社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－CSEC－33 (9)
2006／5／12

中国語迷惑メールにおけるベイジアンフィルタの適用と評価

王 戦 † 堀 良彰 ‡ 櫻井 幸一 ‡

† 九州大学大学院システム情報科学府情報工学専攻 〒812-8581 福岡市東区箱崎 6-10-1
ou@itslab.csce.kyushu-u.ac.jp

‡ 九州大学大学院システム情報科学研究院 812-8581 福岡市東区箱崎 6-10-1
{hori, sakurai}@csce.kyushu-u.ac.jp

あらまし 迷惑メールに対する，ベイズ確率を用いた統計的なフィルタリング (いわゆるベイジアンフィルタ) の研究は以前から行われていたが，2002 年に発表された P. Graham の論文 " A plan for spam " [1] は人々の注目を集め，ベイジアンフィルタを実装したソフトウェアが多数開発されるようになった．

ベイジアンフィルタリング [4][5][8] に対する研究は，日本語と英語の電子メールについては盛んである．しかし，中国語の電子メールに対しては今まで学術的な解析が行われていなかった．そこで本論文では，中国語の電子メールを処理する際のベイジアンフィルタリングのパラメータと迷惑メール判定精度の関係について分析し，パラメータの最適値について考察した．

Application and evaluation of Bayesian filter for Chinese spam mail

Zhan Wang † Yoshiaki HORI ‡ Kouichi SAKURAI ‡

†The Department of Computer Science and Communication Engineering, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan,
ou@itslab.csce.kyushu-u.ac.jp

‡Faculty of Information Science and Electrical Engineering, Kyushu University 6-10-1 Hakozaki, Higashi-ku, Fukuoka, 812-8581 Japan
{hori,sakurai}@csce.kyushu-u.ac.jp

Abstract A statistical filtering based on Bayes theory, so-called bayesian filtering, has been researched for anti-spam through it before. From Graham's thesis in 2002, a lot of spam mail filters based on the Bayesian filtering have been developed and widely applied to the real system in recent years.

The implementation of the statistical filtering corresponding to the e-mail written in English and Japanese has already been developed. On the other hand, the implementation of the statistical filtering corresponding to e-mail written in Chinese is still few. In this thesis, we adopted a statistical filtering called as Bsfilter and modified it to filter out e-mails written in Chinese. When we targeted e-mails written in Chinese for experiment, we analyzed the relation between the parameter and the spam mail judgment accuracy of the filtering, and also considered the optimal value of the parameter.

1 はじめに

近年、インターネットの発展と電子メールの普及に伴い、迷惑メールが増加している。電子メールは、非常に安価に大量に送信できることから、悪質な業者が受信者の望まない広告メールを不特定多数に送信しているためである。迷惑メールに対する技術的対策として近年利用が増えているものの一つに，ベイジアンフィルタリングがある．これは電子メールをその内容から迷惑メールとそれ以外の正当な電子メールに分類する統計的フィルタリング [6] の一種で，過去の電子メールからヘッダや本文等に含まれる単語等をトークンとして抽出し，各トークンの出現確率は全て独立であるという仮定を設けた上でベイズの定理を使用して各トークンに迷惑メール確率を設定し，各トークンの出現という事象（結果）からその電子メールが迷惑メールである，または正当な電子メールであるという原因を推定する手法である．

本稿では、メール内容を利用した統計的手法であるベイジアンフィルタで、中国語を処理する場合の問題に注目する。ベイジアンフィルタリングに対す

る研究は，日本語と英語の電子メールについては盛んである．しかし，中国語の電子メールに対しては今まで学術的な解析が行われていなかった．そこで本論文では，中国語の電子メールを処理する際のベイジアンフィルタリングのパラメータと迷惑メール判定精度の関係について分析し，パラメータの最適値について考察した．

