

НЕМНОГО ИСТОРИИ

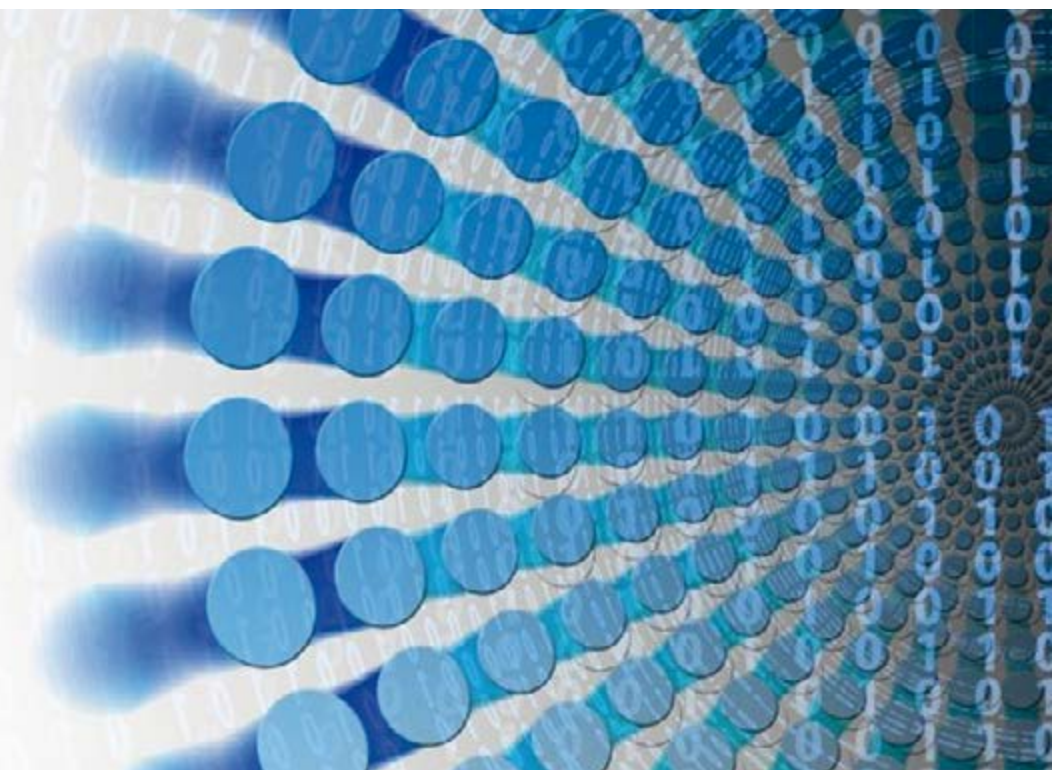
Изначально парсинг использовался для сбора и анализа текстовой информации, но в процессе развития информационных технологий он начал активно использоваться для автоматической обработки больших объемов данных. Сейчас в основном применяют данную технологию для автоматического наполнения сайтов контентом, сбора и анализа цен в интернет-магазинах или сайтах конкурентов, сбора данных из резюме и объявлений, размещенных на специализированных порталах, автоматического сбора отзывов и комментариев, а также упоминаний компании или бренда в социальных сетях или на форумах.

В последнее время технологии, основанные на парсинге, активно применяются также и для сбора персональных данных, и часто сторонки компании или сервисы собирают и контактные данные, такие как телефонные номера, адреса электронной почты и многое другое, относящиеся к персональным данным. Хотя в последнее время активно принимаются законы, ограничивающие сбор и обработку персональных данных.

Но, к сожалению, многих это не останавливает. А как мы знаем, такие данные часто используются недобросовестными компаниями для всякого рода мошенничества.

Сегодня существует достаточно большое количество специализированных программ и даже онлайн-сервисов для парсинга. Приведем лишь некоторые программы для примера.

- Mozenda
- Netpeak Checker
- Netpeak Spider
- Scrapy
- iDatica
- Automation Anywhere
- Octoparse
- BeautifulSoup
- Scraper API
- Webharvy
- ParseHub
- Diffbot
- Content Grabber
- FMiner
- Import.io



- Visual Web Ripper
- И другие...

Кто-то использует готовые продукты, кто-то разрабатывает и внедряет собственные, в зависимости от задач, которые необходимо решать. Но суть всех этих инструментов, специализированных или с широким функционалом, сводится к сбору, сохранению, обработке и анализу информации.

С точки зрения маркетинга, парсинг наиболее востребован для таких задач, как анализ сайта для SEO-оптимизации, сбор текстов и ключевых слов, сбор информации о ценах на продукцию у конкурентов, отслеживание форумов и социальных сетей на отзывы о компании или бренде, автоматическом наполнении товарами каталога или интернет-магазина и т.д.

В рамках данной статьи не будут подробно рассматриваться функциональные возможности или особенности работы разных инструментов и программ, всю подробную информацию можно найти на сайтах этих продуктов, так же как и инструкции и примеры использования.

