УДК 004.65

Прокопенко В.В.,
студент магистратуры
2 курс, факультет «Аппаратного,
программного и математического
обеспечения вычислительных систем»
Российский технологический университет
Россия, г. Москва

ПАРСИНГ КАК ОДИН ИЗ ИНСТРУМЕНТОВ ИНТЕЛЛЕКТУАЛЬНЫХ БАЗ ДАННЫХ

Аннотация: *в данной статье рассматривается использование парсинга в интеллектуальных базах данных для управления ценовыми параметрами. Рассказывается о возможных целях применения парсера о, схеме работы, актуальных способах и законности сбора информации.*

Ключевые слова: *парсинг, интеллектуальные базы данных, ценообразование, СУБД.*

Annotation: *this article discusses the use of parsing in intelligent databases for managing price parameters. It describes the possible purposes of using the parser, the scheme of work, current methods and the legality of collecting information.*

Key words: *parsing, intelligent databases, pricing, DBMS.*

Любой предприниматель устанавливает цену на свой товар и использует ее как средство достижения поставленных целей и один из элементов своей конкурентной политики. Важная роль ценообразования для предприятия или фирмы бесспорна, а современная ценовая политика очень разнообразна.

Существуют два подхода к процессу рыночного ценообразования: установление индивидуальных либо единых цен.

Индивидуальная цена определяется на договорной основе в результате переговоров между покупателем и продавцом, приводящих к согласованию интересов обеих сторон. В случае единой цены все покупатели приобретают товар по одинаковой цене. Установление единых цен для всех потребителей может быть связано с особенностями рынка данного товара либо с технической сложностью и большими дополнительными издержками при дифференциации цен. Единые цены предпочтительно устанавливать прежде всего там, где предприниматель выводит стандартизированный продукт серийного производства на массовый рынок. В этих условиях важно, чтобы массовый потребитель знал цену, мог сравнить ее с ценой конкурирующих товаров и относительно легко принять решение о покупке.

Долгое время цена была важнейшим критерием принятия потребительских решений. Для стран с относительно невысоким уровнем жизни, для более бедных слоев населения, а также применительно к товарам массового спроса это и сегодня справедливо. В последние десятилетия получили широкое развитие и другие, неценовые факторы конкуренции. Однако цена остается существенным элементом конкурентной политики, оказывающим большое влияние на рыночное положение и прибыль предпринимателя.

Ценообразование — единственный элемент системы маркетинга, не сопряженный со значительными расходами (как, например, при организации рекламы). Вместе с тем ценовая политика многих предпринимателей оказывается зачастую недостаточно проработанной и содержит много ошибок. Наиболее часто встречающиеся ошибки: ценообразование излишне ориентировано на издержки; цены недостаточно часто приспосабливаются к изменению рыночных условий; цена рассматривается в отрыве от других элементов системы маркетинга (так называемого маркетингового комплекса); цены недостаточно структурируются по отдельным вариантам продукта и сегментам рынка.

Соответственно база данных которая может содержать данные по всем ценовым показателям о товарах от различных поставщиков могла бы

существенно облегчить задачи поиска и сравнения стоимости, а так же функциональных данных изделий или потребительских услуг. Именно для этих нужд возникла необходимость в написании базы данных ценовых показателей, а так же программных модулей по сбору информации от проверенных поставщиков об актуальных предложениях; модуля противодействия защите сайтов от автоматического сборщика информации. .

Интеллектуальные базы данных

На заре использования компьютеров форматы хранения данных и средства для манипулирования ими изобретались программистами индивидуально для каждого случая. Это неудобно и неэффективно, и вскоре были предложены унифицированные способы хранения данных (модели данных) и разработаны унифицированные системы доступа к данным (системы управления базами данных — СУБД). Рассматривая концепцию базы данных (БД) с самой общей точки зрения, можно отметить, что БД поддерживает три основные группы операций: занесение данных в базу, поиск данных в базе и извлечение данных из базы, причем извлекаются те данные, которые до этого были занесены. Были предложены и используются несколько идей, к которым это ограничение неприменимо. Одной из них является идея интеллектуальной базы данных.

