オンライン広告データ分析と分析基盤

株式会社きざしカンパニー
森下　民平

株式会社きざしカンパニー
稲垣　陽一

株式会社マイクロアド
青井　順一


概要

オンライン広告において、Webメディア閲覧ユーザーの行動分析はビジネス上極めて重要なテーマである。しかし、対象データは膨大なため、効率的な分析は容易ではない。本稿では、オンライン広告の概要、データ分析に有用なページカテゴリ解析、データ分析事例について述べ、次世代分析基盤の主要コンポーネントである高速な行動ログ解析エンジンを紹介する。

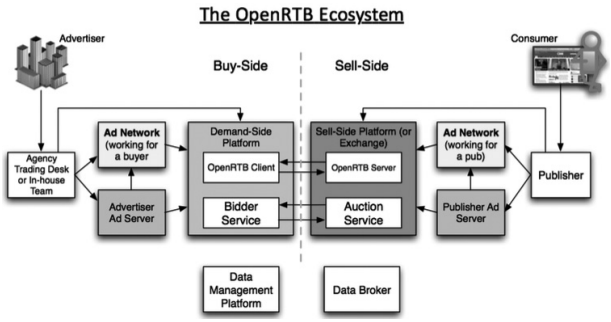
1. はじめに

オンライン広告においては、Webメディア閲覧ユーザー（以下ユーザー）のメディア閲覧行動を分析し、ユーザーの興味や関心といった潜在ニーズを推定することは営業収益に直結する極めて重要なテーマである。ユーザーの潜在ニーズに即して広告を配信できれば、ユーザーの閲覧満足度は高まり、広告クリックや広告主サイトの商品購買が増えることで、広告主の広告費用対効果を高められるからである。

広告に興味を持つユーザーは広告主や広告内容ごとに異なり、ユーザーの興味・関心も日々推移すると考えられる。したがってオンライン広告では、様々な角度からの分析を日々積み重ね、広告主ごとに広告配信に適したユーザー行動推定モデルをリファインしていくことが求められる。

しかしその一方で、分析対象データであるWebアクセスログは短期間で膨大な量が蓄積され、効率的な分析や分析サイクルの短縮化は容易ではない。マイクロアドでは、早くからHadoopなどの大量データの集計・分析基盤を導入し、一定の成果を上げてきた。分析の応答時間をさらに飛躍的に向上させ、分析サイクルの短縮化やサービス開発サイクルを短縮化させるため、現在我々は次世代の分析基盤を構築中である。その核となる構成要素の1つは、ログデータに特化したオンメモリ分散並列解析エンジンで、分析に必要な検索・解析を大幅に高速化するものと期待されている。

図1　オンライン広告RTBの概要（OpenRTB　API仕様［Ope 14］より引用）



本稿では、まずオンライン広告システムの概要を述べる（第2節）。次に、データ分析やデータ解釈に有用なWebページカテゴリ解析サービスについて述べ（第3節）、カテゴリを用いたデータ分析事例と分析に用いる技術を紹介する（第4節）。次に、次世代分析基盤の主要コンポーネントである行動ログ解析エンジンの概要を紹介する（第5節）。最後に、オンライン広告データ分析における我々のアプローチをまとめ、展望を述べる（第6節）。

2. オンライン広告システムの概要

オンライン広告市場は、2012年時点で約1，000億円、2017年の市場予測で3，200億円以上と急速に拡大しており、2017年時点では市場の3分の1がRTB経由の広告になると予想されている［Mar 14］。RTB（Real Time Bidding）［横山 12］は、ユーザーが広告枠を持つWebメディアにアクセスした際に、広告枠のリアルタイムオークションが行われ、オークションの勝者となった広告主が当該ユーザーの閲覧メディア内の広告枠に広告を配信できる、という仕組みのことをいう。RTBにおいて、広告枠を買う側（広告主側）のソフトウェアコンポーネントをDSP（Demand-Side Platform）と呼び、広告枠を売

る側（Webメディア側）のコンポーネントをSSP（Supply-Side PlatformまたはSell-Side Platform）と呼ぶ（図1）。広告配信を受けたユーザーが広告をクリックし、商品購入などの広告効果を達成することをコンバージョン（conversion）あるいは単にCVと呼ぶ。オンライン広告企業は、広告主サイトに埋め込んだHTMLタグとDSPを通じて広告主サイトアクセスログを取得し、SSPを通じてWebメディアアクセスログを取得する。オンライン広告データ分析の主要な目標は、これらのログを用いたユーザー行動分析により、CVする可能性の高いユーザーを識別するCVユーザー予測モデルを作成することである。

