УДК 65.011.56

ПРИМЕНЕНИЕ РЕГУЛЯРНЫХ ВЫРАЖЕНИЙ ДЛЯ СОЗДАНИЯ КАТАЛОГА ИНТЕРНЕТ-МАГАЗИНА

П.М. Пашков, О.А. Печень
Новосибирский государственный университет
экономики и управления «НИНХ»
E-mail: ppm@cn.ru, kamon@smtp.ru

Рассмотрена автоматизированная обработка частично структурированных данных с применением механизма регулярных выражений, показаны возможности такой обработки в применении к потребностям торговых организаций для формирования ценовых предложений. Рассмотрены основные проблемы и задачи при автоматизированной обработке ценовых предложений. Предложена архитектура информационной системы автоматизированного формирования ценовых предложений. Показана возможность использования предложенной системы в других областях для обработки и анализа частично структурированных данных.

Ключевые слова: регулярные выражения, частично структурированные данные, автоматизированная классификация, интернет-технологии, интернет-магазин, прайс-лист.

APPLICATION OF REGULAR EXPRESSIONS FOR CREATION OF THE CATALOGUE OF ONLINE STORE

P.M. Pashkov, O.A. Pechen
Novosibirsk State University of Economics and Management
E-mail: ppm@cn.ru, kamon@smtp.ru

The article considers the automated processing of partially structured data with use of the mechanism of regular expressions, possibilities of such processing in application to requirements of trade organizations for formation of price offers are shown. The main problems and tasks are shown at the automated processing of price offers. Offered the architecture of information system of the automated formation of price offers. Possibility of use of the offered system in other areas for processing and the analysis of partially structured data is shown.

Key words: regular expressions, partially structured data, automated classification, internet technologies, online store, price list.

Развитие интернет-технологий открывает новые возможности по обеспечению бизнес-информацией деятельности управленческих и операционных подразделений предприятий. В настоящее время практически все предприятия и организации представлены в Интернет с помощью своих корпоративных сайтов [2]. Наполнение и функциональность корпоративных сайтов достаточно разнообразны, но, как правило, все сайты содержат информацию для взаимодействия с внешней средой: клиентами, поставщиками и другими заинтересованными лицами. В совокупности эта информация имеет высокую ценность для поддержки маркетинговой деятельности. Однако данные, которые выкладываются на сайтах, имеют различные форматы и относятся к категории слабоструктурированной информации. В связи с этим в последнее время интенсивно разрабатываются информа-

ционные технологии для обработки частично структурированных данных. Одной из таких технологий являются регулярные выражения, которые находят все более широкое применение для ведения бизнес-аналитики в среде интернет [1, 3].

Регулярные выражения – это формальный язык для выполнения операций со строками и подстроками в тексте. Реализации механизма регулярных выражений включаются во многие системы, начиная с момента появления операционных систем семейства Unix. Например, поддержка регулярных выражений есть в языках программирования и средах разработки C++, Delphi, Perl, Java, HTML5, PHP, JavaScript, Python, Tcl, Ruby. В том числе компания Microsoft обеспечивает поддержку регулярных выражений в своих системах и средах, например в Microsoft Office и платформе Net Framework.

Использование регулярных выражений позволяет создавать гибкие шаблоны для поиска подстрок в текстах и выполнения операций копирования, замены, удаления текста. Без этого механизма для обработки текстов часто необходимо программирование сложных функций разбора (парсинга) и обработки текстов.

Например, так выглядят регулярные выражения для поиска и разбора артикула для ноутбуков некоторых производителей: ([HP]\s[a-z]{2}[0-9]{3}[a-z#]+).

Это выражение означает, что система будет искать подстроку, которая содержит:

– обязательно префикс HP и символ пробела либо табуляции;

– минимум две латинские буквы;

– три цифры;

– возможно еще несколько латинских букв.

([Sony]\s[a-z]{4,5}[0-9]{1,2}[a-z]{2,5}) – аналогичный пример для ноутбуков Sony:

– есть слово Sony и символ пробела либо табуляции;

– содержит четыре или пять букв;

– одну и две цифры;

– две или пять букв.

