多属性項目の履歴情報に基づく電子メイル文書のフィルタリング手法

獅々堀 正幹[†] 藤井 誠[††]
安藤一秋[†††] 青江順一[†]

近年，インターネットの普及にともない，大量のメイル文書が氾濫し，各メイル文書の重要度を自動的に判定するフィルタリング処理の実現が望まれている．従来のフィルタリング技術は，出現単語といった単一属性を扱ったものが多く，十分な精度が得られなかった．本論文では，受信済みのメイル文書から送信元，テーマや類型等の多属性項目の組合せから成る構造化知識を獲得し，その知識を用いたフィルタリング手法を提案する．さらに，各属性の意味に従って近似処理を適用すべき属性の順番を定義し，学習データ数が少ない場合にも，ノイズの混入を極力抑えて正確な重要度を算出する．4 人の被験者に対する実験結果から，単語の出現頻度という単一属性のみで判定した重要度よりも，多属性値の組合せを考慮した重要度の方が人手で判定した結果と高い相関を示すことが確認できた．

A Filtering Method for E-mail Documents Using Personal Profiles

Masami Shishibori,[†] Makoto Fujii,[††] Kazuaki Ando[†††]
and Jun-ichi Aoe[†]

Nowadays, E-mail is very often used as the one of communication methods, however a lot of E-mail documents including useless information such as commercial mails is distributed to our computers. In order to solve this problem, we propose a method to filter out unimportant E-mail documents automatically by judging the content of each document. This method extracts the corpus-based knowledge from existing E-mail documents which have been already received by each user. This knowledge consists of multi-attribute items such as the sender, theme and type of each existing document. By using the acquired knowledge, this method can show whether a new E-mail document is important or not for the user. From the experimental results, it is found that this method can calculate more accurate point than traditional methods, which use sole attribute based on the statistics of word occurrences.

1. はじめに

近年のインターネットの普及にともない，電子メイルがコミュニケーションの 1 つの手段として確立されてきた．電子メイルは環境保護の面からもペーパーレス化を促進し，ほとんどの書類が電子メイルを介して配布されつつある．このように，情報の伝達が容易になった反面，各個人向けの計算機にも大量のメイル文書が送信され，重要なメイルと商用メイルに代表される重要度が低いメイルとを自動分類する処理，すなわち，コンテンツベースの情報フィルタリング技術の重要性が高まってきている[1)]．

本論文では，各個人が受信済みのメイル文書内に重要度を判定するための履歴情報が含まれていると考え，各個人があらかじめ優先度付けした既存のメイル文書からプロファイルを作成し，そのプロファイルを用いてフィルタリングを行う手法を提案する．

本手法では，重要度を決定する要因として送信元，勧告・要求・疑問等の文の類型[2)]，文のテーマ，時間的制限（これらを属性と呼ぶ）を取り上げ，既存のメイル文書内に含まれる各属性値とユーザが設定した優先度との組合せから成るプロファイルを作成する．このプロファイルは，ユーザがどのような多属性値の組合せを含むメイル文書に重要性を感じているかを示しており，このプロファイルを用いることで各個人に適応したフィルタリングが可能となる．また，学習データ数が少なく，プロファイルがスパースな場合，類似した属性値に置換する必要がある．本手法では，各属性の意味に従って，置き換えを適用すべき属性の順番を定義し，極力少ないノイズで重要度を近似する．

以下，2 章では，他の研究と比較することにより，本研究の位置づけを行う．次に，3 章において，メイ

† 徳島大学工学部
Faculty of Engineering, Tokushima University
†† 三菱電機株式会社
Mitsubishi Electric Corporation
††† 香川大学工学部
Faculty of Engineering, Kagawa University

ル文書の重要度とは何か，また，いかなる要因から決定されるかを明確にした後，4 章では，プロファイルの作成方法とプロファイルを用いた重要度の算出方法，さらに，プロファイルがスパースな場合の対処方法を説明する．そして 5 章で，本手法の評価を行い，最後に 6 章でまとめおよび今後の課題について述べる．

2. 本研究の位置づけ

インターネットを介した文書を対象とするフィルタリング技術☆は，近年活発に研究報告がなされている．まず，佐藤ら[3]は電子ニュース記事から重要語を検出し，サマリーを自動生成する手法，また，長谷川ら[4]は電子メイル文書から日時・場所等のスケジュール情報を自動抽出する手法を提案している．両手法とも言語特徴から求めた表現パターンとのパターンマッチングを行うルールベースの手法である．メイル文書の重要度を判定する場合にも，事前にルール化した重要な表現パターンとのパターンマッチングにより重要度を算出する方法が考えられる．しかし，重要性の判断基準は各個人ごとに異なり，履歴情報にも左右されるため，一意に決められたルールベースの手法では対応できない．Foltz ら[5]は情報検索手法の 1 つである LSI 法を適用し，電子ニュース記事の類似性を求め，ユーザにとって重要な記事の推定を行っている．しかし，LSI 法自体の計算量が多く，メイル文書のフィルタリングという実時間処理が要求される分野には不向きである．

また，メイル文書のフィルタリングを取り扱った研究も数件発表されている．まず，加来田ら[6]は，受信したメイル文書に対するユーザの行動（参照時間等）と文書内に含まれる単語の頻度に基づいて作成したプロファイルを重要度の判定に用いている．しかしながら，単語という単一属性のみを対象としているために精度が低く，また，学習データ数が少ない場合の対処法についても触れられていない．次に，長谷川[7]は，メイル文書内から抽出した多属性値の内容からメイル文書をランキングする手法を提案している．この手法は，送受信履歴からユーザが重要性を感じている属性に高い重み係数を与え，それら属性の組合せからメイル文書の重要度を計算している．しかしながら，実際のメイル文書では，同じ属性でも属性値の違いにより重要度が異なるため，属性値ごとに重み係数を付与し，それら多属性値の組合せを考慮する方法でなければ，より正確なフィルタリングは行えない．