Более подробно будет рассказано о практическом применении возможностей парсинга для маркетинга и автоматизации процессов сбора и обработки информации.

ПАРСИНГ НА ПРАКТИКЕ

На практике маркетологи часто сталкиваются с проблемами сбора информации о конкурентах (цены, продукты, каталоги и т.д.). Анализ их сайтов и рекламных кампаний, поиск «битых» ссылок и страниц с недостоверной информацией, автоматическое создание базы данных по продуктам или услугам для дальнейшего наполнения своих каталогов, а также другие задачи напрямую связаны со сбором и обработкой информации в интернете.

Следить за конкурентами, уровнем цен и новинками в товарах или услугах можно, конечно, и вручную, просматривая сайты и собирая информацию в таблицы для дальнейшего анализа. Но мы-то живем уже в XXI веке, и автоматизация процессов проникает все глубже и шире в нашу жизнь и работу. Поэтому совсем неэффективно делать работу, которую можно доверить компью-





теру самостоятельно. Тут на помощь приходят специальные программы или сервисы парсинга. Настроив их один раз на сайты или каталоги, которые вам необходимо отслеживать, вы можете выбрать периодичность, с которой нужно обновлять информацию и в каком виде вам ее предоставить. Таким образом вы можете мониторить всю необходимую информацию, получая готовые отчеты с собранными данными, например, в формате электронных таблиц.

Стоит отметить, что этот инструмент также удобен для мониторинга цен у дилеров, если ваша компания придерживается правила единой цены.

Теперь представьте, что вы наполняете свой каталог товарами, у вас есть много поставщиков, а ваш интернет-магазин содержит много тысяч товаров. Конечно, если у вас небольшой каталог или интернет-магазин, то заполнить его товарами можно и вручную, главное, чтобы времени для этого было достаточно. Но гораздо удобнее «спарсить» всю информацию с сайтов поставщиков или интернет-каталогов, таких как Onliner или Яндекс.Маркет, например.

Оперативно можно получить готовые данные в нужном нам формате, а также сверить

эти данные, если мы берем одинаковые данные из нескольких источников, и найти различия или неточности. Далее так же оперативно можно обновлять данные, опираясь на обновления на сайтах поставщиков или интернет-каталоги.

Еще один пример, где активно применяется парсинг, – это социальные сети: отслеживание трендов в конвенте, наблюдение за лидерами мнений или известными блогерами, артистами и т.д. Также вполне реальны:

- автоматический поиск упоминания вашей компании, торговой марки на форумах и в социальных сетях;
- автоматический сбор отзывов, оставленных на популярных ресурсах, о ваших продуктах или услугах и много других задач, на которые способна ваша фантазия.

А что же насчет формирования базы данных лиц, принимающих решения, у ваших потенциальных клиентов или партнеров? Это реализуется в виде сбора информации из архивов резюме на сайтах поиска работы, контактов в социальных сетях и других источников с подобной информацией.

Если ваша компания поставляет оборудование или расходные материалы, пирсингом можно собирать информацию о тендерах и закупках на специализированных площадках, анализировать эту информацию и получать достаточно данных для понимания объемов потребления потенциальных клиентов или приблизительных сроков следующих закупок. Зная объемы потребления расходных материалов, можно рассчитывать объемы выпуска или реализации у компаний, которые вас интересуют.

Это лишь небольшая часть возможных способов применения этой технологии.

А ЭТО ЗАКОННО?

Что же с законностью данной технологии? Сбор информации, размещенной в открытых источниках, не запрещен. Информация, которая размещена публично, доступна и постоянно собирается и обрабатывается. Есть ограничения, связанные с использованием собранной информации, а также с этическими вопросами, особенно что касается персональной информации.

Сегодня многие страны активно внедряют законы, связанные с защитой персональных данных. Это накладывает определенные ограничения как на доступ к данной информации, так и на ее

использование. Также хочется отметить момент, связанный с использованием информации, которая попадает под защиту авторских прав, – тексты, изображения и т.д. Поэтому лучше убедиться, что, например, изображения или статьи, которые вы планируете собрать в автоматическом режиме, пригоднее для дальнейшего использования без нарушения авторских прав.

Все было бы очень просто, если бы только мы знали про такой инструмент, как парсинг, и использовали его. Многие компании защищают свои данные от парсинга, добавляют ограничения на количество запросов, вводят дополнительную авторизацию или подтверждение с помощью капчи и т.д. Но современные программы и сервисы для парсинга умеют обходить многие из таких систем защиты, использовать прокси-серверы или другие инструменты для успешной работы, хотя при этом, конечно, может снижаться их скорость работы. Умеют такие программы и сервисы также подменять user-agent, представляясь для сервера роботами Google или Yandex.

В любом случае парсинг – весьма удобный инструмент в маркетинге, SMM, наполнении сайтов и многих других областях. Согласитесь, это удобно и очень экономит самый ценный и ограниченный ресурс, которым мы располагаем, – время.