Программирование аналогичных по функциональности программных процедур потребует некоторых усилий, превышающих затраты на написание регулярных выражений.

Следует отметить, что реализация механизма регулярных выражений для каждого конкретного продукта может отличаться по возможностям. Например, используемая в Microsoft Office VBA библиотека Microsoft VBScript Regular Expressions должна подключаться в явном виде в настройках VBA, и реализует только основной синтаксис регулярных выражений, но даже в таком виде реализация механизма позволяет экономить усилия при решении прикладных задач.

Очень удобно и эффективно использовать регулярные выражения при обработке частично структурированных данных в прикладных информационных системах. Рассмотрим возможность автоматизации с помощью регулярных выражений критически важного для любой торговой компании вопрос – подготовка ценового предложения потенциальным покупателям или прайс-листа.

Для составления своего прайс-листа любая компания обрабатывает входные прайс-листы своих партнеров – поставщиков, конкурентов. Как правило, все прайс-листы предоставляются в формате Excel различной структуры. Классификация самих предлагаемых услуг и товаров выполняется в рамках общепринятых в данной предметной области соглашений, но тем не менее эти соглашения оставляют такую свободу маневра, что без предварительной обработки сравнить прайс-листы невозможно. Безусловно, крупные компании, занимающиеся разработкой прикладных информационных систем, например 1С, предлагают стандартизированные решения для универсальной классификации и автоматизированного обмена данных, но реально большинство торговых компаний не могут по разным причинам использовать такие решения, это скорее светлое будущее, а как говорят руководители предприятий – «работать надо сейчас». Более того, обрабатывать прайс-листы надо быстро, чтобы успеть дать свое предложение вперед конкурентов, и сама себестоимость такой обработки должна быть невелика.

Поскольку существует значительный спрос на автоматизацию обработки прайс-листов, компании-разработчики ПО давно предлагают свои решения. Например, существуют решения для 1С, интегрируемые в типовые конфигурации, специализированные продукты, например E-Trade PriceList Importer или PDS-Price. Все эти продукты решают задачу загрузки разноформатных прайс-листов и сведения данных по товарам в единую товарную матрицу, а на ее основе формирование сводного прайс-листа поставщиков (и конкурентов для анализа). На основе сводного прайс-листа (товарной матрицы) формируется собственный прайс-лист с необходимыми наценками, сроками и условиями поставки и т.д. Более того, фирма ElBuz (разработчик программ серии E-Trade) предлагает даже услугу ведения готового каталога товаров для небольших компаний. Все эти решения позволяют достаточно хорошо формировать сводную товарную матрицу. Также существует целый ряд небольших, менее мощных, продуктов для решения аналогичных задач:

– программа обработки прайс-листов PDS-Price (http://www.pdsprice.ru/);

– программа анализа прайс-листов TradesMan (http://www.kocheridi.ru/tradesman.htm);

– программа автоматизации обработки прайс-листов PriceMaker (http://www.pricemaker.aloha-system.com/);

– программа анализа Price-Guru FREE (http://www.price-guru.com/).

Следует упомянуть, что многие компании создают собственные средства обработки входных прайс-листов, учитывающие специфику работы компаний, их информационных систем, рынков на которых работают компании.

Анализ таких систем и опыт работы с ними позволил сформулировать наиболее актуальные, по нашему мнению, но еще не решенные задачи, по крайней мере, для сегментов торговли оргтехникой, компьютерами и компьютерными комплектующими, а также бытовой техникой:

1. Каждая компания имеет свой классификатор товаров, который может в той или иной степени отличаться от «общепринятых», более того, иногда необходимо поддерживать несколько классификаторов – например, классификации товаров для оптовой и розничной торговли отличаются прин-

ципиально. Для оптовой торговли важно учитывать производителя в классификации, а для розничной торговли важно учитывать потребительские параметры (например, для мониторов в оптовой торговле классификация идет по производителям, а в розничной – по диагонали монитора). Классификация для интернет-магазина также часто отличается ввиду некоторой специфики интернет-торговли.

