汎用連想計算エンジン GETA を用いたスパム判別に関する研究

石黒 雄輔 (0610006)

北陸先端科学技術大学院大学 情報科学研究科

2008 年 8 月 8 日

キーワード: 統計的手法、スパムフィルタリング、連想計算.

インターネットの普及により、生活の様々な面で利便性が高まってきた。特に電子メールは現在、社会的な活動を支える通信基盤として広く普及している。しかし現在、電子メールの 90%ー95% をスパムメールが占めるといわれ、大きな問題となっている。既存のスパムメールフィルターは、人手で記述したルールに基づくヒューリスティックスフィルターとベイジアンフィルターを用いた手法が代表的である。問題点として、ヘッダー解析規則を全て人手で記述するこの方法では膨大な労力が必要になり、管理上の負担が重くなる。またベイジアンフィルターは精度向上にはトレーニングが必要であり、新たなタイプのスパムメールに対し対応に時間がかかる。そのため現在では、商用のスパムメールフィルタは複合フィルタなどが技術を組み合わせているケースが多い。

本研究では、単純な統計的手法によるスパム判別の可能性を新たな角度から追求し、メールのテキスト (本文と`Subject` の自然言語部分) のみを用いた連想計算によるスパムフィルターの構成法を提案し、そのスパム判別の一次フィルター (または最終段フィルター) としての可能性を探る。その際、最大限の既存ツールの活用を前提とし、高速な汎用連想計算エンジンGETA を用いる。連想計算とは文書と単語の多重集合とみなし、単語の文書中の出現頻度のみで単語群-文書、文書群-単語の間の類似度を計算する。このことにより、文書-文書間や単語-単語間、文書-単語間の様々な連想計算が可能となる。

連想スパムフィルターはメールから単語群抽出の前処理の後、連想計算に閾値判別を組み合わせることで構成する。連想スパムフィルターの可能性を評価するための実験システムは形態素解析器`Mecab` などの前処理機構と連想計算エンジンGETA の 2 つからなる。GETA では単語の文書内での出現頻度は行成分を文書内の単語頻度ベクトル、列成分を単語の文書での出現頻度ベクトルとする行列`WAM` で表現されている。閾値判別機構は現状では未実装であり、閾値判別の可能性を明らかにするため、実験を行い、リファレンススパムメール群に対する単一の類似度（実数値）とメールテキストの異なり単語数に関する線形回帰のみによって推定された単純な閾関数で実験を試みた。その際、実験のデータセッ

トはリファレンススパムメール群 α、β がそれぞれ約 80,000 件, 約 150,000 件、必要なメール群 γ が約 1,000 件を用意した。

実験における比較軸としてはメール集合の類別、各メールの抽象化、評価関数の選択を取り上げた。ここでメール集合の類別とはメール集合の分割を指し、メールの抽象化とは、メール単体の構成要素のうち、どの部分を抽出するか、および抽出された部分に対する処理をいう。メールの類別では、英語と日本語に分類を行い、メールの抽象化では品詞によるフィルタリング (日本語)、重複の有無や記号の保存、除去や本文と `Subject` の切り分けを行った。また評価関数の比較は`TF`,`SMARRT measure` で行った。これらの比較軸で実験を行ったところ、リファレンススパムメール群を日本語メールと英語メールに類別することにより、適切な抽象化のもとでスパムメール群と必要なメール群が顕著に分離する。

連想スパムフィルターのセットアップには膨大な計算を行う必要がある。この処理は二段階あり、メール群の単語頻度情報作成と GETA の要素である`WAM` の作成からなる。前者は現状の実装では重い処理だが、一度行えば新たなリファレンススパムメールの追加処理は漸進的であり、大きな問題点にはならない。また、行列の作成は漸進的にはできないが 10 万件程度のスパムメールであれば、数十秒程度で終了する。これらのセットアップができていれば、入力メールに対する連想スパムフィルターの処理は 0.3 秒以下であり、効率的に行うことができる。

現在の閾関数は、スパムメール群に対する単一の類似度（実数値）に対するメールテキストの異なり単語数との線形回帰のみの単純なものに限られており、より適切な閾関数の構成法と複数の WAM を組みあわせたスパム判別による改善が今後の課題である。