**Выполнение визуального анализа данных**

Результаты, получаемые при анализе данных с помощью методов Data Mining, не всегда удобны для восприятия человеком. Во множестве классификационных или ассоциативных правил, в математических формулах человеку достаточно сложно быстро и легко найти новые и полезные знания. Из-за сложности информации это не всегда возможно и в простейших графических видах представления знаний, таких как деревья решений, дейтограммы, двумерные графики и т. п. В связи с этим возникает необходимость в более сложных средствах отображения результатов анализа. К ним относятся средства визуального анализа данных, которые в зарубежной литературе часто называют термином Visual Mining.

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.

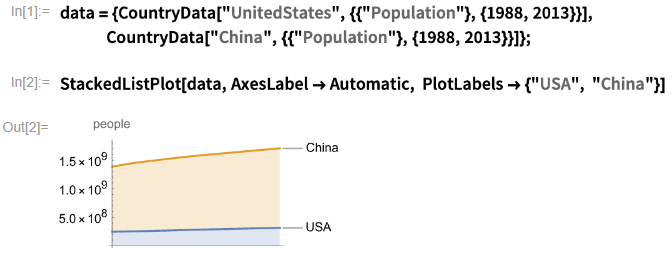
Визуальный анализ данных можно представить как процесс генерации гипотез. При этом сгенерированные гипотезы можно проверить или автоматическими средствами, или средствами визуального анализа.

Визуальный анализ данных обычно выполняется в три этапа:

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

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

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



**Характеристики средств визуализации данных**

Существует достаточно большое количество средств визуализации данных, предоставляющих различные возможности. Для выбора таких средств рассмотрим более подробно три основные характеристики:

1) характер данных, которые нужно визуализировать с помощью данного средства;

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

3) 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. 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:

1) одномерные данные — одномерные массивы, временные ряды и т. п.;

2) двумерные данные — точки двумерных графиков, географические координаты и т. п.;

3) многомерные данные — финансовые показатели, результаты экспериментов и т. п.;

4) тексты и гипертексты — газетные статьи, Web-документы и т. п.;

5) иерархические и связанные — структура подчиненности в организации, электронная переписка людей, гиперссылки документов и т. п.;

6) алгоритмы и программы — информационные потоки, отладочные операции 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, так и более сложные, основанные на сложном математическом аппарате. Кроме того, при визуализации могут использоваться комбинации различных методов. Выделяют следующие типы методов визуализации:

1) стандартные 2D/3D-образы — гистограммы, линейные графики и т. п.;

2) геометрические преобразования — диаграмма разброса данных, параллельные координаты и т. п.;

3) отображение иконок — линейчатые фигуры и звезды;

4) методы, ориентированные на пикселы — рекурсивные шаблоны, циклические сегменты и т. п.;

5) иерархические образы — древовидные карты и наложение измерений.

Пользователь должен иметь возможность работать с образами: видеть их с разных сторон, в разном масштабе и т. п. Для этого у него должны быть соответствующие возможности взаимодействия с образами:

