自然言語処理 133－5
（1999．9．10）

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
徳島大学 工学部 知能情報工学科
E-mail : {bori, aoe}@is.tokushima-u.ac.jp

近年，インターネットが普及し，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．それに伴い，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．メイル文書の重要度を判定する際に，その判断基準は個人毎に異なるため，一意に決められた知識を用いることはできない．そこで，本稿では，重要度判定の基準は各個人が受信済みのメイル文書内に隠されていると考え，受信済みのメイル文書から送信元，文書のテーマや類型等の多属性項目の組み合わせから成る構造化知識を重要度判定の知識として獲得し，その知識を用いたフィルタリング手法を提案する．更に，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．

キーワード：情報フィルタリング，電子メイル文書，プロファイル，知識獲得

A Filtering Method for E-Mail Documents Using Personal Profiles

Masami SHISHIBORI, Makoto FUJII, Kazuaki ANDO and Jun-ichi AOE
Dpt. of Information Science & Intelligent Systems, Faculty of Engineering, Tokushima University
E-mail : {bori, aoe}@is.tokushima-u.ac.jp

Nowadays, E-mail is very often used as the one of the communication methods with the advance of the network technology, however a great deal of E-mail documents including useless information such as commercial mails is distributed to our computers. In order to solve this problem, we propose the method to filter out unimportant E-mail documents from a great of received E-mail documents automatically by judging the content of each document. We considered that existing E-mail documents, which have been already received by each user, include criteria for judging the importance of the new E-mail document. Then, this method regards existing E-mail documents as the corpus and extracts the corpus-based knowledge from them. 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 enables us to give the point which shows whether the new E-mail document is important or not for the user.

Keywords: Information Filtering, E-mail Documents, Profile, Knowledge Acquisition

１．はじめに

近年のインターネットの普及に伴い，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，殆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，しかも，それらのすべてが重要なものばかりではなく，不要なメイルが数多く送信されるため，業務に支障をきたすという状況が発生しつつある[1]．このように，大量のメイル文書が氾濫している現在，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的に分類する処理，即ち，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[2]．

このような問題に対して，電子ニュース記事やメイル文書から各種情報を自動抽出する手法[3],[4]，ニュース記事の自動分類を行う手法[5],[6]，また，ユーザの履歴情報（プロファイル）を用いてメイル文書の処理順序を提示する手法[7],[8]等が提案されている．しかし，現在までのフィルタリング技術は定型的な文書を対象にした研究が多く，また，メイル文書を対象にした研究については，メイル文書特有の形態を生かした解析ができていない点に問題がある．

そこで本稿では，各個人が受信済みのメイル文書内に重要度を判定するのための履歴情報が含まれていると考え，各個人が予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[9]，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に対して頻度集計した結果をプロファイルとする．本手法により作成されたプロファイルは，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により，各個人に適合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，作成したプロファイルがスパースな場合には，類似した単語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ているため，極力少ないノイズで重要度を近似することができる．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つの効果共に，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．

以下，**２．**では，他の研究と比較することより，本研究の位置づけを行う．次に，**３．**において，メイル文書の重要度とは何か，また，如何なる要因から決定されるかを明確にした後，**４．**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，更に，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．**で，本手法の評価を行い，最後に**６．**でまとめ及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[1]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を提案している．長谷川ら[4]は電子メイル文書を対象に，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をルール化し，そのルール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重要度に関するポイント数が付加された表現パターンを事前にルール化し，メイル文書内に含まれる表現パターンを検出することにより重要度を算出する方法が考えられる．しかし，どのような内容が重要かといった判断基準は各個人毎に異なり，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重要度を適切に判定できない．

また，Foltz ら[5]は情報検索手法の一つであるLSI(Latent Semantic Indexing)法[6]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．

更に，加来田ら[7]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）とその文書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．また，学習データ数が少ない場合の近似法についても触れられていない．