2007年設立のマイクロアドは、国内外12,000社以上の広告主、1,000を超えるメディア、6,500万人以上の広告配信可能ユーザー、月間3,000億回以上の広告リクエストを扱うオンライン広告大手で、BLADEと呼ぶDSP、COMPASSやAd-Funnelと呼ぶSSPをそれぞれ提供している。

きざしカンパニーでは、ブロガーの発信情報から話題を抽出し、ブロガーの興味や関心、領域ごとの熟知度を計測するシステムを開発してきた[稲垣 10]。ブロガーの執筆した記事を言語解析し、あらかじめ用意してあるカテゴリ体系に自動分類するものである。我々マイクロアドときざしカンパニーは、オンライン広告向けにこの技術を応用し、ユーザー行動分析に用いている。分析の元データであるログには、ユーザーがアクセスしたURLが記録される。URL数は膨大で、かつ単にURL表現を見るだけではページ内容も一般には不明なため、ユーザー行動分析には、URLは必ずしも適した表現ではない。またURLをそのまま扱っても、予測性能の高いCVユーザー予測モデルは作成できない。そこで我々は、マーケティング担当者が理解しやすいカテゴリ体系にURLの指し示すページ内容を分類し、データ分析に用いている。きざしカンパニーは、カテゴリ解析のほか、機械学習を用いたデータ分析、解析エンジンの研究開発によりデータ分析業務を支援している。次の節では、Webページをどのようにカテゴリ化するかを述べる。

3. ページカテゴリ解析

今回の研究開発では、解析対象となるデータソースは、ブロガーの発信情報（ブログ）ではなく、ユーザーの閲覧行動ログとなる。ブラウザを通してユーザーが閲覧したあれこれのURLを解析し、ユーザーがどのようなトピックに対して興味・関心をもっているかを分析する。この情報を広告推薦の精度向上に役立てようというものである。

URLの指し示す内容のカテゴリ自動分類を行うため、機械学習アルゴリズムとしてSVM（Support Vector Machine）を用いている。教師データとしては、あらかじめトピックカテゴリに分類されたURL数十万件を用いる。

表1　ページトピックカテゴリの例

大分類	中分類	小分類
アート＆エンターテインメント	音楽	クラシック音楽
アート＆エンターテインメント	音楽	ロック
ファッション	服飾雑貨	靴
ファッション	服飾雑貨	帽子
スポーツ	チーム競技	ラグビー
旅行	観光名所	温泉

トピックカテゴリは、大・中・小の3階層で約1,300程度の分類で構成されている。カテゴリの一部の例を表1に示す。次の節では、オンライン広告データ分析の概要と、トピックカテゴリをどのようにユーザー行動分析に用いるかを述べる。

4. オンライン広告データ分析

本節では、オンライン広告のデータ分析事例の一部と分析に用いる技術を紹介する。オンライン広告データ分析の主要課題は、過去のユーザー行動からCVに至るユーザー行動を突き止め、CVユーザーを予測することである。つまり、ユーザーが広告主サイトあるいはWebメディアのどのページにどの程度アクセスしたかを予測の判断材料とし、ユーザーを「CVする」・「CVしない」の2クラスに分類する問題として定式化し、これを解けば良い。過去データには、教師データ、すなわち実際にユーザーがCVしたか否かの情報が含まれる。統計的機械学習（statistical machine learning）では、このような問題をクラス分類問題、判断材料とする量を特徴量（feature）、予測対象の変数をクラス変数と呼ぶ。

オンライン広告では、多数ユーザーに無作為に広告を配信しても、実際にCVするユーザーはごく一部である。このようにクラス分布に大きな偏りがある場合を不均衡データと呼び、そのまま学習してもクラス分類性能は低いが、学習データの正例・負例比率を調節する単純な方法で分類性能を向上させられることが知られている[Hulse 07]。したがって本稿では、学習データの正例負例比率（CVユーザーと非CVユーザーの比率）は1対1にしたものを用いている。

予測モデルの性能評価は、利用可能な履歴データを学習データと予測データ（性能評価用データ）に分割し、学習データを用いて学習したモデルで予測データを予測し、「CVする」・「CVしない」について、予測結果と正解との一致度合いを計測することで行う。予測モデルの性能評価指標には様々なものがあり、たとえば、予測件数のうち正しく予測した件数で表される正確度（accuracy）は、単純で直感的だが、CV予測のようにまれな事象の予測性能評価指標としては適切でない。たとえば予測件数1万件のうち、実際にCVしたのは1件だったとしよう。このとき、常に全件をCVしないと回答する予測モデルの正確度は99.99%と高いが、このモデルは明らかに役に立たない。

