УДК 004.838:004.853:004.855.5:004.738.5

АНАЛИЗ ЭФФЕКТИВНОСТИ РАБОТЫ WEB-САЙТА С ПРИМЕНЕНИЕМ МЕТОДОВ ИАД

В.В. Хайлова[1]

В работе описывается разработанная автором система анализа поведения посетителей web-сайта с использованием методов интеллектуального анализа данных и наглядной интерпретацией полученных результатов. Система снабжена интеллектуальными функциями: кластеризация посетителей относительно выделенного целевого атрибута при помощи правил ДСМ, сгенерированных в программной среде QuDA и оценка качества страниц web-сайта методом нечеткого вывода.

Введение

В настоящее время для анализа поведения посетителей web-сайтов широко применяются системы учета статистики посещений, использующие рейтинговые показатели и критерии, построенные на основе подсчета количества обращений к web-страницам. Они отслеживают базовые числовые параметры, такие как посещаемость сайта за период времени, глубина просмотра сайта, источники входящего трафика и так далее. Эту информацию web-аналитик должен выявить и оценить самостоятельно и принимать решения на ее основе. Недостатками применения рейтинговых показателей являются то, что они не во всех случаях адекватно отражают степень популярности информационного ресурса и не могут помочь в оптимизации структуры web-сайта. Рост объема собираемой информации и увеличивающаяся сложность исследуемых объектов требует автоматизации анализа. Требуются средства для выявления закономерностей в поведении посетителей и классификации объектов сайта.

Более качественную информацию, используемую для повышения эффективности функционирования web-сайта, дает применение методов интеллектуального анализа данных.

Раздел ИАД, прикладной областью которого является Интернет, называется Web mining. Направление Web mining можно определить как

[1] *125993, Москва, Миусская пл. д.6, РГТУ, veritas_vert@mail.ru*

поиск и анализ полезной информации в сети Интернет с применением методов интеллектуального анализа данных. Это направление включает три области исследования: Web content mining, web structure mining и web usage mining [Scime, 2005].

Область Web content mining занимается автоматизированным поиском и извлечением информации из содержимого или описания электронных документов, полученных из различных источников сети Интернет. Web structure mining – это процесс извлечения информации из структуры гиперссылок и перекрестных ссылок, отслеживание дефектов их структуры, анализ связей между ссылками и объектами ссылок. Web usage mining занимается изучением и анализом шаблонов взаимодействия посетителей с web-сайтами и web-сервисами [Cooley at al., 1997].

Данная работа относится к области Web usage mining. Эта область возникла около 10 лет назад, первыми работами можно считать исследования [Chen at al., 1996], [Mannila at al., 1996] и [Yan at al., 1996]. В настоящее время это направление бурно развивается во всем мире, к сожалению, в России она практически не представлена и открытых разработок по ней не ведется.

Можно выделить три основных метода ИАД, применяемых в области web: поиск ассоциативных правил [Elo-Dean at al., 1997], поиск последовательностей в цепочках перемещений (сессиях) посетителей по сайту [Mannila at al., 1996] и кластерный анализ, позволяющий группировать посетителей или структурные элементы сайта, имеющие схожие характеристики [Yan at al., 1996].

В настоящее время уже существует ряд коммерческих систем в большей или меньшей степени использующих методы ИАД. Примеры: Predictive Web Analytics (компания SPSS), HitsIntoLeads (компания NetMining) и другие. Такие системы приобретают все большую популярность. Это связано с тем, что растет потребность в средствах интеллектуального анализа данных, так как они позволяют автоматически извлекать из больших массивов данных полезные знания. Разработанная система также предоставляет такие возможности.