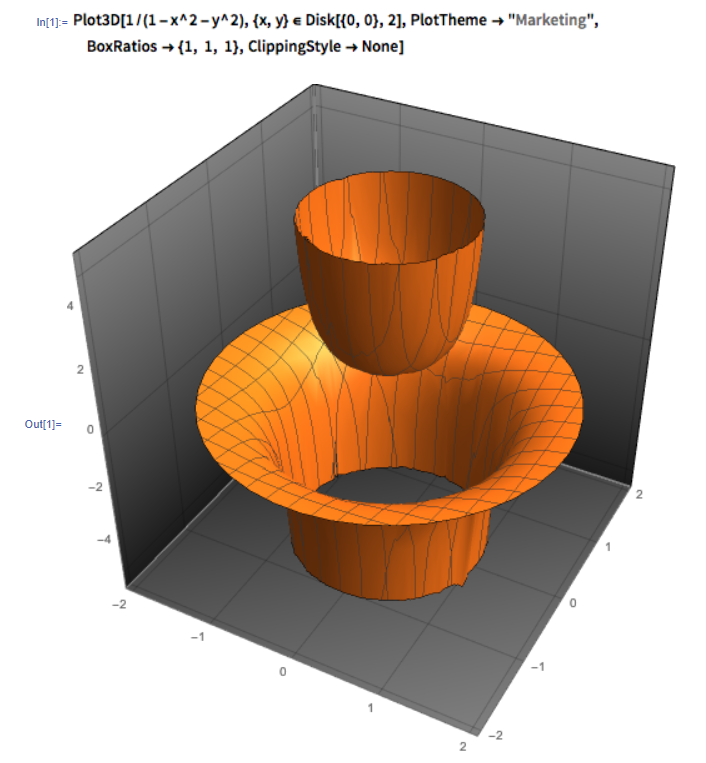
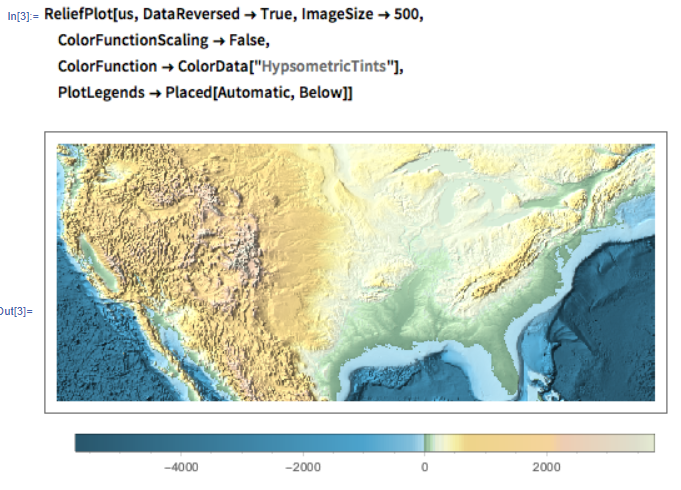
1) динамическое проецирование;

2) интерактивная фильтрация;

3) масштабирование образов;

4) интерактивное искажение;

5) интерактивное комбинирование.

**Методы визуализации**

Основная идея методов геометрических преобразований — визуализировать преобразования и проекции данных в декартовом и в недекартовом геометрических пространствах. К этим методам относятся:

1) точки и матрицы;

2) гипердоли;

3) поверхностные и объемные графики, контуры;

4) параллельные координаты;

5) текстуры и растры.

Матрица диаграмм разброса (Scatterplot Matrix) является комбинацией отдельных диаграмм разброса, что позволяет отображать более одного атрибута. Значения атрибутов отображаются в диагональных ячейках матрицы, а остальные ячейки представляют собой отношения между ними.  
Гипердоли являются модификацией матрицы диаграмм разброса. Основная концепция та же, за исключением того, что в ячейках матрицы отображаются скалярные функции.

На ранних фазах визуального анализа большие величины непрерывных данных могут отображаться с помощью объема. Объемный рендеринг позволяет пользователю видеть внутреннюю часть объемных графиков. Цвета, яркость и полупрозрачность используются, чтобы изобразить различия распределений и значения атрибутов. Подвижность объемных графиков используется, чтобы визуализировать различные их слои.

Еще одним распространенным методом геометрических преобразований является метод параллельных координат. Данный метод предполагает представление атрибутов параллельными линиями на не декартовой плоскости. Данные представляются кривыми линиями, которые пересекают линии атрибутов. Точки пересечений соответствуют значениям соответствующих атрибутов отображаемых данных.

Текстурная и растровая визуализации используют способность человека к преаттентивному восприятию информации. Такой метод в совокупности с различными визуальными свойствами позволяет отобразить большое количество атрибутов.

**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:

1) линейчатые фигуры;

2) "лица Чернова";

3) цветные иконки;

4) глифы и др.

Линейчатая фигура представляет собой иконку с некоторым количеством ветвей (линий).

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

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.



**Иерархические образы**

Иерархические образы используются для отображения иерархий и отношений в данных. Они применяются в следующих методах:

1) иерархические оси;

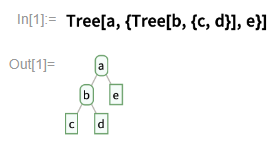
2) наложение измерений;

3) деревья.