☆ 情報抽出もフィルタリングと同じ分野として以後議論を進める．

本論文で提案する手法は，文献 3)，4) と異なり，受信済みのメイル文書から獲得したコーパスベースの知識を用いてフィルタリングを行う．また，多属性値の組合せからプロファイルを作成する点が文献 6) と異なり，多属性値の組合せを考慮して重要度を算出する点が文献 7) と異なる．さらに，文献 7) では，主題となりやすい重要語句が登録された辞書を用いているが，本手法では学習データからそれらの語句を獲得する．なお，文献 6)，7) では，既存のメイル文書に対する優先度をユーザの行動から推定しているが，これについては本論文では論じない．

3. メイル文書の重要度とは

3.1 重要度の要因

本研究では,「他のメイル文書よりも早く読む必要がある」，または,「早急に返事を出す必要がある」メイル文書を重要度が高いと定義する．この定義を前提にすると，重要度は文書のテーマや時間的制限の有無，そして，文の類型等に左右されると考えられる．また，一般文書には存在せず，メイル文書にのみ含まれる特有な情報として，送信元，Cc，引用文と本文の関係，過去のメイルとの関連性等も重要度の判定に有効である．これらの中で今回は，形態素レベルから得られる情報でフィルタリングに有効と思われる項目として，次のような項目を重要度を構成する要因として取り上げる．

(α) メイル文書の送信元；
(β) メイル文書に含まれる文の類型；
(γ) メイル文書に含まれる時間的制限；
(θ) メイル文書のテーマ；

以下，これらを重要度決定のための属性（以後，単に属性）と呼び，各属性の値を属性値と呼ぶ．

各属性値は，以下の方法で取得可能である．まず，(α) についてはメイル文書内ヘッダー From の項目から，また，(β) についてはメイル文書の本文内に含まれる助述表現[8]，(γ) については時間表現[9]や時間を表す副詞から取得可能である．(θ) については，主題となりやすい名詞類を登録した辞書[7]を用いることが考えられるが，汎用性が失われてしまう．また，本来ならば意味解析等の高度な基礎解析技術を導入し，各メイル文書の「テーマ」を特定すべきであるが，実時間処理が困難になる．そこで本手法では，ユーザ主導の方法により「テーマ」を特定する．まず，受信済みのメイル文書が事前にユーザによって類似した内容の文書ごとにまとめられ，メイルボックスの各フォルダに分類されていることを前提とする．受信済みのメイ

ル文書については，各文書が格納されていたフォルダ名を「テーマ」とする．また，新たに受信した入力メイル文書については，文書分類法と同様な方法（詳細は 4.1 節で述べる）で各フォルダとの類似度を求め，最も類似したフォルダの名前を「テーマ」とする．

3.2 重要度の個人差

3.1 節で述べた観点から考えると，メイル文書の重要性の判断基準が個人ごとに異なることは明白である．たとえば，送信元に対して，A 君は［青江先生］からのメイル文書には高い重要度をおいているが，B 君は［青江先生］からのメイル文書にはさほど重要性を感じていないかもしれない．このように，各個人ごとに重要な属性値は異なるため，フィルタリングを行う際には，各個人ごとの知識が必要になる．また，文の類型までも考慮すると，A 君は［青江先生］から届いた［勧告］のメイル文書には高い重要度を感じているが，［依頼］の内容を含むものは重要度が低いかもしれない．一方，B 君は，A 君とは反対の組合せに重要度をおいているかもしれない．このように，各個人ごとに重要な属性値の組合せは異なるため，重要度判定知識として，属性値の組合せの情報を準備する必要がある．

以上，メイル文書の重要度は，各属性値に対して個人ごとに異なった重み付けがされ，多属性値の複雑な組合せにより，重要度が決定されているという前提の下で本手法を提案する．

4. フィルタリング手法

4.1 フィルタリング手法の概要

本研究では，各ユーザが受信済みのメイル文書内に重要度を判定するための知識が存在すると考え，各ユーザの判断基準に従って優先度☆付けされた既存の文書から各属性値を抽出する．そして，抽出した属性値と優先度から成る構造化知識を用いて各個人に適応した重要度を算出する．

図 1 に本手法の概要図を示し，処理の概要を説明する．まず，学習部（図 1 内実線の流れ）では，ユーザによってあらかじめ優先度が付与され，各フォルダに分類された既存のメイル文書を形態素解析する．その後,「類型」および「時間的制限」に対する属性値は表現パターン数が有限個であり，かつ，個人差が少ないため，表現パターンを登録した背景知識を用いて属性値を検出する．また,「送信元」に関しては，検出したメイルアドレスをコード化し，そのコードを属性値

☆ ユーザがあらかじめ既存の文書に対して付与する重要性のレベル値を「優先度」，入力文書に対してシステム側が求めた重要性のレベル値を「重要度」と呼び，双方を区別して用いる．

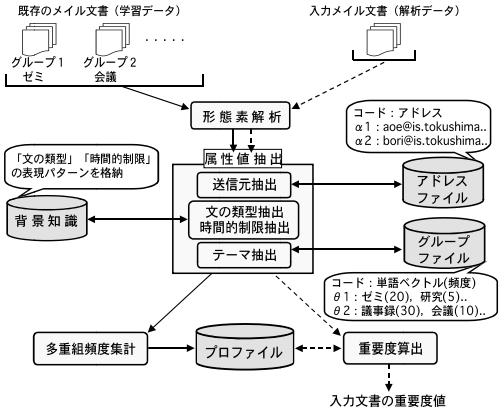

図 1 本フィルタリング手法の概要図
Fig. 1 The outline of this filtering method.

