ネットワーク図によるデータ可視化システムの開発

佐藤 邦弘
(株) 日経リサーチ

１．はじめに

インターネット調査の普及により、サンプルの偏りなどの問題はあるものの、調査を実施しデータを集めることは飛躍的に容易になった。しかしながら、データの収集が容易になる一方で、データの解析については、依然としてクロス集計への依存度が非常に高い。これは解析手法の多くが

① 解析結果の優劣が、解析者の能力に依存する
② 結果を説明する相手に対しても、一定の知識を求める

といった要素を持つために、「解析」と「説明」の両方の要因から、「多くの人が容易に理解でき」、「人によって結果に差が出ない」クロス集計が選ばれやすくなっていると考えられる。

２．目的

本発表の目的は、このような現状に対して、より「解析者に依存しにくく」「結果の理解が容易」なデータを読み取る手段として、クロス集計を基本に、①特徴のある箇所を抽出し、②高い一覧性で可視化するシステムを提案する。

３．手続き

クロス集計に比較的近い分析手法には、クロス集計の中から特徴のあるカテゴリーをみつけだす CATDAP などがある。ここでは、新たに２つのアイデアを導入することで、クロス集計の新しい可視化を試みた。

まず、最初のアイデアとして、可視化の手法にグラフ（ネットワーク図）を用いた。グラフはツリー構造よりも上位概念に位置し、高い自由度をもつ。また、Watts(2003)を参考に、関係性を可視化する方法として２部グラフの形式を用いた。

次のアイデアとして、特徴の抽出にあたり有限情報源に対する情報量を用いた。情報量は、事象が起こった確率をＰとした時、I=－log(P)で表される。これは、確率の小さな事象が起こったことを知らせる情報ほど、情報の価値が高いことを表している。

クロス集計表の価値を情報量で測るにあたり、表１のような２ｘ２のクロス集計を考える。このとき O_{11} のセルがもつ情報の価値を測るということは「『車種Ａのユーザー』という回答にとって、『スキーが趣味』というユーザーの回答は、どの程度価値の高い（独立事象ならば確率的に起こりづらい)情報なのか」ということを意味する。すなわち、N 個のデータの中から、X_1（車種Ａのユーザー数）個のデータを無作為に取り出したとき、そのうち、O_{11} 個のデータが、「スキーが趣味」の属性をもつデータである確率を計算することで情報量が計算される。

表１

		スキーが趣味		
		Yes	No	合計
車種Aのユーザー	Yes	O_{11}	O_{12}	X_1
	No	O_{21}	O_{22}	X_2
合計		Y_1	Y_2	N

数式を作成するにあたり、全ての N 個のデータから、X_1 個のデータを無作為にひとつずつ取り出していくような過程を想定する。このとき、N が全く減らないような無限情報源を仮定すると、情報量は

$$I = -\log(P) = -\log\left\{ {}_{N}C_{O_{11}} \left(\frac{Y_1}{N}\right)^{O_{11}} \left(\frac{Y_2}{N}\right)^{O_{12}} \right\}$$

となる。

一方、N が減るような有限情報源を仮定すると、

$$I = -\log(P) = -\log\left(\frac{{}_{Y_1}C_{O_{11}} \, {}_{Y_2}C_{O_{12}}}{{}_{N}C_{X_1}} \right)$$

となる。

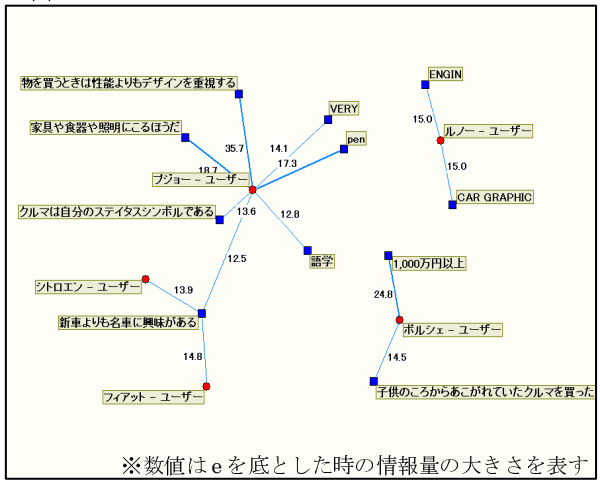
無限情報源を仮定した場合、計算はシンプルだが、Y_1 が小さい数値の場合、確率は実際におこった事象の確率よりも大きくなってしまう。このため、マイナーな雑誌の閲読といったデータの少ない属性情報の特徴は、性別・年代などのデータの多い属性情報に埋もれやすくなる。一方、有限情報源を仮定した場合には、Y_1 が小さい場合には、取り出す過程で Y_1 の数も減少し確率は小さな値となるため、情報量は大きくなる。

日経リサーチでは、以上の２点の特徴をもったネットワーク図の可視化システム“i”s Mining を開発した。その結果を次に紹介する。

３．結果

調査データには、(株) 日経リサーチ で行った自動車のブランド価値を表す「カーブランディング２００７」のデータを用いた。図表は、「プジョー」「シトロエン」「フィアット」「ポルシェ」「ルノー」といった外資系自動車メーカーのユーザープロフィルの特徴を表したマップである。ここから「プジョー」ユーザーの特徴は、『家具・食器・照明にこる』、『性能よりデザイン重視』、『pen を閲読』」など一貫したデザインに対するコダワリが浮かび上がっていることがわかる。次に２つのアイデアの効果をみることにする。まず、2 部グラフで可視化したことの利点として、「新車よりも名車に興味がある」の項目が、「プジョー」「シトロエン」「フィアット」ユーザーの共通項として浮かび上がっていることがあげられる。このことは２部グラフにより、各メーカー間におけるユーザーの関係性の読み取りが容易になっていることを示している。次に、有限情報源を仮定した情報量の利点として、「ポルシェ」に特徴が出ていることがあげられる。ポルシェのユーザーは、サンプル数が 28 と少ないにも関わらず、「子供の頃から憧れていた車を買った」などがユーザーの特徴として高い値を示しており、少ないデータ数でも情報量が特徴をうまく浮かび上がらせていることがわかる。

図 1



※数値はｅを底とした時の情報量の大きさを表す

４．考察

これらの結果は、膨大なクロス集計のセルの中から、情報の価値の高いところだけを取り出し、その表側と表頭を線でつなぎ一覧にしたものと考えることができる。数値はパラメータを持たず一意に決まり、分析者によって変化することはない。このことから、当初の課題に対する回答として、このシステムが一定の有効性をもつことが示唆された。

参考文献

[1] 村上、田村(1998)『パソコンによるデータ解析』, 朝倉書店

[2] Watts, Duncan J. (2003) *SIX DEGREES：THE SCIENCE OF A CONNECTED AGE.* .W. W. Norton & Company.， (辻竜平、友知政樹訳 (2004)『スモールワールドネットワーク』，阪急コミュニケーションズ)