ПРОФАЙЛИНГ ПОЛЬЗОВАТЕЛЕЙ СЕТИ ИНТЕРНЕТ. АЛГОРИТМЫ НЕЧЕТКОЙ ЛОГИКИ КАК ИНСТРУМЕНТ ДЛЯ РЕШЕНИЯ ЗАДАЧ ПРОФАЙЛИНГА

Забродин И.Е.

Томский политехнический университет, Институт кибернетики
zabr.din@gmail.com

Введение

Профайлинг веб-пользователей – это объединение их в группы на основании полученной информации об их поведении. Актуальность данной темы обусловлена тем, что с развитием технологий становится возможным собирать все большие объемы информации о пользовательской активности, но использовать ее без предварительной кластеризации практически бесполезно. Решение задач профайлинга представляет интерес для владельцев коммерческих интернет ресурсов, так как позволяет составить портрет потенциальных клиентов и вести продажи эффективнее, сопоставить информацию о том, кого удалось привлечь на свою веб-страницу, и на кого она действительно ориентирована. Кроме того, разделённая на группы аудитория представляет большой интерес для специалистов по контекстной рекламе, поскольку позволяет организовывать более точный таргетинг. Технология поисковой рекламы или рекламы на тематических ресурсах по сравнению с рекламой на основе профайлинга, не дает четкого понимания о социальном и демографическом положении пользователя (например, о поле и возрасте) и, следовательно, не может быть использована в рекламной кампании с необходимостью таргетинга по возрасту, полу, увлечениям, привычкам. Долгосрочная задача профайлинга гораздо шире, чем построение предположений, основывающихся на прошлом поведении пользователя в Интернет, она состоит в использовании прогнозирующей модели. Теоретически, если возможно проанализировать достаточное количество статистической информации о поведении пользователей в Интернет, то возможно прогнозировать не только где они были, но и куда, скорее всего, они направятся [1].

Источники данных

Источниками данных для кластеризации может служить как прямое предоставление пользователем информации о себе (заполнение анкет, регистрационных форм), так и самостоятельный сбор данных на стороне сервера. В первом случае информации удается собрать гораздо меньше, а также невозможно точно определить ее как достоверную. При самостоятельном сборе данных удается собрать гораздо больше информации как о посетителе в целом, так и о его поведении в рамках конкретной сессии. Это возможно благодаря записям серверных логов, использованию cookies файлов, внедрению в код веб-страницы отслеживающих java-скриптов. Основываясь на данных принципах, для использования в частных проектах были созданы современные комплексные программные продукты, такие как Google Analytics и Яндекс Директ. Кроме того, сбор информации для профайлинга проводится глобально поисковыми и рекламными компаниями с помощью роботов-мониторов [2].

Параметры, характеризующие посетителя

Современные технологии позволяют собрать множество исходных данных о пользователе, таких как время входа и выхода со страницы, ссылку-реферрер, IP-адрес, браузер, траектория движения курсора и многие другие. Анализируя эти параметры, можно получить множество данных для дальнейшей классификации. Например, самые простые расчеты позволят нам вычислить время, проведенное пользователем на странице. Сложные вычисления доходят до того, что мы можем с большой долей вероятности утверждать, мужчина или женщина посетил веб-страницу. Нет смысла использовать сразу все данные, и в разных системах классификации мы получим разный набор параметров, характеризующих посетителя. Но есть ряд самых важных и часто используемых параметров, а именно [3]:

- Географическое положение пользователя
- URL-реферрер
- Средняя длительность сеанса
- Глубина сеанса
- Используемое устройство и браузер
- Страницы входа и выхода
- Пол и возраст

Основные методы кластеризации

На сегодняшний день наиболее распространенными являются следующие подходы к профайлингу пользователей:

- Разведывательный анализ данных
- Эволюционное моделирование
- Нейронные сети
- Ассоциативные правила

Разведывательный анализ данных часто используется как методика предварительной обработки и подготовки записей к более глубокому анализу. Этот вид статического анализа позволяет уменьшить размерность данных, проверить взаимосвязи между переменными, выявить наиболее ценные с точки зрения кластеризации подмножества исходных данных. Результаты анализа отображаются в виде простых графиков и

таблиц для последующего анализа с использованием более сложных методов.

Эволюционное моделирование основано на существовании пары «ситуация – принятое решение» для каждого конкретного случая отнесения пользователей в ту или иную категорию. Данный метод позволяет группировать пользователей довольно точно на более поздних этапах исследования, но при этом требует больших трудозатрат на первом этапе. Зачастую это ручное вмешательство экспертов-аналитиков и создание экспертных систем на основании их работы.