本稿で提案する手法は，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，多属性値の組み合わせからプロファイルを作成する点が文献[7]と異なり，重要度も実時間で算出可能である．更に，文献[8]では，主題となり易い重要語句が登録された辞書を用いていたが，本手法では学習データからそれらの語句を獲得する．尚，文献[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，これについては本稿では論じない．

---

[1] 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．

3．メイル文書の重要度とは

3．1　重要度の要因

本手法の説明する前に，まず，重要度とは何か，また，どのような要因から重要度が決定されるかを検討する必要がある．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．

属性値は各属性毎に，以下のような項目から取得可能である．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[10]，及び（γ）については時間表現や時間を表す副詞から取得可能である．最後に，（θ）の「テーマ」を取得するために，主題となり易い名詞類を登録した辞書[8]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，受信済みのメイル文書が予めユーザによって類似した内容の文書毎に纏められ，かつ，メイルボックスの各フォルダーに分類されていることを前提とする．そして，受信済みのメイル文書については格納されていたフォルダー名を「テーマ」とする．また，重要度が未知の入力メイル文書については，文書分類法と同様な方法で各グループとの類似度を求め，最も類似したグループのフォルダー名を「テーマ」とする．

3．2　重要度の個人差

3．1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毎に異なることは明白である．例えば，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知れない．このように，各個人毎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，各個人毎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知れない．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知れない．このように，各個人毎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は，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毎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元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

3章で述べた定義に基づくと，メイル文書のフィルタリング処理における最も重要な点は，どれだけ各ユーザに適した知識を用いれるか，そして，その知識を如何にして獲得するかであると言える．この点に関して，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のための履歴情報が含まれていると考える．即ち，各ユーザが既に受信しているメイル文書に，各ユーザの判断基準に従った優先度[2]付けを行い，優先度付けされた既存のメイル文書を解析した結果得られる知識を用いれば，各ユーザの好みや履歴情報を反映した重要度が算出可能だと考える．

