3 フィルタリング

IRI コミュニケーションズ
安藤 一憲　ando@iri-com.co.jp

spam メール対策の中にあってフィルタリングはユーザの手の届く範囲に実装されるべき技術である．spam メールの定義は個々人によって微妙に異なっており，その違いを吸収するためには，ユーザがフィルタをコントロールできることが必要になるからである．従来から多く使われているメールサーバに届いたメールを手元に持ってくるための POP3 プロトコルが単一のメールボックスだけを念頭において設計されていることも，フィルタの存在形態に少なからず影響を与えている．spam メールは万人に平等に降るわけではなく，1 日数万通という人から 1 通も来ない人まで非常にバリエーションが広い．自分に 1 日 15 通しか spam メールが来ないから自分のいる組織に対策は必要ないと考えるのは早計で，隣の席の人間が数千通の spam メールを受信している可能性があることを知るべきである．このような状況のもと，ユーザから見た場合に spam メール対策の最後の砦となるのがフィルタである．

フィルタの種類

最近の spam メールフィルタは複合型が多く，いくつかの技術要素が組み合わされてできているケースが多い．各フィルタ製品なりフィルタリングソフトウェアなりを単純には分類できない状況ではあるが，ここではその主な技術要素を分解してみようと思う．

【ベイジアンフィルタ】

ある確率変数のセットが規定された時にその条件下でそれらの確率変数と何らかの因果関係を持つ別の確率変数がどうなるかを扱うのがベイジアンフィルタのもとになっているベイジアンネットワークの考え方である（図 -1）．ベイジアンネットワークは事例（確率変数のセット）による学習が可能であり，画像認識，音声認識，ロボットと応用範囲も広いことが知られている．spam メールフィルタでこの仕組みが応用され始めたのは Paul Graham の「A Plan for SPAM」[1] という文章がきっかけとされているが，この学習可能という性質は前述の「ユーザによる spam メールの判定基準の違い」を吸収する効果を発揮している．経緯はともかく，現在ネットワークユーザの身近で最も役に立っている人工知能分野の成果はベイジアンネットワークかもしれない．

spam メールフィルタでは辞書にある個々の単語の文面への登場が確率変数のセットに対応しており，その条件下で「spam メールかそうでないか」を判定する．計算量の増大を避けるために，辞書の規模を一定にしている実装もあるが，辞書をどのような単語で構成するかによっても判定性能が左右されることになる．

ここまで書くと，「なんだ spam メールフィルタというのはベイジアンフィルタがあれば十分ではないか」と錯覚する方もいるかもしれない．だが世の中はそう甘くはない．spammer は送信元 IP アドレスベースのブロッキングを回避するためにワームを作成して 100 万台以上と言われるマシンを束ねて botNet を形成するくらいしたたかな連中である．

spam メールの文中には「Vi*gra」等，途中にあらぬ文字を混入した単語が散見される．これは人間が読む時には文字を推定することで認識されてしまうが，ベイジアンフィルタは辞書にこの単語の登録がない限りこれを単語として認識することはできない．この手法は「Snow freaking」と呼ばれている．すべて辞書に登録しておけばいいと考える人は結構いそうだが，「Viagra」と認識できそうな文字列だけで 60 京通りにも及ぶという試算も発表されている [2]．その数の文字列を辞書に持つのは現実的ではないので，単語のかわりに正規表現のパター

