ESOMAR “RESEARCH WORLD”: 2024/04/02

合成データ: データに基づく意思決定のゲームチェンジャー

4 月 2 日

著者 Kathryn Topp (CEO of Yabble)

インサイト・テック企業 Yabble の CEO 兼創設者である Kathryn は、リサーチとソフトウェアの分野で 20 年の経験を持つ受賞歴のあるリサーチャーである。

合成データは、インサイト業界で勢いを増し始めています。合成データの作成と使用は、現実世界のデータの不足の増加に対する単なる対応策ではありません。実際、合成データは、大きなデータギャップを埋めるために設計された戦略的なツールです。

合成データは、インサイト業界で勢いを増し始めています。Research World に最近掲載された記事では、合成データが、特にその潜在的な誤用や、元のデータ収集方法に取って代わった場合に引き起こす危険性について、懐疑的な見方をもって検討されました。

この批判は、技術の進歩に対する昔からの懸念を示しています。人間は生まれつき変化を嫌うものです。特に、変更されるのが確立された慣行（この場合はデータ収集）である場合はそうです。この記事の目的は、これらの懸念を掘り下げ、合成データがもたらす潜在的な機会をリサーチャーに理解してもらうことです。

戦略的ツールとしての合成データ

合成データの作成と使用は、現実世界のデータの不足の増加に対する単なる対応ではありません。実際、合成データは、大きなデータギャップを埋めるために設計された戦略的ツールです。たとえば、**「動的な AI 駆動型ペルソナ」**は、実際の顧客とのやり取りを正確に模倣し、従来は広範なフィールドワークを必要としていたインサイトを提供できます。

生成 AI は、調査や人間の参加者だけに頼るのではなく、従来のデータ収集メカニズムで得られるものよりも通常より幅広く、最新かつ正確で膨大なデータセットを活用して、人間のインプットの代わりとして機能します。

この方法により、インサイトを得るための新しい方法の探求が容易になり、質問の範囲が広がります。その結果、バーチャルな聴衆は、複数のソースからのデータを動的に処理することで、さまざまなトピック、市場、人口統計に関する包括的なインサイトを即座に提供します。

これは「フェイクデータ」ではないことを知っておくことが重要です。これは、従来の調査データ、公

開されている統計、トレンドレポートなど、多数の信頼できるソースから収集された非常に現実的で信頼性の高いデータから合成および派生された、AI 生成データです。急速に進化する市場環境では、このようなリアルタイムのデータ合成は有益なだけでなく、ますます必要になっています。

データの豊かさの進化

合成データは私たちを現実から切り離す可能性があると批判する人たちもいますが、彼らは生成 AI が単にデータを蓄積するだけではなく、データに命を吹き込むという事実を見落としています。これにより、時間の経過とともに成長し、適応し、継続的な関連性と正確性を確保できます。また、その正確性は従来のデータに匹敵するだけでなく、従来のデータと競合します。最近の Marketing Week の記事で、マーケティング愛好家の Mark Ritson 氏 は、「AI から得られる消費者データのほとんどは、三角測量すると、主要な人間のソースから生成されたデータと約 90% 類似している」と述べています。

拡張データは、独自のデータセット、学術コンテンツ、リアルタイムの Web 検索などの多様なデータソースを高度な機械学習アルゴリズムと統合することで、インサイトをさらに深めることができます。この方法は、従来の調査や人間のフィードバックを超え、生成 AI を利用してデータを総合的なナレッジ レイクに統合します。このモデルでは、データの融合によって詳細なペルソナを作成し、関連する質問と回答を作成し、重要なインサイトを抽出して、ユーザーに関心のあるトピックのニュアンスに富んだ多次元ビューを提供し、正確で AI 主導のインサイトを提供します。

従来のサンプルの課題への対応

市場調査業界が従来のサンプルパネルでデータ品質の課題を抱えていることは周知の事実です。データ品質は、市場調査のパートナーまたはサプライヤーを選択する際に、断然最も重要な要素です（GRIT レポート 2020）。従来のパネルは、リサーチャーと協力して調査設計の改善をサポートするなど、あらゆる手段を講じてこれらの課題に対処しようとしていますが、調査のユーザーは合成データなどの代替データソースにも目を向けるべきだと考えています。

合成データは、市場調査における従来の調査データによく見られるサンプル品質の問題に対処するのに役立ちます。従来のサンプル データには、あらゆる対策を講じたにもかかわらず、表面的な回答、誤った入力、怠慢な回答者が含まれることがよくあります。これにより、調査結果が大幅に歪んでしまい、データセットの忠実度が低下する可能性があります。

逆に、実際のデータセットから取得され、高度なアルゴリズムで作成された合成データは、よりクリーンで、より制御された一連のインサイトを提供します。調査の回答によく見られるノイズや無関係な情報を最小限に抑え、利用されるデータの質の高さを保証します。リサーチャーは、従来のサンプル データによくある不正確さや表面的な部分に煩わされることなく、精度の高い領域に踏み込むことができます。

この人工的なデータ生成は、主観的な判断ではなく、事前に決められたルールとパラメータに依存するため、人間のデータ収集者が持ち込む可能性のある無意識の偏見を軽減するのにも役立ちます。もちろん、LLM には固有の偏見がありますが、これらは通常識別可能であり、調整できます。調査回答者や

調査設計によって無意識にもたらされるものではありません。さらに、合成データは既存のデータセットのギャップを埋めることができ、より総合的で包括的な母集団の視点を提供します。

調査の革新と効率性

合成データはすでに、リサーチをより安価でスケーラブルなものにし始めています。従来のデータ収集方法では、データの収集、クリーニング、検証という長いプロセスを伴うことが多く、時間とコストがかかります。一方、合成データは、大量のデータを迅速に生成でき、特定のリサーチニーズに合わせて調整できます。つまり、リサーチャーは、現実世界のデータを収集するのにかかる時間のほんの一部で、膨大なデータにアクセスできます。

大規模なフィールドワークや調査を必要とせずに現実世界のシナリオを模倣したデータを作成できるため、リサーチャーはデータ収集ではなく分析と解釈に集中できます。

さらに、合成データは、リサーチの運用効率にとって極めて重要なデータ品質と一貫性の面で大きな利点があります。従来のデータ収集では、不一致やギャップが頻繁に発生し、多くの場合、追加のデータ収集ラウンドや複雑なデータクリーニング手順が必要になります。合成データは、特定の品質基準に準拠し、欠損値や外れ値などの一般的なデータの問題がないようにプログラムできるため、より高いレベルの整合性が保証されます。この整合性により、リサーチャーは作業中のデータを信頼でき、データの検証と前処理に費やす時間が短縮されます。さまざまなシナリオと条件をシミュレートできるため、モデルと仮説のより包括的なテストと検証も可能になり、より堅牢で信頼性の高いリサーチ結果につながります。

合成データの今後の方向性

合成データに対する懐疑的な見方は変革をもたらすテクノロジーに対する自然な反応ですが、特にその正確性、関連性、有用性を考慮すると、こうしたイノベーションを認識し、受け入れることが重要です。

合成データに関する議論が進むにつれ、テクノロジーがデータへのアプローチ方法を進化させていることは明らかです。合成データは、市場調査やその他のさまざまな領域を再定義する可能性を秘めています。責任ある利用と継続的なイノベーションにより、合成データは、デジタル時代の洞察力に富んだ意思決定に不可欠な要素となり、従来のデータ収集方法を補完し、場合によっては強化することになります。

以上