クラス分布に大きな偏りがある2クラス分類問題では、予測

モデルの性能評価指標にAUC[平井 12]を用いるのが一般的である。予測モデルを用いた予測時に、CVする可能性が高いユーザーほどより大きな実数値スコアが出力されるとしよう。あるスコアを閾値としたとき、閾値以上のユーザーをCVユーザー、閾値未満を非CVユーザーと判定すると、その閾値を選択した場合の真陽性率(true positive rate、本当にCVしたユーザーを正しくCVと判定した割合)、偽陽性率(false positive rate、実際にはCVしなかったユーザーを誤ってCVと判定した割合)とを求められる。縦軸に真陽性率、横軸に偽陽性率を取り、閾値をスコアの高い方から低い方へ全ユーザーがCVユーザーと見なされるまで動かしながらプロットしたものをROC曲線(Receiver Operating Characteristic Curve)と呼び、ROC曲線の下側面積をAUC(Area Under ROC Curve)と呼ぶ。閾値とするスコアをs、真陽性率と偽陽性率を返す関数をそれぞれTPR、FPRとすると、AUCは定義より、式(1)で表される。

$$\mathrm{AUC} = \int_{\infty}^{-\infty} \mathrm{TPR}(s)\,\mathrm{FPR}'(s)ds \qquad (1)$$

ランダムな予測結果を返すモデルのAUCは0.5となり、必ず予測を的中させるモデルのAUCは1.0となる。したがって、AUCが大きく1.0に近いほどより良いモデルと評価する。またAUCは、予測データの正例と負例の組合せペアのうち、正例のスコアが負例のスコアより高いものの割合と等価なため、正例の件数をP、負例の件数をNとおくと、式(2)のように表すことができ、ROC曲線を描画しなくても算出することができる[Wu 07]。ただし$I_{s(i,j)}$は正例iのスコアが負例jのスコアより大きいときに1、それ以外では0を返す指示関数である。

$$\mathrm{AUC} = \frac{1}{PN}\sum_{i=1}^{P}\sum_{j=1}^{N} I_{s(i,j)} \qquad (2)$$

ある広告主向けのCVユーザー予測モデルを作成し、そのCV予測性能を示したROC曲線とそのAUCの例を図2と図3に示す。

D個の特徴量からクラス分類を行う問題を考える。入力となる事例ごとの特徴量ベクトル$\mathbf{x}$は$\mathbf{x}=(x_1,\ldots,x_D)^T$で表される実数の$D$次元縦ベクトル($T$は転置)で、各特徴量の重みベクトル$\mathbf{w}$も同様に$\mathbf{w}=(w_1,\ldots,w_D)^T$だとすると、2クラス分類問題を表す線形識別モデルは式(3)で表され、予測時にはyが正なら正例、それ以外なら負例と判断する。学習時にはN件の学習データ$\mathbf{X}=(\mathbf{x}_1,\ldots,\mathbf{x}_N)$のそれぞれについて、当該データが正例か負例かを示す教師データ$t\in\{+1,-1\}$が渡される。学習の目標は、未知の事例特徴量ベクトルが渡されたとき、このクラスを正しく分類する重みベクトル$\mathbf{w}$(とバイアスb)を推定することである。

$$y = \mathbf{w}^T\mathbf{x} + b \qquad (3)$$

式(3)の線形関数で2クラスに線形分離できない問題を解く場合、特徴量の組合せなどを考慮する非線形識別問題として解く方が予測性能は高くなることがある。ただし、どの特徴量が識別に効くかを判断することは線形識別モデルの方が容易であること、大量データの場合は線形識別モデルで充分な性能が出ることが多いこと、非線形識別モデルは学習により時間がかかることが多いことなどから、目的に応じて適した手法を使い分ける必要がある。

クラス分類問題を解く手法はさまざまだが、SVM[Cristianini 05]やRandom Forest[Breiman 01]は広く利用されており、本稿でもこれらを利用している。SVMは、線形空間をより高次元の非線形空間に射影するカーネルトリックと呼ぶ手法を組み合わせることで非線形識別に容易に拡張できる。またSVMは、所与のパラメータのもとでの大域最適解を見つけられるという理論的な利点があること、識別性能が高くlibsvm・liblinearといった実装が公開されたことなどから広く利用されるようになった。Random Forestは、学習データの復元抽出でデータの偏りを減らすバギング(bagging)と複数の決定木(decision tree)の多数決合議でクラス分類を行うアンサンブル学習(ensemble learning)の組合せ手法で、非線形識別を行い、識別性能も高いことで知られる。