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

[2] ユーザが予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．

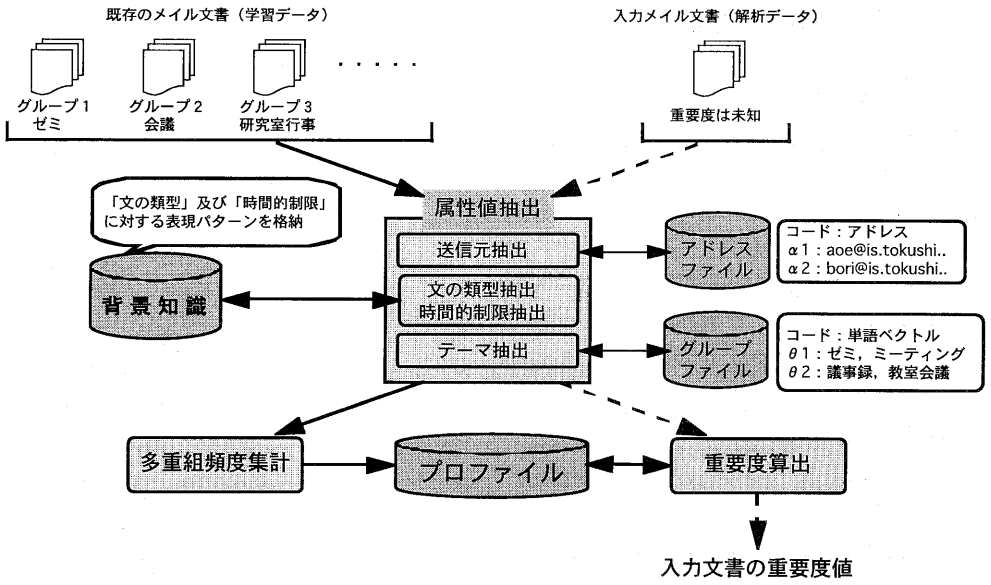

図1　フィルタリング手法の概要図

によって予め優先度が付与され，かつ，内容の類似性により各フォルダーに分類されている．これらの文書を形態素解析し，３．１で述べた取得方法に従って個々の文書に含まれる各属性値を検出する．ここで，「文の類型」及び「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．同様に，「テーマ」についてもフォルダー名をコード化し，そのコードを属性値とする．更に，フォルダーに格納されている全メイル文書に対して，ヘッダー Subject の欄から名詞を切り出し，そのフォルダーを特徴付ける単語ベクトルを作成する．そして，コードと単語ベクトルの対応をグループファイルに記載する．そして，各メイル文書毎に，検出した多属性値と優先度の組み合わせから成る多重組を生成する．すべてのメイル文書に対して多重組を生成した後，それらの頻度を集計した結果をプロファイルとする．尚，背景知識の内容，及びプロファイルの作成方法については４．２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「文の類型」及び「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，入力文書内ヘッダー Subject の欄から名詞を切り出し，その名詞とグループファイルに記載されている単語ベクトルとを比較し，入力文書と最も類似したフォルダー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元にして重要度を確率的に求める．尚，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で述べる．

４．２　学　習　部

４．２．１　背景知識

まず，背景知識には「文の類型」及び「時間的制限」の属性値を検出するための表現パターンが格納されるのであるが，類似したパターンをグルーピングすることで一つの属性値を構成している．グルーピングすることでプロファイルのスパースさを若干回避できる．例として，背景知識の一部を表１に示す．

表1　背景知識の例

属性	属性値	コード	表現例
文の類型	勧告	$\beta 1$	～すること，～するように，～せよ
	依頼	$\beta 2$	～して下さい，～してくれ，～して欲しい，～してちょうだい
	条件付依頼	$\beta 3$	もし～ならば・・・下さい，～だったら・・・してくれ
	勧誘	$\beta 4$	～しましょう，～しよう，～しようよ，～行こう
	疑問	$\beta 5$	～でしょうか，～ですよね，～かな，～か，～?
時間的制限	1週間以上	$\gamma 1$	1ヶ月以内に，二三週間以内に，
	1週間以内	$\gamma 2$	1週間以内に，7日以内に，
※但し，時区間を含まない時間表	3日以内	$\gamma 3$	なるべく早く，忘れないうちに，二三日中に
現は，着信時間を考慮して分類す	24時間以内	$\gamma 4$	出来るだけ早く，早めに，24時間以内に，一両日中に
る	12時間以内	$\gamma 5$	今すぐ，早急に，すぐに，1時間以内に，1時間後に

まず，「文の類型」については文末に現れる助述表現を対象にし，その表現パターンを文献[9],[10]を参考に分類した．次に「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[11]を含む時間表現；

(2) 時区間を含まないが時点[11]は含む時間表現；

(1)に属するものには「1時間以内に」や「1週間以内に」等が挙げられ，これら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．一方，(2)には「明日まで」や「今週中に」等が属するが，これら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．そこで，(2)に属する表現については，メイル文書の着信時間（または参照時間）を考慮し，時区間を求めた後で対応する属性値に分類することにした．尚，「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 $D=\{d1, d2, \ldots, di, \ldots, dA\}$ とすると，各文書は $di=\{xi1, xi2, \ldots, xij, \ldots, xiB\}$ と多重組の集合を持ち，各多重組は $xij=(\alpha l, \beta m, \gamma n, \theta q, \kappa r)$ という多属性値の多次元ベクトル値となる．尚，多重組の構成方法は以下のように1文書毎に行う．

【多重組構成手順】

手順1：「送信元」の属性値 αl を多重組に代入；

手順2：「文の類型」の属性値 βm が複数存在する場合，その数だけ多重組を複製し，βm 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 γn を検出し，検出数だけ多重組を複製した後，γn を代入；