2. В некоторых случаях возникает неоднозначное соответствие товаров. В частности, из-за описаний характеристик товаров, например, один поставщик может указывать цвет изделия, и изделия разного цвета в его прайс-листе считаются разными изделиями (да еще и с разной ценой), а другой поставщик (оптовая фирма) вообще не отличает цвет изделия и считает, что это одно изделие, и какое предлагает реально, из прайс-листа определить нельзя.

3. Прайс-листы, как правило, бывают самых разных форматов, это технически решается не очень сложно. Можно настроить свою систему обработки прайс-листов на формат поставщика. Но иногда поставщик (или конкурент) без предупреждения начинает менять формат своего прайс-листа (в «тяжелых» случаях он может делать это раз в несколько дней). Например, за рассылку прайс-листа у поставщика отвечает много менеджеров, а процесс создания прайс-листа не совсем формализован, цены у поставщика хорошие, и подстраиваться под «несущественные» пожелания клиентов он не хочет. В таких случаях приходится постоянно контролировать обработку меняющегося прайс-листа.

4. Одинаковые товары от разных поставщиков могут отличаться названиями и достаточно сильно, алгоритмы распознавания в этом случае могут быть ориентированы на поиск артикула производителя изделия (иногда он бывает в отдельной колонке, иногда включается в название), либо на уникальный код поставщика, который он указывает в прайс-листе. Могут быть и существенно более сложные алгоритмы установления идентичности товаров.

То есть автоматическое распознавание всех товаров, как правило, невозможно, об этом говорят сами разработчики программ, реально распознается лишь некоторая часть (около 20–30 % товаров), остальное обычно однократно сопоставляется вручную, и система запоминает составление (каким способом делает это конкретная система пока неважно). И тут появляется еще одна проблема – в некоторых областях (например, компьютерное оборудование или бытовая электроника) номенклатура поставщиков обновляется за месяц на 10–20 %. При объеме номенклатуры в 30 000 позиций это означает, что в день в предложениях поставщиков в среднем появляется около 150 новых позиций и около 150 уходит, при наличии 10 поставщиков (у которых позиции появляются неравномерно) в день необходимо делать примерно 1500 сопоставлений, из них автоматически выполняется всего около 300–500.

Следует также понимать, что для корректного составления прайс-листа менеджер должен знать товарные группы, и необходимо для эффективной продажи (например, для интернет-магазина) также составить описание товара для потенциального покупателя. Формирование описаний для каталогов товара и интернет-магазинов вообще отдельная тема, поскольку в ин-

тернет-магазине эффективно продавать товар без описания практически невозможно, и мы рассмотрим эту тематику, видимо, в следующей статье.

В итоге во многих компаниях работают целые отделы, где сидят менеджеры, формирующие прайс-листы и наполнение интернет-магазинов, при этом повторение товаров все равно встречается довольно часто (например, итернет-магазины компаний «Левел» и «Открытые технологии»).

Рассмотрим, что необходимо сделать, по нашему мнению, чтобы повысить качество автоматизированной обработки прайс-листов.

Во-первых, надо определить, что может быть взято за базу для автоматической идентификации товара. Как показывает анализ прайс-листов, можно использовать:

– артикул товара производителя;

– артикул товара поставщика;

– типовые наименования моделей (от производителя).

Артикул товара производителя. В случае бытовой техники работает очень хорошо, например, артикулы выглядят так:

(ELECTROLUX) EWH 30 Centurio, (ELECTROLUX) EWH 50 Quantum Slim, (ARISTON) ABS BLU R 80 V SLIM, (ARISTON) PRO 15 R/3, (ARISTON) ABS VLS PW 50.

То есть видно, что у каждого производителя есть система формирования названий моделей.

Поскольку производитель всегда присутствует в прайс-листе (но иногда только в названии товара, а не отдельной колонкой), то найдя регулярными выражениями производителя, можно описать также на основе регулярных выражений правила разбора названия товара и идентификации.

Для компьютерной техники ввиду большого количества вариаций моделей опять-таки удобно регулярными выражениями разобрать модель и классифицировать ее, приведем пример товарных позиций типичного прайс-листа:

ASUS RT-N65U 802.11n 300+450Mbps dual-band USB3.0 FTP/Media Server GigaLAN

ASUS RT-N66U 802.11n 450Mbps dual-band USB Printer/FTP Server GigaLAN.

