ИСПОЛЬЗОВАНИЕ ВНЕШНИХ СТАТИСТИЧЕСКИХ ДАННЫХ ПРИ ПЕРСОНАЛИЗАЦИИ САЙТА

К.П. ШЕРЕМЕТЬЕВ, *доц. каф. автоматизации и управления МГУЛ, канд. техн. наук,*
А.Г. ЦАРЕВ, *асп. факультета ЭСТ МГУЛ*

Интернет предоставляет большие возможности по сбору и анализу статистических данных сайта. Сбор и анализ статистической информации является неотъемлемой частью стратегии развития сайта и позволяет проводить полноценные исследования рынка покупателей, дает оценку коммерческим действиям и таким образом дает возможность планировать бизнес и претворять в жизнь наиболее удачные решения [5].

Постановка задачи

Целью данной статьи является исследование возможности использования статистических данных поисковых систем для персонализации сайта.

Средства сбора статистических данных

В сети Интернет статистика приобретает статус основного исследовательского инструмента [5].

Существует три общепринятых способа сбора статистических данных.

Внешний счетчик (ВС). Представляет собой специальный скрипт, который загружается у пользователя одновременно с загрузкой странички веб-сайта и передает на сервер статистики информацию об этой загрузке [5]. Услуги пользования внешними счетчиками предоставляют такие статистические сервисы, как Spylog, TopMail, Hotlog и т.д. Большинство внешних счетчиков распространяется на бесплатной основе.

Программа – анализатор лог-файлов (ЛА). Программы, которые по специальным алгоритмам обрабатывают лог-файл – файл сервера, в котором фиксируются все действия посетителей. ЛА распространяются на платной основе. Наиболее популярные из них – это Webtrends, Analog, Netpromouter.

Внутренний счетчик. Представляет собой скрипт, который сохраняет сведения о загрузке страницы в отдельный файл или собственную базу данных сайта. Недостатки внутренних систем зависят от квалификации разработчиков и выделенными на нее средствами. По сравнению с ВС и ЛА внутренние счетчики имеют ряд неоспоримых преимуществ:

1) конфиденциальность собранных статистических данных;

2) практически неограниченное время хранения собранных данных;

3) возможность защиты от полной потери собранных данных;

4) возможность учета полной информации о просмотренной странице;

5) возможность применения полученных статистических данных при персонализации сайта.

Разработанная авторами система персонализации сайта www.aldera.ru [3] основана на внутреннем счетчике. Посредством этого счетчика осуществляется сбор следующих данных:

1) IP – адрес посетителя;

2) идентификатор сеанса;

3) URL-адрес предыдущей страницы (реферер);

4) наименование площадки, с которой осуществлялся переход на страницу; площадка (ПП) – это специальное место на странице сайта, которое отведено под ссылки: прайс товаров, рекламные площадки, персонализированные рекламные площадки;

5) тип содержимого просмотренной страницы: новость, описание товара, статья;

6) заголовок просмотренной страницы (товара, статьи, новости);

7) дата и время перехода на страницу;

8) количество показов страницы в момент перехода;

9) количество переходов на страницу в момент перехода;

10) количество показов страницы в момент обращения к статистической информации;

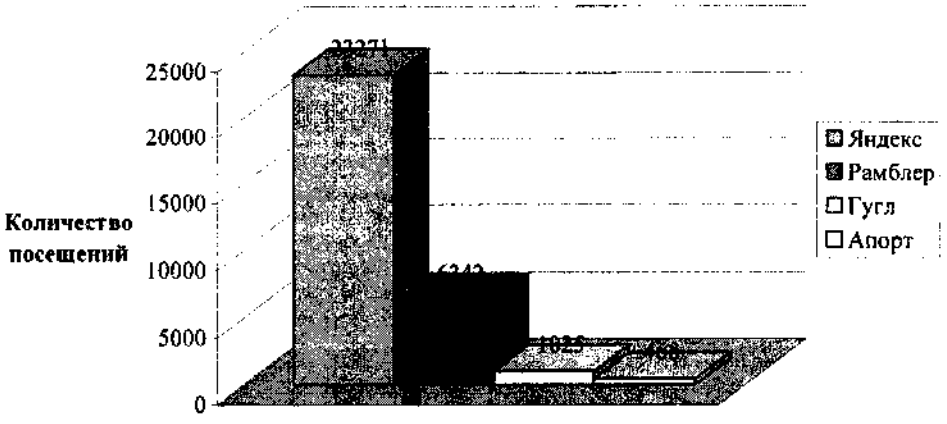

Рис. 1. Посещаемость сайта www.aldera.ru из поисковых систем с 24.03.2005 по 27.01.2006

11) программное обеспечение, которое использовал посетитель при просмотре страницы.