Интеллектуальная базы данных (Intelligent Database) предоставляет эффективный способ хранения, поиска и извлечения большего числа фактов, чем те, которые были изначально загружены в базу.

Общая структура интеллектуальной базы данных представлена на рис. 1.

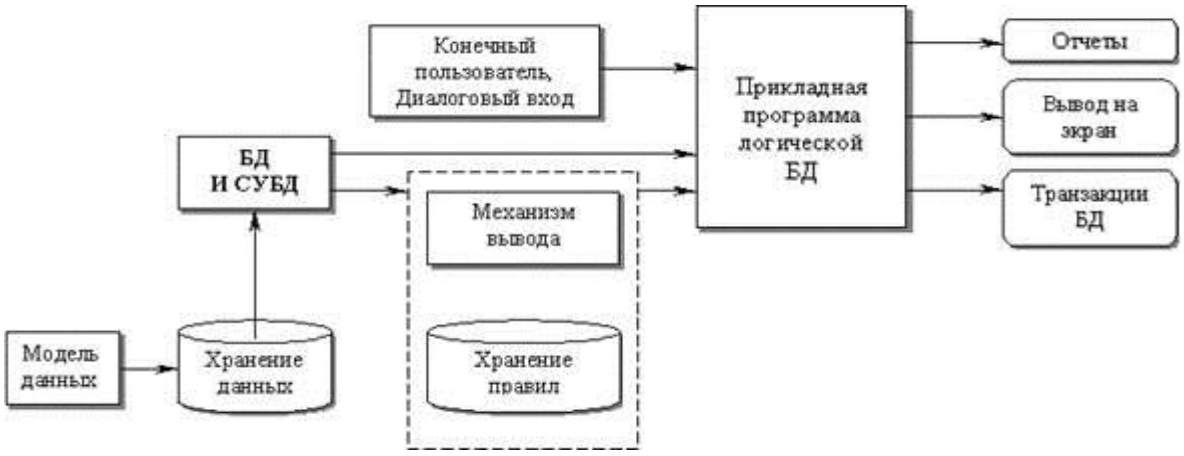

Рисунок 1. Общая структура интеллектуальной базы данных

В экономике интеллектуальные базы данных заняли очень важную роль и позволяют во много раз сократить время обработки задач ценообразования и получения требуемой информации.

На сегодняшнее время найти товар или услугу по сети интернет не составит большого труда от рядового пользователя компьютера, но если требуется найти оптимальную цену или так называемую «золотую середину» между ценой и качеством, то на данную задачу придётся потратить достаточное количество времени, которого зачастую не хватает. Для этих целей актуально создать базу данных, которая с помощью интеллектуальных элементов способна выдать требуемую пользователю информацию в кратчайшие сроки.

Если же с базой данных всё довольно понятно, работает связка запрос-ответ, то что на счёт анализа различных предложений с разных сайтов или определенных, доверенных источников. Рассмотрим же один из модулей, который должен отвечать за наполнение нашей базы актуальными данными и получим ответ на вопрос, стоит ли автоматизировать данную задачу или использовать так называемый ручной труд.

Парсер как один из модулей интеллектуальной базы данных

Парсер (парсинг) товаров – специальная программа (или алгоритм), позволяющая собирать необходимые сведения с заранее определенных интернет-магазинов. Чаще всего их используют при наполнении интернет-магазинов данными и мониторинге цен конкурентов. Что такое термин “парсинг” – это обработка информации в соответствии с определенным алгоритмом. При самостоятельном поиске вам потребуется вручную заходить на каждый предложенный поисковиком сайт в Интернете и собирать оттуда данные, систематизировать и выявляя необходимые. Парсер полностью выполняет все эти процессы.