Артикул товара поставщика. Любой серьезный поставщик также имеет свою систему классификации товаров, и если отношения с этим поставщиком долгосрочные, его прайс-лист стабилен по структуре, то можно для конкретного поставщика также создать набор регулярных выражений, которые позволят классифицировать товар и автоматически обновлять его предложения в сводной товарной матрице.

На начальном этапе только необходимо составить (во многом автоматически, тут помогут регулярные выражения, обрабатывающие артикул товаропроизводителя) соответствие между общей товарной матрицей и прайс-листом поставщика.

Типовые наименования моделей (*от производителя*). Бывают ситуации, когда не всегда указывается производитель и нельзя опереться на коды поставщика, т.е. в прайс-листе находятся «стандартные» сокращенные наименования поставщиков, например,

DWA-127/A1A, DPR-1061, TL-WN721N, TEW-736RE.

Тогда для работы регулярных выражений можно использовать структуру прайс-листа, и по категориям товара определять подходящий набор правил, а уже на их основе классифицировать товар.

Во-вторых, для всех поставщиков надо описать структуры их прайс-листов, так как в силу различий в структуре должны применяться те либо иные наборы регулярных выражений, и необходимо определять для конкретного прайс-листа расположение «полезных» информационных полей и правила «выкусывания» информации.

В-третьих, даже если мы не можем уверенно распознать товар, но мы можем идентифицировать хотя бы товарную группу, то отнеся туда товар и пометив его, мы сократим затраты на ручную обработку и разноску товара.

В-четвертых, подразумевается, что мы имеем некоторую «единую базу товаров», которая включает в себя классификации товаров от обрабатываемых поставщиков, коды и названия товаров производителей, коды поставщиков. Это позволяет использовать такую базу, как «базу знаний» по товарам для классификации в «тяжелых» случаях. Обсуждение организации базы также является темой отдельной статьи.

Следует подчеркнуть экономическую значимость оперативного и качественного формирования прайс-листа. Без каких-либо средств автоматизации (только с использованием средств Microsoft Office) менеджер может обрабатывать в день до 500 позиций по десятку поставщиков. В итоге на средний прайс-лист компьютерной компании в 3000 позиций требуется содержать отдел из 5-6 менеджеров закупа, стоимость вопроса – около 150 000–200 000 руб. в месяц, или 50 руб. на одну товарную позицию. Внедрение средств обработки прайс-листов (E-Trade Price List Importer) позволяет повысить скорость и объем обработки примерно в 10 раз – до 3000 позиций в день. В итоге стоимость поддержки одной позиции прайс-листа удешевляется и составляет около 10 руб. На самом деле компании за счет такого решения расширяют товарное предложение и качественно увеличивают состав прайс-листа – в несколько раз, предлагая позиции «под заказ». Для многих компаний такая стратегия стала ключом к успеху (пример – компания «Открытые технологии», сумевшая сформировать широкое и качественное товарное предложение в своем интернет-магазине и за счет этого шагнувшая на новый уровень развития). Дальнейшее повышение скорости и снижение себестоимости обработки прайс-листов позволит компаниям, раньше других внедрившим подобные решения, не только выживать, но и активно развиваться на высококонкурентном рынке.

Таким образом, разработка системы обработки прайс-листов, которая позволяет поднять качество распознавания товаров с 20–30 %, хотя бы до 60–80 %, позволяет получить конкурентное преимущество компаниям, которые ее начнут применять.

Как может выглядеть архитектура и реализация такой системы, по нашему мнению. Сначала опишем основные необходимые компоненты.

Классификатор товаров. Все товары для повышения точности распознавания желательно предварительно классифицировать. Основой классификации может служить некоторая структура прайс-листа, его деление на категории/группы/классы, каким-либо образом содержащиеся в нем и которым можно извлечь информацию для классификации. Предваритель-

но мы должны создать какую-то свою базовую классификацию (либо несколько классификаций, связав их друг с другом). Далее можно составлять и поддерживать привязки групп/категорий товаров в прайс-листах к классам своего классификатора товаров.