とする．コードとアドレスの対応はアドレスファイルに記載する.「テーマ」についてもフォルダ名をコード化し，そのコードを属性値とする．さらに，フォルダに格納されているメイル文書に対して，*Subject* の欄，および本文から名詞類☆☆を切り出し，そのフォルダに対応するテーマを特徴付ける単語ベクトルを作成する．この単語ベクトルは単語と出現頻度の対から構成され，テーマ（属性値コードで表現）と単語ベクトルの対応をグループファイルに記載する．ただし，*Subject* は本文の「見出し」なので，*Subject* 内に出現した単語の頻度は K 倍し，本文内に出現する単語より頻度の重みを重くする．ここで，K の値を低くしすぎると，*Subject* の内容が活かしきれず，逆に高くしすぎると，本文内に頻出する重要語が無視されてしまう．したがって，K の値は，*Subject* ではカバーしきれない部分を本文でうまく補えるように設定する経験的なパラメータ値とする．以上の手順で検出した多属性値と優先度の組合せから成る多重組を各メイル文書ごとに生成し，それらの頻度を集計した結果をプロファイルとする．なお，背景知識の内容，およびプロファイルの作成方法については 4.2 節で詳しく説明する．

次に，解析部（図 1 内点線の流れ）では，重要度が未知の入力メイル文書に対して形態素解析を施した後，背景知識を参照して「類型」および「時間的制限」の属性値を得る．また,「送信元」を表すメイルアドレスは，アドレスファイルを参照して属性値に変換する．「テーマ」に関しては，入力文書の *Subject*，および本文から切り出した名詞類とグループファイル内の各単

☆☆ 普通名詞のほかに人名，地名，会社名等の固有名詞も含む．

表 1 背景知識の例
Table 1 An example of the knowledge dictionary.

属性	属性値名	属性値	表現例
文の類型	勧告	β_1	～すること，～しておくこと，～できること，～するように，おわすれなく
	義務	β_2	～して貰います，～するべきこと，～する必要があります，義務づけます
	依頼	β_3	～して下さい，～してくれ，～して欲しい，～してちょうだい，お願い致します
	条件付依頼	β_4	もし～ならば … 下さい，～だったら … してくれ
	勧誘	β_5	～しましょう，～しよう，～しましょうよ，～しようよ，～しませんか
	命令	β_6	～しなさい，～してこい，～しろ，～せよ，～すべし
	疑問	β_7	～ですか，～でしょうか，～ありますか，～かな，～ありませんか，～？
	告知	β_8	～連絡いたします，～お知らせいたします，～計画しております
時間的制限	1 週間以上	γ_1	1ヶ月以内に，二三週間以内に，1ヶ月後に
	1 週間以内	γ_2	1 週間以内に，7 日以内に，1 週間後に
※ただし，時区間を含まない時間表現は，送信時間を考慮して分類する．	3 日以内	γ_3	なるべく早く，忘れないうちに，二三日中に
	24 時間以内	γ_4	できるだけ早く，早めに，24 時間以内に，一両日中に
	12 時間以内	γ_5	今すぐ，早急に，すぐに，迅速に，1 時間以内に，1 時間後に

語ベクトルとを比較し，最も類似したフォルダのコードを属性値とする．このようにして入力文書に対する多重組を生成した後，この多重組と同じ組合せの多重組をプロファイル内から検索し，それらにどのような優先度が与えられていたかを基にして重要度を確率的に求める．なお，プロファイルの内容に従った重要度算出方法の詳細は 4.3 節で述べる．

4.2 学 習 部

4.2.1 背 景 知 識

「類型」および「時間的制限」の属性値を検出するための表現パターンを類似したパターンごとに 1 つの属性値としてグルーピングし，背景知識に格納する．背景知識の一部を**表 1** に示す．

「類型」については文末に現れる助述表現を文献 2), 8) を参考に分類した．「時間的制限」については，まず，緊迫度の度合いにより，属性値間の閾値を数量的に設けた．そして，副詞については，各表現がどの程度の緊迫度を有するかを十数人にアンケート調査し，その結果から各属性値に割り振った．また，時間表現に関しては，表現形式の違いにより，以下のような 2 種類に分け，扱い方を別にした．

(1) 時区間[9]を含む時間表現；
(2) 時区間を含まないが時点[9]は含む時間表現；

(1) に属する「1 時間以内に」は，時区間が明確に表示されているため，該当する範囲の属性値に分類する．「明日まで」のように (2) に属する表現は，現時点の時間が分からなければ時区間を求めることができないので，メイル文書の送信時間を基準に時区間を求め，属性値に分類することにした．なお，「次回のゼミまで」のように，[時点] + [名詞] から成る表現については，[名詞] に関するスケジュール情報が別途必要になるため，今回は対象外としている．

4.2.2 プロファイル

まず，学習文書データを $D = \{d_1, d_2, \ldots, d_i, \ldots, d_A\}$ とすると，各文書は $d_i = \{x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{iB}\}$ という多重組の集合を持ち，各多重組は $x_{ij} = (\alpha_l, \beta_m, \gamma_n, \theta_q, \kappa_r)$ という多属性値の多次元ベクトルとなる．なお，多重組の構成方法は以下に従う．

【多重組構成手順】

手順 1： 「送信元」の属性値 α_l を多重組に代入；

手順 2： 「類型」の属性値 β_m が複数存在する場合，その数だけ多重組を複製し，β_m を代入；

手順 3： 手順 2 で助述表現が検出された同一文内で「時間的制限」の属性値 γ_n を検出し，検出数だけ多重組を複製した後，γ_n を代入；

手順 4： 複製された各多重組に「テーマ」の属性値 θ_q と優先度 κ_r を代入；

【手順終了】

また，1 文書ごとに生成された多重組の個数に従って，頻度 $freq(x_{ij})$ を与える．ただし，$1 \leq j \leq B$ とすると，$freq(x_{ij}) = 1/B$ であり，$\Sigma_j freq(x_{ij}) = 1$ とする．ここで，$\cap_{ij} x_{ij} \neq \phi$ であるため，同じ多重組の頻度を集計した結果をプロファイル $P = \{p_1, p_2, \ldots, p_k, \ldots, p_C\}$ とする．ただし，p_k は多重組を表し，$\cap_k p_k = \phi$，$freq(p_k)$ は p_k と同じ内容の多重組 x_{ij} の頻度合計値とする．このプロファイルはユーザがどのような属性値の組合せに重要性を感じているかを示している．なお，内容が $(\alpha_l, \beta_m, \gamma_n, \theta_q, \kappa_r)$ のプロファイル内多重組を以後 $p_{<l,m,n,q,r>}$ と記すことにする．

4.3 解 析 部

解析文書データを $T = \{t_1, t_2, \ldots, t_i, \ldots, t_E\}$ とすると，各文書は $t_i = \{y_{i1}, y_{i2}, \ldots, y_{ij}, \ldots, y_{iF}\}$ という多重組の集合を持つ．ただし，y_{ij} は優先度 κ の値は持たない多次元ベクトルである．解析部では y_{ij} ごとに重要度を算出し，y_{ij} と同じ多重組がプロファイル内に存在する場合は通常処理，存在しない場合は近似処理を行う．以降，各処理の詳細を個別に説明する．

