Webマーケティングデータの施策利用における疎データ問題と解決法の概観

A Survey of Sparse data Handling on Web Marketing Data

内田匠 [1*]　吉田健一 [1]
Takumi Uchida[1]　Kenichi Yoshida[1]

[1] 筑波大学大学院 ビジネス科学研究科
[1] 東京都文京区大塚 3-29-1

Abstract: Web marketing is one of the important company activities. Here, its data is often high dimension and sparse. This sparseness disturbs the analysis of web marketing. To alleviate this problem, various approaches are studied. For example, clustering of users and items are used to improve collaborative filtering results. Predictor variable reduction is used to estimate user's response from web-advertisement. In this paper, we classify purposes of web marketing in three objective (i.e., Customer Attraction, Customer Retention, and Cross-Sales), and survey solutions for sparse data problem for each objective.

1　はじめに

[2] は、Web サーバに蓄積されるデータをマーケティングに利用する目的を Customer Attraction(新規顧客獲得), Customer Retention(顧客維持), Cross-Sales(購買単価促進) の三つに分類した。

マーケティングや CRM という広い概念をカバーする上で、この 3 分類に上手くあてはまらない事例もありえるが、1) Customer Relationship Management(CRM) 分野の論文をまとめたサーベイ研究 [7] においても、Customer Attraction の施策として Direct Marketing を、Customer Retention の施策として Loyalty Program, Cross-Sales の施策として Market basket analysis を上げている、2) 物理的に Web サーバで取得可能なデータに関する研究 [2] なども、この分類に従っている、3) 実際に企業が利用している Web マーケティング施策もこの三つの分類に沿ったものが多く見られる、など、この 3 分類は Web マーケティング施策を分類概観する上で有意義な分類であると考えられる。

本論文ではこの 3 分類に沿って提案されている施策や手法を概観し、疎データに対する解決法を整理する。

2　Webマーケティングと疎データ

Customer Attraction の重要な施策として広告が挙げられる。特に広告のクリック率を予測し自動的に配信を最適化する Computational Advertisement の分野には数多くの研究がなされている (例えば、[1, 3, 9])。ここで重要な点は、数十万種類の検索クエリに応じて、何万の種類の広告から検索結果画面に表示する広告を選ぶ検索連動型広告のデータはほとんどが疎である事である。つまり、Web 広告の掲載データでは一般的なクリック率は数%以下であることが多く、多次元かつ希少発生な現象を取り扱うことになる。従って様々な論文において疎データの問題が言及され、様々な対策が提案されている。

Customer Retention の施策としては Loyalty Program が挙げられる。具体的に言うとポイントシステムであり、特に e コマースを中心に多くの企業が実施している。基本的にはポイント付与による顧客維持の効果を評価し、適切なポイント付与率を調整していくことになるが、疎データの問題について言及している論文は見当たらなかった。これは顧客維持におけるポイントシステムの貢献度という可視化しにくい情報をどのように評価するのかが研究の主眼に置かれている事と、サンプルがポイントシステムの加入ユーザのため何らかの商品の購買履歴があるため、広告データほどには疎になりにくく、問題になることが少ない為だと考えられる。

Cross-Sales の具体的な施策としてレコメンドシステムが挙げられる。全商品アイテムの顧客のユーザの購買履歴データから、ユーザとアイテムの推定購買率などをスコア化し、推定値の高いアイテムを Web サイト上に表示する。この購買履歴データは多次元になる。

