The 40th Annual Conference of the Japanese Society for Artificial Intelligence, 2026

適応的スパムフィルタのための軽量な類似メッセージカウンタ

A Lightweight Similar Message Counter for Adaptive Spam Filtering

三宅 悠介 *1　　栗林 健太郎 *1
Yusuke Miyake　　Kentaro Kuribayashi

*1GMO ペパボ株式会社 ペパボ研究所
Pepabo R&D Institute, GMO Pepabo, Inc.

機械学習モデルによるスパム検知は高精度な判定が可能である一方，新パターンの検知から再学習・反映までの期間，システムは未知のスパムに対して脆弱となる．本研究では，これを補う，頻度変化に応じて判定を調整する適応的スパムフィルタのための軽量な類似メッセージカウンタを提案する．テキストと画像に対する軽量なハッシュ手法と時間窓を用いた近似頻度計算により，マルチモーダルなメッセージに低遅延・省メモリで逐次対応する．スパムキャンペーンのシミュレーション評価の結果，平均 3〜4 件の検出遅延でスパムの亜種を検出可能であり，処理時間は 1 件あたり 36〜42 μ s，メモリ使用量は 4〜7MB で実現できることを示した．

1. はじめに

スパムメッセージは増加の一途を辿っている [Group 25]．その背景として，AI の悪用による言語的・費用的障壁の低下が指摘されている．実際に，生成 AI を活用した巧妙なメッセージの出現が報告されており [Bethany 24]，攻撃手法は高度化している．

このような攻撃の巧妙化に対抗するため，情報システムにおけるスパム検知には機械学習モデルの活用が不可欠となっている．防御側でも AI を活用したアプローチ [OpenAI 25] が提案されており，高度な判定が期待できる．しかし，情報システムの実運用においては，処理時間の増大によるメッセージ遅延や，GPU インスタンスの稼働コストが課題となる．特に，スパムキャンペーンに見られるようなアカウントを跨いだ大量の類似メッセージの送信に対しては，この課題が顕著になる．そこで，IoT 分野でのスパムメール対策 [Ahmed 22] に見られるような軽量かつ低コストな方式が，情報システムの運用において有用な選択肢となる．ただし，軽量な機械学習モデルを採用した場合でも，新たなスパムパターンを検知してから，データ収集，再学習，本番環境への反映といった一連の運用作業には一定の時間を要する．この時間差の間，システムは未知のスパムに対して脆弱な状態となる．

この問題に対し，システムを通過するメッセージの類似性と出現頻度に着目したアプローチが研究されている [Coskun 12, Ali 15]．しかしながら，既存の類似メッセージ検出手法には実運用上の課題が残る．具体的には，テキストと画像が混在するマルチモーダルなメッセージへの対応，そして検出処理自体がシステムのボトルネックとならないための低レイテンシ・低コスト化が求められる．

本研究では，メッセージ機能を有する情報システムの運用維持において，これらの実運用上の要件を満たす軽量な類似メッセージカウンタを提案し，類似メッセージの頻度変化に応じて判定を調整する適応的なスパムフィルタリングを実現する．提案手法の有効性を，スパムキャンペーンのシミュレーション実験による検出性能とパフォーマンスの評価を通じて検証する．

連絡先: Pepabo R&D Institute, GMO Pepabo, Inc., Tenjin, Chuo ku, Fukuoka 810-0001 Japan, E-mail: miyakey@pepabo.com

2. 関連研究

本節では，新パターンへ対応するまでの遅延による損失を緩和するための，新メッセージの判定と頻度計算のアプローチに関する既存研究を通して，提案手法に求められる要件を整理する．機械学習モデルに与える初出のパターンのメッセージを迅速に選択する方式 [Ali 15] では，スパムメッセージの類似性を局所感度ハッシュ法（Locality Sensitive Hashing: LSH）を用いて迅速に判定している．既知のスパムメールとの類似性による判定を行う研究 [Ho 14] でも，軽量さが求められる LSH の一種である SimHash をフィンガープリントとして利用している．しかしながら，これらの方式では，パターンがスパムかどうかを予め判断できるモデル等が仮定されている．

