ОБЩИЕ ПРИНЦИПЫ
ИНТЕГРИРОВАНИЯ ДАННЫХ ИЗ РАЗНЫХ ИСТОЧНИКОВ

Нина Базилевская, Head of Market Science Ipsos в России | Октябрь 2020

GAME CHANGERS

Ipsos

Информационная эра характеризуется наличием огромного количества информации из огромного числа источников. Зачастую, для анализа данных, моделирования и выводов уже недостаточно данных из одного источника, а требуется совмещение данных различного типа из разных, независимых источников.

Кроме того, в последнее время наблюдается тенденция к сокращению количества вопросов в анкете и уменьшению средней длины интервью. Мы стремимся облегчить жизнь респонденту, чтобы длина интервью не отражалась на качестве его ответов. Таким образом, объединение данных из разных опросов или из других источников становится все более востребованной задачей.

Основная цель объединения данных состоит в получении надежной и более полной информации из двух и более источников, основываясь на взаимосвязях между источниками данных.

ПОДХОДЫ К ИНТЕГРИРОВАНИЮ ДАННЫХ

Существуют разные подходы к интегрированию данных на агрегированном и индивидуальном уровне, которые применяются в соответствии с конкретной задачей.

Мы выделяем 4 основных подхода в маркетинговых исследованиях:

- **Слияние данных.** Это интегрирование данных, когда наблюдения из двух и более наборов данных имеют прямую связь между собой. Наиболее типичный пример из нашей сферы – это опрос по базе заказчика. У заказчика есть данные о своих клиентах, например, из CRM-системы, а также есть данные опроса. Таким образом, имея общий идентификационный номер в данных опроса и в данных заказчика, мы можем соединить «объективные» поведенческие данные с опросными данными.
- **Агрегирование и сопоставление.** Агрегирование данных в каждом из рассматриваемых источников по одному основанию и перевод в общий набор данных. Самый распространённый пример – это эконометрическое моделирование/MMM (marketing mix modeling), основанное на анализе временных рядов, где общее основание для агрегирования является распределение по временным периодам.
- **Восстановление данных.** Мы можем восстановить недостающие переменные в наборе данных, зная закономерность между новой переменной и существующими переменными. При этом сама закономерность находится на том наборе данных, который включает в себя все необходимые переменные. Например, мы часто восстанавливаем переменную с сегментами в новом наборе данных, определив закономерности между значениями переменных опроса и сегментами на исходной базе.

- **Фьюжн данных (Data Fusion).** Соединение данных по принципу интерполяции данных по похожим профилям объектов. Сначала определяется «похожесть» профилей респондентов по соц-дем характеристикам (чаще всего) и по ключевым интересующим переменным, например, пользование определенными марками/маркой, затем данные объединяются в общие наблюдения у максимально похожих респондентов. механизм, как кризисы обучают потребителей, фиксировать эти уроки и учитывать их в своих стратегиях развития.

Один из самых интересных примеров Data Fusion — это совмещение данных опроса с профилями социальных сетей.

Выбор того или иного метода объединения данных определяется целью анализа данных – бизнес задачей, а также типом данных и возможностями источников.

Возможно также комбинировать различные подходы совмещения данных. Например, слияние данных по «своим клиентам» и восстановление данных для остального рынка.

ПРИНЦИПЫ ОБЪЕДИНЕНИЯ ДАННЫХ

Несмотря на то, что подходы довольно сильно отличаются по своей технике, есть общие принципы объединения данных:

- **Понимание бизнес задачи.** В целом понимание бизнес задачи — это залог успеха любого исследования и анализа данных. Никакие новые технологичные решения не будут успешны, если они не решают поставленную задачу. Также, если мы хотим интегрировать данные из разных источников, то первое что мы должны понять, для чего это нужно. Что мы хотим получить на выходе? В зависимости от того, что мы хотим получить, прогноз целевой переменной, таргетирование или сегментирование, создается план анализа и выбирается оптимальный подход.

- **Правильное определение объекта анализа.** Например, если основной объект анализа – это человек (сегментация, таргетирование), то наилучший подход к объединению данных – слияние данных, если имеется единый ключ для сопоставления, или Data Fusion, если нет общего ключа для слияния данных. В случаях, когда объектом анализа является марка, наилучшим подходом является агрегирование и сопоставление данных.

- **Формирование гипотез перед началом интеграции данных и анализом.** Формирование гипотез – это необходимая часть любого аналитического плана. Прежде, чем приступить к работе по объединению данных, нам нужно понять какие вопросы мы проверяем. В зависимости от этого определяются критерии для объединения. Особенно это важно в случае сложного объединения, такого, как Data Fusion, когда нам необходимо выбрать, какие общие характеристики будут выбраны как критические (совпадение должно быть 100%), а какие – как дополнительные, а также какой из наборов данных будет «донором», а какой набор будет «получателем» данных.

  Например, если мы изучаем отношение мам к питанию детей путем опроса, и хотим обогатить эту информацию их поведением в социальных сетях из другого набора данных, а также у нас есть гипотеза, что активность[1] мамы в социальных сетях очень сильно связана с возрастом ребенка, то значит наш опрос – это набор «получатель», а данные соц-сетей – «донор», при этом переменная «возраст ребенка» - это критический показатель, а, например, пол ребенка может быть дополнительным показателем.

[1] Под активностью имеется в виду активное обсуждение проблем питания ребенка в социальных сетях

- **Точное знание структуры данных.** Очевидно, перед объединением следует подробно изучить структуру данных обоих дата сетов: названия переменных, тип данных, коды, допустимые значения и т.д., удостовериться, что идентичные переменные не просто по содержанию похожи, но и идентичны по кодировке и типу данных. Например, если используется интервальная шкала по возрасту, то возрастные интервалы и их коды должны совпадать. Кроме того, требуется предварительная обработка данных обоих наборов перед их объединением.

- **Понимание ограничений методов/подходов.** Для валидности результатов необходимо понимать ограничения каждого из подходов и возможности каждого из наборов данных, которые мы собираемся объединить в один.

  При слиянии баз по единому ключу, когда не все наблюдения доступны по ключу, нужно понимать, какие могут быть смещения по распределению ключевых показателей и как их можно скорректировать.

  Если речь идет об агрегировании на основе временных периодов, то важно понимать, что временные периоды должны быть идентичны, и их количество должно быть достаточно для анализа. Также важно проверить волатильность показателей, так как отсутствие волатильности ведет к невозможности выявить корреляции между переменными. Кроме того, важное ограничение вносят кризисные явления или несистемные события в рамках рассматриваемой истории.

  У подхода по восстановлению данных также есть ограничения. Для переноса закономерностей из одного набора в другой нужно убедиться достаточно ли у нас переменных в наборе «получателе» для воспроизведения этих закономерностей, и сопоставимы ли выборки по социально-демографическим и другим критическим параметрам.

  В рамках Data Fusion существуют еще множество подходов к реализации, которые накладывают свои ограничения. Основные ограничения связаны с наличием достаточного количества общих переменных в объединяемых наборах и достаточного количества наблюдений в каждом из наборов.

  Таким образом, задача объединения данных, как и все аналитические задачи, требует понимания бизнес целей, детального плана анализа и понимание ограничений. «Объединение ради объединения» не ведет к успешности решения задачи, так как за любой технической, математической задачей должна быть четко поставленная содержательная цель. Соблюдение основных принципов интегрирования данных из разных источников увеличивает эффективность решения задачи и достижения бизнес цели.

Нина Базилевская, Head of Market Science Ipsos в России | Октябрь 2020