CV予測を判断する特徴量をどのように設計するかは、通常、ドメインの知識を活用した試行錯誤が必要で、オンライン広告も例外ではない。ここでページカテゴリを用いることで、解釈しやすい意味のある単位で分析結果を得られるだけでなく、CV予測性能を向上させることができる。単純に、URLごとの何らかの量(閲覧時間やアクセス回数)を特徴量とすると、異なりURLごとに特徴量が必要となり、特徴量空間が極めて高次元になる。一般に、予測に要する学習データ量は、特徴量空間の次元数に対して指数オーダーで増加し、学習データ量が一定なら、特徴量空間の次元数が最適を超えて高次元になるほど予測モデルの性能は悪化する(次元の呪いと呼ばれる現象)。ページカテゴリを用いてURLを意味のある単位に集約することで、次元の呪いを避けられる。

分析に用いたデータは、全広告主の3ヵ月分のサイトアクセスログから、無作為抽出した1万人と、ある特定の広告主サイトのCVユーザー1万人についての行動ログである。他サイトへのアクセス行動から、特定広告主サイトでCVするユーザーを識別するCVユーザー予測モデルを作成したい。異なりURL約590万件を、大・中・小、約1,300のカテゴリに分類し、各カテゴリの平均閲覧時間を特徴量とした場合のROC曲線とAUCを図2に示す。

平均閲覧時間に加え、各カテゴリの日毎アクセス回数を特徴量に加えた場合の結果を図3に示す。平均閲覧時間のみの場合のAUCが約0.786なのに対し、日毎アクセス回数を特徴量に加えると0.844とより良い予測性能を示していることがわかる。またいずれの場合でも、生成された予測モデルの

図2　ROC曲線とAUC:カテゴリの平均閲覧時間を特徴量とした場合

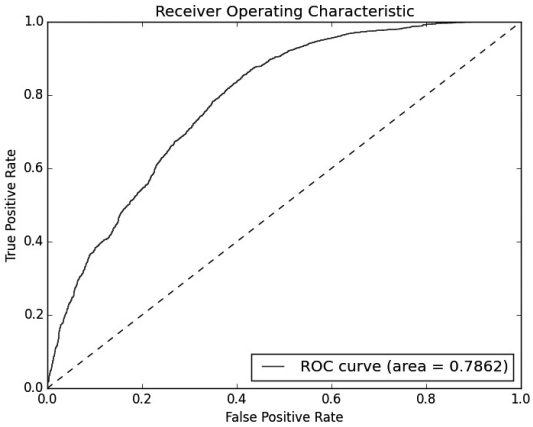
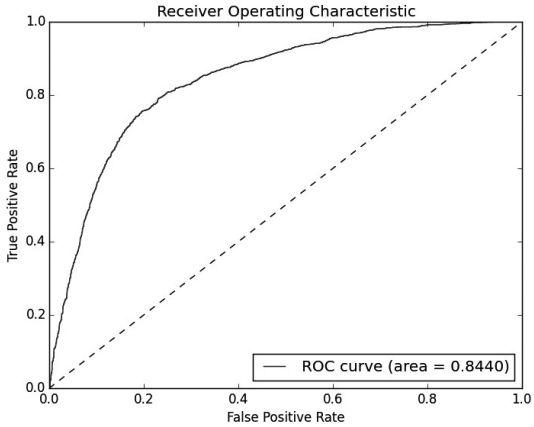


図3　ROC曲線とAUC:カテゴリの平均閲覧時間と日毎アクセス回数を特徴量とした場合。平均閲覧時間のみを特徴量とする場合（図2）よりAUCが高く、より良い予測性能を示している。



特徴量の重みベクトルから予測に重要な特徴量を分析すると、大・中・小いずれのカテゴリ階層も同様に重要なことがわかっている。

5. 行動ログ解析エンジン

現在我々は、次世代のデータ分析基盤を構築中である。本節では、その主要コンポーネントの1つである高速な行動ログ解析エンジンについて述べる。

巨大なデータを分析するため、様々な技法が研究開発されている。その主流の技術は、Hadoopなどに代表されるバッチ型のディスクベース技術である。我々はブログ検索エンジンで培った技術を応用することにより、これらとは異なるアプローチをとっている。