新パターンが常にスパムとは限らない状況では，システムを通過するメッセージの類似性と出現頻度に着目するアプローチも報告されている [Coskun 12]．この方式では，メッセージ本文の n-gram 出現回数を近似的に計算する方式を導入しているが，各 n-gram ごとに複数の bin 参照と閾値判定が必要であり，また画像への適用が困難である．スパムメッセージは画像内にも混入している [Zhang 23]．これらのパターンにおいても同様にカウントできるのが望ましいが，上述の研究は，テキストを想定しているため，画像データに対して類似メッセージ判定の頑健性は期待できない．このような非構造化データの意味的特徴を保持しながら表現する埋め込みモデル [Cheng 23] は，上述の軽量な方式と比べて処理時間や運用コストが大きい．

以上の関連研究の検討から，実運用に適した類似メッセージカウンタに求められる要件を整理する．第一に，テキストと画像の両方に対して，計算コストを抑えた軽量な類似性判定が必要である．第二に，メッセージの到着に応じて逐次的に処理できる必要がある．第三に，長期間の運用においてメッセージパターンの種類が増加しても，メモリ使用量が際限なく増大しない仕組みが求められる．次節では，これらの要件を満たす軽量な類似メッセージカウンタを提案する．

3. 提案手法

本節では，前節で整理した要件を満たす軽量な類似メッセージカウンタを提案する．提案手法は，(1) 逐次的な類似性の判定，(2) 頻度変化に応じた判定の調整，(3) 時間経過への対処，の 3 つの構成要素からなる．図 1 に，提案手法を含むスパム

判定の全体フローを示す．メッセージが到着すると，既存の機械学習モデルによるスパムスコアの算出と並行して，提案手法による類似メッセージの頻度計算が行われる．テキストは TF-IDF によるベクトル化の後，LSH によりハッシュ値に変換される．画像は dHash により直接ハッシュ値に変換される．得られたハッシュ値は，時間窓を用いた近似頻度計算手法である Sliding Window Count-Min Sketch（SW-CMS）に入力され，類似メッセージの出現頻度が推定される．最終的に，機械学習モデルによるスパムスコアと頻度カウントを統合し，スパム判定を行う．以下，各構成要素について詳述する．

3.1 逐次的な類似性の判定

テキストと画像それぞれに対して，軽量かつ逐次処理可能なハッシュ手法を採用する．

テキストメッセージに対しては，ランダム超平面に基づく局所感度ハッシュ法 [Charikar 02] を用いる．メッセージを TF-IDF により語彙集合 V に基づくベクトル $\mathbf{x} \in \mathbb{R}^{|V|}$ に変換する．n 個のランダム超平面 $\mathbf{h}_i \sim \mathcal{N}(0, I_{|V|})$（$I_{|V|}$ は $|V|$ 次元の単位行列）を用いて，各ビットを以下のように生成する．

$$b_i = \begin{cases} 1 & \text{if } \mathbf{x} \cdot \mathbf{h}_i > 0 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

n 個の超平面により n ビットのハッシュ値を得る．同一のハッシュ値を持つメッセージを類似メッセージとみなすことで，部分的な変更を含むメッセージの同定が可能となる．

画像に対しては，差分ハッシュ（dHash）[Hamadouche 21] を用いる．画像を $(s+1) \times s$ のグレースケールにリサイズし，隣接ピクセルの輝度比較により各ビットを生成する．

$$b_{i,j} = \begin{cases} 1 & \text{if } p_{i,j+1} > p_{i,j} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

これにより s^2 ビットのハッシュ値を得る．dHash は画像のリサイズと単純な比較演算のみで計算でき，深層学習ベースの埋め込みと比較して極めて軽量である．テキストと同様に，同一のハッシュ値を持つ画像を類似画像とみなすことで，部分的な編集を含む画像の判定が可能となる．