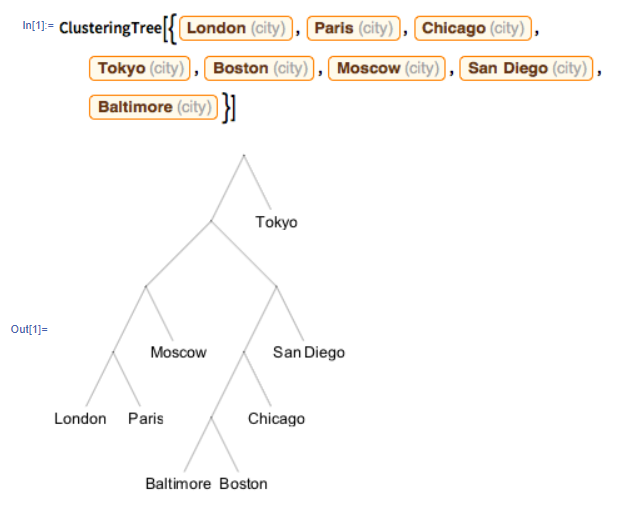
Оси, представляющие каждый атрибут, накладываются горизонтально, при этом первое место в иерархии отводится наиболее изменяемому атрибуту. Такой метод может отображать до 20 атрибутов на одном экране. Для большого количества данных метод может использовать подпространственное масштабирование и, тем самым, походить на древовидную структуру.

Основная идея метода наложения измерений заключается во вставке одной координатной системы в другую. Иными словами, два атрибута формируют внешнюю систему координат, два других атрибута формируют другую систему координат, встроенную в предыдущую, и т. д. Этот процесс может быть повторен несколько раз.

Наглядность данного метода заключается в зависимости от распределения данных внешней системы координат. Поэтому измерения, которые используются для внешней системы координат, должны быть выбраны тщательно. Первыми нужно выбирать наиболее важные измерения.



Пример графической визуализации кластеров в виде дерева представлен ниже.



**Проблемы анализа информации 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 следующие проблемы:

1) небольшой процент действительно нужной информации среди множества ссылок, которые предоставляют поисковые системы;

2) 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.

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.

Поиск ресурсов предполагает поиск различных Web-источников (преимущественно текстовых) по ключевым словам. Данный этап разделяют на два класса: поиск документов и поиск сервисов.

Большинство работ по поиску ресурсов сводится к автоматическому созданию поисковых индексов Web-документов. Для этих целей создавались роботы, индексирующие слова в документах и хранящие вычисленные индексы для дальнейшего их использования при обработке запросов пользователей. Наиболее популярными роботами считаются WebCrawler и AltaVista. Они способны сканировать миллионы документов и хранить индексы слов в этих документах. Существует много различных индексов, которые в настоящее время активно используются.

После того как ресурсы найдены, из них должна быть извлечена информация, подвергаемая анализу. Часто этот этап называют препроцессинг, т. к. он заключается в подготовке найденных ресурсов непосредственно к анализу. Такая подготовка заключается в преобразовании текстов, путем удаления стопслов, стеммингов, извлечением фраз и словосочетаний и т. п. Другими словами, результатом данного этапа должна быть информация, пригодная для анализа.

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.

**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 и извлечение информации. 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. Как уже упоминалось, для решения этой задачи строятся поисковые индексы. Для их построения используются различные методы, включающие в себя моделирование, классификацию и кластеризацию документов, фильтрацию и др. При этом для классификации и кластеризации документов могут быть использованы методы Data Mining. С этой точки зрения можно говорить, что Web Mining является частью технологии IR.

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.

IE основное внимание уделяет структуре текстового документа, пытаясь извлечь из него ключевые понятия. Таким образом, IE и Web Mining может находиться в разных отношениях друг к другу.

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

**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.

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.

