ВЕРОЯТНОСТНЫЕ АЛГОРИТМЫ ОЦЕНКИ МОЩНОСТИ МНОЖЕСТВ ДЛЯ СЕГМЕНТАЦИИ ПОЛЬЗОВАТЕЛЕЙ НА ОСНОВЕ ПРОИЗВОЛЬНЫХ ПРАВИЛ В РАМКАХ РЕКЛАМНЫХ КАМПАНИЙ

В. Э. Базаревский, В. Э. Базаревский
Кафедра программного обеспечения информационных технологий, Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: baz-val@yandex.ru, enigmanx@yandex.ru

В статье частично рассмотрены вероятностные алгоритмы оцентки мощности множеств, для оценки сегментации пользователей. Приведена оценка точности подходов, их плюсы и минусы. В статье приведено комплексное решение, которое частично решает проблемы каждого из подходов и обеспечивает довольно высокий уровень точности при не больших затратах ресурсов.

Введение

Развитие и повсеместное распространение как интернет, так и мобильных технологий в значительной мере изменило характер современных рекламных кампаний. В последнее время можно выделить такие тренды, как показ рекламы конкретной целевой аудитории вне зависимости от места, вместо ее показа на каком-либо целевом ресурсе (интернет-сайте), сегментирование пользователей по их поведению вместо сегментирования в результате опросов и базовых общедоступных данных. Также, одной из тенденций последних лет можно выделить ретаргетинг – стратегию показа рекламы пользователям, которые уже были замечены на том же интернет сайте, просматривающим рекламируемую категорию товаров.

I. Обзор предметной области

Согласно исследованиям, ретаргетинг повышает конверсию рекламных кампаний до 70%. Вместе с тем, ретаргетинг в базовом его виде может вызывать отторжение и недовольство пользователей, а кроме того оперирует только категориями пользователей, которые уже заинтересованы рекламируемой категорией товаров, хотя очевидно множество пользователей со схожими характеристиками является куда более общим. Для повышения качества ретаргетинга (увеличения конверсии, снижения вероятности недовольства пользователей) растет потребность в построении более сложных сегментов пользователей на основе их поведенческих данных. Такие данные можно получить из транзакционных логов современных рекламных серверов, поисковых запросов, аналитических систем сбора информации, логов рекламных интернет-аукционов и т.д. Такая информация деперсонифицирована, но при этом оперирует уникальным, пусть и обезличенным, идентификатором пользователя. Учитывая частоту пользования интернетом современных пользователей, современные провайдеры интернет-рекламы оперируют терабайтами данных в день. При этом, такая информация подвержена очень сильному старению. Согласно статистике, половина собранной таким образом инфомации устаревает в течение первых нескольких дней, таким образом условное время жизни рекламного уникального идентификатора пользователя обычно составляет 14 дней и занимает минимум около 15 ТВ. Современные аналитические базы данных могут обрабатывать подобные объемы данных, однако они как правило достаточно дорогостоящи, оперируют в большинстве статическими данными и рассчитаны на пользование несколькими пользователями, а потом не способны при приемлемой цене/скорости соответствовать заявленным нефункциональным требованиям.

Как правило, подобные задачи решаются созданием специализированной базы данных, не способной рассматриваться как универсальное средство хранения данных, но вместе с тем превосходящей существующие универсальные базы данных в плане цены/качества. Если рассмотреть специфику решаемой задачи, то ее можно разделить на две части:

- Сегментация пользователей на основе заранее созданных правил;
- Оценка размеров сегментов пользователей при создании правил.