3.2 頻度変化に応じた判定の調整

類似メッセージの出現頻度を効率的に計算するため，Count-Min Sketch[Dalloo 24] を採用する．Count-Min Sketch は，d 個の独立なハッシュ関数と幅 w のカウンタテーブル $C \in \mathbb{R}^{d \times w}$ から構成される確率的データ構造である．ハッシュ値 k を持つメッセージが到着した際の加算処理は以下のように行う．

$$\text{Add}(k) : C[i, h_i(k)] \leftarrow C[i, h_i(k)] + 1 \quad \forall i \in \{1, \ldots, d\} \tag{3}$$

頻度の推定は，各ハッシュ関数に対応するカウンタ値の最小値を取る．

$$\text{Est}(k) = \min_{i \in \{1,\ldots,d\}} C[i, h_i(k)] \tag{4}$$

最小値を採用することで，ハッシュ衝突による過大評価を抑制する．この手法により，固定サイズのメモリで無限のパターン数に対応でき，メッセージの到着ごとに定数時間で加算・推定が可能である．この頻度推定により，あるハッシュ値を持つメッセージが過去に一定数以上出現していたかを即座に判定できる．スパムキャンペーンでは類似メッセージが短期間に大量送信される特性があるため，頻度変化に応じて判定を調整することで，未知のスパムパターンに対しても適応的な対応が可能となる．

3.3 時間経過への対処

長期間の運用において古いパターンの影響を排除するため，SW-CMS[Papapetrou 12] を導入する．原論文では指数ヒストグラムを用いた実装が提案されているが，本研究では実装容易性を考慮し，N 個の Count-Min Sketch をリング状に管理する簡易版を採用する．各ウィンドウは W 回の観測で次のウィンドウに切り替える．

頻度の推定時には，全ウィンドウのカウントを合算する．最も古いウィンドウは新しい観測により上書きされるため，メモリ使用量は $N \times d \times w$ で一定に保たれる．これにより，時間経過とともに出現頻度が減少したパターンは自然に忘却され，メモリ効率を維持しながら最近のスパムキャンペーンの検出が可能となる．

4. 評価

本節では，提案手法の有効性を検証するため，スパムキャンペーンのシミュレーション環境における評価を行う．評価の目的は，(1) 類似性判定と頻度計算によるスパムキャンペーン検出性能の検証，(2) 処理時間とメモリ使用量の実用性の確認である．

4.1 評価方法

未知のスパムキャンペーンでは，未知の語彙を含むメッセージや，フィルタ回避のための微小な変更を加えた亜種が大量に送信される．本評価では，これらを模擬したデータセットで検出性能を検証する．

テキストの評価には SMS Spam Collection Dataset[Almeida 11] を使用した．データセットを学習用 95%と評価用 5%に分割した．TF-IDF の語彙構築には学習用データのみを使用し，評価用データに含まれる語彙は提案手法にとって未知となる．評価用データから 10 件のスパムメッセージを原型として抽出し，各原型に対して単語の追加・削除・置換を各 1〜3 単語で行うことで 9 種類の亜種を作成した．これは，実際のスパムで観測される不可視文字挿入や絵文字変更などの微小な変更を模擬している．未知スパムを模すため，追加と置換先の語彙は評価用データの語彙を使用した．

画像の評価には CIFAR-10[Krizhevsky 09] を使用した．CIFAR-10 は 32×32 ピクセルの 10 クラス分類用の小規模画像データセットであり，シミュレーションの迅速な実施のために採用した．dHash は画像を固定サイズにリサイズしてからハッシュ化するため，本評価結果は実運用環境にも適用可能である．テキストと同様に 10 件の画像を原型として抽出し，各原型に 9 種類の編集を適用して亜種を作成した．実際の画像スパムでは QR コードの配置やサイズの変更など局所的な編集が行われるため，これを模擬するノイズを用いた．編集パターンには Gaussian ノイズと Salt/Pepper ノイズを用いた．Gaussian ノイズは σ=0.001, 0.002, 0.004 の 3 段階で画像全体に正規分布ノイズを加算する．Salt/Pepper ノイズは修正割合 0.1%, 0.2%, 0.4%の 3 段階でランダム位置のピクセルを白または黒に置換する．図 2 に 0.4%の編集例を示す．

