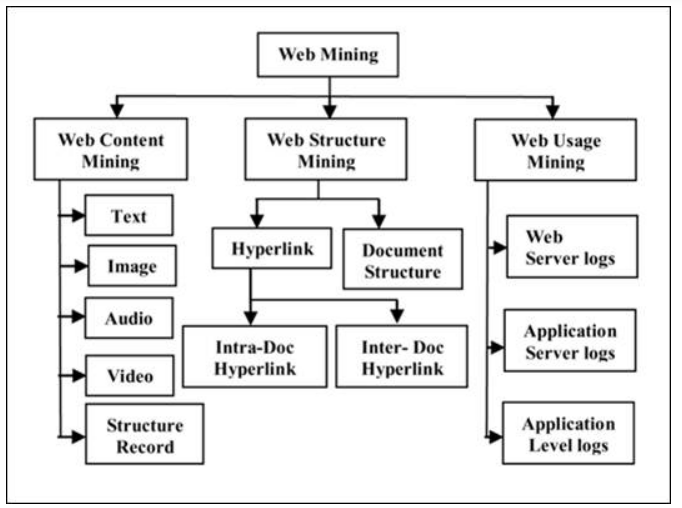
***Этапы Web Mining.***

Технология Web Mining включается в себя следующие этапы:

1. Поиск ресурсов.
2. Извлечение информации.
3. Обобщение.
4. Анализ.

Есть несколько подходов к извлечению данных:

1. Анализ DOM дерева, использование XPath
2. Парсинг строк.
3. Использование регулярных выражений.
4. XML парсинг.
5. Визуальный подход.



***Категории Web*** ***Mining***

***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 (IR) и извлечение информации (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.

***Извлечение Web-контента в процессе инфор-мационного поиска.***

Поиск знаний в сети Интернет является непростой и трудоемкой задачей. Именно это направление Web Mining решает её. Оно основано на сочетании возможностей информационного поиска, [машинного обучения](http://www.basegroup.ru/glossary_ajax/definitions/machine_learning) и Data Mining. Кроме того, Web Content Mining подразумевает автоматический поиск и извлечение качественной информации из разнообразных источников Интернета, перегруженных "информационным шумом". Здесь также идет речь о различных средствах кластеризации и аннотировании документов.

Извлечение знаний заключается в получении знаний из огромных массивов данных. Извлечение данных — это процесс поиска, сбора, хранения информации.Выделяют следующие проблемы работы с информацией из сети Интернет:

1.     Поиск значимой информации.

2.     Выявление знаний из веб-ресурсов.

3.     Персонализация информации.

4.     Изучение индивидуального пользователя.

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

Подход, основанный на базах данных (Database Approach), включает системы:

* многоуровневые базы данных;
* системы web-запросов (Web Query Systems);

Примеры систем web-запросов:

* W3QL (Konopnicki и Shmueli, 1995),
* WebLog (Lakshmanan и др., 1996),
* Lorel (Quass и др., 1995),
* UnQL (Buneman и др., 1995 and 1996),
* TSIMMIS (Chawathe и др.., 1994).

Анализируется содержание документов: находятся схожие по смыслу слова и их количество. Затем решается задача [кластеризации](http://www.basegroup.ru/glossary_ajax/definitions/clustering) или [классификации](http://www.basegroup.ru/glossary_ajax/definitions/classification). Так документы группируются по смысловой близости. Это направление может быть использовано для оптимизации поиска индексированных документов.

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.

В большинстве своем методы применяются в задачах, связанных с выявлением, исследованием или формированием схем DataGuides .

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.



*(Таблица) Методы извлечения Web-контента из слабоструктурированных документов в сохранения в базе данных*

***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.

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

Просмотр страницы— включает в себя одновременное отображение нескольких файлов в браузере пользователя.

Потоки кликов— последовательность открытия Web-стра- ниц (вызываемых кликами).

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

Сессии сервера— набор просматриваемых страниц в рамках одной пользовательской сессии с одного Web-сервера.

Эпизоды— любой семантически значимый поднабор пользовательской или серверной сессии.