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

【手順終了】

また，一文書毎に生成された多重組の個数に従って，頻度 $freq(xij)$ を与える．但し，$1 \leqq j \leqq B$ とすると，$freq(xij)=1/B$ であり，$\Sigma j\ freq(xij)=1$ とする．ここで，$\cap ij\ xij \neq \phi$ 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 $P=\{p1, p2, \ldots, pk, \ldots, pC\}$ とする．但し，pk は多重組を表し，$\cap k\ pk=\phi$，$freq(pk)$ は pk と同じ内容の多重組 xij の頻度合計値とする．また，このプロファイルはユーザーがどのような属性値の組み合わせに重要度を置いているかを示している．尚，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 $p<lmnqr>$ と記すことにする．

4．3　解析部

解析文書データを $T=\{t1, t2, \ldots, ti, \ldots, tE\}$ とすると，各文書は $ti=\{yi1, yi2, \ldots, yij, \ldots, yiH\}$ と多重組の集合を持つ．但し，yij は優先度 κ の値は持たない多次元ベクトル値である．解析部では yij 毎に重要度を算出し，yij 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，処理内容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 $y_{ij}=(\alpha l,\beta m,\gamma n,\theta q)$に対して，$p<lmnq>$にどのような優先度が付与されていたかを，式(1)により条件付確率値として求める．

$$P(\kappa r|\alpha l,\beta m,\gamma n,\theta q)=\frac{freq(p<lmnqr>)}{freq(p<lmnq>)}\cdots(1)$$

但し，

$$freq(p<lmnq>)=\sum_{r}freq(p<lmnqr>)$$

とする．また，多重組 y_{ij} 内に全ての多重組が検出されていない場合，存在する属性値のみを対象に式(1)を適用する．例えば，$y_{ij}=(\alpha l,\beta m,*^{3},\theta q)$の場合，式(1)は次のように変化する．

$$P(\kappa r|\alpha l,\beta m,\theta q)=\frac{freq(p<lmqr>)}{freq(p<lmq>)}\cdots(2)$$

但し，

$$freq(p<lmqr>)=\sum_{n}freq(p<lmnqr>)$$

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

$$R_{ij}=\sum_{r}\{r\times P(\kappa r|\alpha l,\beta m,\gamma n,\theta q)\}\cdots(3)$$

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[4]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，優先度は共に 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．

---

[3] ＊は入力文書に対応する属性値がなかったことを表す．

[4] 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．

ID:(送信元，類型，時間，テーマ，優先度)＝頻度
1:(　青江，勧告，12時間，　0，　3)＝2
2:(　青江，勧告，24時間，　0，　4)＝3
3:(　青江，勧告，12時間，ゼミ，　5)＝7
4:(　青江，勧告，　0，ゼミ，　5)＝3
5:(　青江，勧告，　0，試験，　5)＝2
6:(　青江，勧告，二三日，行事，　5)＝3
7:(　青江，依頼，24時間，ゼミ，　4)＝3
8:(　青江，依頼，　1週間，試験，　2)＝2
9:(　青江，依頼，　0，行事，　1)＝3

図２　プロファイルの例

4．3．2　近似処理

プロファイルがスパースで，入力文書の多重組と同じものがプロファイル内に存在しなかった場合にも，適切なフィルタリングを行えることが必要である．そこで本手法では，同じ多重組が存在しなかった場合，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（ベクトル空間内の各軸）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）

まず，文の類型，テーマ，送信元は文書への興味度（ベクトルの方向）を構成し，時間的制限は緊迫度（ベクトルの大きさ）を表しているため[10]，「時間的制限」の属性値を変更してもノイズが入りにくいと考え，最初に置き換えを行う．「時間的制限」に対する近似は，まず，この属性値を無視した式(2)による重要度を求める．しかしながら，この重要度は γn を全く無視した値であるので，「時間的制限」の変化に対する「優先度」の変化の割合をプロファイル内のデータから求め，先の重要度にこの変化の割合を加味した値を近似値とする．即ち，ベクトルの方向は変えずに，ベクトルの大きさのみを変えて近似する．