Нейронные сети – мощный инструмент анализа, так же требующий тонкой настройки на первом этапе. Кроме того, требуется большой аналитический опыт для интерпретации данных. Так же, в результате работы нейронной сети могут появиться статистически правильные, но бессмысленные семантически данные.

Использование ассоциативных правил – достаточно простой и распространенный инструмент, основанный на принципах нечеткой логики [4].

Нечеткая логика при решении задач кластеризации

Нечеткая логика – математический аппарат, позволяющий относить объекты к тому или иному множеству с определенной долей вероятности. Характеристикой нечеткого множества выступает функция принадлежности (Membership Function). Обозначим через MFc(x), где C – нечеткое множество, X – значение конкретного параметра. Тогда нечетким множеством C называется множество упорядоченных пар вида C={MFc(x)/x}, MFc(x) [0,1]. Значение MFc(x)=0 означает отсутствие принадлежности к множеству, 1 – полную принадлежность.

Проиллюстрируем данный подход на примере понятия «длительное посещение Веб-страницы». В качестве области значений X в данном случае будет выступать время посещения в секундах от 0 до 100 и более. Нечеткое множество может выглядеть следующим образом:

C={0/0; 0,15/20; 0,60/40; 0,70/60; 0,80/80; 0,90/100; 1/100+}.

Таким образом, посещение страницы, продлившееся 60 секунд, принадлежит к множеству «длительное» с вероятностью 80%. В одной ситуации мы можем говорить о том, что пользователь провел на странице достаточно времени для достижения какой-либо цели. В другом – не достаточно. Именно в этом и проявляется нечеткость задания множества [5].

Суть использования данного инструмента в рамках методики ассоциативных правил состоит в том, что на основании начального набора строгих записей и заранее спроектированной базы ассоциативных правил формируется нечеткий логический вывод, показывающий, с какой долей вероятности можно считать пользователя отнесенным к той или иной категории. Для конечного представления данных пользователю аналитической системы часто выносится определенное решение, являющееся склонением нечеткого вывода в строну большей вероятности.

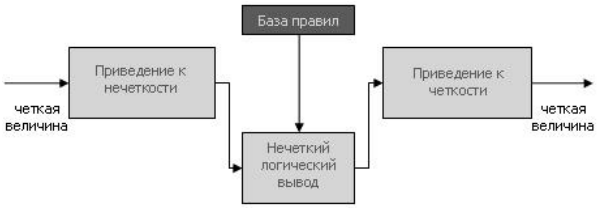

Рис.1. Схема построения нечеткого логического вывода.

Приведем пример использования методики. Краткосрочным посещением принято считать просмотр страницы менее 20 секунд. Предположим, время просмотра страницы Index.html в конкретном случае равно 15 секундам. Значит, это краткосрочное посещение. Если пользователь совершает краткосрочное посещение и страница его визита – главная, то с вероятностью 89% процентов можно отнести его посещение к «отказу» (незаинтересованность в дальнейшем просмотре Веб-ресурса) и считать, что он не вернется на сайт, и с вероятностью 11% утверждать о других факторах, приведших к такому поведению. В рамках данной системы пользователь был отнесен к группе случайных посетителей, не приносящих доход.

Заключение

Таким образом, разработка методов кластеризации с использованием нечеткой логики является перспективным направлением в веб-профайлинге. Настройка таких методов производится проще, чем у аналогов на рынке, а качество им не уступает.

Литература

1. Гребенщиков С. А., Силич В. А., Комагоров В. П., Фофанов О. Б., Савельев А. О. Технология разработки информационной системы поддержки принятия решений для управления проектными работами при обустройстве месторождений // Научно-технический вестник ОАО «НК Роснефть». - 2012. - Вып. 29 - №. 4. - С. 38-42

2. Ganti V., Gerhke J., Ramakrishan R. CACTUS Clustering Categorical Data Using Summaries. In Proc KDD’99. - 1999.

3. Kaushik A. Web Analytics: An Hour a Day. - Sybex - 2007.

4. Clifton B. Advanced Web Metrics with Google Analytics. - Wiley Publishing, Inc. - 2008.

5. Fasel D., Zumstein D. A Fuzzy Data Warehouse Approach for Web Analytics // Visioning and Engineering the Knowledge Society. A Web Science Perspective. - 2009. – С. 276-285