сыщения услугами сотовой связи настанет примерно к концу 2013 г., когда сотовому оператору будет уже практически невозможно увеличивать свою абонентскую базу. Вместе с тем, процедура оценивания коэффициентов модели является более сложной по сравнению с моделью логарифмического тренда, а также коэффициент детерминации существенно ниже, чем в первой модели.

Выбор модели для прогнозирования числа абонентов зависит от множества факторов, включая особенности ретроспективных данных, временные и финансовые ресурсы, которыми обладает аналитик и прочее. Однако несомненным является тот факт, что прогнозирование числа абонентов является неотъемлемой частью анализа при оценке эффективности и рисков инвестиционных проектов на рынке сотовой связи.

Библиографический список

1. Бывшев, В.А. Введение в эконометрию. Ч. 2. / В.А. Бывшев. – М.: Финакадемия, 2003.
2. Трегуб, И.В. Математические модели динамики экономических систем / И.В. Трегуб. – М.: Финакадемия, 2009.
3. Harald Gruber. The economics of mobile telecommunications. Cambridge University Press, New York, 2005.

О СБОРЕ ПОЛЬЗОВАТЕЛЬСКИХ ДАННЫХ В СИСТЕМЕ ПЕРСОНАЛИЗАЦИИ ИНТЕРНЕТ-МАГАЗИНА

А.Г. ЦАРЕВ, *асп. каф. электроники и микропроцессорной техники МГУЛ*

science4me@mail.ru

С момента распространения Интернета по всему миру появилось и быстро развивается новое направление коммерческой деятельности – электронная коммерция. Конкуренция в электронном бизнесе постоянно растет, одновременно с этим усложняется структура и увеличиваются объемы наполнения интернет-магазинов. Увеличение объемов информации несет большую проблему – невозможность ее обработки традиционными, привычными для человека способами визуального анализа [1]. Пользователям становится сложнее ориентироваться, находить и выбирать то, что необходимо. Это особенно актуально для крупных интернет-магазинов. Поэтому задача удовлетворения потребностей конечного пользователя становится наиболее важной и вместе с тем сложной. До недавнего времени эти проблемы решали только путем улучшения навигации и механизмов поиска в интернет-магазинах.

В настоящее время большую популярность приобретают интернет-магазины, реализующие персональный подход к каждому пользователю. Эффективность такого подхода во многом зависит от того, насколько точно и полно собранные данные характеризуют потребности пользователя. В работе рассматриваются вопросы сбора данных об интересах и поведении пользователей в контексте их применения в системе персонализации, а также описывается разработанный автором модуль сбора пользовательских данных (МСПД), интегрированный в систему персонализации интернет-магазина.

Постановка задачи

В работе поставлены следующие подзадачи:

– определить набор пользовательских данных, необходимых персонализации интернет-магазина;

– выявить подходящее средство сбора пользовательских данных для его использования в системе персонализации интернет-магазина.

Набор пользовательских данных, необходимых для персонализации

Основная задача любого средства сбора пользовательских данных состоит в сборе данных, посредством которых система персонализации могла бы эффективно выполнять возложенные на нее функции.

Выделим этапы работы системы персонализации, предшествующие представлению данных конечному пользователю, и на их основе определим необходимые для сбора данные:

1) идентификация пользователей;

2) идентификация сеансов;

3) анализ потребности пользователей;

4) анализ поведенческих характеристик пользователей.

Для идентификации пользователей используют IP-адреса устройств приема-передачи данных и Cookie-идентификаторы – небольшие фрагменты информации, хранящиеся на клиентских машинах [2]. Использовать указанные параметры необходимо совместно, так как возможны ситуации, когда их применение по отдельности может привести к неверной идентификации [3]:

– несколько пользователей на один IP-адрес;

– несколько IP-адресов у одного пользователя;

– один пользователь использует различные браузеры. В таком случае каждому браузеру будет присвоен свой Cookie-идентификатор.

В целях повышения достоверности идентификации можно использовать дополнительную информацию, например программное обеспечение пользователя (агент) [3].

