ホワイトリスト・ブラックリストを用いたスパム検出法

日大生産工(院)　○島崎　啓知
日大生産工　　篠原　正明

1 まえがき

近年、スパムメールの増加が深刻な問題になりつつあります。完全にスパムかどうかを判定するのは難しく、判定ミスを避けることはできません。スパム検出の精度を向上させ、判定ミスを減らすことが重要な課題になっています。

スパムの判定ミスには、false positive（スパムでないのにスパムと判定されるミス）とfalse negative（スパムなのに正規メールと判定されるミス）の2種類あります。

false positiveは、受け取りたいメールが届かなくなる、あるいはスパムの中に正規メールが混じってしまうことにより、間違えて捨ててしまう可能性があり、false negativeより害があると言えます。

本研究では、ホワイトリスト・ブラックリストを用いてスパムを検出する方法を提案します。また、false positiveを減らす効果があるかどうか考察を行います。

2 ホワイトリスト・ブラックリスト

ホワイトリストとは、信頼できるメールサーバーの一覧のことです。メリットはスパムメールを完全に遮断できることです。デメリットは登録をしないと受信できないことです。ブラックリストとは、スパムメールを送り付けてくるメールサーバーの一覧のことです。メリットはスパムメールを概ね遮断でき、ホワイトリストに比べ未知の正規メールを多く受信することができることです。デメリットはスパムメールを完全に遮断できないことと、カテゴリーで一律に遮断すると正規メールを受信できなくなることです。

本研究では、「確実」にスパムでない単語のリストをホワイトリスト、過去に出現した単語のリストをブラックリストと考えスパム判定を行います。

3 スパム判定

新しくメールを受信するとフィルタによって本文は単語に分解され、その単語の中からスパム度が0.5から最も離れている単語をn個（経験則からn = 15）抽出します。その単語の結合確率（以下の数式a）を使って、そのメールがスパムであるかどうか判定します。

$$P_{(M)} = \frac{\prod P_{(w)}}{\prod P_{(w)} + \prod (1 - P_{(w)})} \cdots \text{(a)}$$

このとき

$spam: P_{(M)} \geq 0.9$　、$ham: P_{(M)} < 0.9$

これをPaul Graham方式と言います。
P(w)は単語wのスパム確率であり、以下のように計算されます。

$$P_{(w)} = \frac{b / n_{bad}}{ag / n_{good} + b / n_{bad}} \cdots \text{(b)}$$

$$P_{(w)} = 0.99 \quad (n_{good} = 0)$$

$$P_{(w)} = 0.01 \quad (n_{bad} = 0)$$

ブラックリストに含まれていない単語（未知の単語）は、P(w) = 0.4とします。

b…wがspamに登場する回数
g…wがhamに登場する回数
nbad…wが含まれるspamメールの総数
ngood…wが含まれるhamメールの総数
a…バイアス

false negativeよりfalse positiveの方が「害が大きい」という考えより、バイアス(a = 2)をかけることによってfalse positiveを減らします。

Spam Detection using White List and Black List

Hiroshi SHIMAZAKI and Masaaki SHINOHARA

4　ホワイトリストの利用

新しくメールを受信したときに、本文中にホワイトリストに存在する単語がある場合、スパム度を計算するときにその単語を強制的に使用します。

例えば、「水泳」という単語をホワイトリストに登録しておきます。新しくメールを受信したときに、本文中に「水泳」という単語が存在する場合、スパム度が0.5から最も離れている単語を15個抽出し、さらにホワイトリスト内の「水泳」を利用して、単語16個で結合確率を計算します。

ホワイトリスト内の単語のスパム度はすべてP(W) = 0.01とします。スパム度の小さいものを加えることによって、false positiveを減らす効果があると想定されます。

ホワイトリストを利用した場合と利用しない場合の結果を「例題1と例題2」、「例題3と例題4」で示します。

[例題1]

メールを受信したとき、単語を5個抽出してスパム度を計算する。(n1 = 0.98, n2 = 0.86, n3 = 0.79, n4 = 0.54, n5 = 0.4)
このメールのスパム度P(M)は0.998873となり、0.9以上になるので「スパム」と判定される。(例題終了)

[例題2]

メールを受信したとき、単語を5個抽出した。さらに本文中にホワイトリストに含まれる単語が1個存在していたので、これも合わせてスパム度を計算する。単語n1~n5は例題1と同じとし、n6はホワイトリストに含まれる単語とする。(n1 = 0.98, n2 = 0.86, n3 = 0.79, n4 = 0.54, n5 = 0.4, n6 = 0.01)
このメールのスパム度P(M)は0.899510となり、0.9未満になるので「スパムでない」と判定される。(例題終了)

[例題3]

メールを受信したとき、単語を5個抽出してスパム度を計算する。(n1 = 0.99, n2 = 0.94, n3 = 0.89, n4 = 0.76, n5 = 0.4)
このメールのスパム度P(M)は0.999962となり、0.9以上になるので「スパム」と判定される。(例題終了)

[例題4]

メールを受信したとき、単語を5個抽出した。さらに本文中にホワイトリストに含まれる単語が1個存在していたので、これも合わせてスパム度を計算する。単語n1~n5は例題3と同じとし、n6はホワイトリストに含まれる単語とする。(n1 = 0.99, n2 = 0.94, n3 = 0.89, n4 = 0.76, n5 = 0.4, n6 = 0.01)
このメールのスパム度P(M)は0.996277となり、0.9以上になるので「スパム」と判定される。(例題終了)

5 結果と考察

例題1~4の結果から「ホワイトリストを利用したことでスパム度が減った」と言える。

例題3と例題4の結果から「スパム度がかなり高いものに対して、ホワイトリストを利用してもスパムと判定される」と言える。以上の結果から、false positiveを減らす効果がありそうだが、false negativeが増える可能性もあると考えられる。

例えば、「水泳」をホワイトリストに登録しておくと、本文中に「水泳」が含まれるメールは非スパムと判定されやすくなる。スパムメールに「水泳」が含まれていても同じ効果が出てしまう。しかし、スパム度がもともと高いメールに、ホワイトリストを利用してもスパムと判定されるので、false negativeを抑える効果もあると考えられる。

6 おわりに

Paul Graham方式を改良したGary Robinson方式があります。このRobinson方式が優れている点は、単語wの総出現回数を考慮に入れている点で、総出現回数が少ない場合、単語wの比重が小さくなるようになっているので情報不足を補えます。

Robinson方式を利用した場合、今回の結果とどのように違ってくるのかは今後の課題とします。

「参考文献」

[1] http://akademeia.info/index.php?%A5% D9%A5%A4%A5%B8%A5%A2%A5%F3%A5%D5%A5%A3%A5%EB%A5%BF

[2] http://www.sabamiso.net/yoggy/hiki/?%A5%D9%A5%A4%A5%B8%A5%A2%A5%F3%A5%D5%A5%A3%A5%EB%A5%BF%A4%F2%BB%C8%A4%C3%A4%BFspam%C8%BD%C4%EA%A4%CE%CE%E3