大規模ECサイトの閲覧履歴データの分析

清水　信夫　　データ科学研究系　助教

【はじめに】

近年、計算機科学の発展により、市販のデータベース管理ツールや従来のデータ処理アプリケーションでは処理が困難なほど大規模かつ複雑なデータ集合が多数出現するようになっており、それらはビッグデータと呼ばれ注目されている。ビッグデータから有用な情報を抽出する需要が増大する今日において、データ集合における個々のデータではなく特徴的な属性に関して自然に分けられた集団に着目し、それらが各々もつ特徴およびそれぞれの間の差異を調べる手法の重要性は高まっている。

本研究では、そのようなビッグデータの一つと考えられる、ある大規模ECサイトにおける閲覧履歴データを対象として、各属性ごとに分けられた顧客の集団に着目し、異なる顧客集団間で商品の閲覧および注文行動にどのような差異がみられたかについて分析した。

【データの概要】

本研究において分析対象とするのは主にアパレル系の商品を扱っているECサイト閲覧履歴データである。ここには、2011年9月1日から2013年4月末までにおいて、ECサイトに登録している約10万人の顧客が閲覧した商品と、その商品の注文の有無に関しての情報が各行ごとに記録されており、総数は約6,116万件にのぼる（そのうち商品が注文された総数は約34万件である）。顧客に関する情報としてID・性別・生年月日・登録日・退会日・都道府県・サイトへのアクセス端末種別（PCからかモバイル端末からか）が記録されており、商品に関する情報としては名前・サイズ・色・分類などの情報が定められている。また、これらの商品が売られているショップ（計534店）のIDの情報も設定されており、各顧客はそれらの中からお気に入りショップを登録することが可能である。顧客が商品を閲覧した際においては、商品に関する各種情報とアクセスおよび退出の日時が記録され、お気に入りショップの登録履歴のある顧客についてはその情報も記録されている。このECサイトにおいては、商品によっては定められたセール期間において割引販売が行われており、商品の閲覧および注文日時がセール期間内か期間外かの情報も記録されている。

【分析の目的】

このECサイト閲覧履歴データを分析するにあたり、私達は顧客の年代別もしくはショップ別の商品閲覧傾向の違いや、商品の閲覧日時と注文の有無が顧客集団の属性によりどう関連するかに興味を持った。そこで、ショップの状況把握のために、各ショップごとの商品合計金額の可視化を行った。また各顧客に関して

・商品の注文時期がセール期間内か期間外か

・お気に入りショップが有るか無いか

・アクセス端末はPCかモバイルか

の各属性に注目し、それぞれの属性ごとに区分した集団を考えた。そして、各集団に属する顧客がセール期間内および期間外においてどのような商品注文行動をとるか、および各集団間でどのような差異がみられるかについて分析した。

【商品閲覧数の可視化】

各ショップ別や年代別など様々な属性別の商品閲覧状況の違いを可視化するにあたり、Hadoopを用いて商品閲覧数を集計し、商品の合計金額・品名・各種分類・閲覧曜日・時間・購入者年代など計777項目による表を作成し、Mondrianを用いて可視化を行った。図1は各ショップの商品合計金額のヒストグラムであるが、商品合計金額が1億円以上の大人気ショップは極めて少数で、大半のショップの商品合計金額は低い水準であることが解る。

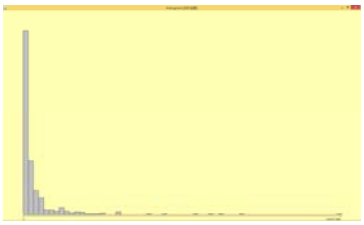

図1: MondrianによるECサイト内の各ショップの商品合計金額のヒストグラム

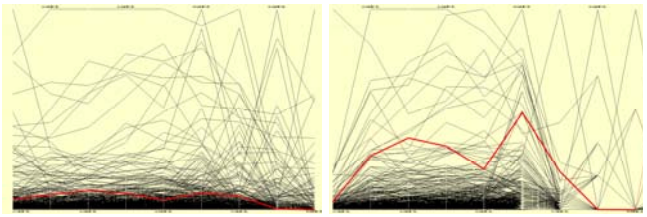

図2: MondrianによるECサイト内の商品閲覧数の年代別集計結果の可視化例
（左：商品購入なし，右：商品購入あり）

図2は、ECサイト内の各ショップにおいて販売されているある商品（この図ではショーツ）についての閲覧数を、商品の購入につながった場合およびつながらなかった場合それぞれにおいて可視化したものである。横軸は顧客の各年代を表し、縦軸はそれぞれの場合の商品の閲覧数を表している（スケールは異なる）。特定のショップにおける商品が各年代において購入の有無によらず閲覧数の上位を占める傾向がみられる中で、赤太線で示した商品の30代顧客における閲覧数に注目すると、他のショップの商品と比べて閲覧のみの数は必ずしも多くないが、購入につながった閲覧数が比較的多めであることから、この商品は30代顧客向けによく売れていることが解る。