Следующий этап – это идентификация сеансов доступа. В качестве наиболее распространенного способа идентификации сеанса применяется принцип на основе ограничений по времени между интервалами обращений пользователя к серверу [9]. Обычно интервал составляет от 20 до 30 минут. Если пользователь совершил обращение по истечении заданного интервала, то считается, что обращение принадлежит новому сеансу. Таким образом, для идентификации сеансов необходимо фиксировать дату и время обращений.

Анализ потребностей пользователей различается в зависимости от применяемого подхода к фильтрации информации [4, 10].

– Контентная фильтрация. Данный подход основан на оценке релевантности страниц сайта потребностям пользователя [5]. Персонализация данных осуществляется по данным текущего пользователя без учета опыта аудитории сайта.

– Совместная фильтрация. Данный подход основывается на анализе обращений и сведений, полученных от всей аудитории сайта, в том числе конечного пользователя [5]. Стоит заметить, что в совместной фильтрации не учитывается текстовое наполнение страниц сайта.

Несмотря на отличия, существующие подходы объединяет использование URL-адресов страниц, к которым обращались пользователи. Для контентной фильтрации характерно использование URL-адресов, к которым обращался конечный пользователь, а для совместной фильтрации характерно использование URL-адресов всей аудитории сайта.

Обычно осуществляется сбор URL-адресов двух типов: страницы, к которой обратились (реквест), и страницы, с которой произошло обращение (реферрер).

Анализ характеристик поведения пользователей проводится на основе временных наблюдений (время посещений страницы, продолжительность сеанса и т.д.), а также интерфейсных индикаторов (ввод данных в поля форм сайта, положение полосы прокрутки и т.д.). Временные наблюдения являются более надежными и простыми в реализации, так как могут быть полностью выполнены на стороне сервера.

Таким образом, для функционирования системы персонализации необходимо обеспечить сбор и хранение пяти типов пользовательских данных:

– IP-адреса компьютера;

– Cookie-идентификатора клиента;

– даты и времени обращения пользователя;

– адреса страницы, к которой обратился пользователь;

– адреса страницы, из которой произошло обращение пользователя.

Стоит отметить, что указанные данные достаточны только при совместной фильтрации. Для контентной фильтрации, помимо указанного набора данных, необходимо обеспечить связь URL-адресов с содержанием соответствующих страниц сайта.

Выбор средства сбора пользовательских данных

В настоящее время в интернете существует два общепринятых средства сбора пользовательских данных [6]:

– внешний счетчик;

– встроенные инструменты Веб-сервера.

Внешний счетчик представляет собой специальный скрипт, который загружается на стороне клиента одновременно с загрузкой странички веб-сайта и передает на сервер статистики информацию об этой загрузке [6]. Услуги пользования внешними счетчиками предоставляют специальные статистические сервисы, такие как Spylog, TopMail, Hotlog, LiveInternet и т.д. В рамках поставленной задачи внешние счетчики имеют существенные недостатки:

– невозможность беспрепятственного обращения к статистическим данным;

– ограниченная разработчиками функциональность;

– возможность отключения или частичного ограничения сбора данных пользователем сайта посредством настроек браузера.

Более надежным средством сбора статистических данных считаются встроенные инструменты Веб-сервера. На сегодняшний день практически все Веб-серверы предоставляют возможность ведения и администрирования протокола обращений пользователей к сайту. Данные обо всех обращениях записываются в обыкновенный текстовый файл (лог-файл). Однако и этот способ непригоден для использования в системе персонализации ввиду большой трудоемкости обработки лог-файла и ограниченной функциональности.

Подходящим решением поставленной задачи может быть разработка и использование МСПД. Данное средство представляет собой скрипт, который сохраняет сведения о загрузке страницы в собственную базу данных интернет-магазина. Недостатки таких счетчиков зависят от квалификации разработчиков и выделенных на нее ресурсов. По сравнению с традиционными средствами МСПД имеет ряд неоспоримых преимуществ:

– возможность беспрепятственного обращения к статистическим данным;

– возможность сбора нетрадиционных данных;

– хранение данных в наиболее удобном для обработки виде.

В соответствии с методом сбора данных МСПД можно разделить на две группы [7]:

– использующие активное профилирование;

– использующие пассивное профилирование.