В созданной системе используются следующие методы интеллектуального анализа:

- ДСМ-метод [Финн, 1983; Финн, 1991] применяется для кластеризации сессий посетителей и выявления возможных причин, по которым посетитель достигал или не достигал выбранных «целей» на сайте.
- Нечеткий вывод [Zadeh, 1965] применяется для классификации страниц сайта. С помощью него оценивается качество страниц на основании текущих данных о посещениях. При этом, автор исходит из того, что по поведению посетителя на сайте можно судить о качестве

навигационной структуры и контента, а так же соответствию представленной информации запросам посетителей.

В качестве информационной базы для разработки и тестирования системы был выбран сайт электронной коммерции www.vertus.ru.

Этот сайт был выбран из тех соображений, что наиболее явный результат от внедрения системы можно получить для сайта, ориентированного на контент или прямые продажи, так как на web-сайтах такого типа проще выделить целевые атрибуты в сессиях посетителей и быстро оценить результат произведенных изменений. Фактически, улучшение качества сайта должно «конвертироваться» в увеличение количества заказов.

1. Анализ сессий посетителей с помощью ДСМ-метода в среде QuDA

ДСМ-метод применяется в данной работе для анализа поведения посетителей web-сайта и выделения среди них групп со сходным поведением относительно выделенных целевых атрибутов. В качестве внешнего программного средства, осуществляющего реализацию ДСМ-метода, используется среда QuDA **[**Григорьев и др., 2002; Grigoriev, 2003**]**.

С помощью созданной в рамках работы системы сбора информации была заполнена база данных о поведении посетителей. На этом информационном массиве были произведены эксперименты по поиску наиболее эффективного способа кластеризации посетителей в программной среде QuDA. Проведенный эксперимент показал, что наилучшие результаты среди всех выбранных для сравнения методов (Ripper, Наивный Байес, и др. [Барсегян и др., 2004]) для кластеризации посетителей по целевому атрибуту показал один из самых распространенных вариантов ДСМ-метода *простой метод сходства с запретом на контрпример*. Этот вариант хорошо себя зарекомендовал в различных экспериментах **[**Кузнецов и др., 1996; Gergely at al., 2007**]**

В процессе исследования автор исходил из того, что на любом web-сайте, особенно на сайтах связанных с электронной коммерцией, можно выделить так называемые «цели». «Цель» - это, например, заполнение посетителем формы отправки сообщения или оформление заказа в Интернет-магазине.

На основании характеристик сессий посетителей (серий из нескольких просмотров страниц сайта одним посетителем) можно делать выводы о причинах достижения или не достижения ими «целевых» разделов.

Процесс анализа начинается с подготовки данных. Из системы выгружаются данные в формате CSV. Каждая строка таблицы представляет собой описание одной сессии посетителя. В столбцах перечислены характеристики сессии (Referrer, число просмотров и

множество разделов сайта). На пересечении столбца и строки автоматически, при генерации данного файла ставится 0 или 1, что показывает, был ли посетитель в разделе или нет. Файл с таблицей импортируется в QuDA.

Выделяется целевой атрибут, в нашем случае это факт посещения страницы «корзина» или «заказ». Относительно целевого атрибута строятся классификационные правила. Были проведены эксперименты с различной детализацией разделов, различными настройками ДСМ-машины и различным количеством исходных фактов. В результате были выбраны параметры, при которых получаются наиболее интересные результаты с точки зрения их практического применения.

На выходе был получен набор правил, которые дают разбиение исходных сессий посетителей на группы относительно целевого атрибута. В контексте web-сайта, можно рассматривать посещение или не посещение каких-либо разделов как вероятную причину достижения или не достижения «цели» - это и является новой полезной информацией для web-аналитика.

На основе экспортированного из QuDA файла формируется отчет, при этом полученные правила обрабатываются и выводятся в наглядной форме.

