И.М. Янников, М.В. Телегина, М.М. Казанцев

АНАЛИЗ МЕТОДОВ АВТОМАТИЗИРОВАННОГО ЗАПОЛНЕНИЯ БАЗ ДАННЫХ И АЛГОРИТМОВ АВТОМАТИЗИРОВАННОГО ПОИСКА ИНФОРМАЦИИ

(Ижевский государственный технический имени М.Т. Калашникова)

В настоящее время сложно переоценить актуальность безопасности потенциально опасных объектов (ПОО). Постоянная угроза террористических актов и усовершенствование методов и способов противоправных действий требуют постоянного совершенствование мероприятий по защите объектов от несанкционированного доступа. Кроме того ПОО и сами по себе являются сложными инженерно-техническими сооружениями, функционирование которых сопряжено с риском возникновения аварий и чрезвычайных ситуаций (ЧС) [1-3].

Основными компонентами системы ИИСБ являются подсистема управления, подсистема сбора, хранения и обработки данных и подсистема информационного обеспечения принятия решений. Именно в подсистеме информационного обеспечения принятия решений формируются оперативные решения, направленные на ликвидацию угрозы [4].

Одним из оперативных решений является совершенствование системы безопасности ПОО, куда относится выбор средств физической защиты на основе имеющейся информации о возможных угрозах на объекте, путях проникновения, характеристик средств защиты. Авторами с помощью MS SQL Server Express была разработана база данных «Средства физической защиты потенциально опасных объектов» [5].

Статичная база данных средств физической защиты, без постоянного актуализации информации имеет очень низкую практическую ценность. В связи с этим особенно остро встаёт вопрос автоматического поиска и обновления информации о существующих на рынке средствах физической защиты.

Для обеспечения надежной физической защиты потенциально опасных и критически важных объектов, применения современных средств физической защиты предлагается реализовать автоматическое обновление информации о средствах физической защиты в базе данных.

Сегодня автоматизированное заполнение баз данных задачи решаются несколькими способами:

1) извлечение данных с веб-страницы;

2) извлечение данных из структурированных файлов, например, прайс-листов;

Извлечение информации с веб-страниц основано на анализе объектной модели документа (Document Object Model). DOM – это не зависящий от платформы и языка программный интерфейс, позволяющий программам и скриптам получить доступ к содержимому HTML-, XHTML- и XML-документов [6]. Данный подход удобен тем, что позволяет извлекать данные любого типа и лю-

бой сложности, а также получать необходимое значение элемента по пути его расположения. Однако у данного подхода есть и свои недостатки:

- требуется вручную задавать расположение нужных элементов на странице;
- при изменении структуры целевого сайта, необходимо заново указывать пути расположения необходимых элементов;
- зачастую DOM-путь сложен и неоднозначен, что затрудняет получение значения элемента.

Наряду с анализом DOM-дерева сайта, в некоторых случаях, используют прямой парсинг строк с веб-страницы. Этот приём применим в случаях, когда информация на сайте размещается по каком-то шаблону. Для извлечения информации здесь гораздо эффективней будет работать именно парсинг строк [7].

Нельзя однозначно выделить подход, который будет 100% применим во всех случаях, поэтому современные библиотеки для парсинга HTML данных, как правило, комбинируют, разные подходы. Например, HtmlAgilityPack позволяет анализировать DOM дерево, а также с недавних пор поддерживается технология Linq to XML. Data Extracting SDK использует анализ DOM дерева, содержит набор дополнительных методов для парсинга строк, а также позволяет использовать технологию Linq для запросов в DOM модели страницы.

Для сравнения рассмотрим некоторые из этих библиотек более подробно.

HtmlAgilityPack – это гибкий HTML - парсер , который строит DOM и поддерживает простой XPATH или XSLT. Это библиотека для .NET, которая позволяет разобрать HTML-файл [5]. Названия методов соответствуют интерфейсам DOM плюс реализована возможность взаимодействовать при помощи методов: GetElementbyId(), CreateAttribute(), CreateElement() и т.д., так что работать особенно удобно, если приходилось сталкиваться с JavaScript. В целом это быстрая, довольно удобная библиотека для работы с HTML. Распространяется по лицензии Microsoft Public License.

Существует более продвинутая версия данной библиотеки, называемая Fizzler. В ней реализована более удобная возможность работы с селекторами CSS. В остальном это тот же самый HtmlAgilityPack, со всеми присущими ей достоинствами и недостатками. Распространяется по лицензии GNU Lesser General Public License.

AngleSharp – современный, практичный, быстроразвивающийся парсер написанный на C#. Его API построен на базе официальной спецификации по JavaScript HTML DOM, что делает его использование для парсинга HTML очень удобным.

Data Extracting SDK – удобный парсер, позволяющий извлекать информацию как из веб-ресурсов, так и из простого текста. Для работы использует библиотеку Microsoft.mshtml для получения DOM-дерева HTML страницы и информации об HTML элементах. Что бывает неудобно в некоторых случаях. Например, когда модель сайта не соответствует COM.

Из сравнительного анализа тестов данных парсеров [6], задачей которых является извлечение адресов из ссылок на странице и данных из таблиц, можно

сделать вывод, что оптимальным выбором сейчас будет AngleSharp, так как он активно разрабатывается, обладает интуитивным API и показывает хорошее время обработки.

Разработанная база данных средств физической защиты потенциально опасных объектов с функциями автоматического обновления данных позволит не просто хранить и обрабатывать большой объем информации, но и обеспечит высокую эффективность применения новых современных средств физической защиты.

Литература

1. Габричидзе Т.Г., Янников И.М. Структура и принцип построения комплексной многоступенчатой системы безопасности КВО (ХОО, ОУХО) // Теоретическая и прикладная экология. – Киров. – 2007. – №2. – С.55–69.

2. Янников И.М., Куделькин В.А., Телегина М.В., Габричидзе Т.Г. Комплексный подход к организации мониторинга защищённости потенциально опасных объектов с использованием ГИС-технологий // Интеллектуальные системы в производстве. – Ижевск: Изд-во ИжГТУ. – 2015. – №3 (27). – С.83–87.

3. Янников И.М., Прокофьев Д.В. Комплексная безопасность потенциально опасных объектов. Предпосылки и принципы её построения // Математические модели и информационные технологии в организации производства. № 1(30) - Ижевск: Изд-во ИжГТУ, 2015. –С. –35-38.

4. Куделькин В.А., Янников И.М. Структурная схема интеллектуальной интегрированной системы безопасности потенциально опасных объектов // Известия Самарского научного центра Российской академии наук. Том 17, №6(2), – 2015.– С. 726 – 728.

5. Янников И.М., Соболева Н.В., Куделькин В.А., Казанцев М.М. База данных средств физической защиты потенциально опасных объектов // Интеллектуальные системы в производстве. – Ижевск: Изд-во ИжГТУ. – 2017. – № (). – С.– (в печати).

6.Document Object Model (DOM). [Сайт]. https://www.w3.org/DOM/ (Дата обращения 21.12.2016).

7.Подходы к извлечению данных из веб-ресурсов. [Сайт]. https://habrahabr.ru/post/99918/ (Дата обращения: 23.12.2016).

8.Html Agility Pack. [Сайт]. http://htmlagilitypack.codeplex.com/ (Дата обращения: 24.12.2016).

9. Распарсить HTML в .NET и выжить: анализ и сравнение библиотек. [Сайт]. https://habrahabr.ru/post/273807/ (Дата обращения: 12.01.2017).