Активное профилирование предполагает сбор личных сведений у пользователя посредством прямого ввода данных: анкетирования и опросов. Обычно осуществляется сбор следующих данных:

– личные сведения о пользователе (пол, возраст, место проживания и т.д.);

– интересующая тематика и информация (указание рубрик, поисковые запросы и т.д.);

– предпочитаемые настройки интерфейса.

В идеальном варианте, когда пользователь готов ответить на все вопросы, активное профилирование может эффективно справиться с задачей фильтрации информации. Однако на практике лояльное поведение пользователя – это большая редкость. Часто пользователи вводят неверную информацию о себе и пренебрегают заполнением анкет и настраиванием интерфейса [8]. Желание пользователей – найти необходимую информацию за наименьший промежуток времени без лишних усилий.

В отличие от активного, пассивное профилирование не обременяет пользователя вводом какой-либо информации и ведет сбор данных скрытно и незаметно для него. Это наиболее продвинутый и сложный способ составления портрета пользователя [7]. Пассивное профилирование может осуществляться как на стороне сервера, так и в связке сервер–клиент. Благодаря пассивному профилированию стало возможным отслеживание изменений интересов и динамики поведения пользователей.

Очевидно, что формирование пользовательской модели на основе одного вы-

бранного метода малоинформативно и недостаточно для эффективной персонализации. Целесообразно использовать совокупность пассивного и активного профилирования, причем в основе такого гибрида должно лежать пассивное профилирование, так как в таком случае уменьшается вероятность предоставления пользователем неверной или некорректной информации.

Модуль сбора данных о поведении и интересах пользователей, интегрированный в систему персонализации интернет-магазина

Разработанный МСПД использует пассивное и активное профилирование и ориентирован на контентно-совместный подход к фильтрации информации. Основная нагрузка по сбору данных ложится на пассивное профилирование. Незаметно для пользователя фиксируются его перемещения по интернет-магазину, заказанные товары, дата и время обращений, поисковый запрос, если пользователь обратился к интернет-магазину через поисковую систему. Активное профилирование используется только при сборе поисковых запросов к форме поиска интернет-магазина.

Ожидания при анализе интересов и потребностей пользователя связаны с собранными URL-адресами обращений пользователей. В том виде, в каком они передаются на сервер, URL-адреса неинформативны и пригодны только для использования в совместной фильтрации, так как в таком случае не требуется семантического анализа поступивших на вход данных. Однако, кроме названия домена и имени файла, анализ URL-адресов может выявить важные сведения о потребностях пользователей.

– Модель среднесрочных интересов. Среднесрочные интересы пользователей проявляются в каждом конкретном сеансе [5] и могут быть определены на основе извлечения поискового запроса из реферрера, если пользователь обратился к интернет-магазину из поисковой системы.

– Модель краткосрочных интересов. Краткосрочные интересы могут меняться при каждом новом обращении пользователя к странице интернет-магазина и определяются по тексту ссылок, к которым произошло обращение, заголовку, содержанию и информационному типу просмотренных страниц (новости, обзоры, описание товаров, заказ товара и т.д.) на основе анализа реквеста.

Помимо этого в URL-адреса гиперссылок всех блоков внедрены специальные идентификаторы, позволяющие определить блок, с которого произошел переход.

В связи с высокими ожиданиями от собранных URL-адресов предъявляются повышенные требования к идентификации сеансов. Так, в процессе эксплуатации описываемого модуля был выявлен существенный недостаток традиционного способа идентификации сеансов на основе IP-адреса и Cookie-идентификатора. При первом обращении нового пользователя к странице интернет-магазина на сервер передается только IP-адрес. Cookie-идентификатор в данном случае можно получить только при следующем обращении к серверу. Как было указано в работе, IP-адрес при новом обращении может измениться, тогда первая страница реального сеанса выпадает из идентифицированного, что недопустимо. Для решения данной проблемы в базе данных была создана таблица пользователей, в которой генерируются Cookie-идентификаторы пользователей. Если пользователь обратился в интернет-магазин первый раз, то есть у него нет Cookie, то сначала в таблице пользователей осуществляется запись нового пользователя, далее значение порядкового номера записи присваивается Cookie-идентификатору и передается в механизм регистрации обращений и браузеру клиента. При разрешении клиентом использовать Cookie-идентификаторы изложенный способ гарантирует верную идентификацию сеанса.

