УДК 336.72

РАЗРАБОТКА СИСТЕМЫ АНАЛИЗА ПОВЕДЕНИЯ ПОСЕТИТЕЛЕЙ ВЕБ-САЙТОВ (ЧЕЛОВЕКА В ВЕБ-СРЕДЕ)

Первунинский С.М., Кудрявцев О.А.
Черкасский государственный технологический университет

Проанализировать потребность в анализе поведения человека в сети Интернет. Разработать модель веб-среды и поведения человека. Выработать критерии оценки результатов анализа поведения посетителя веб-сайтов. Провести обзор готовых решений (Счетчики, Google Analytics, системы веб-аналитики CMS). Принять и обосновать решение по созданию системы анализа поведения. Составить план реализации.
Ключевые слова: пользователи в сети, веб-аналитика, веб-среда, Web-mining.

Постановка проблемы. В электронном бизнесе, точно так же, как и в обычном, исследованием типов пользователей Интернет занимаются очень многие и поэтому можно выделить множество различных классификаций потребителей.

Недавно на свет появилось достойное внимания исследование, посвященное классификации Интернет-пользователей, его авторами стали компании *Booz-Allen & Hamilton и Nielsen // NetRatings*.

В результате было выявлено семь категорий Интернет типов поведения. Пользователи из некоторых категорий оказались хорошими потенциальными покупателями, в то время как другие практически не поддаются традиционным маркетинговым предложениям.

В результате было выявлено семь категорий Интернет типов поведения. Пользователи из некоторых категорий оказались хорошими потенциальными покупателями, в то время как другие практически не поддаются традиционным маркетинговым предложениям.

В ходе исследования была выявлена новая форма сегментации Интернет рынка по степени посещаемости. Новизна состоит в том, что потребители классифицируются не по демографическим признакам, а по тому, как они ведут себя в Интернет. Какова длительность их пребывания в онлайновом режиме, как много времени они проводят на каждой странице, насколько хорошо они знают сайты.

Исследование выявило семь типов пользовательских сессий, причем оказалось, что три из них наиболее привлекательны для онлайнового бизнеса, чем другие. К этим трем относятся сессии с целью развлечений, поиск информации и просто серфинг.

Продолжительность сессий этих типов самая высокая и составляет от 33 до 70 минут, причем на одну страницу тратится от 1 до 2 минут. Такой режим означает, что пользователи склонны засиживаться на одной странице и могут подвергаться влиянию различных сообщений.

Ниже приводится полный список типов пользовательского поведения согласно Booz-Allen & Hamilton и Nielsen // NetRatings.

Анализ последних исследований и публикаций. При написании данной статьи за основу исследований была взята книга Анализ данных и процессов: учеб. пособие / А.А. Барсегян, М.С. Куприянов, И.И. Холод, М.Д. Тесс, С.И. Елизаров. В данной книге описаны принципы web- и process-minig которые были использованы. Так же были исследованы и реазизованы алгоритмы для process-mining (Alpha algorithm).

Выделение нерешенных ранее частей общей проблемы. Всемирная сеть сейчас содержит огромное количество информации, знаний. Пользователи на различных условиях могут просматривать всевозможные документы, аудио- и видеофайлы. Однако это многообразие данных скрывает в себе проблемы, которые могут возникнуть не только при анализе, но и при поиске необходимой информации в Интернет.

1. Проблема поиска нужной информации связана с тем, что пользователь не всегда сразу может найти необходимые ему электронные ресурсы. Лишь небольшой процент ссылок среди предложенных поисковыми системами приводит к требуемым документам. Также труден поиск неиндексированной информации такими средствами.

2. Проблема обнаружения новых знаний. Даже если найдено множество информации, для пользователя извлечение полезных знаний является довольно трудоемкой и непростой задачей. Сюда же можно и отнести сложности, связанные с осмыслением сведений, понятием тех идей, которые были вложены авторами.

3. Проблема изучения потребителей связана с предоставлением пользователю информации, которая оказалась бы ему интересна. Это особенно актуально для электронных торговых порталов, которые могли бы "подсказывать" пользователю при выборе товара.