Каждому слову во множестве ставится в соответствие некоторое свойство. Данное свойство может иметь или логический тип, отражающий наличие или отсутствие слова в тексте, или числовое значение, отражающее частоту появления слова в тексте. Последующая обработка может быть связана с удалением пунктуации, нечастых слов, стоп-слов и др. Уменьшение числа свойств возможно за счет применения различных методов выбора свойств, основанных на расчете следующих метрик:

1) информационного прироста (information gain);

2) полного количества информации (mutual information);

3) перекрестной энтропии (cross entropy);

4) вероятности успешного исхода (odds-ration).

**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:

1) гипертекстовая классификация;

2) классификации и кластеризации;

3) изучение отношений между Web-документами;

4) извлечение шаблонов или правила;

5) поиск шаблонов и слабоструктурированных данных.

**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.

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.

Методы извлечения Web-контента для целей базы данных используют представления, которые отличаются от представлений, используемых для целей информационного поиска. Данные методы в основном используют представления в виде модели объектного обмена.

С помощью Wolfram Mathematica можно получить информацию для заполнения баз данных.





**Исследование использования Web-ресурсов**

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

1) Препроцессинг.

2) Извлечение шаблонов.

3) Анализ шаблонов.

Выделяют следующие типы данных, применяемые в задаче исследования использования Web-ресурсов:

1) использование — данные, которые описывают использование страниц, такие как IP-адресса, ссылки на страницы, а также дату и время доступа к ним;

2) пользовательские профили — данные, которые обеспечивают демографическую информацию о пользователе, а также регистрационную информацию.

Данные об использовании Web собираются в различных источниках, которые можно разделить на следующие основные группы: серверы, клиенты, прокси.

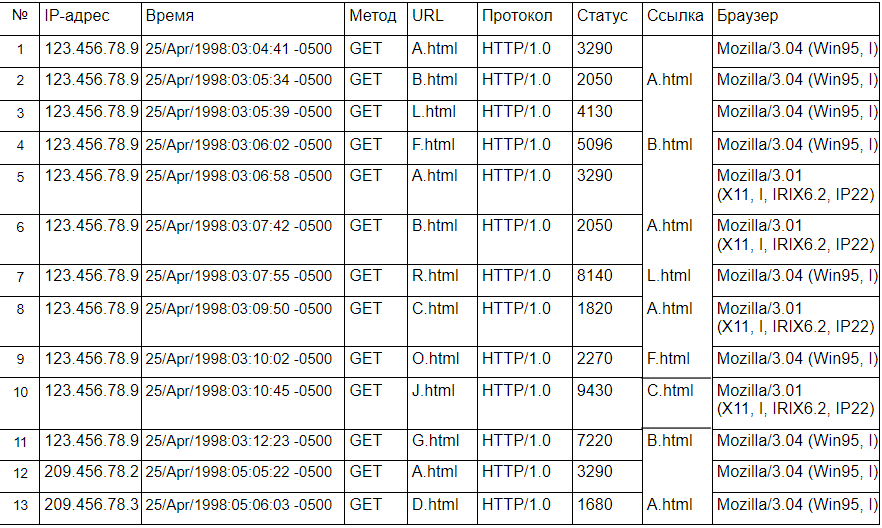
На стороне сервера информация для анализа может извлекаться из логов, трафика сервера, куки-файлов, запросов пользователей и др.

**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.

При условии, если удалось идентифицировать пользователя, 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.

Пример серверного лога приведен ниже.



**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 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 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. Кластеризация страниц позволяет выявить группы страниц с близким по смыслу содержимым.

Методы классификации могут быть использованы для развития профилей пользователей, относящихся к определенному классу или категории.

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.

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.

