スパムメールに対するベイジアンフィルタの効果考察とAHP逆分析

日大生産工(院)　○島崎　啓知
日大生産工　　篠原　正明

1 まえがき

ベイズ定理を用いてスパム（迷惑）メールを検出するベイジアンフィルタの効果を文献調査した。さらに、単語の独立性仮定がベイジアンフィルタの有効性の阻害要因の１つである点の注目し、単語発生の相関性を考慮したAHP逆分析にもとづくベイズアプローチを考案する。

2 ベイジアンフィルタの効果考察

ベイジアンフィルタはベイズの定理をスパムフィルタに適応したものである。具体的には、受信者に多くのメールをスパムと正規メールに分類させ、その中で使用されている言葉をスパム辞書と非スパム辞書に登録していき、それに基づいてフィルタしていくという手法である。

一般的に３ヶ月ほど辞書を蓄積（トレーニング）すると使えるレベルの精度になると言われていた。わかりやすい手法であることもあり、多くの個人用メールソフトに実装された。

2．1 大きな欠点

トレーニング処理の存在が欠点である。ほとんどのソフトは最初から基本的な辞書が組み込まれているが、ソフトをインストールした後に必ずトレーニングが必要である。これは個人にとっても企業にとっても大きな負担になる。

2．2 トレーニングが必要な理由

すべての人や企業に適用できる汎用の辞書が作れないからである。汎用の辞書が最初から用意できれば、個人や企業でのトレーニングのプロセスは不要になり、負担が軽減されるが、それは実現できていない。

なぜなら、ベイジアンの有効性が個人の社会行動や嗜好などの要素に大きく依存しているからである。例えば、ゴルフに興味のない人がゴルフ用品の販売のメールを受信したら、高い確率でスパムであると判断する。一方、受信者がゴルフ好きの人ならゴルフ用品の特売情報を求めている可能性が高いため、その人にとってはスパムではない。よって、この条件を汎用のスパム辞書に登録することはできない。

さらに、ベイジアンフィルタの有効性を決定的に劣化させてしまったのは、スパマーがベイジアンポイズニング（Bayesian Poisoning）という手法を使うようになったことである。これは、非スパム辞書に入っていることが確実であると予測される言葉を大量に付加することによって判定を有利にしてしまう手法である。

また、HTMLや、Lの小文字を数字の1で置き換えるなどの意図的な誤記、スペースや記号を使って単語の判読を難しくするなどの手法が一般的になり、ベイジアンフィルタの精度は大きく劣化してしまった。

Effectiveness of Bayesian Filtering against Spam Mails and its AHP-based Analysis

Hiroshi SHIMAZAKI and Masaaki SHINOHARA

３　スパムメールのフィルタリングに活用する２つの情報

３．１　メールのヘッダー情報（送信元や件名など）

例えば、スパム送信者として有名な送信元をリスト化しておくと、フィルタリング時に、そこから送られてきたメールをスパムと判断する。件名に特定の語句が含まれているメールをスパムとすることも可能である。しかし、これだけだと未知の送信者や件名を持つメールをブロックできないのである。

３．２　メールの本文

本文のテキスト情報を解析して、それがスパムであるかどうかを判断する。メールの本文を解析する手法として最も一般的なのが、ベイズ分類と呼ばれる手法である。どのようなメールがスパムで、どのようなメールが正当なメールかという情報をあらかじめ用意しておく。送られてきたメールに対して、どちらの可能性が高いかを解析して結果を決める。

例えば、Aという単語が含まれるときそのメールがスパムであるか，そうでないかを判断する。まず，スパムメールと正当なメール両方を含むサンプルメールを用意する。それを解析して，各単語の出現頻度を算出する。次にベイズの定理を利用して，Aが含まれるときスパムである確率と，そうでない確率を導き出す。