1.1. Примеры результатов применения ДСМ-метода для оценки эффективности работы сайта Vertus.ru

- Было выяснено, что с поисковой системы «Яндекс» приходит больше целевых посетителей, чем с «Рамблера»: было сформировано правило, где причиной посещения корзины являлся факт просмотра каталога и значение атрибута referrer=yandex.ru.
- Поиск сайта работает неудовлетворительно, посетители не могут найти интересующую их информацию: ДСМ выделил правило, где причиной непосещения корзины было посещение только главной страницы и раздела «Поиск». Можно сделать вывод, что посетитель заходил на главную страницу, пытался воспользоваться поиском, а потом уходил с сайта.
- Большое количество посетителей не устраивает текст в разделе «Доставка». Было выделено правило, где причиной непосещения корзины являлся факт захода в раздел «каталог» и «доставка». Покупатели сначала находят интересующий их товар, а потом смотрят условия доставки, возможно, их не устраивает стоимость этой услуги.
- Были обнаружены определенные сложности взаимодействия посетителей с формой регистрации. Благодаря применению ДСМ-метода, было выявлено большое количество посетителей, заходивших в корзину, но не оформивших заказ. Возможно, что они предпочитали заказать товар

по телефону, и не стали оформлять заказ на сайте из-за сложности интерфейса.

- Были выявлены наиболее популярные разделы каталога и перечень разделов сайта, интересующие фактических покупателей.

Таким образом, ставшую актуальной в наши дни задачу разностороннего анализа сайтов на предмет эффективности их работы, как показал эксперимент, можно достаточно эффективно решать при помощи не применявшегося ранее в данной области ДСМ-метода.

2. Оценка качества страниц сайта с помощью нечеткого вывода

В работе был применен нечеткий вывод для решения задачи классификации страниц web-сайта на основании текущих данных о посещениях. Автор исходит из того, что по поведению посетителя на сайте можно судить о качестве его навигационной структуры и контента, а так же соответствию представленной информации запросам посетителей.

Для оценки качества отдельной страницы были выбраны два критерия (входные переменные): «посещаемость» и «разница между входами и выходами». Для переменной «посещаемость» было выбрано три терма: «плохая», «средняя» и «хорошая». Терм «плохая» описан Z-образной функцией, Терм «средняя» - трапецевидной или π-образной, «хорошая» - S-образной**.** Переменная «Входы и выходы» представляет собой разницу между количеством входов на страницу в начале сессии и выходов из нее в конце за период времени. Для ее оценки были выбраны три терма «плохо», «средне» и «хорошо». Термы заданы соответственно Z, π и S образными функциями. Для выходной переменной «качество страницы» были выбраны три терма: «хорошее», «среднее» и «плохое».

Была построена база правил для выходной переменной «качество страницы». Для каждой конкретной страницы производится фаззификация входных переменных и вычисляются степени их принадлежности к каждому из термов соответствующих лингвистических переменных. Вычисляется степень принадлежности объекта классификации классам из базы знаний. Далее выбирается класс с максимальной степенью принадлежности.

В интерфейсе системы результаты работы нечеткого вывода представлены в виде дерева страниц, каждая страница окрашивается в цвет, соответствующий ее «качеству». Таким образом, администратор сайта может сразу оценить ситуацию за день или любой другой период времени.

3. Практическая реализация системы

3.1. Назначение программного средства

Созданная программа предназначена для решения конкретных задач, связанных с анализом эффективности работы web-сайтов. Система занимается сбором информации о поведении посетителей и предоставляет пользователю отчеты на основе собранной информации. В системе предусмотрены модули для анализа качества страниц сайта при помощи нечеткого логического вывода, а также предусмотрен анализ поведения посетителей с использованием правил, построенных ДСМ-машиной в среде QuDA.