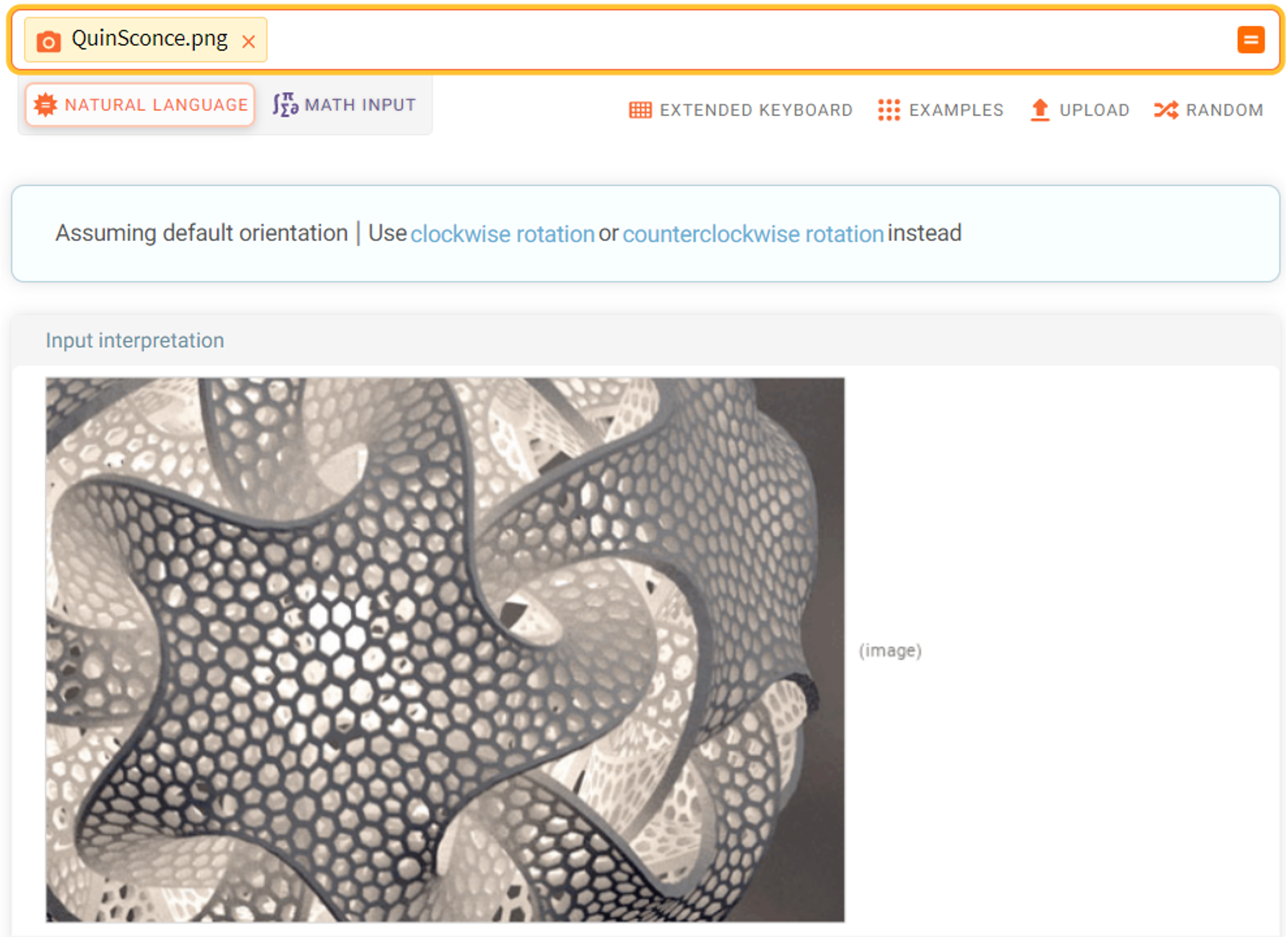
Моделирование использования Web-ресурсов позволяет обеспечить не только теоретическую основу поведения пользователей, но и помочь в предсказании.

