O-021

メール開封率向上を目的とした学習技術の適用
Application of Learning Technology to Improve Email Open Rate

黄 錫明† 中込 健† 井上 豊†
Huang Ximing Nakagome Ken Inoue Yutaka

1. はじめに

昨今、電子メールは生活や仕事における情報伝達手段の 1 つとして、重要な役割を果たしている。その中において、多くの宛先に到達性の高いメールを配信するためのシステムやアプリケーションが多く提供されている。これらのシステムやアプリケーションが企業に利用される背景には、特定電子メール法や個人情報保護法などの法規制、および IP レピュテーションやスパムメールハニーポット等、不特定多数の宛先に大量のメールを配信する際に生じる運用上の困難な課題の存在が挙げられる。

近年、マーケティングの業界では、前述のような情報配信に関わる基本的な課題への対策とは別に、クロスチャネルのデータ連携によって e コマースサイトのコンバージョン率を上げるような取り組みなども進んでいる。本研究では、それらのチャネルの内、重要な 1 つの要素であるメールに関して、如何に開封してもらうかという点に着目し、その課題へのアプローチ方法の提案と検証結果を示す。

2. メール開封時刻の学習

一般的に e コマースサイトからのメール受信者は、常に大量のメールマガジンを受信していることが多く、受信者個人のセグメント情報や興味に基づいて作成したメールを配信したとしても、他のメールに埋もれてしまい、多くのメールが読まれないままになってしまう可能性がある。そこで、メールの開封数を向上するための方法について検討することを目的とし、メール受信者がメール開封する時間の予測可能性を調査するために送信時間と開封時間の相関関係についてデータ学習実験を行った。この実験で用いたトレーニングデータは、メール送信時間や開封時間のリアルデータを参考にしてモデルを作成し、擬似的に生成した。

表-1 トレーニングデータ種別

	説明変数(独立変数)				目的変数
	メール送信時間	メール送信曜日	メール送信月	想定開封デバイス種別	メール開封時間
種類	24	7	12	5	24
タイプ	数値	文字列	数値	文字列	数値
例	13	Monday	2	Android	15

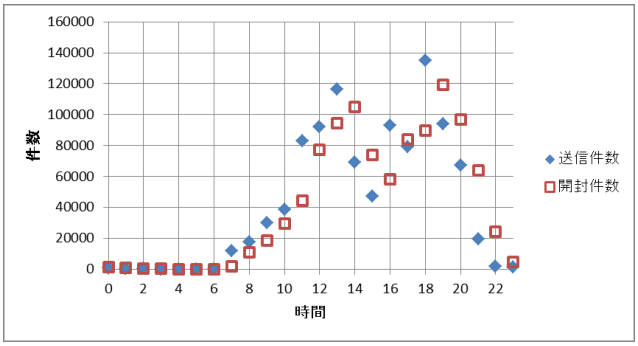

図-1 送信連動開封型データの分布

2.1 トレーニングデータの作成

トレーニングデータの分類は、データ学習を行う上で重要なフェーズである。この段階で参照可能な独立変数を全て学習させようとするとノイズが多くなり、予測精度や収束速度が下がる可能性がある。表-1 に示す通り、独立変数にはメールを開封するという行動と関係性が高いと考えられる「メールの送信時間」「送信した曜日」「送信した月」「送信したメールが開封対象と想定するデバイスの種別」の 4 つを選定した。トレーニングデータの組み合わせは、以下に従い擬似的に生成した。

- メール送信時間 $P_t(n)$：日常の平均的なメール送信時刻に関するリアルデータより算出した送信分布に従い、確率的に時間毎のデータを生成。
- メール送信曜日 $P_w(n)$：一般的な日常のリアルデータ統計では、月曜日から徐々にメール配信数が上昇し、金曜日をピークとして土曜日と日曜日は急激に下がるため、その確率分布に近くなるようデータを生成。
- メール送信月 $P_m(n)$：ある年のリアルデータを元に作成した月毎メール配信の確率分布に従いデータ生成。
- 想定開封デバイス種別 $P_d(n)$：国内における各デバイスの普及率を参考に各デバイスの利用確率分布を作成し、それに従いデータを生成。
- メール開封時間 $O_t(n)$：学習対象となる値であり、リアルデータ開封時刻のデータ分布を参考に生成。