3.2. Описание схемы работы системы

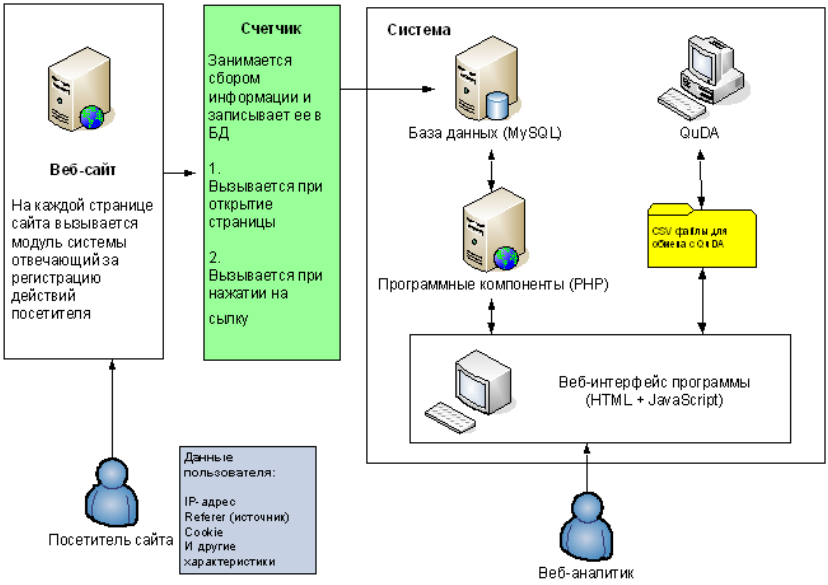

Рис. 1 Схема работы системы

Система представляет собой клиент-серверное приложение (Рис. 1). В качестве клиента выступает web-браузер. Вся программные компоненты и СУБД для хранения данных расположены на web-сервере. Предусмотрена одновременная работа с системой нескольких пользователей.

Для программной реализации системы был выбран язык PHP. Для хранения данных используется СУБД MySQL. Система работает на web-сервере Apache.

Пользовательский интерфейс программы реализован с помощью HTML и JavaScript. Выполнены требования корректного отображения интерфейса системы наиболее распространенными современными браузерами (IE, FireFox, Opera).

Для анализа эффективности работы web-сайта, необходимо установить на его страницы javascript-компонент (счетчик системы). Он используется для сбора информации о посетителях и передачи данных php-компоненту, расположенного на сервере, при каждом запросе посетителя к web-сайту. Php-компонент осуществляет первичную обработку и сохранение данных о посетителях в БД.

Система собирает общую информацию о посещениях сайта; она представляет основные отчеты, характерные для систем сбора информации о посетителях без интеллектуальных функций: число уникальных сессий за день, число загрузок страниц без учета уникальности посетителя, среднее число просмотров страниц в одной сессии, список самых популярных страниц за период времени, источники переходов, отчеты об индивидуальных характеристиках посетителей, первая и последняя страница в сессии, список ключевых слов, по которым посетители находят web-сайт в поисковых системах. Предусмотрен выбор диапазона дат, определяющих временной интервал, за который будет выводиться соответствующая разделу информация. Кроме того, система представляет сводные данные об использовании посетителями блоков навигации на страницах сайта.

Наряду с обычными функциями сбора статистики, система снабжена интеллектуальными функциями: она анализирует качество структурных элементов web-сайта методом нечеткого вывода и отображает результат анализа в наглядной форме, а также формирует файл в формате CSV для импорта в программу QuDA и осуществляет интерпретацию полученных правил.

3.3. Сфера применения результатов работы

Предлагаемая в работе система может применяться для анализа эффективности работы любого web-сайта и оценки качества его структурных элементов. Особенно актуально ее применение для информационно нагруженных сайтов, обладающих сложной структурой.

Пользователями разработанной системы представляются специалисты, занимающиеся поддержкой работы web-сайта, оптимизаторы, web-аналитики, специалисты по Интернет-маркетингу.

