УДК 621.391.2

ИНТЕЛЛЕКТУАЛИЗАЦИЯ ОТОБРАЖЕНИЯ И ОБРАБОТКИ РАЗНОРОДНОГО ИНТЕРНЕТ-КОНТЕНТА С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ DATA MINING

Чикрин Д.Е.
Научный руководитель – д.ф-м.н., профессор Надеев А.Ф.
Казанский государственный технологический университет им. А. Н. Туполева

В рамках данного доклада было разработано ПО, представляющее собой рекомпилятор HTML-страниц со сложной структурой в HTML-аналог, имеющий лишь необходимые детали оформления, либо в WML – страницу, которая может быть загружена на произвольном мобильном терминале, поддерживающем формат WAP.

Результат представляет собой основную смысловую часть исходного документа, а также ссылки на соответствующие по тематике ресурсы в различных поисковых системах (при этом порядок и расположение информационных блоков, а также оформление результата зависит целиком и полностью от предпочтений пользователя, как жестко настраиваемых, так и определяющихся статистически при помощи алгоритмов Data Mining). Благодаря возможности работы в режиме пакетной обработки информации, программа может рассматриваться как крайне полезный инструмент для структурирования и сжатия имеющегося HTML-контента, а также разработки собственных WML-проектов на основе имеющейся HTML-базы. Это подтверждается также следующими результатами работы представленного рекомпилятора:

1). Полученные страницы имеют окончательные размеры в 3-7 раз меньше исходных.

2). Исходный код полученной страницы является хорошо структурированным, при этом является максимально совместимым с любым из существующих браузеров в связи с использованием только стандартных элементов оформления.

3). Представленное ПО позволяет выделить основную часть обрабатываемой страницы в виде, хорошо воспринимаемом с экранов как стационарных, так и мобильных терминалов, с учетом предпочтений пользователя, которые могут быть заданы как автоматически, так и вручную.

4). Встроенное разбиение единой исходной страницы на WML-карты обусловленного размера, а также на под-сайты, необходимое для того, чтобы сотовый телефон либо другое устройство с ограниченной памятью смог воспринять данную информацию. В противном случае приходится вручную разбивать текст на части, что является процессом достаточно трудоемким.

5). Возможность работы в пакетном режиме, при этом программа может быть запущена в виде резидента, что, учитывая высокую скорость обработки (единицы секунд даже на обработку самых сложных и больших страниц) позволяет эффективно применять ее для решения проблемы структуризации и сжатия имеющегося Интернет-контента.

При разработке представляемого программного продукта были использованы методы Data Mining, такие, как кластерный анализ, ассоциативные правила, деревья решений и байесовые сети, а также широкий спектр алгоритмов синтаксической обработки.