Об Одном Подходе к Мониторингу Электронных Библиотек

Рамиз Шыхалиев
Институт Информационных Технологий НАНА, Баку, Азербайджан
ramiz@science.az

***Аннотация* — Мониториг электронных библиотек является очень важным с точки зрения оптимизации и оценки качества контента и сервисов электронных библиотек. Статья посвящена мониторингу использования электронных библиотек посредством анализа веб-трафика электронных библиотек. Для этого предлагается классифицировать веб-трафики пользователей электронной библиотеки.**

Ключевые слова — электронная библиотека, мониторинг, анализ веб-трафика, классификация веб-трафика, кластеризация веб-трафика

I. ВВЕДЕНИЕ

Современные электронные библиотеки являются очень сложными информационными системами. Их архитектуры позволяют иметь доступ к хранилищам различной информации, таких, как текст, аудио, видео и т.д. Электронные библиотеки предоставляют также сервисы по цифровому сохранению документов, управлению распределенными базами данных, гипертекстами, фильтрации, поиску информации [1].

Сегодня электронные библиотеки представляют широкий спектр академических, научно-исследовательских и других ресурсов, к которым можно получить доступ через Интернет [2]. Поэтому электронные библиотеки широко используются в учебном процессе и в научных исследованиях. При этом целью каждой электронной библиотеки является удовлетворение информационных потребностей всех типов пользователей. Однако с ростом количества контента и сервисов, а также количества пользователей, оптимизация и оценка качества контента и сервисов электронных библиотек становятся трудной задачей. Решение этой задачи, невозможно без мониторинга использования пользователями контентов электронных библиотек. При этом мониторинг может быть осуществлен путем анализа веб-трафика электронных библиотек. В результате мониторинга могут быть созданы статистические отчеты по использованию контента и сервисов электронных библиотек, и на основании этих отчетов провайдеры электронных библиотек могут определить наиболее часто используемые контенты и сервисы. В конечном счете эти знания позволят провайдерам электронных библиотек оптимизировать контенты и повысить качество предоставляемых пользователям информационных сервисов.

Целью статьи является мониторинг электронных библиотек посредством анализа веб-трафика. Причем анализ веб-трафика может быть осуществлен как в онлайн-режиме, так и на основании лог-файлов веб-серверов электронных библиотек.

II. МЕТОД АНАЛИЗА ВЕБ ТРАФИКА ЭЛЕКТРОННЫХ БИБЛИОТЕК

Существуют различные подходы к созданию электронных библиотек [3-5]. Анализ этих подходов показал, что информационные сервисы, оказываемые этими электронными библиотеками, основываются на веб-технологии и используют HTTP (Hypertext Transfer Protocol) протокол, то есть эти библиотеки имеют веб-серверы обслуживающие запросы пользователей. Исходя из этого, предлагается обобщенная архитектура мониторинга электронных библиотек (рис.1). В качестве монитора могут быть использованы различные средства мониторинга веб-серверов, например, WebSpy архитектура [6].

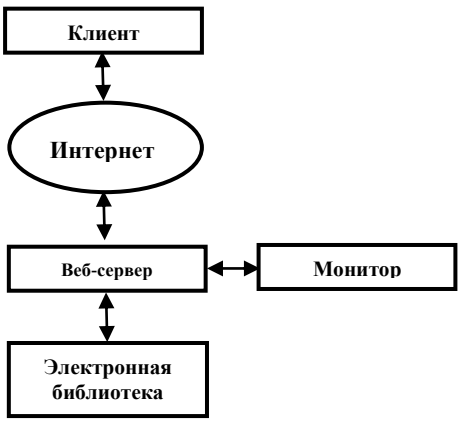

Рис.1. Обобщенная архитектура мониторинга электронных библиотек

Веб-серверы являются средством доступа пользователей Интернета к ресурсам и сервисам электронных библиотек, с помощью которых пользователи могут получить тексты, видео- и аудио- файлы и т.д. При этом для мониторинга использования ресурсов электронных библиотек более подходящим является анализ веб-трафика. Так что веб-трафик может быть использован как один из фактических показателей использования электронных библиотек. Вместе с тем, веб трафик является носителем информации о поведении пользователей и использовании контента электронных библиотек. На основе статистического анализа веб-

трафика можно определить статистические характеристики использования контента электронных библиотек.

Для анализа веб-трафика важным определить характеристики его мониторинга. Эти характеристики могут быть определены посредством исследования параметров (показателей) атрибутов описывающих веб трафик, к которым могут быть отнесены, дата и время, IP-адреса, хиты (hits), типы запросов (по использованию контента и сервисов), объемы переданных байтов и т.д. Таким образом, целью мониторинга веб-трафика является определение даты и времени, IP-адресов, типов запросов (типов контента и сервисов), количества хитов и объемов переданных байтов.