メール開封時間データは、生成モデルとして「①送信連動開封型データ分布」と「② 時間帯別開封データ分布」の 2 種類を作成し、各々のモデルに従って生成したデータを用いて実験を行った。本来、メールは開封されない場合もあるため、未開封メールに関しても考慮する必要があるが、予備実験により未開封データに関しては、学習ノイズとなることが分かっているため、本実験は全てのメールが開封されるという前提で実施した。

① 送信連動開封型データ分布の場合

メール開封時間のピークが、メール送信時間の 1 時間後となるよう、正規分布モデルに従いデータを生成した。ここで、正規分布の分散値は 0.5 と設定した。擬似的に生成したデータにおける送信時間と開封時間の相関係数は約

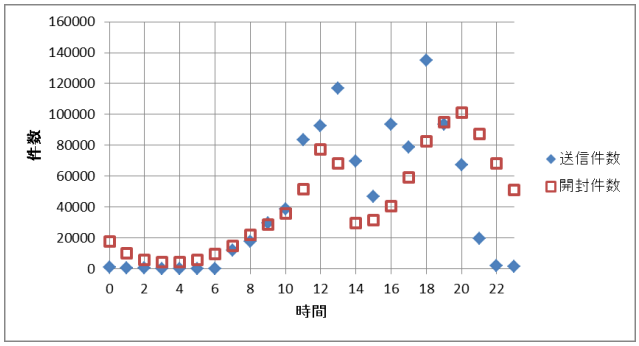

図-2 時間帯別開封型データの分布

†エクスペリアンジャパン株式会社：www.experian.co.jp（2017 年 6 月より チーターデジタル株式会社に商号変更：www.marketinggate.jp）

0.96 であり、図-1 に示す通り、各時間における送信件数と 1 時間後の開封件数は、同じような値になっている。

② 時間帯別開封型データ分布の場合

本分布は開封時刻のリアルデータに従った確率分布に従ってデータ生成する。次の 2 つのパターンに分けてメールの開封時間を算出する。1 つ目はメールが配信された日の 23 時台までに開封される場合であり、この場合は、送信時間から 23 時台までの確率分布を正規化してデータを生成する。2 つ目はメールが配信された日の 24 時までに開封されない場合であり、この場合は、メール配信の翌日 14 時までに開封されるものとした確率分布を正規化してデータを生成する。この時、当日の開封確率を P(t)、翌日の開封確率を P'(t')、メール送信時間を t_0とした場合、当日と翌日に開封される確率の関係は式(1)のように表すことができる。

$$\sum_{t'=0}^{13} P'(t') = 1.0 - \sum_{t=t_0}^{23} P(t) \quad (1)$$

擬似的に生成したデータの送信時間と開封時間の相関係数は約-0.22 であり、図-2 に示す通り、昼および夜がメールの開封件数が多くなる分布であることが分かる。

2.2 線形回帰分析モデルの適用と評価

本実験では、線形回帰分析（二乗損失関数+確率的勾配降下法）による学習モデルを適用した。学習モデルの主なパラメータは表-2 に示すとおりである。学習モデルが収束するためには、十分なトレーニングデータが必要となるため、事前に幾つかの予備実験を行った。その結果を参考にし、収束されるために必要なデータ数およびパラメータ数を調整した。今回の実験では、収束までの学習試行回数は、送信連動開封が 20 回、時間帯別開封が 8 回であった。ここで、適用した線形回帰分析によるメール開封時刻の学習に関する評価を行うため、式(2)に示す RMSE(Root Mean Square Error)を用いてモデルの評価を行った。

$$\mathrm{RMSE} = \sqrt{1/\mathrm{N} \sum_{i=1}^{\mathrm{N}} (\text{予測値} - \text{実値})^2} \quad \text{N: レコード数} \quad (2)$$

