

プロファイリングを行なうメールエージェントの構築

小林晃裕 杉野政代 中川浩一 石井直宏
名古屋工業大学 知能情報システム学科

1 はじめに

近年、計算機技術の急速な発展に伴い、計算機の存在は身近なものとなってきている。特に電子メールシステムは今や我々にとって必要不可欠なものである。電子メールを利用するにあたって、受信したメールを複数のフォルダに分類したいという要求がある。しかしメールの受信量が増えるに従い、この作業はかなり煩雑なものとなる。

そこで本研究ではユーザに代わって電子メールを自動的に分類するインターフェースエージェントを構築した。本システムではエージェントがユーザの作業を観察し、その政策を学習して徐々に分類精度をあげていく学習アプローチをとる。エージェントはメールのヘッダ (To,From,Subject) 及び内容を解析し、学習したプロファイルとの比較によって特徴ベクトルを求めて適切なフォルダを提案する。

2 キーワード抽出

本研究で構築したメールエージェントは、以下の要素に基づいて学習および提案を行なう。

- メールの送信先を示す To フィールド
- メールの発信元を示す From フィールド
- 題名を示す Subject フィールド
- メール本文

Subject フィールドおよびメール本文は、それぞれから抽出したキーワードを用いる。キーワード抽出のために、奈良先端科学技術大学院大学で開発された日本語形態素解析システム『茶筌』[1]を使用している。本研究ではこの『茶筌』を使って、Subject フィールド、メール本文を単語に分割し、そこから名詞や英単語を抽出する。抽出されたキーワードは、学習により重要度、時間情報を付加してプロファイルに加えられる。重要度とはそのキーワードが過去に現われた回数をあらわしている。

A Mail Agent System with User Profile
Akihiro Kobayashi, Masayo Sugino,
Kouichi Nakagawa, Naohiro Ishii
Department of Intelligence and Computer Science
Nagoya Institute of Technology

3 メールエージェントシステムの構築

3.1 プロファイル

メールエージェントはプロファイルに基づいて提案・学習を行なう。プロファイルは各フォルダ毎に用意され、フォルダに分類されたメールの To フィールドと From フィールド、茶筅によって抽出した Subject フィールド及びメール本文のキーワードを保持する。

杉野ら [2] のメールエージェントも同様にプロファイルを利用しているが、フォルダ毎の分類方針の違いをうまく判別できないという問題点があった。分類方針の違いとは、フォルダによって From フィールドでほぼ決まるものや, Subject フィールドと内容で決まるものがあるなど、注目すべきフィールドが異なることである。

そこで本システムでは、そのフォルダへの分類に際して上記の 4 つの要素のどれに強く影響されたかをあらわす特徴ベクトルも保持する。特徴ベクトルによりそのフォルダの分類方針を表すことができる。

3.2 提案

エージェントは新しくメールを取り込むと To フィールド、From フィールドを取得し、茶筅により Subject 行とメール本文からキーワードを抽出する。これらのキーワードと各フォルダのプロファイルとを比較する事によって(図 1)スコア計算を行なう。

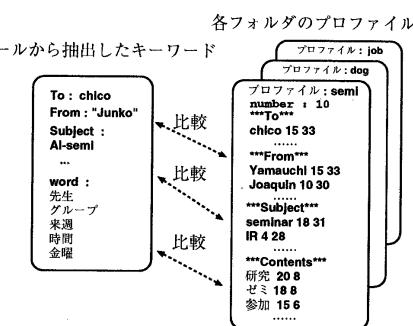


図 1: プロファイルとの比較

このとき、キーワードの重みはプロファイル中における出現確率とする。これによってどのメールにも含まれる事の多い一般的な名詞で必要以上にスコアが高くなる事がなくなる。また、メール量の多いフォルダに他のフォルダが影響される可能性が低くなる。

スコア計算により、To, From, Subject, メール本文の 4 つの要素のスコアで表される特徴ベクトルが計算され

る。この特徴ベクトルを各フォルダに保持されている特徴ベクトルのテーブルと比較する。特徴ベクトルのテーブルには過去の特徴ベクトルとそのフォルダに分類されたかどうかを表すフラグが対になって格納されている(図2)。