実際には，複数の単語の組み合わせなどもっと複雑な情報を利用して解析する。また，使われる情報は必ずしも単語の出現頻度だけとは限らない。メールの構造など，単語以外の情報を使うこともできる。最終的にスパムである確率が一定値を超えたらスパムと判定する。

４　AHPによる逆分析

一部の専門家によれば、ベイジアンフィルタに使われているベイズ定理は「ナイーブベイジアン」と呼ばれ、各単語の出現を独立であると仮定しているため、有効性が一定以上にならない。

独立性を仮定する場合としない場合について、AHP図式（[３]、[４]）を用いた逆推定確率の検討を簡単な例を通して以下に示す。

[例4.1]　１つの単語に注目してスパム語か否かを判定する例：英文メール全体の60%が正常メールで、残り40%がスパムメールであるとしよう。「republic」という単語に注目すると、正常メールの5%に、スパムメールの30%に、単語「republic」が含まれるとする。あるメールに単語「republic」が含まれるときに、そのメールがスパムと判断される確率（スパム度逆推定確率Pr(spam | republic)）は、図４．１のAHP図式において破線矢印で表示した確率をベイズ定理のAHP解釈（[３]、[４]）に従って、以下で計算できる。

メールが単語「republic」を含む確率：

Pr(republic)=0.6×0.05＋0.4×0.3=0.15

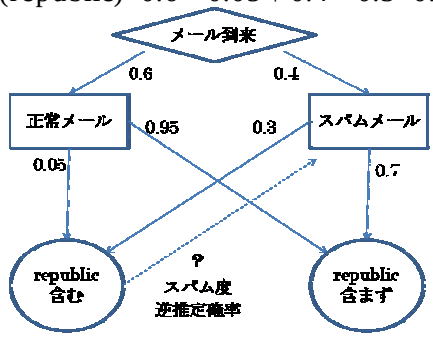

図4.1：単語「republic」に注目したスパム判定のAHP図式

確率フローの局所均衡の仮定より、次式(1)が成立することがベイズ定理として要求される。

Pr(republic)×Pr(spam | republic)
= Pr(spam)×Pr(republic | spam)　…　(1)

従って、スパム度逆推定確率Pr(spam | republic)は以下で計算できる。

Pr(spam | republic)=0.4×0.3 / 0.15=0.8

すなわち、この例では単語「republic」のスパム度は0.8と推定される。

[例4.2]　２つの単語に注目してスパム語か否かを判定する例：「republic」と「strong」の２つの単語に注目して、スパム度を判定する際に考えられる３つのAHP図式を、図4.2(a)、4.2 (b)、4.2 (c)に示す。

図4.2(a)では、republicのスパム度逆推定確率とstrongのスパム度逆推定確率を独立に求め、両者の積により、２つの単語を共に含んだ時にスパムである確率を推定する。

図4.2(b)では､相関性を含んだ形で２つの単語を共に含んだ時にスパムである確率を推定することができる｡図4.2(a)での結果と比較することにより、２つの単語間の相関性を調べることができる。

図4.2(c)では、文献[4]のパス型ベイズ定理を用いて、２つの単語を共に含んだ時にスパムである確率を推定する。

５ まとめ

ベイジアンフィルタの有効性を調査し、有効性阻害要因の１つである単語の独立性仮定を克服する方法として、相関性を考慮できる２種類のAHP図式に基づくベイズアプローチを提案した。具体的な適用例ならびに３つ以上の単語についてのAHP図式などは今後の課題である。

「参考文献」

1) http://www.thinkit.co.jp/free/article/0611/8/4/
2) http://itpro.nikkeibp.co.jp/members/NBY/techsquare/20040528/5/
3) 篠原正明：ベイズの定理とマルコフ連鎖と枝確率フロー平衡、第36回日本大学生産工学部・学術講演会・数理情報部会論文集、pp.41-44(2003.12).
4) 篠原正明,篠原健：ベイズ定理の一般化－多段階逆推定とパス解釈－、第38回日本大学生産工学部・学術講演会・数理情報部会論文集、pp.59-62(2005.12).