Для анализа поведения пользователей кроме временных используются интерфейсные наблюдения – уровень просмотра страницы. Для этого осуществляется сбор следующих параметров:

– высота рабочей области экрана;

– высота документа в браузере;

– максимальная глубина прокрутки при просмотре страницы.

Посредством интерфейсных наблюдений предполагается проводить оценку значимости востребованных страниц и корректировать соответствующим образом профиль пользователей.

В качестве программной платформы системы персонализации в целом и МСПД в частности был выбран язык клиентских сценариев JavaScript, язык серверных сценариев PHP в связке с базой данных MySQL. Такой выбор обусловлен большой распространенностью, высокой производительностью, удобством и простотой использования указанной среды. МСПД функционирует в составе системы персонализации товарных предложений двух интернет-магазинов: krug.ru и aldera.ru. На основе получаемой информации проводится исследование интересов пользователей и поиск факторов, влияющих на их поведение.

Минимальный набор данных, необходимых для персонализации:

– IP-адрес компьютера;

– Cookie-идентификатор клиента;

– дата и время обращения пользователя;

– адрес страницы, к которой обратился пользователь;

– адрес страницы, из которой произошло обращение пользователя.

Для сбора пользовательских данных в системе персонализации интернет-магазина целесообразно использовать МСПД, сохраняющий всю полученную информацию в базу данных интернет-магазина.

Для персонализации целесообразно использовать как пассивное, так и активное профилирование, причем в основе такого метода должно лежать пассивное профилирование.

МСПД осуществляет сбор минимального набора данных, а также данных, характеризующих взаимодействия пользователей с интерфейсом интернет-магазина:

– высота рабочей области экрана;

– высота документа в браузере;

– максимальная глубина прокрутки при просмотре страницы.

МСПД функционирует в составе системы персонализации товарных предложений двух интернет-магазинов: krug.ru и aldera.ru.

На основе получаемой информации проводится исследование интересов и поведения пользователей интернет-магазинов.

Библиографический список

1. Тенденции развития интеллектуальных информационных систем в сети Интернет // Интеллектуальные технологии в образовании, экономике и управлении: Сборник статей 2 Международной конференции. – Воронеж, 2005. – С. 197–198.
2. Лаура Томсон. Разработка web-приложений на PHP и MySQL: Пер. с англ. 2-е изд., испр./ Лаура Томсон, Люк Веллинг. – СПб: ООО «ДиаСофтЮП», 2003. – С. 265–266.
3. Щедрин, А. Основы извлечения знаний из Internet / А. Щедрин // Открытые системы. – 2003. – № 04.
4. Новичихин, А.В. К вопросу об эффективности и проблемах при построении моделей оптимизации Web-сайтов / А.В. Новичихин // Материалы VI Всероссийской объединенной конференции IST/IMS-2003. – Воронеж: ВГУ, 2003.
5. Зайцев, И.Б. Адаптивные гипермедиа издания, интегрированные в Интернет: дис. ... канд. техн. наук: 05.13.06. / Зайцев Илья Борисович. – М.: РГБ, 2005. – С. 37.
6. Место статистики в онлайн-продвижении. URL:http://mediainform.com.ua/rus/forreklama-articles/cat_452-items_143-mode_full-sp_20.html 13/06/2007.
7. Бунин, О. Персонализация сайтов / О. Бунин // Мир интернет. – 2001. – № XII(62).
8. О персонализации веб-сайта. URL: http://www.webmascon.com /topics/ development/ 16a.asp 13.06.2007.
9. Cyrus Shahabi, Farnoush Banaei-Kashani. «A Framework for Efficient and Anonymous Web Usage Mining Based on Client-Side Tracking», Department of Computer Science, Integrated Media Systems Center, University of Southern California, USA, 2001.
10. Herlocker J., Konstan J., Borchers A., and Riedl J. An algorithmic framework for performing collaborative filtering. // In Proceedings of the 22 nd annual international ACMSIGIR conference on Research and development in information retrieval, pages 230-237, 1999.