シミュレーションでは，学習用・評価用データ，原型および亜種（計 100 件）を混合し，無作為に並び替えたメッセージ列を用いた．メッセージ列を 1 件ずつ提案手法で処理し，頻度推定値が閾値 1 以上のハッシュ値が再度確認された時点で検出とした．評価指標には平均検出遅延と偽陽性率を用いた．平均検出遅延は，各原型について最初の亜種が出現してから，その原型に属するいずれかの亜種が初めて検出されるまでに処理されたメッセージ数の平均である．偽陽性率は，学習用データ

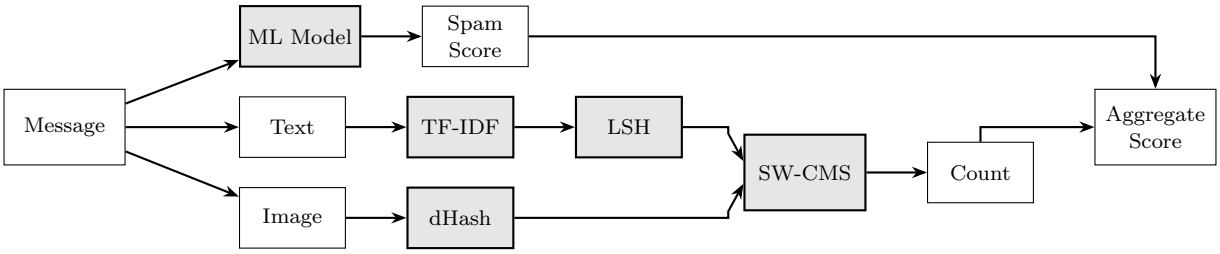

図 1: 提案手法の処理フロー

図 2: ノイズ付与による画像編集の例

表 1: 編集パターン別ハッシュマッチ率

テキスト（LSH）		画像（dHash）	
編集パターン	マッチ率	編集パターン	マッチ率
1 単語追加	86%	Salt 0.1%	69%
2 単語追加	85%	Salt 0.2%	51%
3 単語追加	73%	Salt 0.4%	22%
1 単語削除	52%	Pepper 0.1%	74%
2 単語削除	21%	Pepper 0.2%	56%
3 単語削除	11%	Pepper 0.4%	23%
1 単語置換	52%	Gaussian 0.1%	42%
2 単語置換	33%	Gaussian 0.2%	46%
3 単語置換	13%	Gaussian 0.4%	43%

と評価用データのうち，過去に出現した別のメッセージとハッシュ衝突を起こし誤って検出された割合であり，値が小さいほど良い検出性能を意味する．

パラメータ選定について述べる．LSH のビット数 n は類似判定の粒度と衝突率のトレードオフがあるため，F1 スコアが最大となる $n \in \{8, 16, 32, 64\}$ から $n = 32$ を選定した．SW-CMS パラメータは共通で $d = 64$，$w = 8192$，$N = 3$，$W = 2000$ とした．画像に対しては dHash サイズ $s = 8$ を用いた．シミュレーションは 10 回実施し，平均結果を報告する．

4.2 スパムキャンペーン検出性能

テキストおよび画像に対するスパムキャンペーン検出性能を報告する．平均検出遅延は，テキストで 3.39±0.40 件，画像で 3.58±0.66 件となり，いずれも数件レベルでの迅速な検出が可能であることが示された．偽陽性率は，画像で 0%，テキストで 3.61±0.06%となった．テキストの偽陽性率が 0 でない理由は，TF-IDF ベクトル化において語彙が限定されるため，異なるメッセージ間でハッシュ衝突が発生しやすいためである．特に，本評価で用いた SMS データセットのようにメッセージ長が短い場合，単語数が少ないため TF-IDF ベクトルが同一になりやすく，この傾向が顕著になる．したがって，実運用においては頻度の閾値を調整することで，偽陽性率と検出遅延のトレードオフを制御することが望ましい．

