УДК 002:330.163, 004.031.42

ИЗВЛЕЧЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ С САЙТОВ ЭКОНОМИЧЕСКОЙ СТАТИСТИКИ

Трусов А.Н., Кацуро Д.А., Иванченко П.Ю.

ФГБОУ ВПО «Российский экономический университет», филиал Кемерово, e-mail: mors_kem@mail.ru

В статье рассматривается вопрос о получении необходимой экономической информации в сети интернет, а также о ее структурировании для дальнейшего применения с целью прогнозирования экономического развития предприятий и регионов. Проведен краткий анализ сайтов государственной экономической статистики, а также описана возможность автоматизированной обработки информации с них путем синтаксического анализа. Сформулированы основные требования к программному обеспечению, которое может выполнять функции анализа сайтов. Приведен алгоритм взаимодействия модулей работы программы.

Ключевые слова: экономическая информация, статистика, интернет, синтаксический анализ сайтов, автоматизация

EXTRACTION AND PROCESSING OF INFORMATION FROM THE SITES OF ECONOMIC STATISTICS

Trusov A.N., Katsuro D.A., Ivanchenko P.Y.

Russian State University of Trade and Economics, branch, Kemerovo, e-mail: mors_kem@mail.ru

The article deals with the question of obtaining the necessary economic information in the Internet, as well as its structuring for further use to predict the economic development of enterprises and regions. A brief analysis of the sites of government economic statistics, as well as automated information processing capability by the parser is described. The basic requirements to software, which can serve as the analysis of sites and analyzed the algorithm the program modules interact with each other. The algorithm of interaction between modules of the program is described.

Keywords: Economic information, statistics, online parsing websites, automation

В современном информатизированном мире понятие экономической информации очень важно. Еще более важно правильное использование такой информации для решения задач прогнозирования экономического развития предприятий и региона в целом [2].

Экономическая информация – это информация об экономических отношениях и процессах общественного воспроизводства. Она используется в системе управления народным хозяйством, наряду с такими видами информации, как научно-техническая, социально-политическая, и отражает взаимосвязи между вещественными элементами производства. Всякая информация характеризуется двумя основными признаками – количеством и качеством. Качественный признак позволяет классифицировать её по признакам знаний, функциями управления и т.д. Количественный признак позволяет выяснить единицы измерения, на основе которых можно установить ее объем и трудоемкость получения, а также технические средства для передачи, сбора, хранения и фиксирования, технологию обработки.

Экономическая информация окружает нас везде и очень быстро распространяется по сети интернет. Правительственные и государственные органы предоставляютобъемную информацию пользователямс помощью таких интернет ресурсов, как, например, «Федеральная служба государственной статистики» [5]. Большие объемы экономической информации делают актуальным вопрос качественности, достоверности, и, главное, структурированности экономической информации, для использования ее на практике, в частности, с целью прогнозирования деятельности предприятий, регионов, стран и т.д. [3].

Анализ сайтаэкономической статистики [5] позволилвыделить следующие информационные разделы:

Основные социально-экономические показатели.

Экономическая ситуация.

1. Производство товаров и услуг;
2. Потребительский рынок;
3. Оптовая торговля;
4. Институциональное преобразование;
5. Цены;
6. Финансы;
7. Социальная сфера.
8. Уровень жизни;
9. Рынок труда;
10. Образование;
11. Заболеваемость;
12. Правонарушения;

13. Демографическая ситуация.

Следует отметить, что на различных сайтах информация представляется поразному. Одни и те же показатели могут быть представлены в разных форматах, в таблицах, которые отличаются составной структурой и уровнем подробности.

Решением описанной проблемы является разработка специальных автоматизированных средств получения, обработки и представления экономической информации в структурированном виде. Например, к таким средствам можно отнести программы статистической обработки или интеллектуального анализа данных, которые, при помощи технологии синтаксического анализа сайтов (парсинга), могут обрабатывать необходимую информацию с нескольких сайтов, при этом, не дублируя ее и аккумулируя в отдельные файлы определенной структуры.

Такой программный продукт должен представлять собой платформу, реализующую базовые функции по управлению заданиями:

1. Синтаксический анализ сайтов (парсинг).

2. Скачиваниенеобходимой информации.

3. Структурированное хранение документов.

От разработчика требуется написание программного продукта синтаксического анализа данных (парсеров) сайтов государственной статистики. В ходе проведенного нами анализа были сформулированы следующие основные требования к указанному программному продукту:

1. Возможность гибкой настройки загрузки необходимой экономической информации.

2. Независимость загрузки страниц и их обработка, возможность повторной обработки ранее скаченных страниц.

3. Поддержка процесса разработки программы синтаксического анализа интернет страниц.

4. Возможность дополнения данных, полученных при обработке информации с интернет страниц, с целью выравнивания содержимого.

5. Продолжение процесса загрузки страниц после остановки.

6. Корректная обработка изменений данных полученных с интернета и введенных вручную.

7. Одновременная работа сразу с несколькими сайтами и наборами правил.