Поиск информации. Для нахождения необходимой информации пользователи обычно пользуются поисковыми ресурсами. При этом часто используются простые запросы по ключевым словам. Результатом выполнения запроса является список страниц, отсортированный по некому индексу релевантности, описывающему степень совпадения результата с запросом. Однако существующие поисковые механизмы обладают недостатками. Основным из них является низкая точность результата, вызванная недостаточным учетом семантических связей и контекста найденных в тексте выражений. Индексация интересующих сегментов сети с использованием интеллектуального анализа данных, применяющего алгоритмы математической лингвистики и обработки естественных языков, является перспективным направлением Web Mining в области

поиска информации. Интересный подход описан в статье Anupam Joshi, "Improving Web Search Engine Results Using Clustering".

Анализ структуры сегмента сети. Этот метод заключается в анализе структуры ссылок между различными веб-страницами, внутренними и внешними сайтами в выделенном сетевом сегменте. Появление этого метода было вызвано необходимостью решения задач, возникающих при анализе социальных сетей или специфических областей человеческой деятельности или знаний, например, в анализе цитирования авторов. Результатом такого анализа может служить выявленный набор специфичных страниц следующих типов:

· хабы – из такой страницы ссылки идут на наиболее значимые ресурсы в данной области знаний или на "знакомства" с наиболее значимыми пользователями социальной сети;

· авторитеты – страницы, на которые ссылаются большое количеством авторов по данной тематике или пользователи социальной сети, к "дружбе" с которыми стремится большое количество пользователей.

Топология структуры ссылок представляется в виде направленного графа с помеченными узлами в соответствии с их функциональной классификацией и дугами с весами, описывающими, например, частоты переходов по ссылке. Для моделирования топологии веб-ссылок используется несколько алгоритмов, например HITS (Jon M. Kleinberg, "Authoritative sources in hyperlink environment").

Выявление знаний из веб-ресурсов. Эта задача пересекается с уже описанной проблемой поиска информации. Только здесь у исследователя уже имеется набор веб-страниц, полученных в результате запроса. Далее требуется произвести их обработку с точки зрения автоматической классификации, составления оглавлений, выявления ключевых слов и общих тем. Выявленные знания могут представляться в виде деревьев, описывающих структуры документов или в виде логических и семантических выражений. Решение части этих проблем предлагает Text Mining – технология автоматического извлечения знаний в больших объемах текстового материала, основанная на сочетании лингвистических, семантических, статистических и машинных обучающихся методик (/go.asp?url=-3D-41-52-17-22-18-3E-34-CB-A3-D6-2C-32-9D-83-9F-88-DA-CF-55-59-6F-C5-A8-73-04-43-10-83-27-69-E9-96-42-55-74-98-06-FA-1B Soumen Chakrabarti "Data mining for hypertext", Helena Ahonen-Myka, "Finding co-occuring text phrases by combining sequence and frequent set discovery").

Персонализация информации. Персонализация веб-пространства – задача по созданию веб-систем, адаптирующих свои возможности (навигация, контент, баннеры и другие рекламные предложения) под пользователя на основании собранной и проанализированной информации о пользовательских предпочтениях.

Классическим примером может являться ресурс /go.asp?url=-3D-41-52-17-22-18-3E-34-CB-A3-D6-20-3A-9B-98-80-93-80-C9-48-1A-32-9C на котором один раз заказав дорогую книгу в твердом переплете, пользователь начинает регулярно получать предложения о покупке подарочных изданий по схожей тематике. Другой пример – на основании анализа корзин заказов пользователя ему предлагаются товары, которые он никогда не заказывал, но которые входят в корзины других покупателей, схожих с ним по транзакционному поведению.

Для анализа информации о пользователе следует в наименьшей степени использовать декларируемую о себе информацию, а скорее основываться на стойких шаблонах его "поведения" в сети – последовательности кликов внутри ресурса, переходах на другие под-ресурсы, периодах сетевой активности, осуществляемых покупках и т.д. См. B. Masand, Redwood, "Web Usage Analysis and User Profiling", Miha Gr?ar, "User profiling: Web usage mining".

Поиск шаблонов в поведении пользователей. Эта задача связана с предыдущей, но ее целью является не адаптация ресурса к предпочтениям индивидуальных пользователей, а поиск закономерностей в шаблонах взаимодействия пользователя с веб-ресурсом с целью прогнозирования его последующих действий. Анализируемые действия пользователей могут включать не только переходы по ссылкам, но и отправку форм, прокрутку страниц, добавление в избранные страницы и т.д. Найденные шаблоны используются в дальнейшем для оптимизации структуры сайта, изучения целевой аудитории и для прямого маркетинга.

Разработано множество подходов к решению задачи по выявлению знаний из шаблонов навигации пользователей (Jose Borges и Mark Levene "Data Mining of User Navigation Patterns", A.G. Buechner "Navigation Pattern Discovery from Internet Data").