В первую очередь, целью парсинга является ценовая "разведка", ассортиментный анализ, отслеживание товарных акций. “Кто, что, за сколько и в каких количествах продаёт?” – основные вопросы, на которые парсинг

должен ответить. Если говорить более подробно, то парсинг ассортимента конкурентов или того же Яндекс.Маркет отвечает на первые три вопроса.

С оборотом товара несколько сложней. Однако, некоторые компании которые открыто предоставляют информацию об ежедневных объемах продаж, заказах, или остатках товара, на основе которой не сложно составить общее представлении о продажах, иногда данные сведения могут быть искажены для повышения спроса или скрытия информации от реальных остатках. Смотрим, сколько было товара на складе сегодня, завтра, послезавтра и так в течении месяца и вот уже готов график и динамика изменения количества по позиции составлена (оборачиваемость товара фактически). Чем выше динамика, тем больше оборот.

Можно, конечно, сослаться на перемещение товаров между точками. Но суммарно, если брать, например, Москву — то число не сильно изменится, а в существенные передвижения товара по регионам верится с трудом.

С объемами продаж ситуация аналогична. Есть, конечно, компании, которые публикуют информацию в виде много/мало, но даже с этим можно работать, и самые продаваемые позиции легко отслеживаются. Особенно, если отсечь дешёвые позиции и сфокусироваться исключительно на тех, что представляют наибольшую ценность.

Во-вторых, парсинг используется для получения контента. Многие зацикливаются на том, что парсинг – это именно воровство контента, хотя это совершенно не так. Парсинг – это всего лишь автоматизированный сбор информации, не более того. Например, парсинг фотографий, особенно с “водяными знаками” – это чистой воды воровство контента и нарушение авторских прав. Потому таким обычно не занимаются (в своей работе большинство ограничивается сбором ссылок на изображения, не более того, но иногда требуется отследить наличие видео на товар и дать ссылку и т.п.).

Рассмотрим также сбор описания книг, например, с популярных книжных порталов. Здесь уже ситуация не так однозначна с правовой точки зрения. С одной стороны, использование такого описания может нарушать авторское

право, особенно если описание каждой карточки с товаром было нотариально заверено (что слишком сомнительно — ведь может и не быть заверено, исключение — небольшие ресурсы, которые хотят затаскать по судам воров контента). В любом случае, в данной ситуации придётся сильно "попотеть", чтобы доказать уникальность этого описания. Чтобы данных проблем не было, некоторые разработчики используют синонимайзеры, которые так или иначе меняют текст на примерно похожий, как в лучшую сторону так и в худшую.

Ещё одно из применений парсинга довольно оригинально – “самопарсинг”. Это парсинг собственного ресурса, преследуя несколько целей. Для начала – это отслеживание того, что происходит с наполнением сайта или базы данных: где битые ссылки, где описания не хватает, дублирование товаров, отсутствие иллюстраций и т.д. Полчаса работы парсера — и вот готовая таблица со всеми категориями и данными. Удобно! “Самопарсинг” можно использовать и для того, чтобы сравнить остатки на сайте со своими складскими остатками, есть такой вариант использования для отслеживания сбоев выгрузок на сайт. Ещё одно применение “самопарсинга”, с которым мы столкнулись в работе — это структурирование данных с сайта для выгрузки их на маркетплейс. Пользователю так проще было сделать, чем вручную этим заниматься.

Также парсятся объявления, например, на актуальных площадках частных объявлений. Цели тут могут быть как перепродажи баз риелторам или туроператорам, так и откровенный телефонный спам, ретаргетинг и т.п. В случае с площадками для объявлений это особенно явно, т.к. сразу составляется таблица с телефонами пользователей, несмотря на то, что некоторые площадки подменяют телефоны пользователей для защиты пользователей и публикует их в виде изображения, но от поступающих звонков все равно никуда не уйти.