次に，送信元のアドレスには職業，地位等の点で類似性があること，更に，「送信元」に関しては，属性値のグルーピングをしておらず，属性値間の類似性が高いことから，２番目に置換を行う．「送信元」の近似は以下のように行う．まず，送信元 αL が指定され

た多重組 y_{ij} = ($\alpha L, \beta M, \gamma N, \theta Q$)に対して，$p<LMNQ>$ が存在しなかった場合，送信元を置換した（$\alpha Lrep, \beta M, \gamma N, \theta Q$)から求めた重要度を近似値とする．但し，

$$Lrep = \max_{l}\{sim(p < Lmnqr >, p < lmnqr >)\}$$

とする．ここで，$sim(a, b)$はベクトル空間 a, b の相関係数を求める関数とする．このように「送信元」の近似は，「文のテーマ」「文の類型」「時間的制限」に対して，最も類似した優先度が付けられている送信者に置換する．

上記の近似に対しても，置き換えるべき多重組が見つからなかった場合，「テーマ」に関する近似を「送信元」と同様な方法で行う．尚，文の類型については，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．

5．評　価

本手法の有効性を確認するため，学習データ数による解析度[5]，及び解析精度の変化に関する実験を４人の被験者に対して行った．

実験に用いた学習データの内訳を表２に示す．表２において，学習文書数の括弧内の値は各文書の合計サイズ（単位は byte），送信元数の括弧内の数値は１送信元当たりの学習文書数，フォルダー数の括弧内の値は１フォルダー当たりの学習文書数をそれぞれ示す．多重組数は各学習文書から得られた多重組の数である．また，「文の類型」については310 個（属性値数は9），「時間的制限」については 43 個（属性値数は 5）の表現パターンを登録した背景知識を用いた．これらのデータを用いて各プロファイルを作成した後，学習データとは別に用意した解析データ（A,B,C は各 30 通，D は 20 通）の重要度を求めた．尚，優先度，重要度共にランク数を 5（ランク５が最高値）とした．稼働マシンは Power Macintosh 8500/200MHz である．

まず，解析度を評価するために，学習データ数と解析度の関係を通常処理及び各近似処理毎に求めた結果を図３のグラフに示す．尚，紙面の都合上，図３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．図３より，「送信元」の近似が解析度を大き

表２　学習データの内訳

被験者	学習文書数	送信元数	フォルダ数	多重組数
A	248 (658)	59 (4.2)	7 (35.4)	335
B	234 (501)	57 (4.1)	12 (19.5)	297
C	150 (233)	7 (21.4)	6 (25.0)	213
D	80 (90)	20 (4.0)	5 (16.0)	102

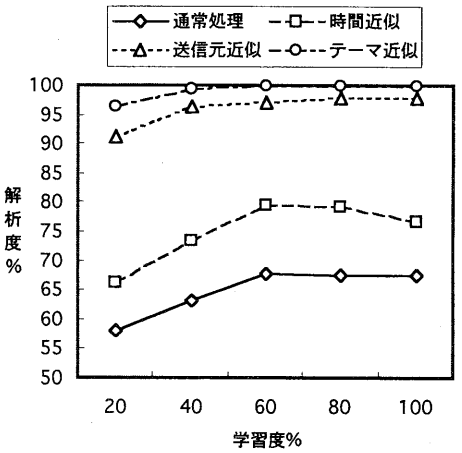

図３　学習データ数と解析度の関係

く向上させていることから，この属性値をグループ化すれば，通常処理の解析度も向上すると予測される．

次に，解析精度を評価するために，通常処理及び全近似を適用した近似処理の２種類について，学習データ数と平均誤差の関係を求めた（図４）．ここでの誤差はシステムが算出した重要度と各ユーザが判定した重要度との差分を示す．尚，これも４人の平均値である．図４より，両処理共に学習データの増加と共に，精度が高まっている．また，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，しかも，学習度が100%でのノイズが 0.018 と非常に少ないことから，本近似手法が有効であることが分かる．