特徴ベクトル				
To	From	Subject	Contents	Flag
0.8	0.55	0.11	0.34	1
0.8	0.0	0.08	0.12	-1
⋮	⋮	⋮	⋮	⋮

↓
そのフォルダに
分類されたか
1: 分類された
-1: 分類されなかった

図2: 各フォルダにおける特徴ベクトルテーブル

エージェントは各フォルダ毎に、距離重み付き K 近傍アルゴリズム (Distance-Weighted k-Nearest Neighbor, 以下 k-NN) により特徴ベクトルのテーブルから最も類似したベクトルを K 個選択し、以下の計算を行なう。

$$Score_{folder_i} = \sum_{j=1}^K flag * distance_j$$

これによって得られた $Score_{folder_i}$ でのスコア $Score_{folder_i} (-1 \leq Score_{folder_i} \leq 1)$ が最も高くなるフォルダを選択し、分類すべきフォルダとして提案を行なう。

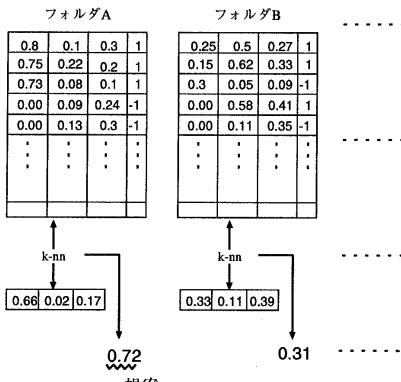


図3: k-NNによる特徴ベクトルの比較

3.3 学習

メールエージェントは、メールをフォルダ A へ分類することになったとき、提案の成功/不成功に関わらずフォルダ A のプロファイルを更新する。このとき、メールに含まれていたキーワードが既にプロファイルにあった場合、出現回数をインクリメントし時間情報を更新する。

また、以下の場合に特徴ベクトルの学習を行なう。

1. エージェントがフォルダ A を提案したがユーザがこれを拒否して他のフォルダを指示した場合
2. エージェントが他のフォルダを提案したがユーザがフォルダ A を指示した場合。

3. エージェントがフォルダ A を提案し、ユーザもこれを受け入れたが、 $Score_{folder_A} < 0$ であった場合

4. エージェントがフォルダ A を提案せず、ユーザもこれを受け入れたが、 $Score_{folder_A} > 0$ であった場合

1,3 の場合には特徴ベクトルにフラグとして -1 を、2,4 の場合には 1 を付加してテーブルに追加する。これによりフォルダ毎の分類方針の違いを吸収できるだけでなく、ユーザの分類方針の変更にも自動的に追従することができる。また、学習が収束するにつれ学習データの増加量は減少する。

4 評価実験

以下の実験結果は 7 個の分類方針の異なるフォルダを用意してエージェントに学習・提案を行なわせたものである。

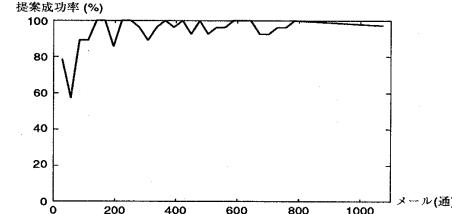


図4: 提案成功率

この結果では、最初に一旦提案成功率が落ちているが、これは最初にプロファイルの学習を行い、その後で特徴ベクトルの学習を行なうため、学習の初期段階においては負事例が不足するからである。その後は高い提案成功率(平均 96.2%)を示している。

5 まとめ

本研究では計算機を使用するユーザをサポートするインターフェースエージェントとして、メールの分類を自動的におこなうメールエージェントを構築した。構築したシステムでは、プロファイルとの比較とそれによって得られた特徴ベクトルの比較の 2 段階の比較を行なうことによって高い分類精度を得ることができた。またユーザの分類方針の変更にも柔軟に対応できることが期待できる。

今後の課題としては、処理の高速化とユーザーインターフェースの向上が挙げられる。

参考文献

- [1] "日本語形態素解析システム『茶筌』 version 2.0 使用説明書 第二版", <http://cactus.aist-nara.ac.jp/lab/nlt/chasen/bib.html>
- [2] 杉野, 中川, 石井, 稲波: "キーワード抽出に基づくメールエージェントの研究", 1999 年電気関係学会東海支部連合大会, p312