文章解析を活用したメールにおける情報支援

山口　重也†　諏訪　敬祐†

†武蔵工業大学大学院　環境情報学研究科

1. はじめに

電子メールはパーソナルコンピュータからモバイル機器まで、多くの情報端末から利用されるコミュニケーション手段である。その利用は私的な場面から会社などの重要な場面まで幅広い状況で利用されている。しかし、電子メールには悪質な迷惑メール(以下スパム)が氾濫しており、社会問題になっている。そのため、多くの電子メール受信システムではフィルタリング手法によるスパムの自動分類と、それを基にしたスパムの非表示化や自動削除などの対策[1]が行われている。

しかしながら、スパムメールのフィルタリングのみでは、図 1 で示すように、スパムではないものの重要度が低い電子メールやフィルタを回避してきたスパムメールなどにより重要なメールが画面外へ隠されてしまい、情報量が低下してしまう問題がある。

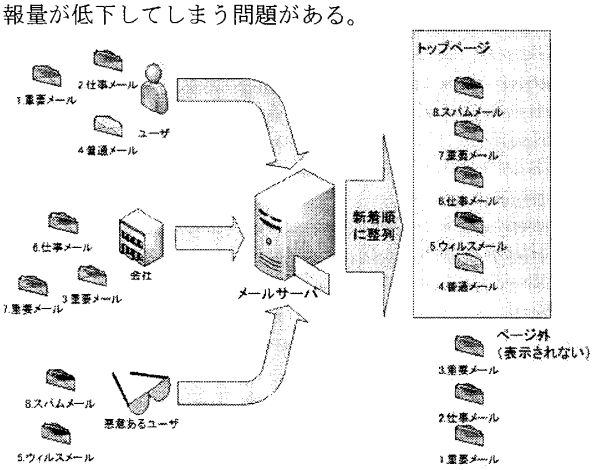

図 1　新着リスト方式におけるメールの流れ

そこで本研究では、フィルタリング手法だけではなく、情報推薦の技術[2]にも着目した。メールに記載されている文章を解析し、重要なメールをメール閲覧画面の上位に推薦することにより、従来のメール閲覧画面に比べて情報量が多いメール閲覧画面の提案を行う。これにより、ユーザのメール閲覧行動の情報支援を実現する。

2. 提案する情報支援システム

2.1 提案するシステムの概要

提案するメール閲覧システムにおけるメールの流れを図 2 に示す。本システムでメールサーバに届いたメールを呼び出すと同時に、その件名や本文を分析することにより「重要」、「仕事」、「私事」、「普通」、「スパム」の 5 種類へ自動的に分類を行う。この分類に従ってメールの閲覧画面を構築することにより、メールの内容に応じた場所への配置と装飾を行う。これにより、メール閲覧画面の情報量を向上させ、ユーザのメール閲覧行動を支援する。

Information Support by Analyzed E-mail Text
† Shigeya YAMAGUCHI, Keisuke SUWA
Graduate School of Environmental and Information Studies, Musashi Institute of Technology

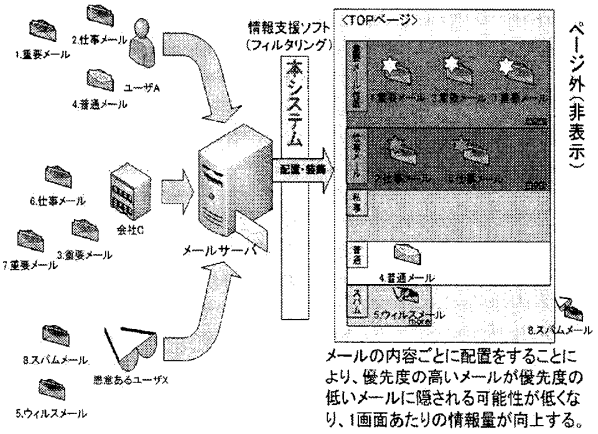

図 2　提案方式におけるメールの流れ

2.2 システム構成

作成したシステムの構成を表 1 に示す。本システムはメールを受信するための外部のメールサーバと、メールサーバにアクセスしてメールを取得し表示を行う本システム(ローカルサーバ) の 2 つで構成されている。

表 1　システム構成

メールサーバ	外部のメールサーバ
本システム (ローカルサーバ)	Apache2.2, PHP5, PostgreSQL8 を利用したメール受信システム
開発言語	HTML,PHP5

3 処理の前提

3.1 メールの分類

本研究ではメールを次の 5 種類に分類を行った。

表 2　メールの分類区分

分類	分類基準
重要	緊急連絡、要返事案件、請求書、登録通知など
仕事	会社、学校、研究関係などのメール
私事	メールマガジン、家族・友人からなどのメール
普通	各分類に分類できなかったメール
スパム	迷惑メール、ウィルス、フィッシング

3.2 フィルタ

上記のメール分類を自動的に行うため、PHP5 の正規表現による文字列マッチング機能を利用した。メールに記載されている件名と本文をマッチング処理し、各分類に含まれやすいと考えられる単語を識別した。そして、メールの重要度や仕事度などを単語に応じてポイント処理を行いメールの分類傾向を計算し、メールの自動分類処理を行った。

3.3 画面デザイン

メールの分類を基に、従来手法より情報量が多くなるメール閲覧画面の設計を行った。上部から順に、重要と判断されたものは最上部に 7 件を赤い背景で強調して表

示する。仕事と判断されたものは 5 件を濃緑色の背景で、私事と判断されたものは 3 件を緑色背景で、普通と判断されたものは強調なしで 3 件を表示、スパムと判断されたものは最下部に 2 件のみを灰色背景で表示をするように設計した。この表示方式を提案方式と呼ぶ。