***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.

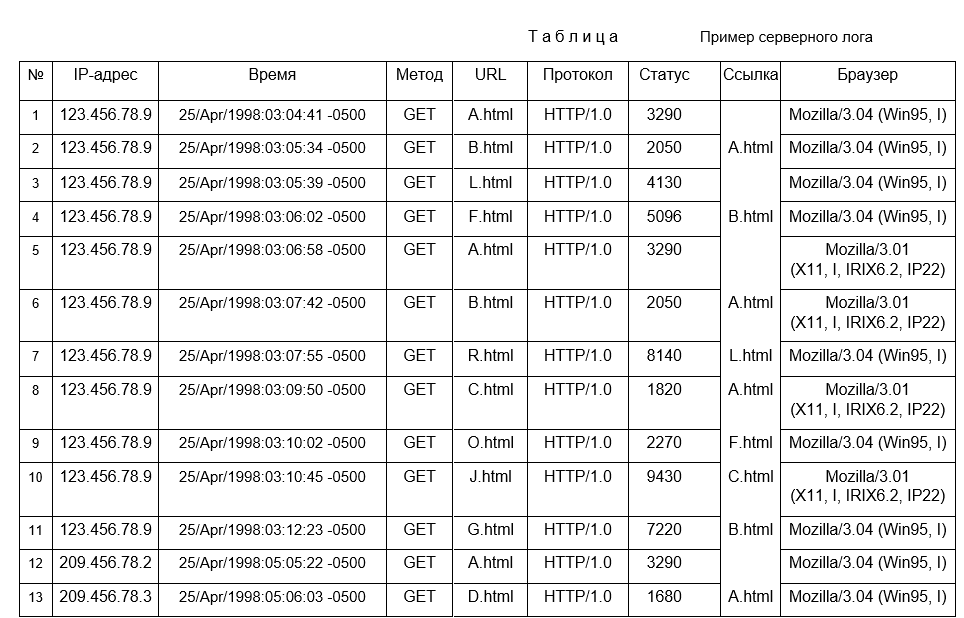
- Один IP-адрес — несколько серверных сессий. Интернет-провайдеры обычно имеют пул прокси-серверов, доступных пользователям. В одно и то же время пользователь может просматривать страницы одного сайта с различных прокси-серверов.

- 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 взаимодействие с сервером через некоторый промежуток времени. Если ID сессии встраивается в каждый URI, то определение сессии устанавливается контентным сервером. В этом случае запрашиваемый контент хранится в поле запроса и сохраняется в логе на сервере. Однако в связи с тем, что для каждой сессии сервер может иметь переменные состояний, в которых хранит некоторую информацию о сессии, то не вся она доступна из URI. Также использование кэша и связанное с этим недохождение запросов до сервера усложняет выделение сессий. Эта проблема может быть решена только мониторингом запросов от клиентской стороны к кэшу. Ссылочные поля запросов могут быть использованы, чтобы определить, когда были просмотрены страницы из кэша.



В таблице приведен пример серверного лога (первая колонка добавлена для удобства), на котором можно увидеть описанные ранее проблемы. В данном примере без использования информации о куки-файлах, встроенной в запрос ID сессии, или информации со стороны клиента не удастся определить, что строки 12 и 13 принадлежат одной сессии, т. к. они различаются IP-адресами. 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 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. Такой вид получаемых знаний может быть весьма 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. Эти страницы могут быть связаны не напрямую друг с другом (т. е. не иметь прямых ссылок друг на друга). Например, алгоритмом Apriori может быть обнаружена корреляция между пользователем, посетившим страницу с электронной продукцией, и пользователем, просматривающим страницу со спортивным оборудованием. Наличие или отсутствие таких правил в области бизнеса и маркетинга может помочь Web-дизайнерам перестроить Web-сайт. Ассоциативные правила могут также служить в качестве эвристических правил, по которым выполняется упреждающая выборка документов, для уменьшения времени ожидания загрузки страниц с удаленного сайта. 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.  
Кластеризация страниц позволяет выявить группы страниц с близким по смыслу содержимым. Эта информация полезна для поисковых машин и персональных ассистентов. В обоих случаях пользователям могут предлагаться гиперссылки в соответствии с их запросом и историей запрашиваемой информации. 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. Для этого могут быть использованы любые методы классификации: деревья решений, метод Байеса, SVM и др. Построение модели зависимостей также может широко использоваться применительно к Web. Целью такого анализа является разработка модели, включающей в себя наиболее значимые зависимости между разными переменными в области Web. Так, например, может быть построена модель зависимости между этапами, которые проходит посетитель, и фактом совершения покупки в интернет-магазине (т. е. модель, отличающая случайного посетителя от потенциального покупателя). Существует несколько вероятностных методов обучения модели, которые могут быть использованы для построения модели поведения пользователя при просмотре Web, включая скрытые модели Маркова (Hidden Markov Models) и Байесовские сети доверия (Bayesian Belief Networks).  
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.