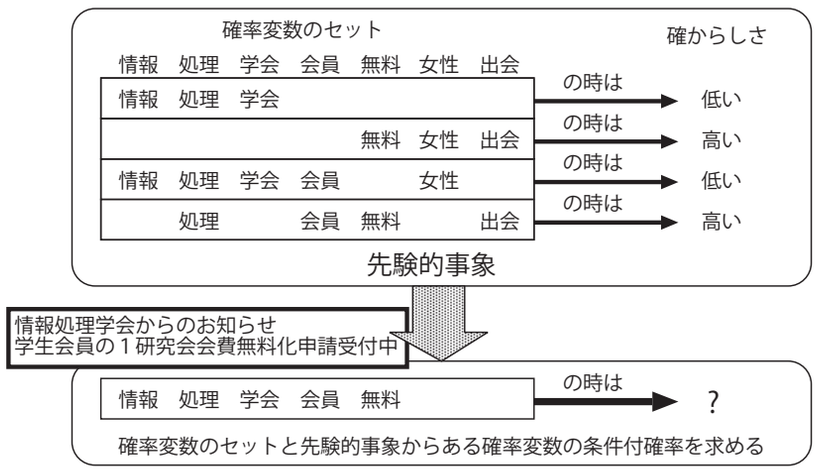

図 -1　ベイジアンネットワーク

ンを辞書に持たせるという手法が検討されている．別の見方をすれば，それだけ人間の単語認識の能力は奥が深く，認識という分野の研究成果もそこまでは及ばないとも言える．単語を単語として認識させない手法はほかにもあり，たとえば HTML のコメント文で単語を分断する手法もよく見られる．さらには単一の文字でバナーを作って商品名を表示する例もある．spam メール対策でベイジアンフィルタの歩んでいる道はかなり険しい．

また，spam メールの中にランダムな単語列を含ませるという方法も散見される．これは，ベイジアンフィルタは辞書に登録された単語の出現傾向で判定をしているので，そもそもの情報にノイズを入れてしまえというアプローチである．判定が狂うという側面もあるが，そういうランダム文字列を含む spam メールを spam メールとして学習させるとベイジアンフィルタの辞書がダメになるという 2 次被害が出る．ランダムな単語列を判定して取り除くアプローチも登場しているが，ランダムな文字列でなくとも，ニュース記事の最後に URL を 1 行記載した spam メールさえ出現してきている．そういう意味では spammer はベイジアンフィルタを十分に研究し警戒している感がある．

ここまでに述べた辞書に登録された単語の出現傾向だけから判定を行うベイジアンフィルタは「ナイーブな」ベイジアンフィルタと呼ばれている．

【ヒューリスティックフィルタ】

メールの本文から個々の観点で特徴抽出をし，それを積み上げて spam メールかどうかを判定するフィルタである．前述のようなベイジアンフィルタとの融合型もあるが，他の統計手法を利用したものもいくつか出てきている．「spam メールの抽出に都合の良いヒューリスティックの選択方法」は「合わない奴は捨てる」か「変数変換して無理やり合わす」とかいろいろあるらしい．「統計的ではない」という理由でその辺の手法の定式化は放置されているような気もする．ちなみに前述の理由でどのような特徴を抽出しているかは隠蔽されたままであるが，たとえば，ベイジアンフィルタと連動して辞書に登録されていない文字列の出現を特徴の 1 つとして扱うヒューリスティックフィルタもあるかもしれない．アイディアはいくらでも出てくるがフィルタの性能はその取捨選択に大きく左右されるものと思われる．「ヒューリスティックの観点を外そうとすればするほど普通のメール文面から離れていくので検出が容易になる」という指摘もある．だが，各製品がどのような特徴抽出をしているかはプロプライエタリの壁に阻まれて具体的に知ることは困難である．

【パターンマッチフィルタ】

これもヒューリスティックの一部と捉えることが可能かもしれないが，統計手法を介さずに，単純にパターンにマッチしたものを叩き落とす／素通しするタイプのフィルタをパターンマッチフィルタという．正規表現を積み上げるタイプの製品が何種類か出ているが，問題はその正規表現を誰が書くのかである．サイト，ドメイン，ユーザの 3 層でそれぞれ設定が可能なフィルタも存在するが，ベースになるパターンをベンダが提供したとしても，結局はユーザや管理者が適切な正規表現を書けるかどうかに最終的な性能が依存する．