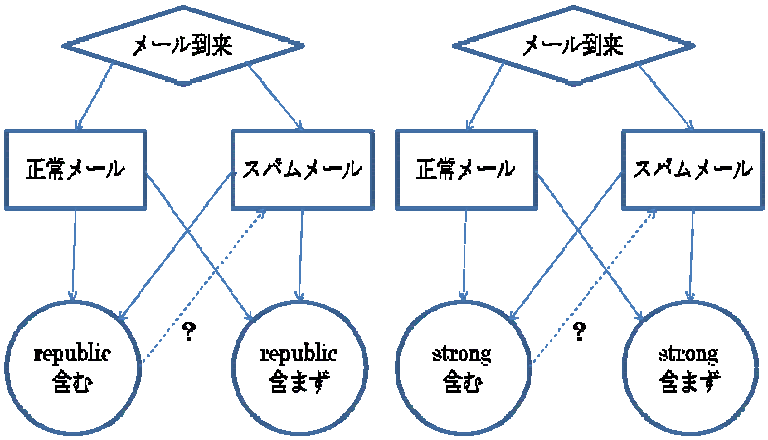

図4.2(a)：２つの単語の発生が独立としたAHP図式

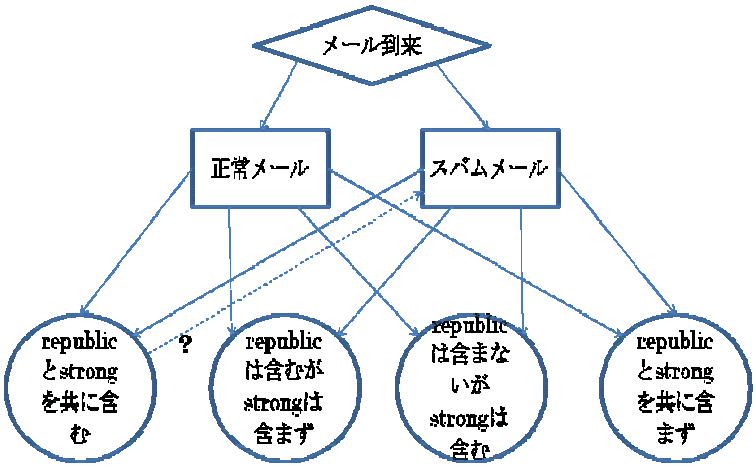

図4.2(b)：２つの単語の発生を同時的にとらえ相関性を考慮するAHP図式

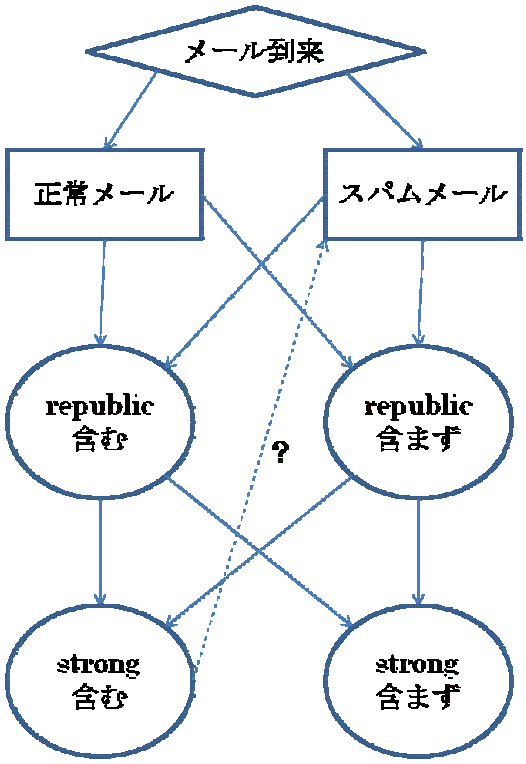

図4.2(c)：２つの単語の発生を直列的にとらえ相関性を考慮するAHP図式