2 ベイジアンフィルタ

統計的なフィルタリングは導入の簡単さと精度が比較的良いという特徴から近年人気を集めている．統計的フィルタリングは，Naive Bayes, Support Vector Machine(SVM), Boosting, Markov Chain などの手法を用いて，過去に存在した正当な電子メールや迷惑メールから抽出した特徴と比較し，判定対象の電子メールが迷惑メールかどうかを判定するものである．このうち，特にナイーブベイズ (Naive Bayes) による手法はベイジアンフィルタリングと呼ばれ，その実装としての迷惑メールフィルタはベイジアンフィルタと呼ばれている．ベイジアンフィルタリングは，原理が単純であること，送信者の協力が不要であるなどの点で比較的導入が容易であり，各利用者が自らの基準で迷惑メールかどうかを指示できること等の特徴を持ち，近年数多くの実装が開発されている．

ベイジアンフィルタリングにおいては，過去の電子メールについての学習データをコーパス (corpus) と呼ばれるデータベースに格納する．コーパスには学習した電子メールに現れたトークンとそのトークンの正当な電子メール，迷惑メールのそれぞれにおける出現回数，学習した正当な電子メールおよび迷惑メールの数が格納される．トークンとは電子メールから抽出された要素のことで，英語などの言語では空白や記号によって単語間を区切るので，これらの言語で記述された電子メールに対するベイジアンフィルタリングでは単語をトークンとして使用する．

電子メールの判定時には，判定対象となる電子メールから抽出した各トークンについて，コーパスから取得した，過去の正当な電子メールおよび迷惑メールにおけるトークンの出現頻度についての情報から，そのトークンを含む電子メールが迷惑メールである確率（トークンの迷惑メール確率）を計算する．そして，抽出されたトークンの中にトークンの迷惑メール確率が 0.5 から遠く離れているトークンを選択する [1]。そして，選択したトークンの迷惑メール確率をもとに，判定対象となる電子メールの迷惑メール確率を計算する．そして，正当な電子メールと迷惑メールを区分する閾値 (threshold) が設定され，電子メールの迷惑メール確率が閾値を上回った場合に迷惑メールと分類し，閾値を下回った場合には正当な電子メールと分類する．

多くの実装では迷惑メール確率の計算式として，Graham によって提案された方式が用いられている．

Graham が用いた方式では，まずトークンの迷惑メール確率を

$$p(w) = \frac{b/n_{bad}}{bias \times g/n_{good} + b/n_{bad}}$$

- $g = 0, b > 0$ のとき $p(w) = 0.99$，$g > 0, b = 0$ のとき $p(w) = 0.01$，$g = b = 0$ のとき $p(w) = 0.4$
- b : 迷惑メールにおけるトークン w の出現回数
- g : 正当な電子メールにおけるトークン w の出現回数
- n_{bad} : 学習した迷惑メールの数
- n_{good} : 学習した正当な電子メールの数
- $bias$: 誤検出を減らすために設定される係数，Graham は 2 を採用

によって計算する．そして，電子メールの迷惑メール確率を

$$\frac{p(w_1)\cdots p(w_M)}{p(w_1)\cdots p(w_M) + (1-p(w_1))\cdots(1-p(w_M))}$$

- $w_1, w_2, \cdots$: 各トークンに対する迷惑メール確率を 0.5 から遠い順に並べたもの．最も遠いものを w_1 とする
- M : 判定に用いるトークンの数，Graham は 15 を採用

とする．電子メールの迷惑メール確率が閾値 t を上回った場合に迷惑メールと分類し，閾値を下回った場合には正当な電子メールと分類する．

3 中国語の対応

中国語の電子メールを受信する環境においては，英語の電子メールのみを想定して開発されたベイジアンフィルタを利用しても，迷惑メールの判別がうまく行えないことがある．このため，中国語の電子メールを受信する環境においては，中国語の電子メールに対応するための作業を行うか，中国語の電子メールに対応したベイジアンフィルタを使用する必要がある．

フィルタリングの開発を進める上で，ベイジアンフィルタをはじめとする統計的フィルタリングの重要性は研究者間では認識され，英語や日本語に対す

るベイジアンフィルタの理論を採用した spam 判定ツールは多数作成，そして公開または市販され，多くの研究者がその恩恵に預かっている．ところが，中国語に関しては状況が異なる．bsfilter [2] のように誰でもフリーにダウンロードできるツールは，筆者の調べるところではまだ公開されておらず，まだ十分にツールが整備されているとは言えない．