Каталог товаров. Собственно каталог с данными товаров – содержит все необходимые атрибуты товара (уникальный код, название, артикул производителя, ссылка на класс товара в классификаторе, возможно ссылки на технические характеристики и описание).

База предложений. Содержит записи о товарных предложениях. Запись товарного предложения включает в себя уникальный код товара, код поставщика, цену от поставщика, дату регистрации ценового предложения (дату прайс-листа), возможно условия поставки (наличие на складе, возможность предварительного заказа, сроки доставки), условия гарантии на товар, условия возврата товара и другие дополнительные характеристики.

Модуль импорта (*парсинга*) *прайс-листов*. Компонент для разбора информации прайс-листов, включающий в себя механизм распознавания и сопоставления товаров на основе регулярных выражений для пополнения базы данных. Управление механизмом распознавания должно выполняться с помощью хранимых в базе данных настроек на конкретный вид (конкретную структуру и формат) обрабатываемых файлов. В настройках должны быть описаны расположение значимых полей и их содержание (например, поле артикула, поле названия и т.д.), формат прайс-листа – наличие заголовков групп и их расположение, правила пропуска некоторых разделов (например, раздел контактной информации). Часть операций определения структуры обрабатываемого файла может выполняться самим механизмом распознавания (так работают программы серии E-Trade).

Модуль формирования товарного предложения. Компонент, формирующий итоговый прайс-лист на основе базы предложений и набора заданных ограничений. Заметим, что модуль сам по себе может быть достаточно сложным, так как следует учитывать не только цену, но и ряд других факторов. К таким факторам относятся – скорость получения товара (например, покупатель интернет-магазина не всегда готов ждать доставку товара две недели, даже если цена товара будет дешевле), стоимость доставки, гарантийные обязательства, возможность получения товара на условиях «под реализацию», возможность возврата товара. Это самые важные критерии, в конкретном случае могут учитываться и другие факторы работы с поставщиками.

Модуль управления системой. Компонент включает в себя средства управления системой.

Реализация системы, по нашему мнению, будет наиболее эффективна на основе интернет-технологий, поскольку это дает:

– высокую гибкость при разработке,

– построение простого интерфейса для работы пользователей,

– независимость от конкретных технических средств и платформ,

– возможность предоставить доступ к системе для любого пользователя/компании в режиме сервиса,

– удобство и простоту стыковки практически с любым интернет-магазином.

Тестовый вариант системы разрабатывается на связке средств PHP-MySQL с применением фреймворка Yii.

Необходимо отметить, что применение предлагаемой технологии автоматизированной обработки частично структурированных данных и соответствующих программных инструментов будет эффективным во многих других областях.

Если к системе добавить компонент, который вместо файлов будет обрабатывать указанные наборы страницы сети интернет, мы дополнительно расширим возможности системы.

Например, поскольку в среде интернет размещается полная информация о состоянии приемной кампании, то возможен постоянный и оперативный автоматизированный сбор данных о состоянии приемной кампании в вузах с сайтов образовательных учреждений при проведении маркетинговых исследований и формировании аналитических материалов.

Литература

1. *Гойвертс Я., Левитан С.* Регулярные выражения: сб. рецептов / пер. с англ. СПб.: Символ-плюс, 2010. 608 с.
2. *Султанова Е.С., Пашков П.М.* Пути построения системы управления корпоративным веб-сайтом // Вестник НГУЭУ. 2014. № 1. С. 312–320.
3. *Фридл Дж.* Регулярные выражения / пер. с англ.; 3-е изд. СПб.: Символ-плюс, 2008. 464 с.

Bibliography

1. *Gojverts Ja., Levitan S.* Reguljarnye vyrazhenija: sb. receptov / per. s angl. SPb.: Simvol-pljus, 2010. 608 p.
2. *Sultanova E.S., Pashkov P.M.* Puti postroenija sistemy upravlenija korporativnym veb-sajtom // Vestnik NGUJeU. 2014. № 1. P. 312–320.
3. *Fridl Dzh.* Reguljarnye vyrazhenija / per. s angl.; 3-e izd. SPb.: Simvol-pljus, 2008. 464 p.