ユーザーの行動ログの基本構造は、（時間、主体、行動種別）という3項から構成される。このトリプルそれだけをソースデータとし、またユーザー行動の特徴的パターンを発見することだけをターゲットとする。汎用的なマイニング機能を捨て、特化した目的にフォーカスすることで、従来困難であった準リアルタイムなマイニングを目指している。具体的には、時間軸に最適化された索引構造を構成し、また人軸にそったクラスタ分散方式を採用することで、コンパクトなメモリとアルゴリズムを実現する。

分析者（例えばマーケッター）は、特定のユーザー集合（自社商品・サービスの購入者）がどのようなクラスタに分類され、それぞれのクラスタはどのような特徴を有しているのかを見い出さなければならない。このためには諸々のパラメータをいじりながら、何度もコンピュータ解析を繰り返す必要がある。しかし一回の解析に数時間や数日がかかるとしたら、思考は分断され、閃きや開眼には至り難い。これが、準リアルタイムの解析を追求する理由である。インタラクティブな分析を可能とする高速な行動ログ解析エンジンの研究開発を現在、進めている。

6. おわりに

本稿では、オンライン広告データ分析の概要と我々の取組みを紹介した。我々のアプローチのユニークな部分の1つは、Webページのカテゴリ解析である。オンライン広告データ分析では、マーケティング担当者が理解しやすく、かつ予測性能の高いユーザー行動予測モデルを得るために、URLを意味のあるカテゴリに自動分類することが有用である。本稿では、カテゴリ解析と機械学習によりユーザー行動予測モデルを作成する分析例を紹介した。データ分析業務では、さまざまな切り口の特徴量を設計し、より良い予測モデルを求めて分析を繰り返す必要があり、分析サイクル短縮は大きな課題である。この分析をインタラクティブに実行するには、応答時間の長い既存の分析環境では力不足である。分析処理の応答時間を飛躍的に向上させるために、我々は、アクセスログに特化したオンメモリ分散並列処理型の行動ログ解析エンジンの研究開発を行い、次世代データ分析基盤の主要コンポーネントとして構築中である。今後、分析サイクルと広告サービスのリリースサイクル短縮に大きく貢献すると期待している。

データ分析に有用な技術は、今後も積極的に取り込み、より良いオンライン広告を実現したい。たとえば機械学習分野の近年の大きな話題は、深層学習（deep learning）［岡谷 15］だろう。深層学習は、特徴量の自動学習と高い予測性能から大きな注目を集めており、オンライン広告への応用もどこまで広まるか興味深い。

本稿で紹介した我々のアプローチが、データ分析に取り組む方々の参考になれば幸いである。

参考文献

［Breiman 01］ Breiman, L.: Random Forests, *Machine*

Learning, Vol. 45, pp. 5−32（2001）

［Cristianini 05］ Cristianini, N. and Shawe-Taylor, J.：サポートベクターマシン入門（大北剛訳）、共立出版（2005）

［Hulse 07］ Hulse, J. V., Khoshgoftaar, T. M., and Napolitano, A.: Experimental Perspectives on Learning from Imbalanced Data, in *Proc. of the 24th Int'l Conference on Machine Learning*, pp. 935−942（2007）

［Mar 14］ 2014年アドテク広告市場は2，258億円/2015年にはスマホ広告がPC広告を上回る見込み、http：//markezine.jp/article/detail/20826（2014）、（2014−09−03付け）

［Ope 14］ OpenRTB API Specification Version2.2、http://www.iab.net/media/file/OpenRTBAPISpecificationVersion2_2.pdf（2014）、（2015−12−06アクセス）

［Wu 07］ Wu, S., Flach, P., and Ferri, C.: An Improved Model Selection Heuristic for AUC, in *ECML PKDD '07: Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases*, pp. 478−489, Springer-Verlag（2007）

［稲垣 10］ 稲垣陽一、中島伸介、張建偉、中本レン、桑原雄：ブロガーの体験熟知度に基づくブログランキングシステムの開発および評価、情報処理学会論文誌：データベース、Vol. 3、No. 3（TOD47）、pp. 123−134（2010）

［横山 12］ 横山隆治、菅原健一、楳田良輝：DSP/RTBオーディエンスターゲティング入門、インプレスR&D（2012）

［岡谷 15］ 岡谷貴之：深層学習、講談社（2015）

［平井 12］ 平井有三：はじめてのパターン認識、森北出版（2012）