*連絡先：(筑波大学大学院 ビジネス科学研究科)
(東京都文京区大塚 3-29-1)
E-mail:skillful.boy@gmail.com

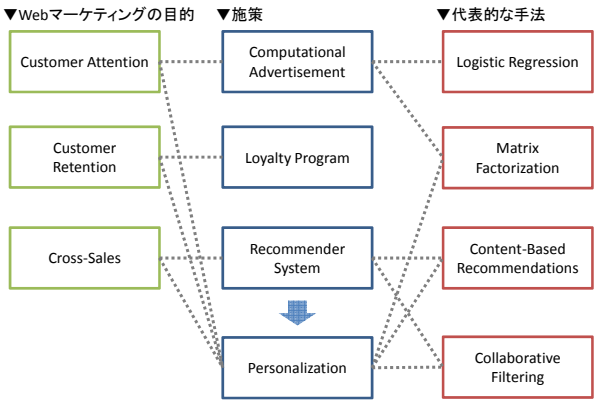

図 1: Web マーケティングの目的・施策・手法

ほとんどの Web サイト訪問ユーザは商品を買わないか買っても数個以内になるので疎データになる。レコメンドシステムに関する論文にも疎データへの対処について言及するものが多い (例えば、[8, 10])。

レコメンドシステムについては今日では Personalization の分野として研究が盛んになってきている [6]。この分野では、単に Cross-Sales だけを目的とせずユーザ一人一人に最適な Web 表示を実現することで Customer Attention, Customer Retention, Cross-Sales を同時に改善するための包括的な施策として扱われている。利用されている手法には Collaborative Filltering や Association rules などが挙げられる。しかし、適応範囲はより広くなり購買履歴だけではなくなっている。そのユーザが次に見るであろう URL を予測してナビゲーションの改善を実現するなど、Web サイト全体を改善する手法となっている。

これまでの議論をまとめると図 1 のようになる。以下では、疎データが大きな課題となっている Computational Advertisement と Personalization について整理する。

3 Computational Advertisement の疎データ問題

Computational Advertisement(CA) は機械学習の技術を応用してインターネット広告の収益化とその広告効果を改善するため研究分野である。対象となる広告として検索連動型広告とディスプレイ広告に大きく分かれ、ユーザからの広告クリック率や広告から Web サイト訪問後の購買率の予測モデルが多く提案されている。

検索連動型広告のクリック率推定において重要な課題になるのは、数の少ない検索クエリ (ロングテール) である。原則として、(クエリ, 広告) のセットごとに予測クリック率を計算する問題である。しかし、検索クエリは多様であり広告主の数に比例して掲載候補広告の数は増えていく。特にロングテールにおいて掲載実績データはほとんどない。

3.1 [5] による分類

この疎データの問題に対して、[5] によると以下のようなアプローチが取られている。

- Feature-based modeling
- Similarity-based collaborative filtering
- Matrix factorization

Feature-based modeling は、検索クエリと掲載候補広告を変数で表現することで、数万種類以上のクエリと広告を数十程度の次元に圧縮するアプローチである。例えば、検索クエリや広告文言に対して何らかの判別ロジックを施し、抽出されたワードカテゴリやフレーズやトピックなどを変数として扱う。これにより、学習データの次元数は判別した変数の数に圧縮される。次元圧縮した学習データに対してロジスティック回帰などを施してクリック率を予測する [9]。このアプローチについては、我々も実際の Web マーケティングデータを用いた提案 [11] を行っており、3.2 で概説する。

Similarity-based collaborative filtering は、掲載実績データのない広告や、ロングテールの検索クエリに対してクリック率を予測する際に、十分に実績データのある類似する広告とクエリを見つけ出し、その実績データに基づいてクリック率を予測する。類似の広告とクエリを見つけ出す際には、k-nearest neighbor などが使われる。これについては、後述の Personalization において詳しく説明する。Matrix factorization は、検索クエリと掲載候補広告の掲載実績データを基に、SVD をベースにした潜在因子行列を推定するアプローチである。これについても、後述の Personalization において詳しく説明する。

3.2 ポアソン分布を仮定した誤差推定

我々は、広告配信対象ユーザの年齢・性別・興味関心などの変数に基づいて、ポアソン回帰による商品購買率分析を行った [11]。この中で疎データの問題に対処するため、AIC を基準としたステップワイズ法による変数削減を行った。