4.3.1 通常処理

通常処理では，多重組 $y_{ij} = (\alpha_l, \beta_m, \gamma_n, \theta_q)$ に対する重要度を $p_{<l,m,n,q>}$ に付与されている優先度を基にし，式 (1) に従った条件付確率値として求める．

$$P(\kappa_r \mid \alpha_l, \beta_m, \gamma_n, \theta_q) = \frac{freq(p_{<l,m,n,q,r>})}{freq(p_{<l,m,n,q>})} \quad (1)$$

ただし，

$$freq(p_{<l,m,n,q>}) = \sum_r freq(p_{<l,m,n,q,r>})$$

とする．また，多重組 y_{ij} 内に未検出の属性値があった場合，存在する属性値のみを対象に式 (1) を適用する．たとえば，$y_{ij} = (\alpha_l, \beta_m, *, \theta_q)$[☆] の場合，式 (1) は次のように変化する．

$$P(\kappa_r \mid \alpha_l, \beta_m, \theta_q) = \frac{freq(p_{<l,m,q,r>})}{freq(p_{<l,m,q>})} \quad (2)$$

ただし，

$$freq(p_{<l,m,q,r>}) = \sum_n freq(p_{<l,m,n,q,r>})$$

とする．式 (1) での確率値は優先度 κ の各ランクごとに求められるため，式 (3) で補正した値を多重組 y_{ij} の重要度 R_{ij} とする．

$$R_{ij} = \sum_r r \times P(\kappa_r \mid \alpha_l, \beta_m, \gamma_n, \theta_q) \quad (3)$$

また，最終的に解析文書 t_i の重要度 R_i は，R_{i1}〜R_{iF} の平均重要度として求める．

解析内容を**図 2** に示すプロファイル例[☆☆]を用いて説明する．まず，多重組 $y_{ij} =$ (青江, 勧告, ∗, ゼミ) に対して，プロファイル内で 3，4 の多重組が参照され，それらの合計頻度が 10，優先度はともに 5 であるので，重要度 5 である確率が 100%，$R_{ij} = 5$ となる．また，(青江, 勧告, ∗, ∗) と「テーマ」が検出できなかった場合，1〜6 の多重組が参照され，合計頻度が 20，優先度 5 が 3〜6（合計頻度 15）なので，重要度が 5 である確率が 75%，同様に，重要度 4 が 15%，重要度 3 が 10%，$R_{ij} = 4.65$ となり,「ゼミ」内容の文書より低い重要度が算出され，プロファイル内容に則した結果を得られる．

4.3.2 近似処理

入力文書の多重組と同じ多重組がプロファイル内に存在しない場合，プロファイル内に存在する他の多重組に置き換えて重要度を近似的に計算する．ここで問題になるのは，どの属性に対し，どの属性値に置き換えるかという点である．各属性には，それぞれ異なった特徴があり，置換によるノイズ混入量に違いがある．本手法では，各属性の特徴を考慮し，置き換えを適用する属性の優先順位を次のように定義する．

時間的制限（γ）⇒ 送信元（α）⇒ テーマ（θ）

以下，各種近似処理について，個別に説明する．

●「時間的制限」に対する近似処理

「類型」「テーマ」「送信元」のように文書の興味を表す属性よりも,「時間的制限」のような興味に対する緊急性を表す属性[6]を変更した方がノイズが入りにくいと考え，最初に置き換えを行う．以後，この近似処理を時間近似処理と呼ぶ．

多重組 $y_{ij} = (\alpha_{l1}, \beta_{m1}, \gamma_{n1}, \theta_{q1})$ に対して，時間近似処理では，まず，多重組 $(\alpha_{l1}, \beta_{m1}, *, \theta_{q1})$ の重要度（これを基準重要度と呼ぶ）を式 (2) により計算する．なお，この計算で用いたプロファイル内の多重組に含まれる時間属性値の平均を γ_{ave} とする．ここで，基準重要度は γ_{n1} を無視した値であるので，以下の式で補正した値を近似値とする．

$$\text{近似値} = \text{基準重要度} + (\gamma_{n1} - \gamma_{ave}) \times \frac{d\kappa_r}{d\gamma_n}$$

ただし，$d\kappa_r/d\gamma_n$ は，時間属性値の変化に対する優先度の変化率を示しており，同じ興味を表す多重組の対の集合から値を推定する．つまり，α_l，β_m，θ_q の値が同じで時間属性値が異なる多重組の対，$p_{<l,m,n',q,r'>}$ と $p_{<l,m,n'',q,r''>}$[☆☆☆]を検索し，時間属性値に対する優先度の変化値 $(\kappa_{r'} - \kappa_{r''})/(\gamma_{n'} - \gamma_{n''})$ を求める．この変化値をプロファイル内のすべての多重組について求め，変化値が 0 以外のものの平均値を変化率 $d\kappa_r/d\gamma_n$ とする．なお，この変化率は，解析処理以前に，プロファイルが作成された時点で求めておく．

多重組 $y_{ij} =$ (青江, 勧告, 3 日以内 (γ_3), 会議) に対する時間近似処理を図 2 のプロファイル例を用いて説

ID:(送信元，類型，	時間的制限，テーマ，優先度)＝頻度
1:(青江，勧告，	0 (γ_0)， 会議， 3)＝ 2
2:(青江，勧告，	1 週間以内 (γ_2)， 会議， 4)＝ 3
3:(青江，勧告，	1 2 時間以内 (γ_5)， ゼミ， 5)＝ 7
4:(青江，勧告，	0 (γ_0)， ゼミ， 5)＝ 3
5:(青江，勧告，	2 4 時間以内 (γ_4)， 試験， 5)＝ 2
6:(青江，勧告，	3 日以内 (γ_3)， 行事， 5)＝ 3
7:(青江，依頼，	1 2 時間以内 (γ_5)， ゼミ， 4)＝ 2
8:(青江，依頼，	1 週間以上 (γ_1)， ゼミ， 2)＝ 3

図 2 プロファイル例
Fig. 2 An example of the profile.