この原因のひとつは，中国語解析の困難性であると考える．ベイジアンフィルタリングでは各トークンに対してトークンの迷惑メール確率を計算する．英語の電子メールの場合，単語と単語の間は空白または記号で区切られているため，空白や記号を利用して機械的にトークンの抽出を行うことで，単語をそのままトークンとして用いることができる．それに対して中国語の電子メールの場合，単語と単語の間には空白や記号が含まれないため，そのままではトークンの抽出を行うことができない．同様の問題は日本語，韓国語をはじめとする他の言語においても発生すると考えられる．

中国語の電子メールにおけるトークンを抽出するには主に二つの方法 [7] が知られている．一つは形態素解析システムを使って「単語/の/区切り/が/明白」のように意味のある単語単位で分割する．もう一つは，n-gram と呼ばれる方法．2〜3 文字ずつをひとかたまりにして「単語」「語の」「の区」「区切」「切り」のように分割する．

形態素解析システム [9] を利用すると，検索ノイズを減らせる．分割された語が意味を持つ単語になっているので，部分的に文字が一致しているだけの意味のない語が検索対象外となるからである．長い単位で語を区切れるので，インデックスのサイズも小さくできる．半面辞書に登録されていない語は正しく区切れず，検索漏れが発生する．特に，中国語は文字種が単語分割のための大きな情報を持つ日本語とは異なり，ほぼ単一文字種 (漢字) である．さらに，複数品詞をもつ語が多いため品詞付与も容易ではない．

こうした問題は n-gram の手法では発生しないため，検索漏れは基本的に起こらない．その代わり，検索ノイズが増えるのがデメリットである．例えば「ことが」で検索すると「ことがら (事柄)」,「みことが (尊が)」のように見つかってしまう．　ここで，ベイジアンフィルタとして ruby を用いた実装のひとつである bsfilter を使用し，中国語の電子メールに対応できるように修正を加えて使用した．迷惑メール確率の計算式には Graham 方式を使用した．

修正したところは主に下記である

- 中国語の文字 (簡体字) を表す文字コード GB2312 を中国語 cn database に入れる
- 中国語のトークンの抽出は bigram を使用する
- cn のデータベースのコードは utf-8

この bigram は下記のような規則でトークンを抽出する

- 孤立した漢字はそのまま 1 つのトークンとして抽出する
- 連続する漢字については 1 文字目と 2 文字目，2 文字目と 3 文字目というように
  隣接する 2 字の漢字をそれぞれ一つのトークンとして抽出する
- 英単語については空白などで区切られた 1 単語を 1 つのトークンとして抽出する

たとえば，“我，是一个学生”(私は学生である) に対して，“我”，“是一”，“一个”，“个学”，“学生”のように分割される．

4　実験

4.1　実験対象

実験対象として，2005 年 5 月から 2006 年 1 月までの九ヶ月間に使用しているメールアドレスに到着した正当な電子メールおよび迷惑メール，および筆者が所有する四メールアドレスに到着した迷惑メールを使用した．また，統計的なフィルタリング手法では，分類に用いる学習データの量が少ないと十分な精度が出ないという特徴があるため，中国の反 spam 組織 [3] が提供する評価用のデータセットも使用した．今回は中国語メールのみを実験対象として，対象となった電子メールは 2158 通で、正当な電子メール 1012 通、迷惑メール 1146 通であった．こちらの電子メールは複数のメールアドレスから収集したため，Received，To など一部のメッセージヘッダについて，すべての電子メールが同一のメールアドレスに到着したものであるかのように整形を行うことで，電子メールの収集元が判定時の確率計算に用いられることのないようにした．

4.2　実験手順

実験の手順は以下の通りである．

(1) 正当な電子メール・迷惑メールを明示した電子メールを一定量ベイジアンフィルタに学習させる．(初期学習)