ユーザーの購買行動は、商品やサービスを Web 上で閲覧し、購買や申込みなどを実際に行うか行わないかの二値で計測される。従ってその計測数 (以下コンバージョン数と記す) は発生確率 p_i の二項分布に従うと考える事ができる。計測されたコンバージョン数を c_i、商

品・広告の接触数を n_i とすると、計測 CVR は c_i/n_i となる。ここで接触数とは、購買履歴データでは各商品の詳細ページの閲覧数、広告データでは各広告のクリック数がそれぞれ該当する。

この計測 CVR が発生確率 p_i の二項分布 Binominal(p_i) からたまたま計測された値であると考えると、p_i の真の値の推定値 $\widehat{p_i}$ は c_i/n_i となり、二項分布の確率 p_i の推定値である CVR の標準誤差は計測値に含まれる誤差に相当すると考えられる。ここで二項分布における c_i の平均と分散は、定理より以下で求められる。

$$E[c_i] = n_i p_i \tag{1}$$

$$Var[c_i] = n_i p_i (1 - p_i) \tag{2}$$

式 (2) より計測 CVR($\widehat{p_i}$) の分散は以下で求められる。

$$Var[\widehat{p_i}] = Var[\frac{c_i}{n_i}] = \frac{n_i \widehat{p_i}(1-\widehat{p_i})}{n_i^2} = \frac{\widehat{p_i}(1-\widehat{p_i})}{n_i} \tag{3}$$

式 (3) の平方根が計測 CVR である $\widehat{p_i}$ の標準誤差となり、この値が小さいほど計測 CVR に含まれる誤差が小さくなると考えることが出来る。

[11] では、ある CVR 評価セグメント i で計測された CVR に含まれる標準誤差 E_i を以下のように定義した上で、計測された CVR を評価する上での指標値とした。

$$E_i = \sqrt{\frac{\widehat{p_i}(1-\widehat{p_i})}{n_i}} \quad (\text{ただし}, \widehat{p_i} = c_i/n_i) \tag{4}$$

上記の定式化のもとで、CVR 評価のための計測データが疎となっている問題に対して、既存の CVR 評価セグメントをまとめることで評価セグメントを再定義し、セグメント内のデータ量を増加させることで解決を試みる。

具体的には既存の評価セグメントはセグメント間で共通の属性情報を持つ。例えば広告システムのデータであれば、広告配信対象のユーザーの性別、年齢、居住地などが属性情報に該当する。既存の評価セグメントはこの属性の組み合わせでグループ分けされており、例えば女性× 20 代×東京在住でひとつの広告となり、これが CVR 評価セグメントとなる。提案手法では、不要な属性を削除することでセグメント分割を抑制し、大きなセグメントに再定義しなおすことで、疎データの状況を改善する。

例えば、既存の広告セグメントが { 女性, 男性 } × {20 代, 30 代, 40 代 } × { 東京, 大阪, 京都 } の属性で構成されていたとする。この既存広告セグメントの数は 2*3*3=18 個となる。この時、属性の中から CVR 評価に与える影響の小さい属性を除外しても CVR 評価に大きな影響を与えない。例えば、CVR 評価への影響が小さい属性が { 女性, 男性, 30 代, 40 代 } だったとする。

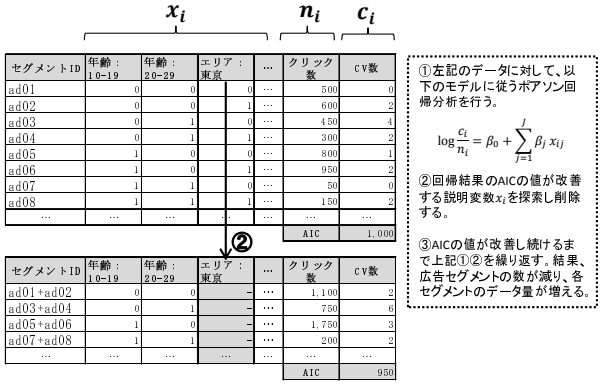

図 2: ステップワイズ法による広告セグメントの再定義方法

この時、CVR 評価のための広告セグメントは {- } × {20 代, その他 } × { 東京, 大阪, 京都 } と再定義できる。この再定義された広告セグメントの数は 1*2*3=6 個となる。既存の広告セグメント数 18 個から減少し、再定義セグメントの個々のデータ量は増加する。