☆ ∗は入力文書に対応する属性値がなかったことを表す．
☆☆ 図 2 内の属性値は，理解しやすいように属性値名を記している．
☆☆☆ ただし，時間属性値の添字の関係は，$n' > n''$ とする．

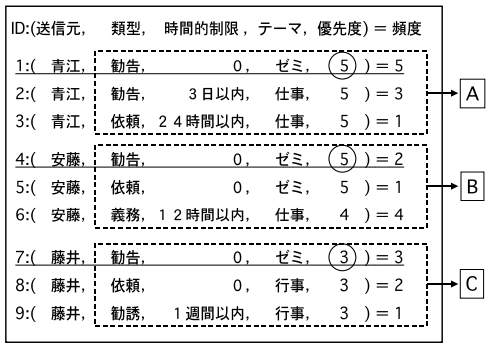

図 3 送信元近似に対するプロファイル例
Fig. 3 An example of the profile for the approximation to the mail sender.

明する．なお，説明の都合上，$\gamma_n = n$ として計算を行う．まず，(青江, 勧告, ∗, 会議) に対して，1，2 の多重組から基準重要度 $= 3.6$，$\gamma_{ave} = 1$ が得られる．また，1，2 の多重組対（変化値 $= 0.5$），および 7，8 の多重組対（変化値 $= 0.5$）から，変化率 $= 0.5$ が求まり，結局，式 (4) より，$3.6 + (3-1) \times 0.5 = 4.6$ が近似値として得られる．

●「送信元」に対する近似処理

時間近似処理を適用しても重要度が算出されなかった場合☆1,「送信元」に対する近似処理（以後，送信元近似処理と呼ぶ）を行う.「送信元」を表すアドレスには，一般的に職業，地位等の点で類似性があり，アドレスファイル内に類似性の高い属性値が存在する可能性があることから，2 番目に置換を行う．

本手法では，類似した「送信元」から届いた同じような内容のメイル文書に対して，ユーザはよく似た優先度を付与していると考える．そこで，同一の「テーマ」「時間的制限」「類型」から成る多重組に，よく似た優先度が付与されている「送信元」ほど，類似していると判断する．たとえば，**図 3** において,「送信元」ごとに A，B，C の多重組群を比較すると,「テーマ」「時間的制限」「類型」が同一の多重組 1，4，7 に付与されている優先度の値から,「青江」が「藤井」よりも「安藤」の方に類似していると判断する．

多重組 $y_{ij} = (\alpha_L, \beta_M, \gamma_N, \theta_Q)$ に対して，上記の方法に従って送信元を α_{Lrep} に置換し，$(\alpha_{Lrep}, \beta_M, \gamma_N, \theta_Q)$ から求めた重要度を近似値とする．なお，$(\alpha_{Lrep}, \beta_M, \gamma_N, \theta_Q)$ に対する多重組がプロファイル内に存在しなかった場合には，この多重組に時間近似処理を適用し，重要度を計算する．以後，送信元近似処理とは，これら一連の処理を総称する．

●「文のテーマ」に対する近似処理

送信元近似処理を行っても重要度が算出できなかった場合☆2,「テーマ」に対する近似処理（以後，テーマ近似処理と呼ぶ）を送信元近似処理と同様な方法で行う．まず，テーマを置換した $(\alpha_L, \beta_M, \gamma_N, \theta_{Qrep})$ から近似値を算出する．同じ多重組が存在しない場合は，この多重組に対して時間近似処理,「送信元」の置換,再度時間近似処理という順で重要度の算出を試みる．なお,「類型」については，言語学的に各属性値間に類似性が少ないことから，今回は近似処理の対象外とした．また，3 種類の近似手法を順次適用し，すべての近似処理に失敗した場合，重要度は算出できないため，優先度区間の中間値を重要度とする．

5. 評　　価

本手法の有効性を確認するため，学習データ数と解析精度の関係，各属性の有効性，および従来手法との比較に関する実験を行った．以下，実験に用いたデータ内容を明示した後，各種実験結果を個別に示す．

5.1 実験データの内容

実験データについては，元データとして既存のメイル文書☆3150〜280 通を各被験者ごとに準備した．そして，各フォルダ内の文書数に比例して無作為に抽出した各 30 通のメイル文書を評価用データとし，残りの文書を学習用データとした．学習用データの内訳を**表 2**☆4に示す．一方，評価用データに付与されている優先度の平均値と分散は，それぞれ 3.52 と 1.34（双方とも 4 人の平均値）であり，学習用，評価用の優先度ともに極端に偏った分布にはなっていない．また，表内の被験者 A と B は研究室内の教職員，被験者 C と D は研究室内の学生のデータである．さらに，背景知識として,「文の類型」については 282 個（属性値数は 8),「時間的制限」については 48 個（属性値数は 5）の表現パターンを登録した．これらのデータを用いて各プロファイルを作成した後，評価用データの重要度を求めた．なお，優先度，重要度ともにランク数を 5（ランク 5 が最高値），稼働マシンは Power Macintosh 8500/200 MHz である．

また，4.1 節で述べた経験的なパラメータ値 K を決

☆1 $y_{ij} = (\alpha_{l1}, \beta_{m1}, *, \theta_{q1})$ に対する基準重要度や $d\kappa_r/d\gamma_n$ が求められなかった場合を示す．

☆2 α_{Lrep} が存在しなかったり，$(\alpha_{Lrep}, \beta_M, \gamma_N, \theta_Q)$ に対する重要度が算出されなかったりした場合を示す．

☆3 各被験者により，優先度が付与され，類似した内容の文書ごとに各フォルダに分類されているものとする．

☆4「多重組数」は各学習文書から得られた多重組の個数を示す．

表 2 学習用データの内訳
Table 2 The contents of the learning data.