4. Перспективы продолжения работы по данной теме

1. Расширить возможности системы с тем, чтобы она могла собирать детальную информацию по отдельным информационным блокам web-сайта (текст, новость и пр.).
2. Построить систему нечеткого вывода для оценки качества отдельного информационного блока, взяв в качестве входных параметров качество страницы, количество кликов на этот блок, время просмотра страницы.
3. Создать модуль системы, задачей которого будет мониторинг популярности ключевых слов, заданных для текстов сайта в основных поисковых системах. Можно выделить следующие критерии популярности:
    a. Объем материалов по ключевому слову в основных поисковых системах.
    b. Число запросов в месяц по заданному слову.
    c. Позиции сайта в поиске по заданному слову.

    Этот модуль позволит получить более достоверную оценку о причине популярности или непопулярности отдельной страницы.
4. Разбить посетителей на группы, например «Покупатель», «Искатель», «Читатель», «Случайный» и, применяя ДСМ-метод, решить задачу определения класса посетителя по его поведению на сайте.
5. Добавить программную реализацию ДСМ-метода в систему. Это позволит отказаться от импорта файлов в среду QuDA, что упростит процедуру анализа.
6. Систему классификации страниц сайта методом нечеткого вывода можно дополнить дефаззификацией выходной переменной, тогда качество страниц будет иметь численное представление. Руководствуясь этим числом, система может выполнять какие-либо действия. Например, страницы с качеством ниже 10% предложить на удаление или реорганизацию.

5. Итог

Апробация созданной системы показала ее работоспособность и соответствие предъявляемым к ней требованиям. В дальнейшем предполагается развивать эту систему.

Список литературы

[Chen at al., 1996] Chen M.S., Park J.S., and Yu P.S. Data mining for path traversal patterns in a Web environment // In Proceedings of the 16th International Conference on Distributed Computing Systems, стр. 385-392, 1996.

[Cooley at al., 1997] Cooley R., Mobasher B. and Srivastava J. Web Mining: Information and Pattern Discovery on the World Wide Web, 1997.

[Elo-Dean at al., 1997] Elo-Dean S. and Viveros M. Data mining the IBM official 1996 Olympics Web site // Technical report, IBM T.J. Watson Research Center, 1997.

[Gergely at al., 2007] Gergely T., Anshakov O., Finn V., Kuznetsov S. Cognitive research: Formal approach. // Series: Artificial Intelligence – Springer-Verlag, 2007

[Grigoriev, 2003] Grigoriev P. QuDA, a Data Miner's Discovery Environment // Technical Report. – Technische Universität Darmstadt, 2003.

[Mannila at al., 1996] Mannila H. and Toivonen H. Discovering generalized episodes using minimal ocurrences // In Proc. of the Second Intel Conference on Knowledge Discovery and Data Mining, стр. 146-151, Portland, Oregon, 1996.

[Scime, 2005] Scime A. Web mining: applications and techniques, 2005.

[Yan at al., 1996] Yan T., Jacobsen M., Garcia-Molina H., and Dayal U.. From user access patterns to dynamic hypertext linking // In Fifth International World Wide Web Conference, Paris, France, 1996.

[Zadeh, 1965] Zadeh L. Fuzzy sets // Information and Control. — 1965. — №8. — P. 338-353.

[Барсегян и др., 2004] Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Minimg. – СПб.: БХВ-Петербург, 2004.

[Григорьев и др., 2002] Григорьев П.А., Евтушенко С.А. ДСМ-рассуждение как средство интеллектуального анализа данных. Результаты тестирования на наборах данных UCI, 2002.

[Кузнецов и др., 1996] Кузнецов С.О., Финн В.К Об одной модели обучения и классификации, основанной на операции сходства // Обозрение прикладной и промышленной математики. 1996. Том 3. Выпуск 1.

[Финн, 1983] Финн В.К. О машинно-ориентированной формализации правдоподобных рассуждений в стиле Ф.Бэкона – Д.С. Милля // Семиотика и информатика. – 1983. – Вып. 20.

[Финн, 1991] Финн В.К. Правдоподобные рассуждения в интеллектуальных системах типа ДСМ // НТИ. Сер. 2. - 1991. - № 15.