表-3 に送信連動開封と時間帯別開封の RMSE による評価結果を示す。この RMSE による評価数値が小さいほど、予測誤差が少ないものと見なされる。実験結果では、目的変数（メール開封時間）に対して、説明変数（独立変数）の相関係数が高い送信連動開封の方が RMSE の値が低く、予測誤差が少ないという結果になった。逆に時間帯別開封の場合は、送信連動開封と比較して、RMSE の値が 8 倍程度となっており、予測誤差が大きいことが分かる。

表-2 学習モデルパラメータ

	送信連動	時間帯別
作成後のモデルサイズ(B)	4032	3840
パラメータの数	42	40
収束までの学習回数	20	8
学習率	1.0	100.0
正則化	L1(1e-4)	L1(1e-4)
トレーニングデータレコード数	100万	100万
データをトレーニング用と評価用の割合	7:3	7:3

表-3 RMSE による評価結果

	送信連動	時間帯別
RMSE	0.7389	5.6171

2.3 考察

本実験では、対照的な 2 つのデータ分布モデルを用いた。開封件数は同じように推移しており、ピーク時間も近い時間帯を示しているが、相関係数は送信連動開封の方が 0.96 と高い値となっている。送信連動開封にて学習した場合は、連動特性を的確に捉え、高い精度で予測できたと言える。

一方で、時間帯別開封は-0.22 となっており、相関係数が低い値を示している。リアルデータを模した時間帯別開封にて学習した場合は、実際は開封率が低い昼の開封ピークと夜の開封ピークの中間点である 15 時前後を開封ピークと予測した。この理由としては、翌日開封されたデータはノイズとして認識された事が考えられ、このことにより夜の予測結果が下振れした可能性がある。この傾向は 2 つのデータ分布モデルに共通したことではあるが、相関係数が低い時間帯別開封に関してのみ、日中の時間帯に影響を与えたのではないかと思われる。

以上より、線形回帰分析を用いた開封予測を行う場合には、説明変数（独立変数）と目的変数（従属変数）の相関性に注意し、相関性の高いリアルデータに近い分布モデルを準備する必要があるということが分かった。

3. おわりに

本研究では、メール開封時刻の予測学習への適用可能性検証を目的として、メール送信情報やメール開封時間のリアルデータを参考にして作成したトレーニングデータにて、そのデータを線形回帰分析にて学習し、メール送信時間とメール開封時間の相関関係を調査した。この実験により、説明変数（独立変数）と目的変数（従属変数）の相関係数が高いほど、より正確な学習が実施できることが分かった。

本実験にて用いた線形回帰分析での学習結果より、単純なパターンでのメール開封時刻の予測は可能であり、利用者がメールを閲覧する可能性の高い時刻を俯瞰的に推測することは、人の行動パターンや状況を学習することによって実施できる可能性はあると認識できた。しかし、線形回帰分析では、説明変数を多く選択すると学習ノイズが大きくなり、収束しないという欠点があるため、より精度が高く複雑な学習を行うためには、他の学習アルゴリズムを適用する必要があると思われる。さらに、相関係数を高めるためには、どの配信メールに対して行われたメール開封なのかを紐付けて、リアルデータを収集し、その履歴を大量なデータとして蓄積する必要があると考えている。

前述のような学習の仕組みをメール配信システムに適用することにより、個別のメールアドレス毎、あるいは特定集団毎にメールを閲覧する可能性が高い時刻を予測し、メールを開封する可能性が高い時刻に送信するというような機能を検討することも可能であると考えている。また、個人がメールを開封する時刻に合わせてメールを配信し、メール閲覧時にメールボックスの上位に表示することで、メールの開封率を上げるような手法も検討したい。これにより、利用者毎に有用なオプトインメール[1]を適切なタイミングで配信するという方法も提案できるのではないかと考えている。

参考文献

[1] 猪内 学, 丹羽 清, “オプトインメールに関する実証的研究 : 配信形式・コンテンツ構成・インセンティブが広告効果に与える影響”, 経営情報学会誌 11(3), 65-79, (2002-09).