УДК 004.852

ПРОГРАММНЫЙ КОМПЛЕКС СБОРА, ХРАНЕНИЯ И АНАЛИЗА ИНФОРМАЦИИ ИЗ ОТКРЫТЫХ ИСТОЧНИКОВ СЕТИ ИНТЕРНЕТ

Пастухов Н.А. (Военно-космическая академия А.Ф. Можайского)
Научный руководитель – к.т.н. Менисов А.Б.
(Военно-космическая академия А.Ф. Можайского)

В данном исследовании разработан подход к выявлению закономерностей между смысловым содержанием текста HTML-блока документа и его обозначением в разметке страницы. Результатом исследования является прототип программного комплекса по сбору, хранению и анализу информации из открытых источников сети Интернет, которая необходима для оценивания информационной обстановки. Цель исследования – создание алгоритмической базы, которая позволит производить разработку унифицированных web-скраперов, ориентированных на web-документы схожей структуры и не привязанные к конкретному источнику сети Интернет.

Введение. В настоящее время сбор информации из различных источников сети Интернет является важным процессом, который напрямую влияет на эффективность любой организации. На данный момент эта задача решается путем создания web-скраперов, уникальных для каждого конкретного ресурса. Расширение списка отслеживаемых ресурсов требует привлечения работы программистов по созданию web-скраперов для каждого нового источника. Однако, в большинстве случаев отслеживанию подвергается список ресурсов схожей структуры, а информация, извлекаемая из каждого источника, однородна по своей природе.

Основная часть. Таким образом, гипотеза исследования заключается в следующем: задача распознавания заданной информации в HTML-структуре нового источника может быть решена как задача распознавания именованных сущностей текстов естественного языка (NER) без привлечения программиста, а на основании распознанной информации может быть составлена инструкции по работе web-скрапера нового источника. Дальнейшее применение алгоритмов суммаризации текста к собранным данным позволит создать систему анализа потоков информации в реальном времени, поступающих из неограниченного числа источников схожей структуры.

Выводы. Практическая значимость заключается в возможности применения полученных результатов для дальнейшего развития систем анализа естественных языков, а также для построения других систем анализа больших потоков данных в реальном времени.