В качестве примера удобно рассмотреть алгоритм скачивания и разбора информации с любого сайта экономической статистики [5], на котором представлена сводка экономических данных за определенный период времени. Непосредственно сам разбор не представляет трудностей, но при реализации такого программного продукта целесообразно иметь ввиду наличие следующих основных модулей (функциональных блоков):

1. Подготовка (Initialization, I): аутентификация, получение данных, распознавание пользователя как человека и т.д.;

2. Загрузка страницы (Download, D);

3. Проверка корректности загрузки (Check, C);

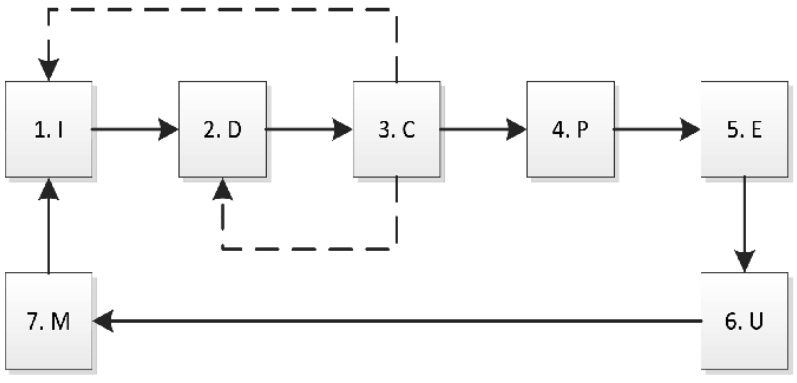
4. Разбор страницы согласно правилам (Parse, P);

5. Извлечение данных из разобранной страницы, возможно, с использованием дополнительных справочников и данных (Extraction, E);

6. Обновление данных в БД (Update, U);

7. Контроль (Monitoring, M): обработка данных других модулей, ведение журналов, предоставление информации о текущем статусе.

На рисунке кратко представлена схема взаимодействия перечисленных модулей, предложенная в [6].



Взаимодействие модулей между собой

Особое внимание следует уделить блокам I и II. В блоке подготовки (I) выполняются все операции, которые необходимы для начала работы, а также для обработки «нестандартных» страниц, которые имеют уникальное построение. Возможно, что перед получением страниц с данными необходимо выполнить безопасную загрузку определённой страницы для получения информации и доказать, что страница посещается человеком. Здесь же определяются и создаются потоки и очередь для конкретного сервера. Следует отметить, что при работе данного модуля необходимо принимать решения о приостановке загрузки, а также её возобновлении в случае возникновения ошибок. В рамках системы данный блок должен быть уникален для каждого сайта, поскольку, например, повторная аутентификация может привести к сбросу текущей сессии.

Модуль проверки (C) необходим для оперативной оценки корректности загрузки конкретной страницы. Несмотря на то, что, по времени, он работает непосредственно сразу после блока загрузки (D), но часто имеет индивидуальную реализацию для каждого сайта. Корректная страница не означает, что там находится именно то, что нужно, поскольку сайт может сообщать об ошибке авторизации, необходимости пройти аутентификацию, повторить попытку через некоторое время, невозможности найти подходящую страницу и т.д. В качестве критериев распознавания можно использовать сокращённый модуль разбора (P), но дополнительно необходимо научиться обрабатывать и сообщения об ошибках.

Целесообразно отделить друг от друга блок разбора (P) и блок экстракции (E),так как целью разбора является анализ страницы и выделение фрагментов с данными, а экстракция уже привязывает данные к конкретным объектам. Разбор работает на уровне HTML ограничен конкретной страницей, а блок экстракции работает уже с данными и может воспользоваться информацией с ранее загруженных страниц.

У блока обновления (U) задача осложняется тем, что данные надо объединять, причём они могут содержать различное число полей и быть корректными лишь на определённые моменты времени. Целью модуля контроля (M) является не только получение информации о текущем статусе и процессе выполнения операций, но и выявление сбоев, а также и возможности перезапуска с любого этапа [6].

Таким образом, из всего вышесказанного можно сделать вывод о необходимости создания и использования специальных автоматизированных средств получения, обработки и предоставления экономической информации в структурированном виде, что дает возможность занесения и дальнейшего ее использования в пакетах финансового анализа [1,4] деятельности предприятий, отраслей, регионов и других крупных экономических систем.

Список литературы

1. Конструктор и решатель дискретных задач оптимального управления («Карма»). Программа для ЭВМ: Свидетельство о регистрации в Роспатенте № 2008614387 от 11.09.2008. Правообладатели: А.В. Медведев, П.Н. Победаш, А.В. Смольянинов, М.А. Горбунов.

2. Медведев А.В. Моделирование стратегии социально-экономического развития региона на основе мезоэкономического подхода и оптимизационной математической модели / А.В. Медведев // Вестник Красноярского госуниверситета. Серия «Физико-математические науки». – 2006. – № 1. – С. 208-214.

3. Медведев, А.В. Поддержка принятия решений при управлении экономикой региона. Монография / А.В. Медведев. – Кемеровский государственный университет. – Кемерово. – 2011. – 106 с.

4. Программа разработки бизнес-плана и оценки инвестиционных проектов [Электронный ресурс]. URL: http://www.expert-systems.com/financial/pe/.

5. Федеральная служба государственной статистики [Электронный ресурс].URL: http://www.gks.ru.

6. Хабрахабр: Архитектура интеллектуального Интернет-паука [Электронный ресурс].URL: http://habrahabr.ru/post/194914.