С точки зрения применения алгоритмов интеллектуального анализа данных при поиске шаблонов пользовательского поведения чаще всего используются следующие методики:

· Кластеризация – поиск групп похожих посетителей, сайтов, страниц и т.д.

· Ассоциации – поиск совместно запрашиваемых страниц, заказываемых товаров.

· Анализ последовательностей – поиск последовательностей действий. Наиболее часто применяется вариант алгоритма apriori, разработанного для анализа частых наборов, но модифицированного для выявления частых фрагментов последовательностей и переходов.

Особенно интересен подход кластеризации последовательностей – поиск групп пользователей со схожими последовательностями действий. На первом этапе в этом подходе выделяются последовательности классифицированных действий пользователя, например, в рамках одной сессии. Затем подсчитываются частоты переходов между различными действиями для составления Марковской цепи заданного порядка. На заключительном этапе полученные Марковские цепи кластеризуются для выявления групп с похожими частотами переходов. Для прогнозирования следующего действия пользователя сначала на основании истории его действий в рамках сессии определяется группа, к которой он принадлежит с наибольшей вероятностью. Затем определяется действие, которое выполняется с наибольшей вероятностью в этой группе с учетом последних действий данного пользовате-

ля. Для реализации такого анализа можно, например, использовать алгоритм Microsoft Sequential Clustering, входящий в Microsoft Analysis Services 2005/2008. Недостатком алгоритма Microsoft является то, что до настоящего времени реализован алгоритм, использующий Марковские цепи только первого порядка.

В качестве примера применения метода анализа последовательности действий можно привести задачу по оптимизации рубрикации одного книжного интернет-магазина, проведенную компанией spellabs. Была выявлена группа, состоящая из пользователей, переходящих долгими путями по ссылкам на книги из разных рубрик и заказывающих в конечном итоге "изотерическую" литературу, до этого отдельно не выделенную в рубрику. Так была выявлена неучтенная целевая аудитория и оптимизирована структура сайта.

В бизнес-аналитике Web Mining решает следующие задачи:

· описание посетителей сайта (кластеризация, классификация);

· описание посетителей, которые совершают покупки в интернет-магазине (кластеризация, классификация);

· определение типичных сессий и навигационных путей пользователей сайта (поиск популярных наборов, ассоциативных правил);

· определение групп или сегментов посетителей (кластеризация);

· нахождение зависимостей при пользовании услугами сайта (поиск ассоциативных правил).

Список литературы:

1. Markov Z., Larose D.T. Data-mining the Web: uncovering patterns in Web content, structure, and usage. – John Wiley & Sons Inc., 2007.
2. Анализ данных и процессов: учеб. пособие / А.А. Барсегян, М.С. Куприянов, И.И. Холод, М.Д. Тесс, С.И. Елизаров. – 3-е издание перераб. и доп. – СПб.: БХВ-Петербург, 2009.

Первунінський С.М., Кудрявцев О.А.
Черкаський державний технологічний університет

РОЗРОБКА СИСТЕМИ АНАЛІЗУ ПОВЕДІНКИ ВІДВІДУВАЧІВ ВЕБ-САЙТІВ (ЛЮДИНИ В ВЕБ-СЕРЕДОВИЩІ)

Анотація
Проаналізувати потреби в аналізі поведінки людини в мережі Інтернет. Розробити модель веб-середовища і поведінки людини. Виробити критерії оцінки результатів аналізу поведінки відвідувача веб-сайтів. Провести огляд готових рішень (Лічильники, Google Analytics, системи веб-аналітики CMS). Прийняти і обгрунтувати рішення по створенню системи аналізу поведінки. Скласти план реалізації.
Ключові слова: користувачі в мережі, веб-аналітика, веб-середовище, Web-mining.

Pervuninsky S.M., Kudriavtsev O.A.
Cherkasy State Technological University

DEVELOP A SYSTEM FOR ANALYZING THE BEHAVIOR OF VISITORS TO WEBSITES (THE PERSON IN THE WEB)

Summary
To analyze the need for analysis of human behavior on the Internet. Develop a model of the web environment and human behavior. Develop criteria for evaluating the results of website visitor behavior analysis. Take a look at the finished solutions (Counters, Google Analytics, CMS web analytics). Accept and justify the decision to create a system for analyzing behavior. Make a plan of implementation.
Keywords: users in the network, web analytics, web environment, Web-mining.