被験者	A	B	C	D
学習文書数	250	230	150	120
合計サイズ（Kbyte）	655	501	233	152
送信元数	59	57	7	28
フォルダ数	7	12	6	5
多重組数	310	291	203	161
優先度の平均値	3.3	3.14	3.63	3.46
優先度の分散	1.27	1.31	2.11	2.07

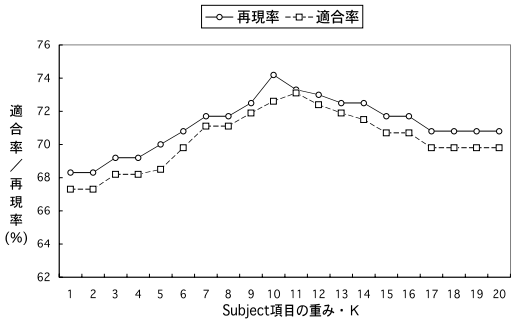

図 4 パラメータ K と「テーマ」検出精度の関係
Fig. 4 Relationship between the parameter K and the detection accuracy of the attribute “Theme”.

定するために，パラメータ K の値を 1〜20 まで 1 刻みで変化させながら学習用データに対する単語ベクトルを構成し，評価用データに対する「テーマ」の検出精度（被験者 4 人の平均）を求めた（**図 4**）.「テーマ」の検出に最適な K の値は，文書量や内容等によって変化すると考えられるが，今回の実験では評価用データに最適化した値として $K = 10$ を用いた．

5.2 学習データ数と解析精度の関係

まず，解析度☆1を評価するために，学習データ数と解析度の関係を**図 5** に示す．なお，図 5 は 4 人の解析度の平均値であり，横軸は学習文書数の割合を示す．また，グラフ内の各線は通常処理，および各近似処理の累積効果☆2を表す．図 5 より，送信元近似処理が解析度の向上に効果的であることが分かる．これは,「送信元」の属性値数が他の属性に比べて多く，また，グループ化もされていないので，学習度が増すに従って，プロファイル内の置換可能な多重組の個数が増加するためである．このことから「送信元」の類似した属性値をグループ化できれば，通常処理の解析度も向上すると予測される．

また，全近似処理を施しても解析度が 100%に達し

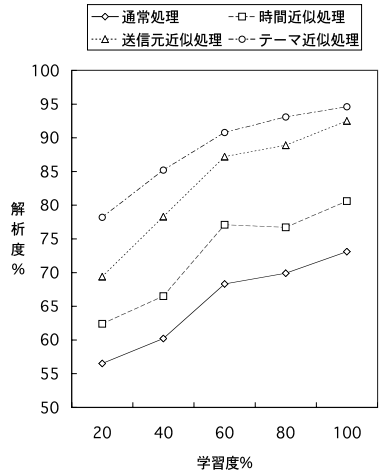

図 5 学習用データ数と解析度の関係
Fig. 5 Relationship between the number of learning data and the recall values.

ていない点に関して，各被験者ごとに解析度を分析してみたところ，被験者 C と D に対しては 100%の解析度を得られたが，A（86.0%）と B（94.0%）についての解析度が低く，特に，テーマ近似処理に対する解析度の伸び率が他の被験者に比べて極端に低かった．これは，被験者 C と D の実験データが狭い範囲（研究室内）のメイルであったのに対して，被験者 A と B のものは広範囲（研究室外）のメイルであり，かつ，送信元に依存した分類☆3が行われていたのが原因である．このような学習用データに対してテーマ近似処理を適用すると，各フォルダ間にまたがって出現する送信元が少なくなるため，各フォルダ（各テーマ）間の類似性が測りにくくなる．また，たとえ類似したテーマを特定できたとしても，そのテーマに対するフォルダ内には対応する送信元が存在しないため，近似処理は行えなくなる．そこで，このような広範囲のメイルに対しては，メイルの内容に即したより詳細な分類を行う必要があり，また，そのような分類が行えれば，より解析度が向上すると推測できる．なお，今回の実験で重要度が算出できた文書数を調べたところ，全近似処理を適用し，かつ，学習度 100%で 97%☆4の評価用文書の重要度が算出できていた．

次に，解析精度を評価するために，学習データ数と平均誤差の関係を求めた（**図 6**）．ここでの誤差は，本手法により算出した重要度とユーザが判定した優先度

☆1 重要度が算出できた入力文書内多重組の割合を示す．
☆2 グラフ内の「テーマ近似処理」の線は,「通常処理＋テーマ近似処理」ではなく,「通常処理＋時間近似処理＋送信元近似処理＋テーマ近似処理」という累積を意味している．
☆3「研究室内」「研究室 OB」「就職関係」等という分類
☆4 図 5 では，全近似処理を適用し，学習度 100%での値が 95%弱であるが，計算の単位が多重組か文書かの違いにより，このようなズレを生じる．

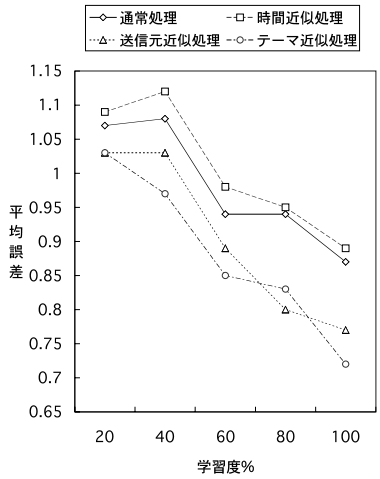

図 6 学習用データ数と平均誤差の関係
Fig. 6 Relationship between the number of learning data and the average precision values.

との差分を示す．なお，これも 4 人の平均値であり，グラフ内の各線は通常処理，および各近似処理の累積効果を表す．図 6 より，全処理ともに学習度が増加するに従い，精度が高まっている．時間近似処理のグラフは，通常処理に追従する形状になっており，本近似アルゴリズムに従った結果となっている．また，送信元，およびテーマの両近似処理については，いずれも通常処理の精度を上回っている．これは，両近似処理により，通常処理では解析できなかった有効な多重組が解析可能になったことに加え，これらの多重組からノイズの混入が極力抑えられた正確な重要度が算出できたことを示しており，本近似手法が有効であることが分かる．

5.3 各属性の有効性

本手法による解析結果の詳細として，評価用データに対する属性値の検出精度を**表 3** に示す．「類型」については，過剰な検出をしている部分もあったが，高い検出精度が得られた．「時間」については［時点］＋［名詞］が多く含まれていたために精度が低い．また，「テーマ」に関しては，内容的に似たフォルダが学習用データ内に存在している場合は，「テーマ」の取得に失敗するケースが多々見られた．本手法のように，単語の出現頻度という単一属性を用いている限り，検出精度には限界がある．今後「テーマ」の検出精度を上げるためには，より高度な基礎解析技術を導入する必要がある．また，これらの検出率が向上すると，より精度の高い重要度が算出可能である．