Законность автоматического сбора информации

В российском законодательстве нет статьи, запрещающей парсинг. Запрещен взлом, DDOS, воровство авторского контента, а парсинг – это ни то, ни другое, не третье и, соответственно, он не запрещен.

Некоторые люди воспринимают парсинг как DDOS-атаку и относятся к нему с сомнением. Однако, это совершенно разные вещи, и при парсинге программист, напротив, стараемся как можно меньше нагружать целевой сайт и не навредить бизнесу. Как в случае со здоровым паразитизмом – объект не должен пострадать, чтобы не пострадал паразит.

Обычно парсят крупные сайты, из топа 300-500 сайтов России. На таких сайтах высокая посещаемость, как правило, несколько миллионов в месяц, может даже и больше. И на таком фоне парсинг одного товара в секунду или в две практически незаметен, нет смысла чаще парсить, 1-2 секунды не товар - это оптимальная скорость для крупных сайтов. Соответственно, и намека на DDOS-атаку в наших действиях нет.

Парсинг – это лишь сбор того, что пользователи могут своими глазами увидеть на сайте и скопировать к себе руками. Таким образом, под статью об авторском праве могут попасть лишь действия с уже собранной информацией, т.е. действия владельца собранной информации. Простым языком человек это делает долго медленно и с ошибками, а парсер – быстро и не ошибается. Что же делать, когда речь касается сбора данных с крупных международных торговых площадок? Человеку просто не под силу такая задача, и парсинг – единственный выход.

Актуальные способы парсинга

К актуальным способам можно отнести уже готовые сервисы по парсингу различных сайтов на предмет товаров или различной информации, но зачастую они либо стоят дополнительных средств, что может быть не так уж и затратно, в случае не частых запросов , но если рассматривать частые запросы для сохранения базы данных в максимально актуальном виде, то затраты весьма возрастут и есть смысл в разработке или написании собственного алгоритма на языке высокого уровня.

Схема работы парсера

Первым делом рассматривается исходный код страницы, программа проходит по нему, как по обычным словам, и находит некоторые соответствия,

которые записаны в ее программный код. Она сравнивает их, сопоставляет и сохраняет то, что нужно вам по определенным условиям. Последний шаг – сохранение в удобном формате данных. То есть какие-то программы или скрипты будут сохранять в SQl, какие-то – в XML, кто-то – в обычном TXT либо в табличном документе. Схема работы представлена на рисунке 2. Резльтат работы предоставлен на рисунке 3.