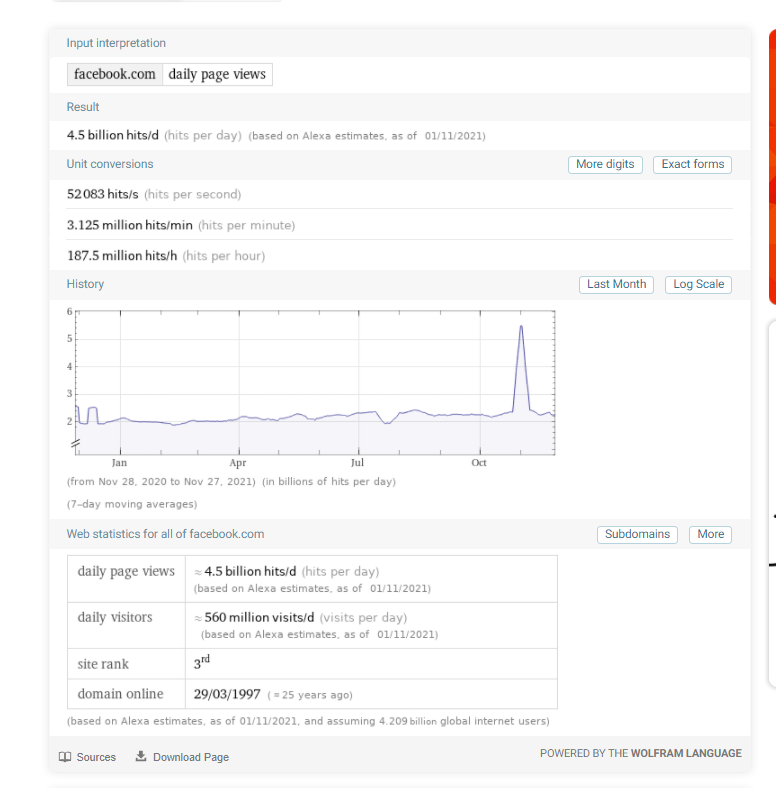
***Этап анализа шаблонов и их применение***  
Последним этапом в исследовании использования Web-ресурсов является анализ извлеченных шаблонов. **Целью анализа является отфильтровать наиболее интересные шаблоны и отбросить ничего не значащие шаблоны**. Методология анализа во многом зависит от области 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 для применения к нему OLAP-операций. Методы визуализации, такие как **раскрашивание или графическое изображение шаблонов**, могут выделять характерные шаблоны или тренды в данных. Контент или информация о структуре может быть использована для фильтрации страниц, используемых определенным образом, содержащих информацию определенного типа или страницы, имеющие определенную структуру гиперссылок.

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

1. источнику данных: сторона сервера, сторона клиента и прокси;
2. типу данных: структуры, контент и информация об использовании;
3. количеству пользователей: однопользовательские и многопользовательские;
4. количеству сайтов: один или множество сайтов;
5. области применения.

Выделяют следующие области применения систем анализа использования Web-ресурсов.

1. Персонализация
2. Улучшение систем
3. Модификация сайтов



1. Бизнес-интеллект