Эти данные используются для анализа поведения пользователя на сайте.

Метод персонализации, основанный на внешней статистике

Существующие программные продукты в области персонализации используют статистику, собранную и локализованную посредством только своих собственных средств. Такой подход не позволяют учесть общую модель поведения пользователей в Интернете. Для построения такой модели необходимо использовать внешние источники. Так как основная часть посетителей большинства сайтов приходит при помощи поисковых систем (ПС) [1], именно они являются наиболее важными источниками. Для использования данных ПС предлагается:

1) выбрать ПС, данные которой будут использованы;

2) выявить список соответствующих запросу товаров (ССЗТ) из поискового запроса посетителя;

3) определить рыночную заинтересованность товаров (R);

3.1) определить спрос на товары из ССЗТ;

3.2) определить количество конкурирующих сайтов;

3.3) рассчитать рыночную заинтересованность товаров по формуле

$$R = S/N_k,$$

где S – спрос на товар;

N_k – количество конкурирующих сайтов по товару;

4) рекомендовать товары из ССЗТ, ранжированные по убыванию рыночной заинтересованности.

После перехода на интересуемый товар алгоритм рекомендации не изменяется за исключением второго пункта: для выявления ССЗТ вместо поискового запроса будет использоваться наименование текущего товара.

Выбор используемой поисковой системы

Для выбора ПС рассмотрим статистику сайта www.aldera.ru:

Из рис. 1 видно, что посещения с Яндекса являются наиболее частыми. Поэтому под внешним источником статистических данных будем понимать поисковую систему Яндекс.

Выявление искомых товаров из поискового запроса

Посещение из поисковой системы несет в себе формализованный пользователем поисковый запрос (ПЗ). Посредством сравнения ПЗ с представленными на сайте товарами предлагается определить список искомых товаров.

Алгоритм определения списка искомых товаров.

1. Выделяем ПЗ из реферера.

2. Сравниваем каждое слово из ПЗ, кроме общеупотребительных и стоп–символов, со словами из наименования товаров. Просматриваемый товар в сравнении не участвует.