Следует отметить, что как правило такие сегменты пользователей создаются до начала рекламной кампании, либо даже в ее процессе, но сегментирование пользователей на основе правил как правило не обязательно делать в режиме реального времени, таким образом, допустимо когда процесс сегментирования занимает до нескольких часов после создания правила. Существующие распределенные системы обработки больших данных, такие как Hadoop, Spark, Storm, Flink справляются с такой задачей в общем виде, хоть и не обеспечивают режима реального времени. Вместе с тем, оценка размеров сегментов пользователей является важной функ-

циональностью при создании сегментов пользователей, а так же анализе рекламных кампаний. При этом такая оценка представляется практически нереализуемой на основе существующих аналитических баз данных, так как они оперируют реляционным представлением данных, тогда как для такой оценки необходимо представление данных на основе теории множеств.

II. Существующие подходы оценки размера множеств

В настоящее время, существует несколько конкурирующих подходов для оценки размера множеств по произвольному правилу, имеющих свои достоинства и недостатки. При том, как правило из требований: динамичности данных, минимизации издержек хранения данных и точности оценки такие подходы реализуют только два из трех требований жертвуя третьим. Среди них можно выделить:

1. Оценка размера множеств на основе битовых масок и семплирования данных по произвольной хэш-функции, имеющей близкое к равномерному распределение (высокое качество и режим реального времени при высоких издержах на хранение данных в оперативной памяти, до 500 гигабайт, отстутствие горизонтального масштабирования);
2. Оценка размера множеств на основе вероятностных структур данных (hyperloglog и minhash. Режим реального времени, низкие издержки на хранение, хорошее горизонтальное масштабирование, проблемы с точностью в случае когда размер множеств, ущаствующих при построении правил превышает более чем в 100 раз размер результирующего множества);
3. Оценка размера множеств на основе сжатых битовых масок и succinct data structures (хорошая производительность, низкое потребление ресурсов хранения, невозможность динамического обновления множеств, на основе которых производится построение сегментов, что ведет к деградации точности оценки при построении сегментов).

В силу того, что часто ни один из вышеописанных подходов не решает всех трех проблем: высокой точности, хорошего горизонтального масштабирования, режима реального времени, предлагается альтернативный подход, который бы соответствовал всем трем характеристикам в достаточной мере, пусть и был бы хуже по отдельным из них.

Таким образом, в качестве базовых стоить взять подходы 2 и 1. Недостатком первого подхода является низкая точность в том случае, когда размер результирущего множества много меньше размеров составляющих его множеств (в частности, при пересечении множеств большего размера). Происходит это из-за относительного характера ошибки вероятностных структур данных. Так относительная ошибка в 2% при оценке множества A из 1000000 элементов приведет к ошибке в 2000% при оценке множества в 1000 элементов, если оно построено путем усечения множетсва A. Недостатком же подхода 1 является необходимость разреженного хранения всей битовой маски базового сегмента даже если его размер минимален. При том, что только малая часть таких сегментов имеет действительно большой размер (около 1%). Вместе с тем, игнорирование “большого хвоста” малых по размеру сегментов может значительно понизить качество оценки размеров пользователей из-за того, что очень часто сегменты как раз строятся из малого размера базовых сегментов.

1. P. Clifford and I. A. Cosma. A statistical analysis of probabilistic counting algorithms. Scandinavian Journal of Statistics, pages 1–14, 2011.
2. J. Lumbroso. An optimal cardinality estimation algorithm based on order statistics and its full analysis. In Analysis of Algorithms (AOFA), pages 489–504, 2010.
3. R. Pike, S. Dorward, R. Griesemer, and S. Quinlan. Interpreting the data, parallel analysis with Sawzall. Journal on Scientific Programming, pages 277–298, 2005.
4. K. Aouiche and D. Lemire. A comparison of five probabilistic view-size estimation techniques in OLAP. In Workshop on Data Warehousing and OLAP (DOLAP), pages 17–24, 2007.
5. Z. Bar-Yossef, T. S. Jayram, R. Kumar, D. Sivakumar, and L. Trevisan. Counting distinct elements in a data stream. In Workshop on Randomization and Approximation Techniques (RANDOM), pages 1–10, London, UK, UK, 2002. Springer-Verlag.