ОБ ОДНОМ МЕТОДЕ ДЛЯ АНАЛИЗА WEB-ТРАФИКА

Ф. Ф. Юсифов

Институт Информационных Технологий НАН Азербайджана
Азербайджан, Баку
farhadyusifov@gmail.com

В данной работе рассматривается возможность извлечения данных методами интеллектуального анализа данных (Data mining) из регистрационных журналов (лог файлов) Web-серверов. С этой целью построена модель для анализа активности пользователей с использованием статических данных Web-сервера.

В связи с увеличивающимся ростом популярности WWW, Web-сайты играют важную роль для передачи знании и информации пользователям. Обнаружение и обработка информации пользовательских образцов является важной задачей для определения эффективной маркетинговой стратегии и оптимального использования Web-сервера [1]. В настоящее время большинство созданных инструментальных средств для анализа Web-трафика обеспечивает сбор статистической информации без извлечения полезных знаний для менеджеров Web. Задача анализа полезной информации становится более стимулирующей, когда объем Web-трафика огромен и продолжает расти [2,3]. Также можно отметить что, WWW непрерывно растет с информационными ресурсами от Web-серверов и числа запросов от пользователей. Предоставление Web администраторам необходимой информации о поведении доступа пользователей стало потребностью и служит улучшению качества действий обслуживания Web-ресурсов.

Наблюдая за трафиком на сервере, мы определяем, сталкиваются ли посетители с его ошибками, по какой ссылке они попали на него и т.д. Если посетитель был переадресован к нам машиной поиска, мы увидим, какой именно машиной и какие ключевые слова при этом использовались. Такая информация может оказаться очень кстати при планировании рекламной кампании или при разработке нового контента [2,3].

В данной работе применен метод кластеризации для анализа Web-трафика и изучение активности пользователей. С этой целью использованы статистические данные активности каждого пользователя ежесуточно в течение недели. Используя эти данные можно построить математическую модель:

$$user_i(w_{i1},w_{i2},w_{i3},w_{i4},w_{i5},w_{i6},w_{i7}),\ i=\overline{1,n};\quad User=\bigcup_{i=1}^{n} user_i$$

где $user_i$ - i-ый пользователь, w - вес активности i-го пользователя по дням неделя.

Можно отметить, что существует большое число методов кластеризации [1,4]. Из них наибольшее распространение получили различные модификации метода ***K***-средних. Используя метод кластеризации ***K***-средних, можно разбить на кластеры пользователей по активности дня. Таким образом используя эту методику можно анализировать активность пользователей и принимать оптимальные решения. Популярность WWW продолжает увеличиваться и поэтому есть потребность в разработке методики и инструментальных средств для обнаружения и обработки информации.

ЛИТЕРАТУРА

1. Wanga X., Abrahamb A., Smitha K. A. Intelligent web traffic mining and analysis. Journal of Network and Computer Applications, 2004, vol. 28, pp. 147–165
2. Рабин Д. Изучайте журналы посещений. Сети и системы связи №1 (121), http://ccc.ru/magazine/depot/05_01/read.html?0201.htm, 2005
3. Iváncsy R., Vajk I. Frequent Pattern Mining in Web Log Data. www.bmf.hu/journal/Ivancsy_Vajk_5.pdf, 2006
4. Дюк В.А., Самойленко А.В. Data Mining. Учебный курс. СПБ: "Питер", 2001