さらに，どのような属性が重要度の算出に有効であるかを評価するために，属性数および属性の組合せを変化させて相関係数を測定した．測定結果を**表 4** に示す．なお，この実験では，属性値の検出精度を 100%にしなければ正確な評価が行えないため，属性値の検出率が低かった「時間」については対象外とし，また，「テーマ」については，評価用データにあらかじめテーマに関するタグを付与し，作為的に検出率を 100%として実験を行った．表 4 より，ほとんどの被験者について「テーマ」が考慮された際の精度が良く，また，すべての属性を考慮した場合が最も精度が良い．このことから，「テーマ」が重要度の算出に有効な属性であること，および，多属性情報を考慮した処理の有効性が分かる．ただし，被験者 D については，「送信元」が考慮されなければ精度が悪くなり，特に「テーマ」を考慮するとノイズを生じている．このように，解析に適した重要度の組合せは揺れがあるので，個人ごとに最も有効な属性の組合せをプロファイルから学習する機能も今後加えていきたい．

表 3 評価用データに対する各属性値の検出精度
Table 3 A detection accuracy of each attribute value from the analysis data.

被験者	テーマ（%）		類型（%）		時間（%）	
	再現率	適合率	再現率	適合率	再現率	適合率
A	66.7	62.5	100	82.5	61.1	100
B	73.3	71.0	95	80.6	47.5	100
C	80.0	80.0	100	81.7	44.4	100
D	76.7	76.7	94.7	82.2	50.0	100
平均	74.2	72.6	97.4	81.8	50.8	100

5.4 従来手法との比較

多属性情報を用いることによる本手法の有効性を示すために，比較実験を行った．比較手法としては，優先度 p に対する単語ベクトル $\mathbf{W}_p$ を作成し，評価用データから作成した単語ベクトルと各 $\mathbf{W}_p$ との類似度をベクトル空間法[10]により求め，最も類似した $\mathbf{W}_p$ の優先度を重要度とする手法を用いた．i 番目の学習データ d_i に対する単語ベクトル $\mathbf{X}_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{id})$（ただし，$x_{ij}$ は単語の頻度☆とする）から $\mathbf{W}_p$ は構成され，その構成方法により，次の 2 種類の手法を比較手法として採用した．

まず，$\mathbf{W}_p$ の j 番目の要素に対する重みを w_{pj}，事例を学習する前の w_{pj} の状態を w_{pj}^{old}，学習後の状態を w_{pj}^{new} とし，*Rocchio* のアルゴリズム[11]に従った以下の式 (4) により，w_{pj}^{new} を計算する方法を用いた．

$$w_{pj}^{new} = \alpha w_{pj}^{old} + \beta \frac{\sum_{d_i \in D_p} x_{ij}}{|D_p|} - \gamma \frac{\sum_{d_i \notin D_p} x_{ij}}{n - |D_p|} \quad (4)$$

ただし，D_p は学習用データ内で優先度 p が付与され

☆ 単語の頻度としては，$tfidf$ 値[10]を用いる．

表 4 属性数と相関係数の関係
Table 4 Relationship between the number of attribute and the correlation coefficients.

	単一属性			二属性			三属性
被験者	送信元	類型	テーマ	送信元+類型	類型+テーマ	送信元+テーマ	送信元+類型+テーマ
A	0.58	0.48	0.51	0.55	0.57	0.67	0.68
B	0.49	0.29	0.54	0.52	0.55	0.63	0.65
C	0.42	0.27	0.58	0.56	0.60	0.66	0.73
D	0.56	0.14	0.44	0.66	0.21	0.53	0.57
平均	0.51	0.30	0.52	0.57	0.48	0.62	0.66

表 5 各手法による相関係数
Table 5 Correlation coefficients of each method.

被験者	*Rocchio*	*Widrow–Hoff*	本手法
A	0.58	0.60	0.58
B	0.42	0.46	0.57
C	0.32	0.34	0.53
D	0.29	0.30	0.38
平均	0.40	0.43	0.52

た文書の集合，$|D_p|$ はその集合内の文書数，n は全学習用データ数を表し，パラメータ値 α，β，γ は文献 11) から 8，16，4 を用いた．

また，*Widrow–Hoff* のアルゴリズム[12]を用い，以下の式 (5) により，最適化を行いながら w_{pj}^{new} を計算する方法を 2 番目の比較手法として用いた．

$$w_{pj}^{new} = w_{pj}^{old} - 2\eta(\mathbf{W}_p^{old} \cdot \mathbf{X}_i - y_i)x_{ij} \qquad (5)$$

ただし，y_i は，$d_i \in D_p$ ならば，$y_i = 1$ となり，逆に，$d_i \notin D_p$ ならば，$y_i = 0$ となるラベル値を意味する．また，パラメータ値 η は 1 とした．

評価方法としては，本手法，および比較手法により算出した重要度とユーザが判定した優先度との相関係数（1 に近いほど相関が高い）を求めた．なお，信頼性を高めるため，10 回交差検定☆を行った．10 回の計測結果の平均を**表 5** に示す．表 5 より，単語という単一属性情報のみしか利用しない比較手法よりも本手法がユーザの判断とかなり高い相関を持つことが確認できた．一方，比較手法がすべての評価データの重要度を算出したのに対して，本手法では多重組の平均解析度が 96.4%，文書単位では 98.5%であり，重要度が算出できなかった多重組☆☆が若干存在した．このことから，本手法が比較手法よりも多くの学習データを必要とすることもうかがえる．