【商品注文者のセール参加傾向の比較】

ECサイトに登録している顧客が商品を閲覧した総数に対する注文総数の割合は約0.6%である。商品注文による利益を増やす上で、顧客が各属性ごとにどのような商品注文傾向を示しているかを理解することが重要となる。そこで、本研究ではMySQLにより購入履歴に近い閲覧履歴を同定し、注文から5時間以内の直近閲覧レコードと注文データを結合した。その上で、商品を注文した顧客に限定して(1)お気に入りショップの有無 (2)アクセス方法の違い(PCからかモバイル端末からか) の2種類の属性を用いて4分割し、各集団ごとにセール期間内および期間外それぞれの商品注文数に注目してRにより分析を行った。

		お気に入りショップの有無				全体	
		有り		無し			
アクセス方法	PC	67,090	比率 0.595	91,482	比率 0.553	158,572	比率 0.570
		112,701		165,712		278,413	
	モバイル	10,543	比率 0.466	21,216	比率 0.519	31,759	比率 0.501
		22,608		40,823		63,431	
全体		77,633	比率 57.4%	112,698	比率 0.648	190,331	比率 0.567
		135,309		205,535		341,844	

図3:各属性における注文数（左側上段：セール期間内，左側下段：セール期間外）
（比率：各属性におけるセール期間内注文数とセール期間外注文数の比（オッズ））

図3は各属性の組み合わせ（計4通り）におけるセール期間内および期間外の商品注文数、およびそのオッズについて示した図である。この結果から、セール期間内の注文率はお気に入りショップの有無を問わずモバイル端末からアクセスしている顧客層よりも**PCからアクセスしている顧客層の方が高め**となっていることが解った。ただし、アクセス方法ごとにみた場合のセール期間内の注文率は、**PC**からアクセスしている顧客の中ではお気に入りショップが**有る**顧客層の方が高い一方、**モバイル**端末からアクセスしている顧客の中ではお気に入りショップが**無い**顧客層の方が高いという全く逆の傾向が現れた。

次に、1回あたりの注文における商品価格が集団間でどのように異なるかに注目した。そこで、4通りの組み合わせにおいて注文商品価格の分布の自然対数の平均値および標準偏差を図4に示す。なお、$e^{8.1375} = 3420$（円），$e^{8.6666} = 5806$（円）である。

		お気に入りショップの有無			
		有り		無し	
アクセス方法	PC	8.1375	0.8565	8.0505	0.8586
		8.6560	1.0065	8.4864	1.0469
	モバイル	8.2476	0.8354	8.0857	0.8983
		8.6666	0.9331	8.5658	0.9788

図4:各属性における1回あたりの注文商品価格の対数値の
平均額（左側）および標準偏差（右側）
（各上段：セール期間内，各下段：セール期間外）

これらの値に関して各組み合わせごとにセール期間内と期間外の間で検定を行うと、いずれの場合でも平均値および標準偏差それぞれで有意に差がみられた。この結果はそれぞれの場合でデータ数が非常に多いため当然といえる。

その上で、各属性ごとの1回あたりの注文商品価格の平均値について調べると、セールの有無やアクセス方法の差異によらず、お気に入りショップが無い顧客層よりも**お気に入りショップが有る顧客層の方が高く**、またセールの有無やお気に入りショップの有無によらずPCからアクセスしている顧客よりも**モバイル端末からアクセスしている顧客の方が高め**になっていることが解った。

【おわりに】

ビッグデータの解析において、個々のデータに対してではなく自然に分けられる少数の各集団の存在を前提とした上で、それらの特徴を記述し、各集団間の差異を見出すことは非常に重要である。本研究においては各ショップの商品合計金額のヒストグラムおよび年代別の商品閲覧数について可視化を行うことで各ショップの状況や特徴的な商品の閲覧状況の例を示した。また商品のセール期間が注文時期にどのように影響を与えているかについて、顧客を属性別に分割して各集団間の比較結果を示したが、今後は各集団における少数の統計量を用いた解析を行うことによって、より詳細な分析結果を得た上で可視化を行うことが期待される。なお、この研究は中野純司氏（統計数理研究所）・山本由和氏（徳島文理大学）・藤原丈史氏（東京情報大学）・本多啓介氏（統計数理研究所）との共同研究であり、分析した閲覧履歴データは経営科学系研究部会連合協議会主催、平成25年度データ解析コンペティションで提供されたデータである。