次に，スパムキャンペーンの亜種検出に重要なハッシュの頑健性を評価する．ハッシュマッチ率は，元のコンテンツに編集を加えた際に同一のハッシュ値が得られる割合であり，この値が高いほど亜種の検出が容易となる．表 1 に編集パターン別のハッシュマッチ率を示す．テキストでは，単語追加では 73〜86%の高いマッチ率を示した一方，単語削除・置換では操作数の増加に伴いマッチ率が低下した．TF-IDF ベクトルにおいて未知語彙の追加はベクトルに影響を与えないため，未知語彙による亜種に対して提案手法は有効である．一方，既知語彙の削除・置換はベクトルを変化させるためマッチ率が低下するが，これらは既存フィルタが学習済みのパターンであり，機械学習モデルによる判定が期待できる．

LSH による類似性判定の特性をより詳細に理解するため，埋め込みモデル [Vera 25] を用いた追加実験を行った．埋め込みベクトルに対して LSH を適用した結果，平均コサイン類似度 95.67%と高い意味的類似性を示すペアにおいても，LSH マッチ率は 20%に留まった．この結果から，意味的な類似性を捉える埋め込み表現においても微小な変化が LSH ハッシュ値の不一致を招くことが示され，類似性判定の精度低下の主因が LSH の離散化特性にあることが示唆された．

画像では，Salt/Pepper ノイズではノイズ率の増加に伴いマッチ率が低下した．Gaussian ノイズでは全体的にマッチ率は低く，ノイズ率による変化は小さかった．dHash は隣接ピクセル間の輝度の大小関係に基づくため，Salt/Pepper のような局所的で極端なピクセル変更は影響が限定的である一方，Gaussian のような全体的なノイズは多くの箇所で大小関係を変化させうる．QR コード変更のような局所的な改変に対しては，dHash の特性上，検出に有利である．

以上の評価から，提案手法は類似メッセージの頻度変化に応じて判定を調整することで，未知のスパムキャンペーンを数件レベルで迅速に検出でき，機械学習モデルの更新が間に合わない間も適応的にスパムに対応可能であることが示された．

表 2: パフォーマンス評価結果（10,000 件）

処理ステップ	テキスト	画像
ベクトル化/ハッシュ化	74.93 ms	182.93 ms
LSH 変換	35.82 ms	-
SW-CMS 操作	252.90 ms	238.36 ms
合計	363.64 ms	421.30 ms
1 件あたり	36 μs	42 μs
メモリ使用量（理論値）	6.93 MB	4.00 MB

4.3 パフォーマンス評価

4.1 節のパラメータ設定を用いて，パフォーマンス評価を行った．テキストおよび画像それぞれ 10,000 件に対して，頻度計算と推定を実施し，処理時間を計測した．計測は 10 回実施し，平均値を算出した．実行環境は Apple M3 Pro，メモリ 36GB，Python 3.14.2 である．

表 2 に結果を示す．処理時間についてはテキスト，画像ともに 1 件あたり 36〜43μs と高速であった．Web サービスにおいて望ましいレスポンス時間は 100ms 以内とされており [Google 20]，提案手法による処理遅延の影響は軽微である．なお，SW-CMS の処理時間はウィンドウ数 N にも依存する．これは頻度推定時に N 個のウィンドウに対してカウンタの参照が発生するためである．

メモリ使用量は LSH 超平面と SW-CMS のカウンタから算出した理論値である．テキストでは約 6.93MB，画像では SW-CMS のみで 4.00MB となる．いずれもメッセージ件数に依存しない固定サイズであり，逐次処理において蓄積によるメモリ増加が発生しない点が本方式の利点である．

これらの結果から，提案手法は既存のスパムフィルタに追加的な機構として組み込む際に，処理遅延や運用コストへの影響を最小限に抑えられることが示された．

5. おわりに

本報告では，機械学習モデルの再学習と反映までの時間差を緩和するため，適応的スパムフィルタのための軽量な類似メッセージカウンタを提案した．提案手法は，テキストと画像に対する軽量なハッシュ手法と時間窓を用いた近似頻度計算を組み合わせることで，マルチモーダルなメッセージに低遅延・省メモリで逐次対応する．

