ベイズ統計によるスパムフィルタの設計

○清水　秋穂、中嶋　卓雄（九州東海大学）

1. はじめに

電子メールは現代において欠かすことのできない重要な通信手段であるが、これを悪用した迷惑メールの増加が電子メールの持つ問題のひとつとなっている。

“スパム（spam)”と呼ばれるこれらのメールは、送信先のユーザーが必要としているか否かに関わらず無差別に、そして大量に送りつけられる。そのため、ユーザーは大量に受信したメールの中からスパムと必要なメールとを選別しなければならない。スパムへの対策として、メールアドレスやキーワードを指定することで受信拒否を行えるメールフィルタ（スパムフィルタ）が開発されたが、送信元アドレス、メールのヘッダ部分などは改ざんされている場合も多く、受信拒否の設定したフィルタによる効果は完全ではなかった。また、単純なキーワード指定では、指定したキーワードが含まれるメールは全てスパムであると判断されるため、必要なメールであっても誤って受信拒否を行ってしまう場合があった。

本論文では、原因を確率的に推定することが可能な、ベイズの定理を用いたスパムフィルタの設計を行う。

2.ベイズ的アプローチ

2.1　ベイズの定理

ベイズの定理とは、ある結果が得られた時その結果を反映したもとにおいて、事後確率を求める定理である。事象$H_1, H_2, H_3, \cdots$のいずれかの原因で事象Aが得られた時、Aが発生した原因がH_iである確率$P(H_i/A)$は以下の通りに示される。

$$P(H_i \mid A) = \frac{P(H_i)P(A \mid H_i)}{\sum_j (P(H_j)P(A \mid H_j)} \quad (1)$$

$P(H_i)$は事象H_iが発生する確率（事前確率）である。$P(A/H_i)$は事象H_iが発生したうえで事象Aが発生する場合の条件付確率であり、原因H_iによる事象A発生の尤もらしさ（尤度）を表す。

2.2　ベイジアンフィルタリング

ベイズの定理では結果から、それを得るための原因を推定することが可能である。ベイズの定理を応用したベイジアンフィルタリングでは、受信したメールに含まれる単語からそのメールがスパムであるか否かを推定する。受信したメールを解析し、スパムであるか否かの判定を行い、その判定の正当性をチェックすることで学習を行う。

従来のメールフィルタリングではキーワード指定によりスパムか否かの判定を行っていた。そのため、実際にはスパムではないメールであっても、指定されたキーワードを含んでいる場合にはスパムであると判定されてしまうことがあった。しかし、ベイジアンフィルタリングではキーワードの指定は行わず、これまでに受信したスパムから統計的に解析を行うため、誤った判定を行う可能性は低くなる。

3.スパムフィルタの設計

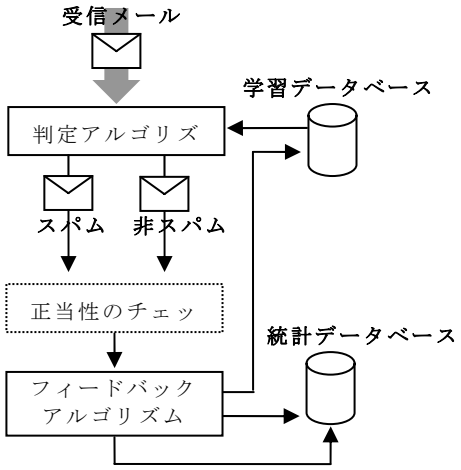

図１　スパムフィルタ処理アルゴリズム

図１にフィルタ処理アルゴリズムの概要を示す。学習用 DB を利用して判定アルゴリズムによりスパムであるか否かを自動的に判定し、さらに学習効果を高めるために、その正当性を人間によってチェックし、その結果をフィードバックアルゴリズムにより、学習ＤＢに反映させる処理アルゴリズムとなっている。また、システムの統計データを計測、集積するための統計 DB も用意する。

3.1　初期化

まず、学習用 DB に初期値を与えて、判定アルゴリズムを動作させる。受信したメールがスパムである確率を $P(S)$、スパムではない確率を $P(non\text{-}S)$とし、それぞれに初期値 0.5 を与える。また、受信したメールの総数を $total$、受信したスパムの数を $spam\text{-}total$ と