【協調型フィルタ】

spam メールのディジタルシグネチャ等の情報を多くのサイトで共有し，同じ文面，同じ種類のものが届い

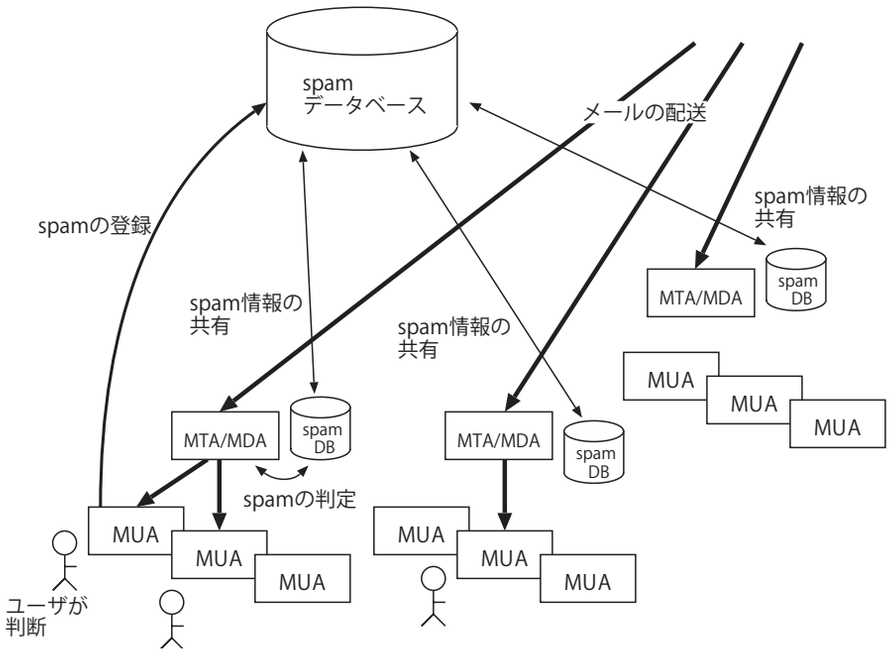

図 -2　協調型フィルタ

た際に spam メールを排除しようとするフィルタである（図 -2）．オープンソースの spam フィルタとして有名な SpamAssasin に含まれている Razor という仕組みなどがこれに該当する．個人レベルの spam メールの定義の違いを吸収することはできないが，同時に大量に配送される spam メールに対しては，ピックアップから情報の共有までが十分な速さで行われれば非常に有効な対策となる．最近ではヒューリスティックな解析結果を共有したり，ヒューリスティックの構成そのものを共有するタイプのものもあるらしい．かといって，個人の spam メールの定義を反映しているベイジアンフィルタの辞書を単純に共有するのはあまり良いアイディアではない．中には 135 万人超のユーザコミュニティから情報を収集する協調型フィルタ製品も存在するようである．

【自動確認付きホワイトリスト】

メールの送り主に対して「本当にメールを送りたいならもう 1 度送ってね」というメールを返し，それに返答した送り主だけを送信許可リストに登録し，次からは普通に送信を許可する仕組み．メーリングリストのアドレス登録確認の手続きに類似している．単純で効果は抜群だが，メールで本人確認をする外部のサービスなど，相手がどのアドレスからメールを送ってくるか分からない場合には何らかの救済措置が必要になる．

【Channeled Address】

あて先アドレスごとに専用の自分のアドレスを用意することで spammer によるアドレスの流通自体を無意味にする手法．別のアドレスからある自分のアドレスへの spam メールが来ても受け付けない．自分のメールアドレスに付加情報を埋め込むわけだが，米国では AT&T が特許を持っており，日本にはずばりそのものに対応する特許は見当たらず，かわりに暗号を埋め込むものについての特許がいくつか存在し，昨年 11 月から今年 3 月まで NTT が「privango」という名称で公開実験をしていたものがこれに相当する．わざわざ暗号を使って認証しなくても，あて先と自分のどのアドレスが対応するのかの対応表がきちんと管理されていればフィルタとして実装可能である．ホワイトリストの進化形と見ることもできる．

■ フィルタの進化

【ベイジアンフィルタと日本語】

英語対応で作られたベイジアンフィルタに日本語の文章を処理させると，漢字 1 文字を「単語」として扱うようになっているものも，そもそも扱えないものもあり，それらの場合には判定の精度はあまり期待できない．日本語の文章を単語に分解するには結局日本語文字列を単語に分離する仕組みと日本語の辞書が必要になる．実際に日本語の辞書を使って文章を単語に分解してベイジアンフィルタを構成している例もある．メールの文章をベイジアンフィルタで処理する場合はこの辞書の言語依存の問題のほかにも，MIME で base64 や

quoted-printable でエンコードされている文書のデコードや，HTML で書いてある場合には実体参照（entity）の復号などが前処理として必要になる．MIME multipart を使っている場合には再帰的に MIME パートの入れ子構造に従ってデコードが必要になる場合がある．さらには，MIME のメールの文章は ISO-2022-JP 以外の異なる文字コードで書かれている場合があるので，文字コードの変換が前処理に必要になることもあろう．問題はこれらの処理をきちんとやらないと，せっかく日本語の辞書を積んで正しく単語として扱えたとしても，ちっとも効果が上がらないことにある．世間で評価の高いベイジアンフィルタのソースコードを読んでみると，実はこれらの前処理がしっかりしていることが多い．ただし，手の込んだ処理は同時に負荷の原因にもなっている．