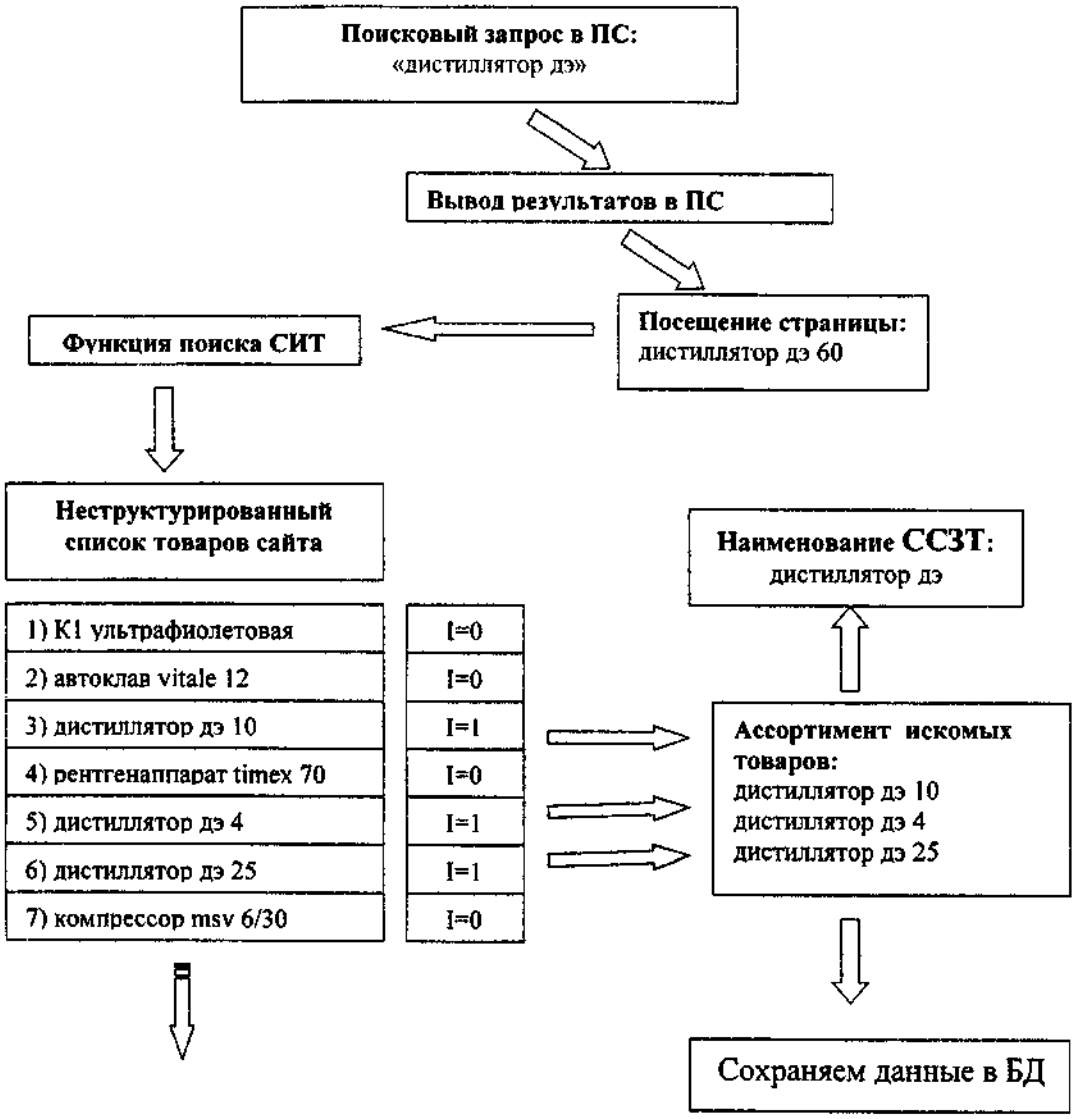

Рис. 2. Схема определения списка искомых товаров

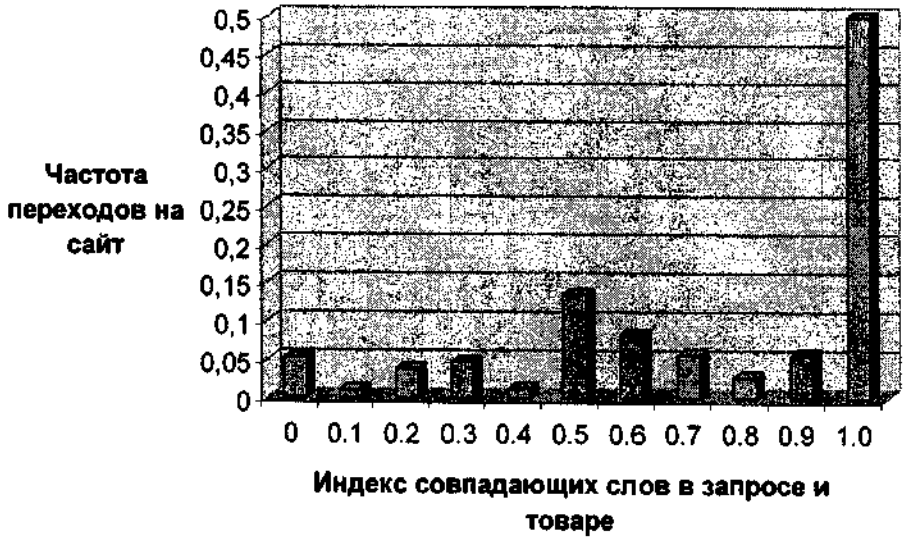

Рис. 3. Гистограмма частоты переходов на сайт в зависимости от количества совпадающих слов в запросе с наименованием товара (данные с 27.10.2005 по 27.01.2006)

4. Для каждого товара рассчитывается индекс совпадающих слов (I)

$$I = Ez/Ni,$$

где Ez – количество слов в запросе, которые совпадают с наименованием товара;

Ni – количество слов в запросе.

5. Товары, у которых Индекс совпадающих слов максимален, заносятся в ССЗТ.

6. Совокупность совпадающих слов принимается за название ССЗТ.

Необходимость использования предложенного метода подтверждается обработкой статистических данных.

Из рис. 3 видно, что пик переходов на сайт достигается при полном вхождении запроса в наименование товара. Поэтому предлагается сузить список предлагаемых товаров до товаров, наиболее схожих с запросом.