3.4 情報量の定義

本研究において、情報量とは次のように定義した。メールが閲覧に表示されているときの、分類に応じた点数。この点数は下表に従って点数化した。なお、表示されていない場合や分類ミスが発生した場合、情報量のロスが発生したものと考え減点がされる。

表 3 ポイント表

配置先/分類	重要	仕事	私事	普通	スパム
重要配置	11	-4	-6	-7	-12
仕事配置	-4	7	-2	-4	-8
私事配置	-6	-2	5	-2	-6
普通配置	-7	-4	-2	3	-4
スパム配置	-12	-8	-6	-4	1
非表示	-11	-7	-5	-3	-1

4 実証実験

実際に各分類 10 件、総計 50 件のサンプルメールをメールサーバへランダムに送信し、従来の新着リスト表示方式と提案画面における情報量の変化を計測する実験を 5 回行った。その表示結果を図 3 と 4 に示す。なお、各画面の表示件数は 20 件である。

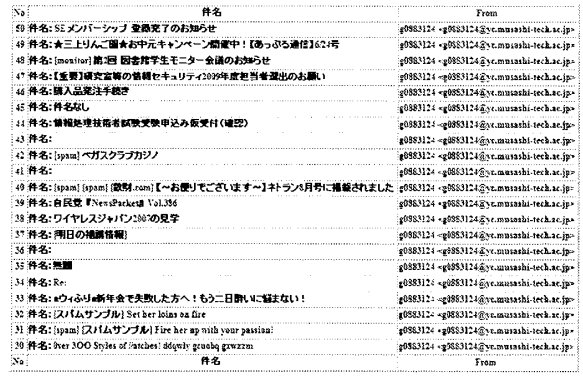

図 3 リスト方式の閲覧画面

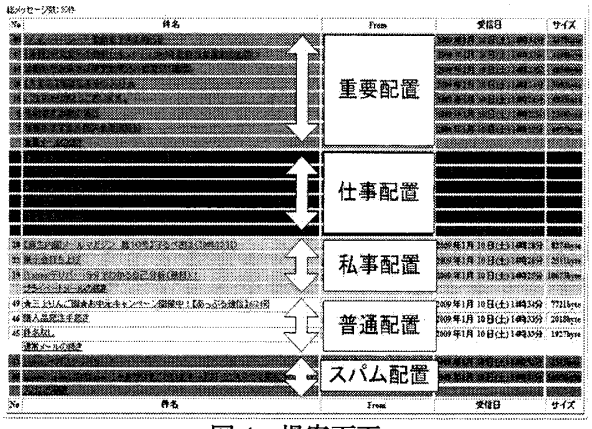

図 4 提案画面

表 3 の基準に従い、各画面の情報量を集計した結果を表 4 に示す。従来のリスト方式の画面における情報量は平均で-50.8 であったが、提案画面における情報量は平均で -16.4 であった。情報量の平均は提案手法の方が 34.4 上回る結果となった。

表 4 情報量の評価結果

実験回	1 回目	2 回目	3 回目	4 回目	5 回目	平均
リスト方式	-40	-66	-82	-24	-42	-50.8
提案方式	-12	-4	-17	-24	-25	-16.4
点数変化	+28	+62	+65	+0	+17	+34.4

5. 考察

実証実験の結果から、メールの自動分類が正常に行われ所定の位置へ推薦が行われたため、提案方式の方が従来方式に比べて 1 画面あたりの平均情報量が多くなることが分かった。従来のリスト方式では、メールの到達順序によっては多くの重要なメールが非表示になってしまったり、重要ではないメールが画面の大部分を占めたり、点数の変動が大きかった。

一方、提案方式ではフィルタによるメールの分類に基づいた表示が行われたため、画面あたりの情報量がリスト方式より高い値で安定したと考えられる。フィルタの自動分類の精度が 70%の正答率であったため、分類区分を間違えたメールが誤った場所に多く表示された場合でもリスト方式に比べて数値変化の幅が小さい値であった。

これらのことから、文章解析を活用したメールの情報推薦により従来のリスト表示より情報量の多いメール閲覧画面の構築をすることができたと考えられる。

6. おわりに

今回の実験で、フィルタリングによるメールの自動分類とそれを基にした情報推薦を行うことにより、従来の新着リスト方式より安定して高い情報量を持つメール閲覧画面の提案を行うことができた。従来のようにスパムメールだけではなく、その他のメールをフィルタリングし適切な配置へ表示を行うことにより 1 画面あたりの情報量が向上することを示せた。

今後の課題としては、実験の試行回数が 5 回と少なかったため乱数が偏っており、重要なメールの多くが最初に送信されるなど提案方式に有利な送信順序になった可能性がある。そのほかにも、送信したメールサンプルがフィルタにとって分類しやすい形式であったためフィルタリングが高精度で行われ、結果として提案手法が高評価になった可能性があり、検討が必要である。

将来的には現在固定されているメールの分類と配置を一般化して、ユーザが任意でメールの分類と配置場所・表示件数を設定し、自由なデザインで表示すること可能にすることや、メールの内容に応じたアドバイスを表示すること、ユーザの行動から学習するフィルタを構築するなど、より総合的な情報支援システムの実現を目指す。

参考文献

[1] 佐々木稔, 新納浩幸, 文章分類を用いたスパムメール判定手法, 情報処理学会研究報告, 2004-FI-76(11), 2004-NL-163(11)

[2] Noriko OTANI, Fumiaki ITOH, Shogo SHIBATA, Takaya UEDA, Yuji IKEDA, An Information Retrieval System based on Personal Viewpoints for Everyday Use, IEEE vol.1,pp.397-404, 1998 http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=725876