(2) 残りの電子メールを 1 通ずつベイジアンフィルタに判定させ，計算された電子メールの迷惑メール確率を記録する．

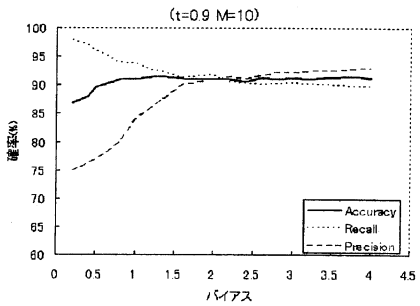
(3) 2 で記録した迷惑メール確率を集計し，2 章で説明した閾値 t，バイアス bias，判定に用いるトークンの数 M の三つのパラメータ [10] を変動させながら，再現率 (Recall. やってきた spam のうち spam と判定できた割合)，適合率 (Precision. spam と判定したうち本当に spam であるものの割合)，正確率 (Accuracy. spam と正当なメールをどれだけ正しく振り分けられたかというフィルタの性能を表す) を測定する．

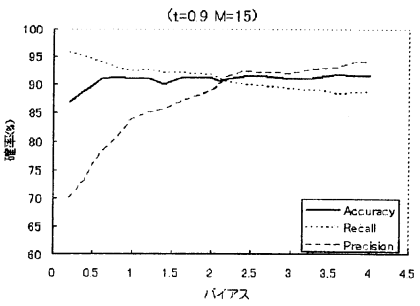
(4) 上記の実験を 15 回試行し，測定値の平均を実験結果とする．

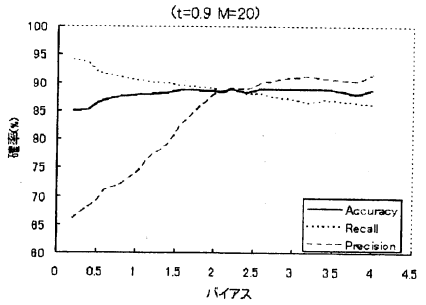
5 結果

5.1 バイアス bias を変動させた場合

各トークンの迷惑メール確率 p(w) を計算する際のバイアス bias を 0.2〜4.0 の間で変動させ，閾値 t を 0.9 とし、判定に用いるトークンの数 M=10，M=15，M=20，M=30 の四つの条件に対して実験を行った．その結果を図 1 に示す．







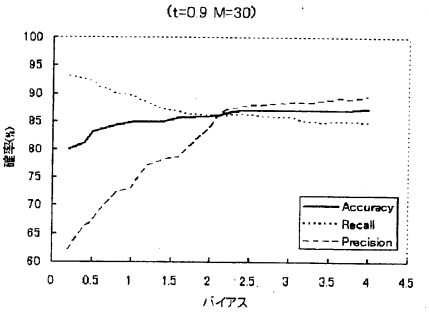


図 1: M と判定精度の関係

Graham のトークンの迷惑メール確率計算式からみると、バイアス bias を大きくすると適合率が高くなり，再現率が低くなっている．バイアス bias を小さくした場合はその逆となっている．

見逃しが発生した場合は誤検出と異なり，単に受信者に迷惑メールが表示されるだけであるから，受信者が手動で隔離または削除を行うだけでよい．見逃しは迷惑メール対策においてはそれほど致命的ではなく，極論すれば到着する迷惑メールのうち半分について見逃しが発生したとしても，全く迷惑メール対策をとらない場合に比べれば一定の範囲で有効である．しかし，あまり見逃しが増えると受信者が手動で処理を行う負担が大きいため，電子メールの利便性の為には，出来るだけ見逃しも少ないことが望ましい．

正確率のバイアス bias に対する分布についてだが，4 条件とも bias < 1 の範囲ではやや低く、1 < bias < 4 辺りで高くなっている．しかし適合率が bias < 1.5 の範囲では，小さすぎる．誤検出が起きると正当な電子メールが迷惑メールとして隔離され，あるいは破棄されてしまう．誤検出は見逃しに比べはるかに重大な判定ミスであることから適合率が小さすぎるとスパムフィルタとして使うことができない。また、1 < bias < 4 の範囲では，正確率には大きな

差はなく，適合率を優先的に高くしたいということを考慮すると bias=2.2 くらいが最適値であると思われる．

5.2 閾値 t を変動させた場合

バイアス bias=0.5, 1.0, 1.5, 2 の四つの条件に対して，閾値 t を 0 から 1 まで変化させて，適合率，再現率，正確率を計算した．ここでは，判定に用いるトークンの数 M は 15 とおく．測定した結果を図 2 に示す．