CVR 評価への影響の小さい属性情報を削除するには、計測 CVR を目的変数、広告属性を説明変数とする回帰分析を行い、回帰結果に対してステップワイズ法による変数選択を行う (図 2)。

変数選択の結果、CVR 評価に対して影響の少ない属性を取り除き (例えば「エリアが東京であるか否か」をセグメントに利用する属性から削除し)、複数のゼグメントを 1 つに纏め (図 2 では ad01 と ad02 セグメント等を 1 つにまとめている)、再定義セグメントの個々のデータ量を増加させる。

この際の回帰分析としてはポアソン回帰分析を採用した。利用したポアソン回帰式を以下に示す。

$$logCVR_i = log\frac{c_i}{n_i} = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} \tag{5}$$

(5) 式にある β は回帰係数を表し、x_{ij} は CVR 評価セグメント i のもつ属性情報 j のダミー変数である。これは説明変数を線形一次結合させたポアソン回帰分析である。この回帰分析の結果生成されたモデルに対し、AIC を基準としたステップワイズ法で変数選択を行う。

実際のマーケティングデータに適用し、この変数削減により推定された購買率の標準誤差率が 80.97%だったものが 64.5%までに改善できた。

4 Personalization の疎データ問題

Personalization の手法を大別すると Content-Based Recommendations(CBR) と Collaborative Filtering(CF)

がある。前者はそのユーザが過去に選択したアイテムに基づいてレコメンドされるのに対し、後者は同じ嗜好をもったユーザ群が選択したアイテムに基づいてレコメンドされる。

CBR にはユーザが選択したことのない属性しか持たないアイテムをレコメンドすることが出来ないと言う欠点がある。疎になっている領域については、そもそもレコメンドの対象外となってしまう。これは一人のユーザにとって、多くのアイテムが疎 (つまり未体験) になる Web マーケティングにおいて大きな欠点である。この欠点を補ったのが CF である。これは類似した他のユーザのデータを利用するため前述の欠点を克服できる。広く使われている方法として k-nearest neighbor がある。

4.1 Collaborative Filtering

以降では、Web 通販サイトを想定し、購買履歴データからユーザへどの商品をレコメンドする際の CF のロジックを確認していく。u をターゲットユーザ、v を他のユーザとし、各商品 i についてそれぞれのユーザの購買数を r_u,r_v とすると、u と v の相関性は以下の数式で表現される。

$$s(u,v) = \frac{\sum_i((r_{u,i} - \overline{r}_u)(r_{v,i} - \overline{r}_v))}{\sqrt{\sum_i((r_{u,i} - \overline{r}_u)^2)}\sqrt{\sum_i((r_{v,i} - \overline{r}_v)^2)}} \quad (6)$$

ここから、k-nearest neighbor で選ばれた neighbor 集団 V 内の各ユーザ v について、商品 i のターゲットユーザ u に対するレコメンドスコアは以下のように定義することが出来る。

$$p(u,i) = \overline{r}_u + \frac{\sum_v s(u,v)(r_{v,i} - \overline{r}_v)}{|\sum_v s(u,v)|} \quad (7)$$

以上は、ユーザ間の類似性を利用した user-base のものであるが、ほぼ同様の理論をアイテム間の類似性に応用した item-base も存在する。CF によりユーザが選んだことのないアイテムをレコメンドすることが可能になり、疎データの問題が緩和されたといえる。これはクラスタリングを施すことで、CBR では探査不可能だった疎の領域についても計算可能になったことによる。