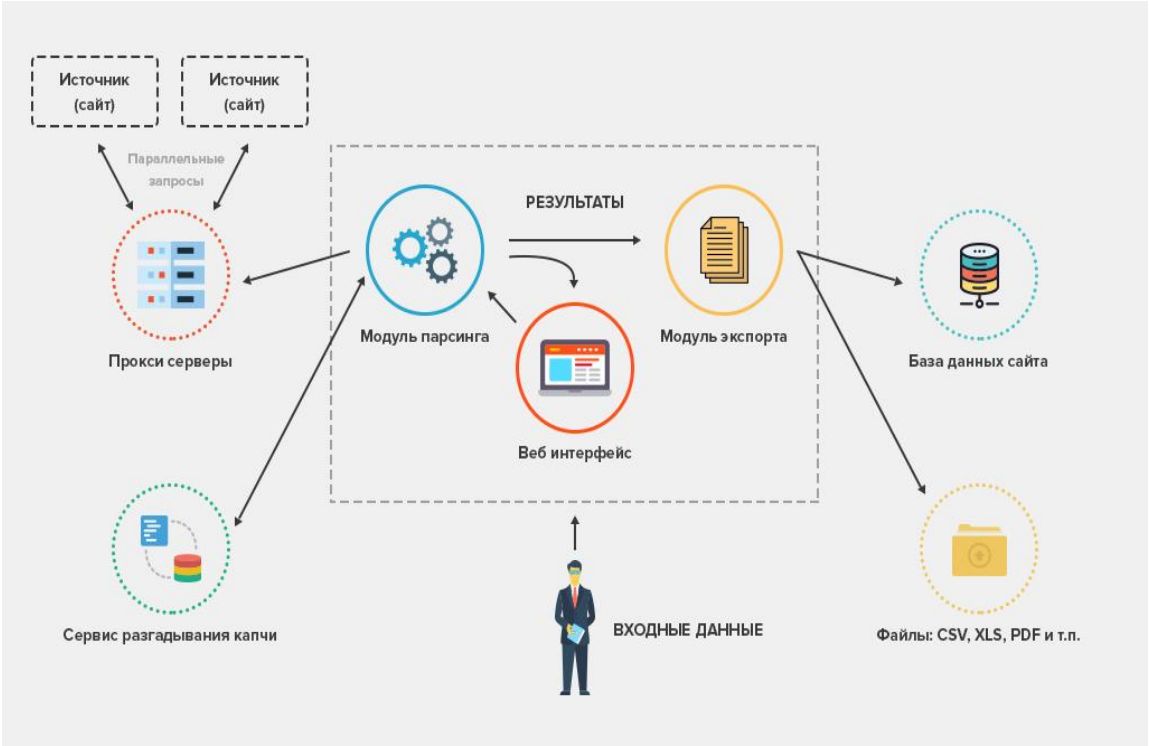

Рисунок 2. Схема работы парсера.

	Наименование	Марка	Страна	Производитель	Описание	Цена
2	Творог "Агуша" классический 4,5% с 6 месяцев	Агуша	Россия	ОАО "Вимм Билль Данн"	Продукт из натуральных компонентов.	28,80
3	Молоко "Агуша" стерилизованное с витаминами	Агуша	Россия	ОАО "Вимм Билль Данн"	Молоко Агуша стерилизованное с	40,40
4	Вода "ФрутоНяня" питьевая негазированная	ФрутоНяня	Россия	ОАО "Прогресс"	Вода ФрутоНяня питьевая	287,00
5	Вода "ФрутоНяня" питьевая негазированная	ФрутоНяня	Россия	ОАО "Прогресс"	Вода Фрутоняня негазированная	162,00
6	Творог "Агуша" детский фруктовый Яблоко-	Агуша	Россия	ОАО "Вимм Билль Данн"	Творог Агуша детский фруктовый с	35,30
7	Кефир "Агуша" для детского питания с 8 месяцев	Агуша	Россия	ОАО "Вимм Билль Данн"	Кефир Агуша для детского питания с 8	28,30
8	Вода "ФрутоНяня" питьевая негазированная	ФрутоНяня	Россия	ОАО "Прогресс"	Вода Фрутоняня питьевая,	194,00
9	Йогурт "Агуша" Клубника - банан 2,7% с 8	Агуша	Россия	ЗАО "ВБД"	Йогурт Агуша 2,7% с клубникой и	44,00
10	Вода "Святой Источник" Спортик природная	Святой Источник	Россия	ООО "Аква Стар"	Природная питьевая артезианская	238,00
11	Творог "Агуша" детский фруктовый "Груша" 3,9%	Агуша	Россия	ОАО "Вимм-Билль-Данн"	Агуша - забота о здоровье и развитии	35,60
12	Творог "Агуша" детский фруктовый Черника 3,9%	Агуша	Россия	ОАО "Вимм Билль Данн"	Творог Агуша детский фруктовый с	35,30
13	Кондиционер "Lenor" для белья детский	Lenor	Россия	ООО "Проктер энд Гэмбл -	Кондиционер для белья Lenor детский	299,00
14	Молоко "Агуша" источник кальция с 8 месяцев	Агуша	Россия	АО "Вим-Билль-Данн"	Молоко стерилизованное	30,20
15	Клей-карандаш "NoName" Gingko 36 г	NoName	Китай	Чжэцзян Цзиньюнь Каунти Синс-	Клей-карандаш Gingko легко склеивает	36,00
16	Йогурт "Агуша" Яблоко-груша 2,7% с 8-ми	Агуша	Россия	ОАО "Вимм Билль Данн"	Комфортное пищеварение.	45,60
17	Биотворог детский Тёма 5% классический с 6	ТЕМА	Россия	АО "Данон "	Обогащенный бифидобактериями для	30,30
18	Бумага туалетная "Mon Rulon" влажная детская	Mon Rulon	Россия	ООО"Авангард"	Создана специально для детей из	56,20
19	Смесь "Агуша"-2 кисломолочная для детей с 6	Агуша	Россия	ЗАО "ВБД"	Смесь Агуша-2 кисломолочная	38,60
20	Гель для мытья детской посуды "Ушастый нянь"	Ушастый нянь	Россия	ОАО "Невская косметика"	Гипоаллергенность и	80,40
21	Пюре "ФрутоНяня" гипоаллергенное из яблок	ФрутоНяня	Россия	ОАО "Прогресс"	Пюре яблочное натуральное для	36,90
22	Биокефир "Агуша" с 8 месяцев 3,2% 204 г	Агуша	Россия	ОАО "ВБД"	Биокефир Агуша обогащенный	29,30
23	Творог "Агуша" Персик с 6 месяцев 3,9% 100 г	Агуша	Россия	ОАО "Вимм Билль Данн"	Творог Агуша "Персик" 3,9% для	35,60
24	Сок "ФрутоНяня" яблоко осветленный без	ФрутоНяня	Россия	ОАО "Прогресс"	Сок Фрутоняня яблочный без	23,50
25	Компот "Агуша" процеженный Яблоко-Курага-	Агуша	Россия	ОАО "Вимм Билль Данн"	Напиток сокосодержащий "Компот	24,90
26	Сок "Сады Придонья" Яблоко прямого отжима	Сады Придонья	Россия	ОАО "Сады Придонья"	Сок Яблочный прямого отжима	21,60
27	Сок "Сады Придонья" Яблоко прямого отжима с	Сады Придонья	Россия	ОАО "Сады Придонья"	Сок яблочный прямого отжима.	64,70
28	Зубная паста "R.O.C.S." Kids малина и клубника	R.O.C.S.	Россия	ООО "ЕВРОКОСМЕД-Ступино"	Зубная паста R. O. C. S. Kids малина и	139,00