正当な電子メールと迷惑メールを区分する閾値 t を大きく取ると，検知される迷惑メールが減少し，逆に小さく取ればフィルタで遮断されるメールが増加する．つまり，t を大きく取ると再現率が減少し適合率は増加する。

Graham のトークンの迷惑メール確率計算式からみると，バイアス bias を大きく取ると単語全体の迷惑メール確率が低くなるので，メール全体の迷惑メール確率が低くなり，メールが迷惑メールとみなされにくくなるから、バイアス bias を大きく取ったほうが再現率が低くなり，適合率は高くなる．

正確率は，bias=2 のときは t=0 付近で，bias=1.5 のときは t=0.1 付近で，bias=1.0 のときは t=0.1, t=0.9 付近で，bias=0.5 のときは t=0 付近で最小となっているが，それ以外はほぼ一様に高くなっている．図から見ると、$0.1 < t < 0.9$ の範囲で正確率の変動幅はかなり小さいことがわかった．閾値 t は 0 と 1 付近を避けて取りさえすれば，正確率に大きい違いはないと思われる．しかし，再現率よりも適合率を優先的に高くしたいと考えるならば t は大きめに取ればよく，Graham の方式で t=0.9 と設定したのは妥当な判断であると思われる．

5.3 トークン数 M を変動させた場合

bias=1,2 の 2 条件に対し電子メールの迷惑メール確率を計算する時に用いるトークン数 M を 1〜100 の間で変動させ，閾値 t を 0.9 とし，実験を行った．その結果を図 3 に示す．

M に対する正確率の分布は 2 条件ともに同じ傾向を示しており，$M < 10$ くらいの範囲では急激に上昇していくが，M=20 前後で最大となり，それ以後は緩やかに低下していく．正確率が最大となるのは，bias=2 のときは M=20〜22 付近，bias=1 のときは M=12〜14 付近である．

迷惑メール判定に使われるトークンが少なすぎると，5,6 語の迷惑メール確率が高いトークンがメール中に存在するだけで，多くの正当な電子メールが

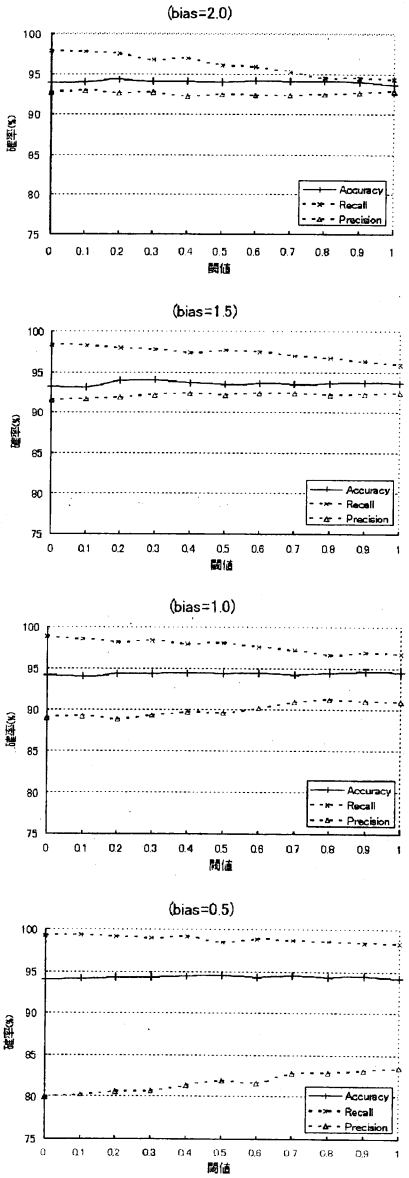


図 2: 閾値と判定精度の関係

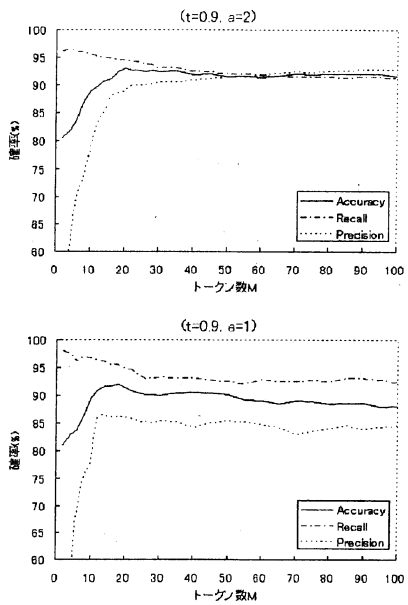


図 3: トークン数と判定精度の関係

迷惑メールと誤って判定されてしまう．したがって，Mを小さくしすぎてはいけない。しかし，あまりMを大きくし過ぎると，Mを大きくするにつれ迷惑メール判定に特徴的な単語の重みが薄れていき，正当なメール，迷惑メールのどちらにも出てくるようなまったく特徴的でない単語まで考慮するようになるのである．