4.2 Matrix Factorization

Web Personalization における疎データ問題に対してのもう一つの解決法として Matrix Factorization(MF) がある [4]。これは Netflix Prize contest での映画のレコメンド部門でグランプリを受賞した手法である。この手法は映画視聴サイトの会員ユーザが視聴した映画につけた 1〜5 点の評価スコアデータ (例えば、表 1 のようなデータ) を基に、各会員がまだ視聴していない映画をレコメンドする。世の中の映画コンテンツの膨大であり、一人あたりのユーザが視聴できる数には限界があるため、ユーザ×映画の評価スコアデータ R_{um} が疎データとなっている事への対策に特徴がある。

表 1: ユーザ×映画の評価スコア行列 R_{um} の例

	Movie1	Movie2	Movie3
User1	5		
User2		4	
User3		2	5
User4	1		4
User5	4		
User6	3	2	

MF のモデルは式 8 で表すことができる。

$$r_{um} = \mu + b_{user}(u) + b_{movie}(m) + a_{um} \quad (8)$$

このモデルでユーザ u の映画 m への評価スコア r_{um} は、全体の平均評価スコア μ、そのユーザがつける評価スコアの全体平均からの乖離 $b_{user}(u)$(ユーザ個人の辛辣度)、その映画が獲得した評価スコアの全体平均からの乖離 (作品の一般的な魅力度)$b_{movie}(m)$ と、SVD(特異値分解) をベースにした映画とユーザの相性推定値 a_{um} の 4 項から成り立つ。しかし、疎データの問題について着目すると重要なのはこの 4 項のうち、SVD を利用したユーザ×映画の相性値 a_{um} の推定になる。

説明のために、$base_{um} = \mu + b_{user}(u) + b_{movie}(m)$ を用い、式 8 を以下のように変形整理する。また、$base_{um}$ の推定計算については、a_{um} の推定の前に実データから事前に求める必要がある。

$$a_{um} = r_{um} - base_{um} \quad (9)$$

a_{um} のユーザ×映画行列 A の推定モデルを考える。まず、ユーザと映画の相性を決定する潜在因子 f_1,f_2 を仮定する。ここで、潜在因子の次元数を 2 としているがこれは任意の数に設定できる。次に、ユーザ×潜在因子の行列 F_u と映画×潜在因子の行列 F_m を想定し、この内積で A を表現できるとする。つまり行列式にすると式 10 となる。

$$A = F_u \times F_m^T \quad (10)$$

これを観測されたユーザ×動画の評価スコアデータ R_{um} から $base_{um}$ を引いて求めた行列 A について適応し、潜在因子行列 fu, fm の各要素の値を推定していく。このモデル式 10 を図式にしたものが図 3 である。

A	m_1	m_2	m_3
u_1	0.00		
u_2		0.00	
u_3		-0.58	0.58
u_4	-0.88		0.88
u_5	0.00		
u_6	0.21	-0.21	

$=$

F_u	f_1	f_2
u_1	fu_{11}	fu_{21}
u_2	fu_{21}	fu_{22}
u_3	fu_{31}	fu_{23}
u_4	fu_{41}	fu_{24}
u_5	fu_{51}	fu_{25}
u_6	fu_{61}	fu_{26}

$\times$

F_m	m_1	m_2	m_3
f_1	fm_{11}	fm_{21}	fm_{31}
f_2	fm_{12}	fm_{22}	fm_{32}

図 3: ユーザ×映画の相性行列の推定モデル

式 10 に基づいて、潜在因子行列 F_u,F_m の各要素を推定していく際には、何らかの誤差関数を設定してそれを最小化する。ここで疎（欠損部）を最適化計算において無視する事が、この方法の大きな特徴である。上図では、例えばユーザ u_2 は映画 m_1 を視聴していないため評価スコアデータが存在しないので誤差を計算することが出来ず、推定計算が不可能になる。そこで、この欠損部分の最適化計算を省略することでこの問題を回避する。例えば誤差二乗和を用いる際、評価スコアデータが存在するユーザ u と映画 m の集合を S とすると、以下の指式を最小化することで k 次の潜在因子の各要素を推定する。

$$min: \sum_{(u,m)\subset S} (a_{um} - \sum_{k} fu_{uk} \times fm_{mk})^2 \qquad (11)$$