Рисунок 3. Пример файла в формате CSV полученного в результате парсинга

Полученный файл вносится в таблицу базы данных и ведётся дальнейшая обработка данных.

Итоги

Подводя итог можно сказать, что парсер – это отличный инструмент для баз данных с помощью которого можно получать данные для дальнейшей работы с ними. Имея множество применений, этот инструмент лучше всего себя раскроет в решении вопросов ценообразования и ценовых показателей, для которых цена является важным параметром и важна актуальность данных. База данных которая имеет в себе ценовые показатели от множества поставщиков и производителей сможет существенно облегчить процесс поиска и подбора необходимых товаров, а так же даёт возможность решения вопроса «золотой средины цены»

Библиографический список:

1. Макконнелл, С. Совершенный код.Мастер-класс/С. Макконнелл - 2-е изд.-СПБ.: БХВ-Петербург, 2017.-896с.

2. 10 инструментов, позволяющих парсить информацию с веб-сайтов, включая цены конкурентов + правовая оценка для России. [Электронный ресурс]. URL:https://habr.com/ru/post/340038/ (дата обращения: 25.05.2020).

3. Парсинг. Что это и где используется. [Электронный ресурс]. URL: https://ipipe.ru/info/parsing (дата обращения: 27.05.2020).

4. Развиваем интернет магазин: что нужно знать о мониторинге цен конкурентов?. [Электронный ресурс]. URL:https://www.plerdy.com/ru/blog/monitoring-cen/ (дата обращения: 01.06.2020).