Секция «Вычислительная математика и кибернетика»

Формализация информационных Интернет источников для автоматизации сбора данных

Найденов Никита Анатольевич
Аспирант
Учреждение Российской академии наук Вычислительный центр им. Дородницына РАН, Теоретические основы информатики, Москва, Россия
E-mail: naidyonov@gmail.com

Все данные, представленные в глобальной сети Интернет, можно назвать неструктурированными, ввиду индивидуальности и специфичности архитектуры каждого ресурса. В основном, такие данные – это HTML страницы, т.е. текстовые структуры. В настоящее время, в связи с постоянным ростом информации во всемирной паутине, необходимо развитие технологий, позволяющих использовать ее для решения различных производственных задач предприятий и организаций, вследствие чего активно развивается область анализа текстовых данных и неструктурированной информации. Очень актуальной является задача предварительной обработки данных, которую можно разделить на 3 этапа: консолидация, трансформация и очистка [1]. Самым трудоёмким этапом является консолидация данных, которая включает в себя сбор данных. Если исследования ведутся с большой выборкой, то для того, чтобы накопить достаточное количество материала, могут уйти недели или месяцы кропотливого труда. Данная работа посвящена разработке методов автоматического сбора информации из открытых интернет источников.

Основную задачу проведенного исследования можно описать следующим образом: требуется собрать данные с открытых новостных источников в сети Интернет за определенный период времени. Если информационный ресурс предоставляет данные в формате RSS [2], то можно воспользоваться инструментами по обработке такой ленты новостей. Однако не все сайты располагают таким форматом. Существует большое количество технологий и механизмов обработки ресурсов как с RSS лентой, так и без нее, но все они требуют настройки определенных параметров для каждого источника.

Цель проведенного исследования заключается в том, чтобы предложить и исследовать метод автоматической формализации информационных источников, которые не предоставляют данные в формате RSS. В таком случае не требуется какая-либо дополнительная настройка для сбора данных с ресурса. Такой подход был исследован экспериментально более чем на 40 новостных источниках, были проанализированы точность и полнота метода формализации ресурсов.

Результатом данной работы можно считать метод автоматической формализации новостных источников данных.

Литература

1. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям // Спб.: Питер, 2009 - 624 с

2. Определение RSS: http://ru.wikipedia.org/wiki/RSS