また，解析精度の評価として，システムが算出した重要度とユーザが判定した重要度との相関係数を求めた．比較手法として，文書内に現れる全単語の頻度から重要度を求めた場合（tf），及び tf/idf を計算し，重要語の頻度から重要度を求めた場合(tf/idf）について相関係数を求めた．実験結果を表３に示す．表３より，本手法がユーザの判断とかなり高い相関を持ち，単一属性を扱う比較手法よりもいい結果が得られた．

[5] 重要度が算出できた解析文章内多重組の割合を示す．

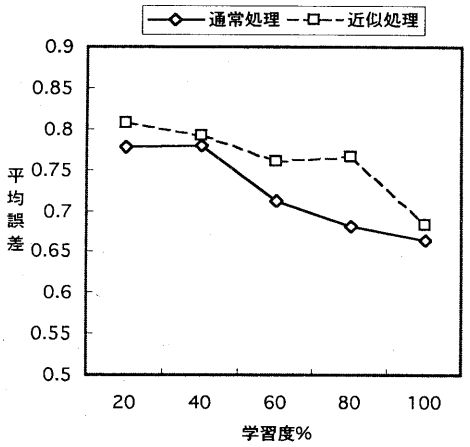

図４　学習データ数と平均誤差の関係

表３　各手法による相関係数

被験者	本手法	tf/idf	tf
A	0.57	0.41	0.38
B	0.51	0.19	0.16
C	0.66	0.39	0.30
D	0.78	0.29	0.27

表４　各属性値の検出精度

被験者	多重組数	テーマ(%)	類型(%)	時間(%)
A	62	59.2	66.4	61.1
B	42	75.0	67.1	37.5
C	57	53.3	81.8	44.4
D	32	54.5	68.4	33.3

本手法による解析結果の詳細として，解析データからの属性値の検出精度を表４に示す．「類型」についてはある程度の検出率を得ているが，「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，今回は Subject 内の名詞のみを対象に「テーマ」を特定しているために検出率が低いが，本文内から得られる重要語も考慮すると検出率が上がると思われる．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．

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 370.5 bytes），解析時間は約 1.2 秒，但し，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 4.6Kbytes である．

６．まとめ

本稿では，受信済みのメイル文書から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．今後は，背景知識の充実化は勿論のこと，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．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柔軟に対応できるように改良する必要がある．

参　考　文　献

[1] R. J. Hall, "How to Avoid Unwanted Email," Commun. ACM, Vol. 41, No. 3, pp.88-95 (1998).

[2] 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).

[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，“電子メールコミュニケーションにおけるスケジュール情報抽出”，情報処理学会自然言語処理研究会資料，NL123-10，pp.73-80 (1998).

[5] P. W. Foltz and S. T. Dumais, "Personalized Information Delivery: An Analysis of Information Filtering Methods," Commun. ACM, Vol. 35, No. 12, pp.51-60 (1992).

[6] P. G. Young, "Cross-Language Information Retrieval Using Latent Semantic Indexing," A Thesis Presented for the Master of Science Degree, The University of Tennessee, Dec. (1994).

[7] 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”，情報処理学会第 57 回全国大会，2F-2，pp.3-336-3-337 (1998).

[8] 獅々堀正幹，藤井誠，安藤一秋，青江順一，“各個人のプロファイルを用いたメイル文書のフィルタリング手法”，信学技報データ工学研究会資料，DE98-31，pp.9-16 (1998).

[9] 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.88-89 (1993).

[10] 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).

[11] 原裕貴，北上始，中島淳，“時間概念の表現とデフォルト推論”，人工知能学会論文誌，Vol. 3，No. 2，pp.216-223 (1988).