評価では，テキスト・画像ともに平均 3〜4 件の検出遅延でスパムキャンペーンの亜種を検出可能であり，処理時間とメモリ使用量も実用的な水準であることが示された．一方で，テキストでは既知語彙の削除・置換，画像では全体的なノイズのように，ハッシュ値への影響が大きい編集に対してはマッチ率が低下するという制約も明らかになった．この制約は，軽量なハッシュ手法における分離性能と類似性検出のトレードオフに起因する．そのため，今後は，軽量性を維持しつつ，多様な編集パターンに対してもより頑健な類似性判定を実現するハッシュ手法について研究を進める．

参考文献

[Ahmed 22] Ahmed, N., Amin, R., Aldabbas, H., Koundal, D., Alouffi, B., and Shah, T.: Machine learning techniques for spam detection in email and IoT platforms: analysis and research challenges, *Security and Communication Networks*, Vol. 2022, No. 1, p. 1862888 (2022)

[Ali 15] Ali, S.-H.-A., Ozawa, S., Nakazato, J., Ban, T., and Shimamura, J.: An online malicious spam email detection system using resource allocating network with locality sensitive hashing, *Journal of intelligent learning systems and applications*, Vol. 7, No. 02, pp. 42–57 (2015)

[Almeida 11] Almeida, T. and Hidalgo, J.: SMS Spam Collection, UCI Machine Learning Repository (2011), DOI: https://doi.org/10.24432/C5CC84

[Bethany 24] Bethany, M., Galiopoulos, A., Bethany, E., Karkevandi, M. B., Vishwamitra, N., and Najafirad, P.: Large language model lateral spear phishing: A comparative study in large-scale organizational settings, *arXiv preprint arXiv:2401.09727* (2024)

[Charikar 02] Charikar, M. S.: Similarity estimation techniques from rounding algorithms, in *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pp. 380–388 (2002)

[Cheng 23] Cheng, D. Z., Wang, R., Kang, W.-C., Coleman, B., Zhang, Y., Ni, J., Valverde, J., Hong, L., and Chi, E.: Efficient Data Representation Learning in Google-scale Systems, in *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 267–271 (2023)

[Coskun 12] Coskun, B. and Giura, P.: Mitigating sms spam by online detection of repetitive near-duplicate messages, in *2012 IEEE International Conference on Communications (ICC)*, pp. 999–1004IEEE (2012)

[Dalloo 24] Dalloo, A. M., Humaidi, A. J., Al Mhdawi, A. K., and Al-Raweshidy, H.: Approximate computing: Concepts, architectures, challenges, applications, and future directions, *IEEE access*, Vol. 12, pp. 146022–146088 (2024)

[Google 20] Google, : Measure performance with the RAIL model — web.dev (2020)

[Group 25] Group, A.-P. W.: Phishing Attack Trends Report – 2Q 2025 (2025)

[Hamadouche 21] Hamadouche, M., Zebbiche, K., Guerroumi, M., Tebbi, H., and Zafoune, Y.: A comparative study of perceptual hashing algorithms: Application on fingerprint images, The 2nd International Conference on Computer Science ' s Complex Systems and their Applications (2021)

[Ho 14] Ho, P.-T., Kim, H.-S., and Kim, S.-R.: Application of sim-hash algorithm and big data analysis in spam email detection system, in *Proceedings of the 2014 Conference on Research in Adaptive and Convergent Systems*, pp. 242–246 (2014)

[Krizhevsky 09] Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)

[OpenAI 25] OpenAI, : Introducing gpt-oss-safeguard — OpenAI (2025)

[Papapetrou 12] Papapetrou, O., Garofalakis, M., and Deligiannakis, A.: Sketch-based querying of distributed sliding-window data streams, *Proc. VLDB Endow.*, Vol. 5, No. 10, p. 992–1003 (2012)

[Vera 25] Vera, H. S., Dua, S., Zhang, B., Salz, D., Mullins, R., Panyam, S. R., Smoot, S., Naim, I., Zou, J., Chen, F., et al.: Embeddinggemma: Powerful and lightweight text representations, *arXiv preprint arXiv:2509.20354* (2025)

[Zhang 23] Zhang, Z., Damiani, E., Hamadi, H., Yeun, C., and Taher, F.: A late multi-modal fusion model for detecting hybrid spam e-mail, in *International Journal of Computer Theory and Engineering*, Vol. 15, pp. 76–81, IACSIT Press (2023)