MF はこのようにして、高次元かつ疎データであった A を、k 次元の行列の内積で表現することで疎データの問題を解消しており、現在では Personalization 施策に多く利用されるようになっている。また、実際には式 11 に正則化項を追加したものが一般的に利用されている。

5 結論と今後の研究について

本論では Web マーケティングデータの施策利用における疎データ問題についての解決法を概観してきた。施策の目的は大きく Customer Attention, Customer Retention, Cross-Sales の三つに分けられ、それぞれに対応する様々な施策が提案されている。その中でも Computational Advertisement と Personalization において特に疎データが課題となりやすい。

各施策を実現する上で、疎データを解決するために様々な手法が提案されており、その中で一般的な手法を確認した。その結果、どの手法についても疎データ解決の方法には二つの要素があるように考えられる。それは「次元圧縮」と「クラスタリング」である。

Computational Advertisement の Feature-based modeling では検索クエリと配信候補広告に何らかの判別器を施して変数抽出し回帰を行っていた。この抽出された変数について変数選択をすることで推定値の信頼誤差率を改善できることを紹介した。また、Personalization の手法である MF も同様に SVD を使った次元圧縮によって疎データを解決できる。CF についてはユーザ×商品の複雑データに対して k-nearest neighbor によってユーザもしくは商品をクラスタ化した。これにより同じクラスタ内の他のユーザの購買履歴を参照し、データ量を増やすことで疎データを緩和しようとしていると言える。

今後の研究方針としては、引き続き Web マーケティングにおける疎データの問題についてアプローチしていきたい。施策だけではなく分析業務についても考察範囲を広げることも重要である。ソーシャル･メディアの分析やユーザの行動解析など分析業務の目的は多岐に渡る。この分野においても疎データ問題を分析し、解決法を整理していくことは有意義である。また、Computational Advertisement や Personalization といった代表的な施策ではなく、例えばメールマガジン配信やメディアポートフォリオ分析などの施策について、疎データの問題を確認し、上述の解決法を適応させていく上での知見について調査していくのも有意義である。引き続き、Web マーケティングについてデータに基づいた施策実現について研究を進めていきたい。

....

参考文献

[1] Deepak Agarwal, Rahul Agrawal, Rajiv Khanna, and Nagaraj Kota. Estimating rates of rare events with multiple hierarchies through scalable log-linear models. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 213–222. ACM, 2010.

[2] Alex G Büchner and Maurice D Mulvenna. Discovering internet marketing intelligence through online analytical web usage mining. *ACM Sigmod Record*, Vol. 27, No. 4, pp. 54–61, 1998.

[3] Patrali Chatterjee, Donna L Hoffman, and Thomas P Novak. Modeling the clickstream: Implications for web-based advertising efforts. *Marketing Science*, Vol. 22, No. 4, pp. 520–541, 2003.

[4] Y. Koren. The bellkor solution to th netflix grand prize. 2009, 2009. http://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf.

[5] Bing Liu. *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media, 2007.

[6] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. Automatic personalization based on web usage mining. *Communications of the ACM*, Vol. 43, No. 8, pp. 142–151, 2000.

[7] Eric WT Ngai, Li Xiu, and Dorothy CK Chau. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, Vol. 36, No. 2, pp. 2592–2602, 2009.

[8] Alexandrin Popescul, David M Pennock, and Steve Lawrence. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pp. 437–444. Morgan Kaufmann Publishers Inc., 2001.

[9] Matthew Richardson, Ewa Dominowska, and Robert Ragno. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*, pp. 521–530. ACM, 2007.

[10] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Application of dimensionality reduction in recommender system-a case study. Technical report, DTIC Document, 2000.

[11] Takumi Uchida, Koken Ozaki, and Kenichi Yoshida. Toward a faithful bidding of web advertisement. In *HCI in Business*, pp. 112–118. Springer, 2014.