し、それぞれに初期値0を与える。スパムの判定に用いる閾値（threshold）θには0.5の値を与える。

3.2 判定アルゴリズム

構文解析により、メールの本文中に含まれる名詞・動詞を抽出し、それらの単語の使用頻度を調べる。

スパムの判定は閾値θと受信したメールのスパムらしさを比較することで行われる。受信したメールm_iのスパムらしさ$q(m_i)$は、以下の式によって求められる。

$$q(m_i) = \sum \frac{f(w_i)q(w_i)}{\log(n)} \qquad (2)$$

ここで、$f(w_i)$は単語w_iがメール本文中に現れる頻度を、$q(w_i)$は単語w_iのスパム度数を表す。nはメール本文の単語数である。

$q(m_i)$を閾値θと比較する。$q(m_i) \geqq \theta$のときメールm_iはスパムであると判定され、$q(m_i) < \theta$のときメールm_iはスパムではないと判定される。ここでは、一例として、単に単語の頻度情報のみを$q(m_i)$の計算式に入れているが、この部分はtf・idfなどの意味評価手法により補正すれば、さらに精度の向上が望めると思われる。

3.3 フィードバックアルゴリズム

ベイジアンフィルタリングにおいて“学習”はフィードバックアルゴリズムにより行われる。

受信したメールm_iのスパムらしさ$q(m_i)$と閾値θを比較し、メールm_iがスパムであるか否かの判定が得られるが、ユーザーによる正当性のチェックにおいて判定のミスが発見された場合、閾値の補正が必要になる。判断ミスの原因は2種類存在し、スパムと判定されたが、実際はスパムでなかった場合、およびスパムでないと判定されたが、スパムであった場合である。スパムであると判定されたメールが実際にはスパムではなかった場合、$q(m_i) < \theta$に近づくように閾値θを増加させるか、または単語のスパム度数が減少する計算式で値を修正する。また、スパムではないと判定されたメールが実際にはスパムであった場合$q(m_i) \geqq \theta$に近づくように閾値θを減少させるか、単語のスパム度数を増加させる。

メールがスパムである確率$P(S)$、スパムではない確率$P(non\text{-}S)$の再計算を以下の式を用いて行う。

$$P(S) = \frac{spam\_total}{total} \qquad (3)$$

$$P(non-S) = \frac{non-spam\_total}{total} \qquad (4)$$

$total$は受信したメールの総数、$spam_total$は受信したスパムの数、$non\text{-}spam_total$は受信したスパムではないメールを示す。

単語w_iのスパム度数$q(w_i)$は、単語w_iを含む場合にメールm_iがスパムである確率$P(S/w_i)$に等しく、ベイズの定理により以下の通りに表される。

$$P(S \mid w_i) = \frac{P(S)P(w_i \mid S)}{P(S)P(w_i \mid S) + P(non-S)P(w_i \mid non-S)} \qquad (5)$$

$P(w_i/S)$はメールがスパムである場合に単語w_iが含まれる確率、$P(w_i/non\text{-}S)$はメールがスパムでない場合に単語w_iが含まれる確率である。これらの確率は、以下の式によって求められる。

$$P(w_i \mid S) = \frac{M_{spam}(w_i)}{spam\_total} \qquad (6)$$

$$P(w_i \mid non-S) = \frac{M_{non-spam}(w_i)}{non-spam\_total} \qquad (7)$$

$M_{spam}(w_i)$は単語w_iを含むスパムメールの総数、$M_{non\text{-}spam}(w_i)$は単語w_iを含むスパムメールの総数である。これらは統計DBに格納される。

4.おわりに

本論文ではベイズ統計を用いたスパムフィルタの設計について述べた。

ベイジアンフィルタリングでは判定と判定結果の修正を繰り返すことでスパム判定の正確性を向上させることができる。こうして得られた高い精度を持つベイジアンフィルタリングの学習内容を複数のユーザーが共有することで、利用初期からスパム判定を正確に行うことができる。しかしながら、どのような内容のメールをスパムであると判断するかは個人によって差があるため、多数のユーザーでベイジアンフィルタリングを共有することには問題がある。

さらに、実際の動作環境において、個々のパラメータの初期値の設定に関する問題、および学習アルゴリズムへのフィードバックするパラメータの修正値の問題など、検討すべき点は多い。今後は、実際に実装させ、各パラメータの調整を具体的に検討していきたい。

問い合わせ先：中嶋 卓雄
〒862-8652 熊本市渡鹿 9-1-1, 九州東海大学
応用情報学部 情報システム学科
tel: 096-386-2837
E-mail：taku@ktmail.ktokai-u.ac.jp