調査データと大規模データの融合への取り組み

～大規模データの個人プロフィール推定～

株式会社ビデオリサーチ　IT・技術推進局　IT2 部ロジックグループ
片柳　伊佐

1.　会社紹介とはじめに

ビデオリサーチでは、テレビ視聴率やラジオ聴取率に代表されるメディア接触測定や、マーケティング調査、広告効果測定調査など多種多様な調査を行なっています。テレビ視聴率調査は 1962 年から開始しており、その他にもマイナーチェンジを経ながら 20 年、30 年と継続して行なわれている調査もあります。

調査データはいわば「集めるデータ」で、サンプル数には限りがありますが、取得項目は調査目的に応じて設計されます。また、サンプルの基本属性が明確で、データの欠損も基本的にはありません。

一方で、近年では Web サイトの閲覧ログデータや、コンビニ・スーパーなどの会員カードによる個人単位の購買データ、交通系カードの移動データなどの大規模データの活用も盛んになってきています。こちらは「集まるデータ」とも言えるもので、レコード数は膨大なものの必要な項目しか取得していないことが多く、データ元のユーザーやカード保有者のプロフィール情報がわからない場合もあります。

こうした大規模データの不足情報を補う方法の一つとして、ビデオリサーチでは調査データと大規模データのデータフュージョンに取り組んでいます。またそれに先がけて、調査データ同士のフュージョンにも取り組み、既に実用化しています。本講演では、調査データ同士のフュージョンから調査データと大規模データとのフュージョンまでのシステム開発の流れを、背景を交えながらご紹介します。

2.　汎用的データフュージョンシステムの開発

ビデオリサーチでのデータフュージョンへの取り組みは、2000 年代前半に特定時点の特定調査データ同士のフュージョンから始まりました。

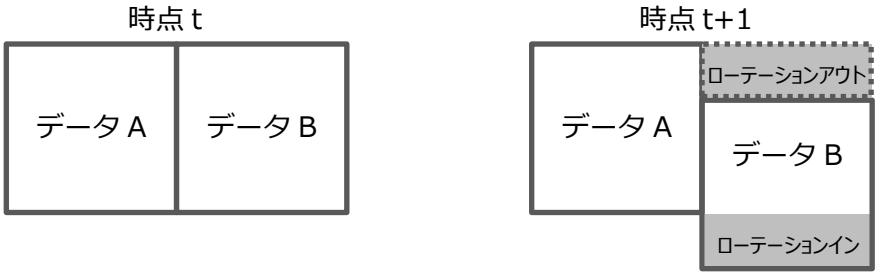
その後、アメリカでフュージョンによるデータ統合・活用が進んでいることなどを見据え、いかなるデータにも対応できる汎用的なデータフュージョンシステムの開発に着手したのが 2007 年です。このとき、NTT データ数理システム社にご相談し、Numerical Optimizer をベースにシステムを開発していただきました。

ビデオリサーチの調査サービスの多くは、無作為抽出で行なわれる市場の代表性のあるデータです。調査データ同士を融合した後も元々の結果を保持することが望まれるため、元データの平均と分散を保持するように類似サンプル同士をマッチングする「制約付き統計的マッチング」を採用しました。（詳細

は数理システムユーザーコンファレンス 2010 でご報告しています。）

3. 調査の実運用に対応するためのシステム機能拡張

「制約付き統計的マッチング」はフュージョン実施時点でデータに存在するサンプルが漏れなくマッチング対象になり、マッチングされたペアがウエイト値を持つことで元データの情報を保持することができます（下図の時点 t)。しかし、定期的にサンプルローテーションがあるパネル型の調査では、ローテーション後にはマッチング相手がいなくなったり、新たに加わったサンプルがマッチングされていない状態になります（時点 t+1)。ローテーション後も元のデータを保持するためにはその度にフュージョンし直すことになりますが、サンプルローテーションが頻繁にある調査では運用上の負担になりかねません。



こうした課題に対応するために、2015 年にデータフュージョンシステムの機能拡張を行ないました。上記の図でいうと時点 t で作成した類似サンプル判別モデルを使って時点 t+1 のローテーションインサンプルのマッチング相手を特定するという方法です。これは「制約付き統計的マッチング」を活用した「制約なし統計的マッチング」といえます。「制約なし」なので元の情報は完全には保持されませんが、高い精度で再現できることを確認しています。

ビデオリサーチではこの手法を使って、ローテーションがある既存の調査データ同士のフュージョンを行い、新たなデータサービスとして展開しています。

4. 調査データと大規模データのフュージョンへの対応

さらに、新たに取り組んでいるのが調査データと大規模データのフュージョンです。前述のように様々な事業者による多様なデータが流通し始めている中、取得項目が限定されている大規模データに調査データを融合することで、データをリッチ化したいというニーズが出てきています。

データフュージョンもそれに応える手法の一つです。しかし、2015 年までに開発したデータフュージョンシステムはフュージョン対象データに 1 万レコードまでという制限があったため、2016 年からは大規模データとのフュージョン方法の研究・実装に取り組んでいます。

機能としては、従来は類似サンプルを厳密解で判別していたところを、数十万、数百万というレコード数のデータを高速にマッチングするために、近似解で判別するという方法を取っています。研究フェ

ーズでは、「制約付き統計的マッチング」とそれを活用した「制約なし統計的マッチング」の両方について、従来の厳密解と新たな近似解でのサンプルマッチングでフュージョン精度比較を行い、近似解でも厳密解と遜色ない精度が得られることを確認しています。この研究・実装過程においても、NTT データ数理システム社に多大なご協力をいただいています。

　この開発により、ビデオリサーチでは１万サンプル程度の調査データ同士、および調査データと大規模データのフュージョンに対応可能になりました。今後、これを活用してさらなるサービス展開をはかって行きたいと考えています。