Определение рыночной заинтересованности предлагаемых товаров

Для определения рыночной заинтересованности на товары необходимо определить их спрос и количество конкурентов. Спрос на товар определяется количеством запросов к ПС Яндекс, а количество конкурентов – количеством сайтов, которые предлагают товар.

Поисковая система Яндекс посредством сервиса «Яндекс.Директ» отображает количество запросов по словам и фразам.

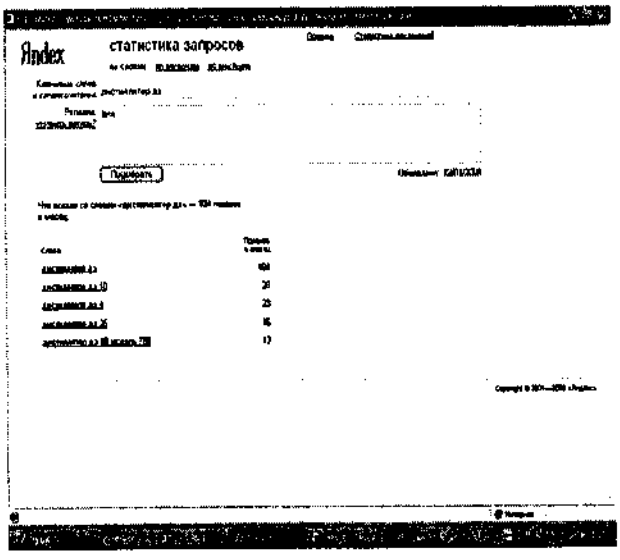

Рис. 4. Страница «Яндекс.Директ» с запросом «дистиллятор дэ»

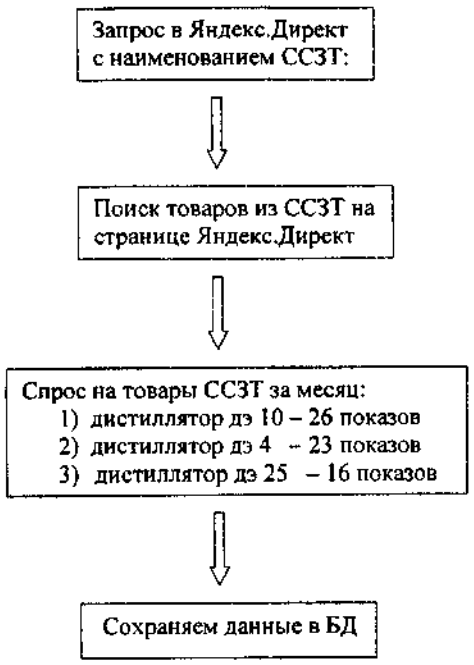

Рис. 5. Схема определения спроса на предлагаемые товары

Таким образом, определение спроса на товары сводится к загрузке их наименования ССЗТ в сервис «Яндекс.Директ» и обработке загруженной страницы.

Алгоритм определения спроса на товары из ССЗТ:

1. Загружаем страницу «Яндекс.Директ» с наименованием ССЗТ.

2. В загруженной странице ищем строки с наименованием товаров из ССЗТ.

3. Определяем количество запросов каждого товара.

4. Сохраняем данные в базе данных (БД).

Таким образом получаем спрос на предлагаемые товары. На величину спроса влияют такие факторы, как знание посетителя о модельном ряде искомого товара, его популярность в близком посетителю обществе и т.д., поэтому нулевое значение спроса в «Яндекс.Директ» не означает отсутствие необходимости в данном товаре. Если у всех товаров из ССЗТ значение спроса равно нулю, то для последующей сортировки приравниваем спрос одной второй.

В результате поискового запроса посетителю выдаются страницы с ссылками, заголовками и аннотациями найденных документов. Однако наибольший интерес представляет первая страница результатов.