Используя характеристики веб-трафиков электронных библиотек, мы можем классифицировать веб-трафики пользователей по используемым контентам и сервисам. В общем, целью классификации веб-трафика является отображение веб-трафиков пользователей в определенных типах контентов и сервисов.

Формально задача классификации веб-трафика электронных библиотек определяется следующим образом. Пусть T_{web} веб-трафик электронной библиотеки, состоящий из n веб-трафиков пользователей, то есть $T_{web} = \{t_{web1}, t_{web2}, \dots, t_{webn}\}$, где веб-трафик каждого пользователя t_{webi} характеризуется k множеством атрибутов $\{x_{i_1}, x_{i_2}, \dots, x_{i_k}\}$ и множеством m классов, то есть типами запрашиваемых контентов $C = \{c_1, c_2, \dots, c_m\}$. Требуется определить такое отображение $f: X \to C$, которое обеспечило бы соответствие веб-трафика каждого пользователя t_{webi} только одному классу.

Для классификации веб-трафика электронных библиотек наиболее эффективными являются методы машинного обучения (МО) [6], являющиеся важной частью искусственного интеллекта. Способность методов МО непрерывно получать новое знание или преобразовывать структуры знания, облегчающие их использование, позволит использовать эти методы для классификации веб-трафика пользователей электронных библиотек.

Процедура МО может быть разделена на две части: создание модели классификации и, собственно, классификация. Методы МО делятся на методы обучения с учителем и обучение без учителя [7, 8]. Обучение с учителем создает структуры знаний, которые используются для отнесения новых образцов в заранее определенные классы. Обучение машины проводится с представлением к ее входу наборов типовых примеров, которые принадлежат заранее определенным классам. Результатом процесса обучения является построение модели классификации на основе анализа и обобщения представленных образцов. Фактически обучение с учителем создает модель взаимосвязи входа и выхода, т.е. осуществляет отображение набора входных атрибутов на выходные классы.

Использование в классификации веб-трафика электронных библиотек метода обучения без учителя может предоставить определенные преимущества. Такая модель позволяет идентифицировать новые классы, то есть новые области знаний, и группировать их в новый кластер, тогда как модели, использующие методы обучения с учителем, могут только идентифицировать веб-трафики, для которых созданы обучающие примеры, и не могут обнаружить новые области знаний [9]. При классификации веб-трафика методы без учителя не нуждаются в начальной ручной разметке входных данных. Будучи основанными на подобии между классифицируемыми объектами, в качестве входных данных в этих методах используют статистические характеристики потока веб-данных. Поэтому для создания модели классификации веб-трафика электронных библиотек в качестве метода обучения без учителя предлагается использовать кластеризацию.

Заключение

В статье предложен подход к мониторингу электронных библиотек. Мониторинг может быть применен для оптимизации и оценки качества контента и сервисов электронных библиотек. Для мониторинга использования электронных библиотек предложена обобщенная архитектура мониторинга электронных библиотек.

Предложенный подход мониторинга электронных библиотек основан на анализе веб-трафика пользователей. Для этого предлагается классифицировать веб-трафики пользователей по запрашиваемым контентам и сервисам. Результаты мониторинга позволят провайдерам электронных библиотек оптимизировать контенты и повысить качество предоставляемых пользователям информационных сервисов.

ЛИТЕРАТУРА

[1] B. Lodha, A. Galundia, P. Bhandari “Digital library: The new mantra of information infrastructure,” Pacific Business Review International, 2013, vol. 5 no. 11, pp. 29-35.

[2] C. Aijun, Z. Zhaozhong, M. Lu “The Tendency of Electronic Resources Development by Analyzing the Statistics of Three Consecutive Years of Electronic Resources Expenditures in Chinese and American Academic Libraries,” Journal of Academic Libraries, 2012, vol. 30, no. 1, pp. 55-58.

[3] Alexandria Digital Library. www.alexandria.ucsb.edu

[4] California Digital Library. www.cdlib.org

[5] California Digital Library Technical Architecture and Standard 2002. www.gseis.ucla.edu/~howard/Classes/208-s00/CDL/CDL-Arch-031000.pdf

[6] M.M. Thirukonda, S.A. Becker “WebSpy: An architecture for monitoring web server availability in a multi-platform environment,” Informing Science, 2002, vol. 5 no 4, pp. 175-187.

[7] N.J. Nilsson, Introduction to Machine Learning, http://robotics.stanford.edu/people/nilsson/MLDraftBook/MLBOOK.pdf

[8] M. Dunham, Data Mining: Introductory and Advance Topics. Prentice Hall, New Jersey, 1st edition, 2003.

[9] J. Erman, A. Mahanti, M. Arlitt “Internet Traffic Identification using Machine Learning,” Proc. of the Global Telecommunications Conference, 2006, pp. 1-6.