

近未来金融システム創造プログラム第 12 回講義レポート

第 12 回は、東京大学大学院工学系研究科の和泉潔教授から「テキストマイニングの可能性」という題目で講義があった。本日の講義は金融と技術の各論の第 4 回で、テキストマイニングなどの AI 技術やオルタナティブ・データを金融において活用する事例などを先生にお話しいただいた。

オルタナティブ・データへのシフト

GDP や既存の経済指標はデータが発表されるまで 2, 3 ヶ月かかることが多い。だが、クレジットカードの決済データを用いれば、より詳細に早く経済動向を読み取れる。あるいは決済データでなくても、新宿への来街者数から消費動向の予測を行うこともできることがわかった。最近はこのようなオルタナティブ・データ（代替データ、非伝統的データ）が使われるようになっており、オルタナティブ・データを提供するビジネスや調査も増加している。従来のデータ分析においては、タイムラグが生じたり、複雑な要因で変動するデータの分析が難しかったりするという問題があった。しかし、オルタナティブ・データの登場により、複雑な変動要因を持つ指標をスピーディに分析することが可能になった。オルタナティブ・データの使用にはいまだ問題点もあるが、オルタナティブ・データの活用を目指す風潮は高まってきている。

オルタナティブ・データの活用事例としては、会社名のテレビでの露出回数と株価の相関関係や、特許の価値と株価の相関関係を明らかにすることなどが挙げられる。オルタナティブ・データは、海外でも政府機関や民間機関によって活用されるようになってきている。しかし、入力データを公表している機関が発表を取りやめにしたことで、そのデータを活用した統計の発表も終了してしまった事例がある。これは、公的統計としてオルタナティブ・データを活用する上では重大な問題である。

他にも、オルタナティブ・データの活用にはいくつか問題点がある。オルタナティブ・データの使用方法に関する明確なルールがないことや指標としての脆弱性がなどである。また、オルタナティブ・データを特に公的な分野で活用する際には問題がある。データの公開が停止する可能性があること、データにバイアスがかかっている可能性があること、政府機関におけるデータを活用できる人材の育成が困難であることなどである。

それらの問題を解決しオルタナティブ・データを活用するために、金融機関は昨今、大学などアカデミックの分野と協力しつつ、オルタナティブ・データを活用できる人材を社内で

育成するようになってきている。

オルタナティブ・データとテキストマイニング

データマイニングは、オルタナティブ・データの問題点を解決する可能性がある。データから因果関係を読み取り、ある事象の原因や波及効果を知ることは、政策などを考える上で有用である。テキストデータには他のデータとは異なり、人間が認識した因果関係に関する情報が含まれている。そのため、テキストマイニングを使用することで、因果関係をデータとして取得することが可能になるかもしれない。しかし、オルタナティブ・データの更新頻度が高かったとしても、そこから因果関係や波及効果を発見し、ニーズに適した情報の提示を行うまでに時間がかかってしまえば意味がない。

そこで、大規模データを効率的に活用するために、データの事象間の因果関係を明らかにし、データとデータを結ぶ理論が必要となる。その理論を構築する過程で、人間の因果の認識が反映された自然言語データの分析が使用される。自然現象の分野とは異なり、経済現象においては、相関＝因果でないため、相関関係が見られた事象間に因果関係が存在するかどうかは人間の認識を軸として確認する必要があるのだ。自然言語における因果関係には、コンピュータでも分析することのできる因果の他、コンピュータによる分析の難しい社会的因果が含まれている。テキストマイニングによって、これらの因果を決算短信などのテキストデータから多層的にAIが見つけることが可能になった。

このように、目的とするデータを得るために、大規模データから適切なオルタナティブ・データを選択し、適切なプログラムで処理し、遅延なく提供できる技術の開発が今進んでいる。

Q&A

Q1. $A \rightarrow B \rightarrow C$ という3層構造の因果推論を統計的推論から行くと、データが足りなかった場合、 $A \rightarrow C$ というように中間層の因果関係が見落とされる可能性がある。テキストデータを元にした推論では、このような問題は生じないのか。

生じる。

Q2. 人間の認識は人によって異なることもあり、平均したら因果が見いだせなくなる可能性があると思うが、このような問題は生じうるか。

自然言語における因果は、統計的因果と異なり、有為な因果を提示しているのではなく、因果である可能性のあるものを全て提示している。そのため、提示された因果が有為であるかは、統計的手法などによって確認する必要がある。

Q3. 時間と共に人の認識が変化し、発見された因果が時間と共に変化する可能性があるのではないか。

因果は、時間と共に間違いなく変わっていく。

Q4. 公的統計を即時性のあるものに改善するのではなく、オルタナティブ・データが必要とされる理由は何か。

確かに、公的統計の改善により、スピード感は改善されるが、指標間・データ間の関連性を見つける上ではオルタナティブ・データが必要とされる。

Q5. オルタナティブ・データから得られる因果関係が信頼されるようになり、その因果を事前に知った人が逆張りして行動することで、結果がカオスになることはないか。

そのような状況は起こり得るが、複雑な因果を公的機関以上に速く読み取るシステムを個人が持つことは難しいと思われる。