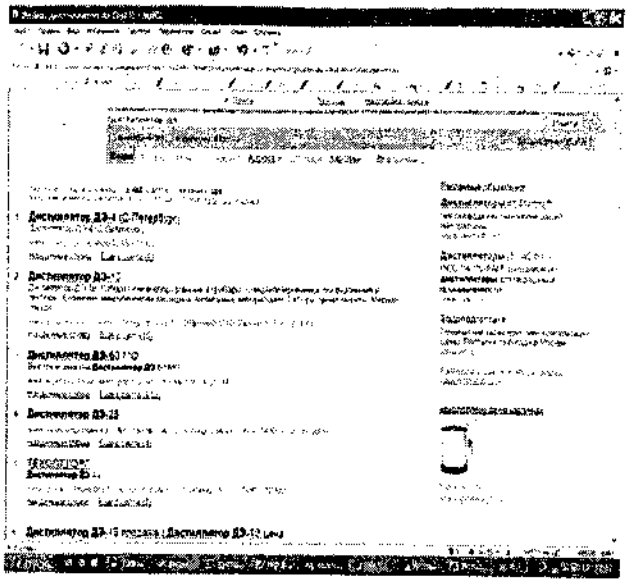

Рис. 6. Страница ПС Яндекс с результатами запроса «дистиллятор дэ»

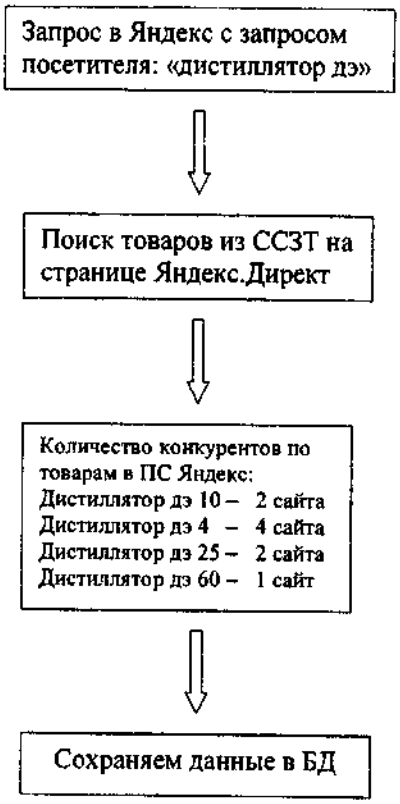

Рис. 7. Схема определения количества конкурентов на товары из ССЗТ

По данным исследовательской компании CyberAtlas, на вторую страницу переходит лишь 24 % посетителей [8]. Поэтому определять количество конкурентов будем по первой странице результатов ПЗ.

Алгоритм определения количества конкурентов на товары из ССЗТ.

1. Загружаем страницу «Яндекс» с ПЗ.

2. В загруженной странице ищем строки с наименованием товаров из ССЗТ.

3. Определяем количество сайтов для каждого товара.

4. Сохраняем данные в БД.

После того как определены спрос и количество конкурентов, определяется рыночная заинтересованность на товары. Высокий спрос и небольшое количество конкурентов повышают вероятность благоприятного для владельца сайта, посещения страницы с описанием товара. Поэтому предлагается рекомендовать товары в порядке убывания рыночной заинтересованности.

Выводы

Приведенный в статье метод позволяет учесть общую модель поведения пользователей в Интернете.

На основании полученных данных подтверждается возможность использования статистических данных внешних источников при персонализации сайта, а также необходимость осуществления дальнейших исследований по данному направлению.

Библиографический список

1. Робин Ноблес Эффективный Web-сайт: учеб. пособие / Робин Ноблес, Керри-Лэй Греди.. – М.: Изд-во ТРИУМФ, 2004, 560с.
2. Шереметьев, К.П. Система персонализации данных для сайтов электронной коммерции / К.П. Шереметьев, А.Г. Царев // Вестн. Моск. гос. ун-та леса – Лесной вестник. – 2005. – №6(42). – С. 172–175.
3. Царев А.Г. «Интернет-магазин для стоматолога. Медицинский алфавит» //Медицинский алфавит. Стоматология. – 2005 – № 3(46). – С. 16-17.
4. Комплексный метод оценки эффективности Интернет-рекламы в коммерческих организациях. www.dis.ru 01/12/2002.
5. Место статистики в онлайн-продвижении URL: http://www.startpromo.ru/ modules/ sections/ index. php?op=viewarticle&artid=50 16/12/2005
6. Программа Site Statistics – объективная Интернет-статистика URL: http://netpromoter.ru/ bulkpromoter/ 16/12/2005
7. Статистика web-сайта, подходы и выбор. URL: http://www.codenet.ru/webmast/html/stat.php 22/12/2005
8. Зачем Вашему корпоративному сайту высокий рейтинг в поисковых системах? URL: http://www.marketlist.ru/why.php 21/01/2006.