また，本手法と同じく多属性情報を扱う文献 7) の手法と比較する．比較手法では，学習事例から属性と属性値の重みを計算しているが，属性値の重みを属性ごとに個別に計算しているため，属性値間の（組合せ）関係が学習時に破棄されてしまう．その結果，入力文書の解析時に属性値間の関係という重要な情報を用いずにフィルタリングを行うことになる．たとえば，(青江, 勧告, 5)☆☆☆(青江, 依頼, 1) (安藤, 勧告, 1)(安藤, 依頼, 5) のように属性値間の関係により優先度が変化する場合，比較手法ではすべての属性値の重みが均一化された 3 になってしまい，学習事例に則した解析が行えない．これに対し，本手法では，属性値間の組合せに重みを与えるため，学習事例に忠実な解析結果が得られる．一方，比較手法のように重回帰分析により属性ごとの重み係数を計算すると，5.3 節で述べたような個人ごとに最も有効な属性の組合せを選択できる可能性も含んでいるため，今後，このような機能は本手法に追加する必要がある．

最後に，各種処理の時間効率についてであるが，平均学習時間は約 11.5 秒（平均学習文書サイズは 385 Kbytes），解析時間は約 1.2 秒，ただし，双方とも形態素解析（辞書は主記憶上）時間も含んでいる．また，プロファイルの平均サイズは約 4.6 Kbytes である．

6. ま と め

本論文では，受信済みのメイル文書から多属性項目から成るプロファイルを作成し，入力メイル文書の重要度を求める手法を提案した．本手法を用いることにより，各ユーザが重要性を感じている内容を含むメイル文書から提示したり，極端に重要性の低いメイル文書を排除することも可能である．これら 2 つの効果とともに，一般業務の促進を図るものであり，社会的効果も非常に高い．

今後は，背景知識の充実化はもちろんのこと，今回は対象外としたメイル文書特有な属性に着手し，フィルタリング精度の向上を目指す．なお，今回の実験では，従来手法との評価に際し，従来手法のパラメータ（α，β，γ，η）を固定値に設定したが，あらかじめ予備的な実験を行い，パラメータを最適値に設定したう

☆ 実験データを 10 個のサブセットに分割し，9 個のサブセットを学習用，残りの 1 個を評価用にする実験を 10 通りのすべての組合せで実行する方法．

☆☆ 重要度が算出できない多重組については，重要度を 3 とした．

☆☆☆ 簡単のため,「送信元」「類型」「優先度」から成る多重組とする．

えでより正確な評価を行う計画である．また，本研究のように，ユーザのプロファイルを作成し，それを利用する場合，ユーザの視点の変化に対応できるような改良が必要である．さらに，各個人ごとに重要度の算出に適した属性の組合せをプロファイルから学習する機構についても積極的に組み込んでいきたい．

参 考 文 献

1) 森田昌宏，速水治夫：情報フィルタリングシステム，情報処理，Vol.37, No.8, pp.751–758 (1996).
2) 高野敦子，柏岡秀紀，平井　誠，北橋忠宏：対話における文脈の定型化と文脈処理の枠組，情報処理学会論文誌，Vol.34, No.1, pp.88–98 (1993).
3) 佐藤　円，佐藤理史，篠田陽一：電子ニュースのダイジェスト自動生成，情報処理学会論文誌，Vol.36, No.10, pp.2371–2379 (1995).
4) 長谷川隆明，高木伸一郎：文書構造の認識と言語の特徴の利用に基づく電子メールからのスケジュールと ToDo の抽出，情報処理学会論文誌，Vol.40, No.10, pp.3694–3705 (1999).
5) Foltz, P.W. and Dumais, S.T.: Personalized Information Delivery: An Analysis of Information Filtering Methods, *Comm. ACM*, Vol.35, No.12, pp.51–60 (1992).
6) 加来田裕和，角　隆一：電子メール利用履歴に基づいた処理順序取得システム，第 57 回情報処理学会全国大会論文集，2F-2, pp.336–337 (1998).
7) 長谷川隆明：送受信履歴と情報抽出に基づく電子メールの個人適応型ランキング，自然言語処理研究会資料，NL132-3, pp.17–24, 情報処理学会 (1999).
8) 首藤公昭：文節構造モデルによる日本語の機械処理に関する研究，福岡大学研究所報，pp.1–121 (1980).
9) 溝渕昭二，住友　徹，泓田正雄，青江順一：日本語時間表現の一解釈法，情報処理学会論文誌，Vol.40, No.9, pp.3408–3419 (1999).
10) Salton, G.: *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley (1988).
11) Buckley, C., Salton, G. and Allan, J.: The Effect of Adding Relevance Information in a Relevance Feedback Environment, *Proc. 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.292–298 (1994).
12) Lewis, D.D., Schapire, R.E., Callan, J.P. and Papka, R.: Training Algorithms for Linear Text Classifiers, *Proc. 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.298–307 (1996).

(平成 11 年 11 月 19 日受付)
(平成 12 年 6 月 1 日採録)

獅々堀正幹（正会員）

平成 3 年徳島大学工学部情報工学科卒業．平成 5 年同大学院博士前期課程修了．平成 7 年同大学院博士後期課程退学．同年同大学工学部知能情報工学科助手．現在同大学工学部知能情報工学科講師．博士（工学）．情報検索，文書処理，自然言語処理の研究に従事．情報処理学会第 45 回全国大会奨励賞受賞．電子情報通信学会，言語処理学会会員．

藤井　誠

平成 10 年徳島大学工学部知能情報工学科卒業．平成 12 年同大学院博士前期課程修了．現在三菱電機電力・産業システム事業所に勤務．

安藤　一秋（正会員）

平成 6 年徳島大学工学部知能情報工学科卒業．平成 11 年同大学院博士後期課程修了．博士（工学）．現在香川大学工学部信頼性情報システム工学科助手．情報検索，自然言語処理の研究に従事．電子情報通信学会会員．

青江　順一（正会員）

昭和 49 年徳島大学工学部電子工学科卒業．昭和 51 年同大学院修士課程修了．同年同大学工学部情報工学科助手．現在同大学工学部知能情報工学科教授．この間コンパイラ生成系，パターンマッチングアルゴリズムの効率化の研究に従事．最近，自然言語処理，特に理解システムの開発に興味を持つ．著書「Computer Algorithms – Key Search Strategies」,「Computer Algorithms – String Matching Strategies」（IEEE CS press）．平成 4 年度情報処理学会「Best Author 賞」受賞．工学博士．電子情報通信学会，人工知能学会，日本認知科学会，日本機械翻訳協会，IEEE，ACM，AAAI，ACL 各会員．