【ヒューリスティックフィルタとベイジアンフィルタの融合】

前述の「ナイーブな」ベイジアンフィルタは spammer の数々の妨害手段の前にすでに突破されてしまっている．だが，ベイジアンフィルタにはもう可能性がないのかと言えばそうではない．ベイジアンフィルタの扱う確率変数には離散的であること以外に特に大きな制限はない．ということは，ヒューリスティックで抽出したいろいろな特徴もベイジアンフィルタの内部で確率変数として扱うことが可能である．現在のベイジアンフィルタは確率変数として辞書の単語の出現傾向だけを扱うのではなく，ヒューリスティックを確率変数として含める拡張をもしてきていると思われる．だが，どこを見て判定しているのかを明かした途端にそこにノイズを入れられることが容易に推察できるためか，オープンソースのフィルタリングソフトウェアを除くと，多くの製品ではベイジアンフィルタであること以上の情報開示をしていない．特徴抽出の部分は隠蔽されたままである．

【MUA の実装】

いまどきの MUA は，ホワイトリスト機能もパターンマッチ機能も，ベイジアンフィルタも搭載している．実装の差は多少あるものの，ユーザの手元にある MUA にはフィルタ技術が集積されている．それもこれも，spam メールの定義が人によって違うことをカバーしようとした結果に見える．だが，扱うメールの数が多い場合には MUA の負荷がかなり気になる．筆者は 1 日 400 ～ 800 通のメールを受信しており，トラブルに備えて 3 種類の MUA 環境でそれを処理しているが，最近特に気になっているのはノート PC 上の MUA で，POP サーバからメールを取得する際に 1 通あたりほぼ 1 秒程度の時間を消費している．すでにフォルダに大量のメールが溜っていることもあるが，やはりウイルスフィルタの処理と MUA によるパターンマッチのフォルダ振り分け処理と spam 判定をやっている処理の重さが，ノート PC の非力な環境で顕在化しているように見える．このような環境はどこの組織にもありそうだが，休み明けの朝，仕事を始めようとしている時の大きなタイムロスに閉口してメールを処理するメインの環境を切り替えるに至っている．高機能なフィルタリング機能を備えた MUA の実用上の課題としてもうしばらく見守る予定である．

■ コンテンツフィルタのこれから

研究ベースでは単語ではなくマルコフ過程で生成した単語列に対して条件付き確率のネットワークを適用したマルコフィアンフィルタの spam メールに対する有効性も調べられているが，結果を見るといまひとつである．ヒューリスティックをベースにベイジアンではない統計手法を組み合わせた製品も出てきており，コンテンツフィルタはまだまだこの先どうなるかは見えないが，せっかくなのであえていくつか予想を立ててみよう．spam メール対策にもいろいろあるが，現状フィルタの機能はクライアント側に集中して実装されているように見える．サーバとクライアントの能力バランスは太陽の活動周期と同じくらいの長周期で振動しているように見えるので，今後はサーバ側に実装するフィルタ機能が少しずつ伸びてくることが予想される．もちろん，クライアント側の負荷の増大がこの動きを加速する可能性はある．ユーザ数の多いメールで顕在化した問題点は，よりユーザ数の少ないアプリケーションで次々に顕在化していくだろう．spam メールに対するコンテンツフィルタは対症療法であり，そこに本質的な問題解決を求めるのは間違いである．だが，その QOL（Quality Of Life）を向上させる効果は否定し難いものがある．spam メールが問題化しなかったら，ベイジアンフィルタがこれだけ有名になることもなかったかもしれない．実は筆者も十数年前には統計力学を少々かじっていた人間である．この分野からどのような面白い技術の応用が出てくるかはとても楽しみなことである．今後も spam メールがこれ以上ひどい状況にならないよう祈りながら見守っていきたい．

参考文献
1）Graham, P.: A Plan for Spam, http://www.paulgraham.com/spam.html (2002).
2）Oliver, J.: Using Lexigraphical Distancing to Block Spam, MIT SPAM Conference (2005).

（平成 17 年 6 月 16 日受付）