**Этап анализа шаблонов и их применение**

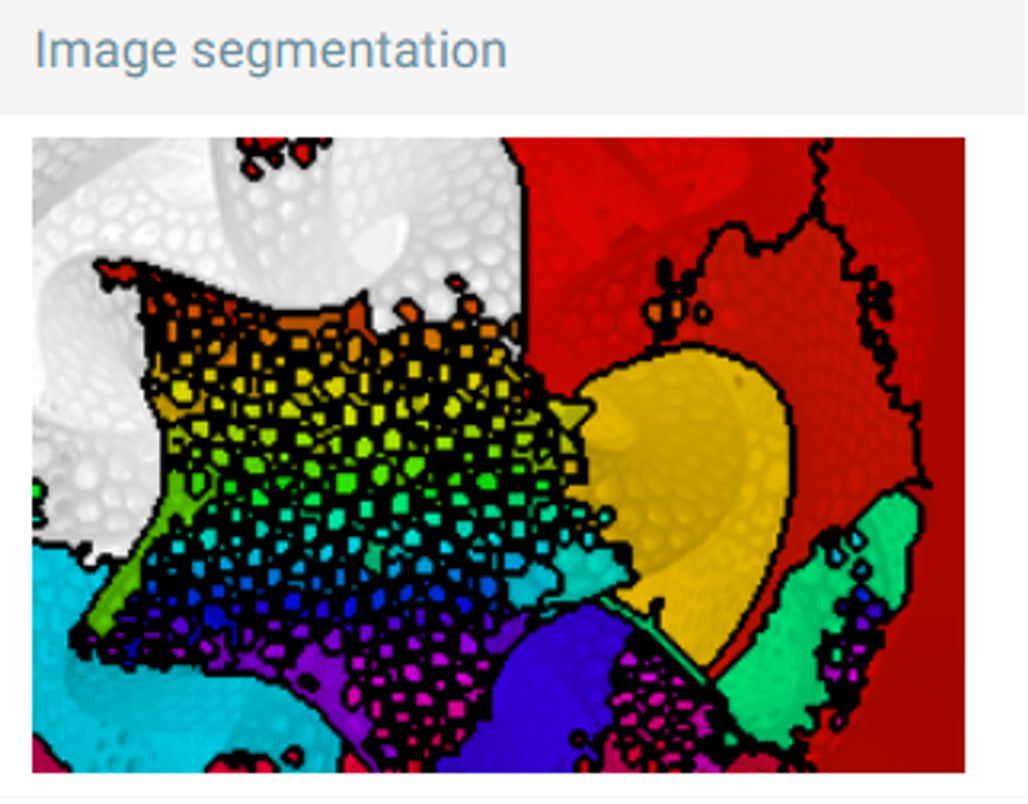
Последним этапом в исследовании использования Web-ресурсов является анализ извлеченных шаблонов. Целью анализа является отфильтровать наиболее интересные шаблоны и отбросить ничего не значащие шаблоны. Методология анализа во многом зависит от области применения, в которой он выполняется. Более общей 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.

Пример работы нейронных сетей. Процедура колоризации изображения.

Исходное изображение представлено ниже.

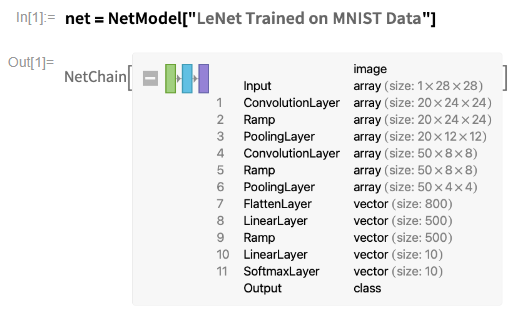


Результат колоризации изображения представлен ниже.

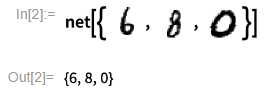


Как можно увидеть из фото, нейронная сеть обучена недостаточно хорошо, чтобы выдавать оптимальные результаты.

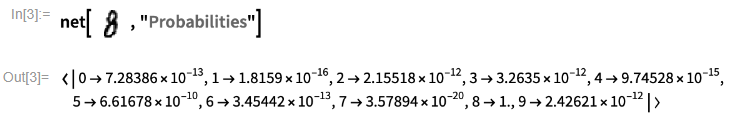
Пример работы нейронной сети по идентификации введенных вручную цифр приведен ниже.



Вводим рукописные «6, 8, 0». Нейронная сеть верно идентифицирует данные символы и выводит их в стандартном виде.



Также нейронная сеть способна выдать вероятности интерпретации введенного символа в качестве члена одного из кластеров. Как можно увидеть, вероятность интерпретации введенной цифры «8» максимальна, т.к. нейронная сеть была обучена на выборке данных, включающих данный экземпляр.



Пример получения информации о химической структуре кофеина представлен ниже.