bias=2の時，$M > 30$の範囲における正確率は緩やかな低下に現れている．bias=1の時，$M > 20$範囲にも正確率は緩やかな低下に現れている．

以上より，bias=2のときはM=25，bias=1のときはM=20くらいを取るのがMの最適値としてよいであろう．これは，正確率が安定して高い範囲であることと，優先的に高くしたい適合率の上昇傾向が収まってきた直後であることを基準として判断した結果である．Grahamの方式のM=15という設定では，迷惑メール確率を計算する際に使うトークン数としては少し少なすぎるという結果となった．

ここで考えられるのはMの最適値は言語の種類，抽出手法や受け取るメールに含まれる単語の平均数によって変化してくるのではないかということである．

6 おわりに

本論文では，中国語の電子メールに対応する統計的フィルタリング (ベイジアンフィルタリング) の実装が無い現状に対して，既存の日本語用の統計的フィルタリング bsfilter を使用し，中国語に対応するための修正を加え，ベイジアンフィルタリングで中国語を処理する場合の問題について実験・考察を行った．実験では，中国語電子メールを処理する際に，ベイジアンフィルタリングのパラメータの変動とともに，判定精度の各値が互いに関係を持ちながら変動した．これらの値から，誤検出の数を抑えながらより多くの迷惑メールを検出できる値をより良いと考えると，抽出したトークンのうち実際に迷惑メール判定に用いるトークン数は25前後が最適で，Grahamの方式のトークン数15という設定では少いという結果が得られた．

今後の課題としては，複数の言語が混在する環境においてベイジアンフィルタリングのパラメータの最適化について検討していきたい。また，ベイジアンフィルタを使用した場合，中国語の電子メールは英語電子メールより検出精度が低いという問題がある．今回実験用のベイジアンフィルタリングはもともと日本語に対応するための実装に修正を加えて使用した．より合理的に修正を行い，トークンの抽出手法の変更，またはホワイトリストなどの手法との併用などの方法を使うと，中国語の電子メールにより良い精度を期待できるだろうと考えられる．

参考文献

[1] P.Graham, A plan for spam

[2] bsfilter, http://bsfilter.org/

[3] CCERT Data Sets of Chinese Emails , http://www.ccert.edu.cn/spam/sa/datasets.htm

[4] G.Robinson, A statistical approach to the spam problem. Linux Journal,Vol.107, 2003

[5] P.Graham, Better bayesian filtering. 2003 Spam Conference

[6] Le Zhang,Jingbo Zhu,Tianshun Yao, An Evaluation of Statistical Spam Filtering Techniques, ACM Transactions on Asian Language Information Processing, Vol.3, No.4, pp. 243-269, Dec 2004

[7] Sun Maosong; Shen Dayang; Huang Changning, CSeg Tagl.0: A Practical Word Segmenter and POS Tagger for Chinese Texts, A97-1018, A Digital Archive of Research Papers in Computational Linguistics

[8] Johan Hovold, Naive Bayes Spam Filtering Using Word-Position-Based Attributes, Second Conference on Email and Anti-Spam ,CEAS 2005

[9] 岩永学, 田端利宏, 櫻井幸一, 迷惑メール対策におけるベイジアンフィルタ実装例の比較, 暗号と情報セキュリティシンポジウム 2004(SCIS2004),Vol.2, pp.1025-1028,(2004).

[10] 福泰樹, 松浦幹太, ベイジアンフィルタを用いた迷惑メールフィルタリングの最適化. 暗号と情報セキュリティシンポジウム 2005(SCIS2005),Vol.1, pp. 199 - 204,(2005).